ISSN 2279-9362

Collegio Carlo Alberto

Theory and learning protocols for the material tempotron model

Carlo Baldassi
Alfredo Braunstein
Riccardo Zecchina

No. 393
December 2014

Carlo Alberto Notebooks

www.carloalberto.org/research/working-papers

Theory and learning protocols for the material tempotron model

Carlo Baldassi,[1,2] Alfredo Braunstein,[1,2,3] and Riccardo Zecchina[1,2,3]

[1]*DISAT and Center for Computational Sciences, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

[2]*Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy*

[3]*Collegio Carlo Alberto, Via Real Collegio 30, 10024 Moncalieri, Italy*

Abstract

Neural networks are able to extract information from the timing of spikes. Here we provide new results on the behavior of the simplest neuronal model which is able to decode information embedded in temporal spike patterns, the so called tempotron [1]. Using statistical physics techniques we compute the capacity for the case of sparse, time-discretized input, and "material" discrete synapses, showing that the device saturates the information theoretic bounds with a statistics of output spikes that is consistent with the statistics of the inputs. We also derive two simple and highly efficient learning algorithms which are able to learn a number of associations which are close to the theoretical limit. The simplest versions of these algorithms correspond to distributed on-line protocols of interest for neuromorphic devices, and can be adapted to address the more biologically relevant continuous-time version of the classification problem, hopefully allowing for the understanding of some aspects of synaptic plasticity.

Contents

I. INTRODUCTION

Recent studies concerning learning processes in neural circuits have highlighted the role of spike timing and synchrony (e.g. in sensory systems [2–5]), leading to a view of the learning devices as a class of time coincidence detectors of a limited number of spikes (at least under certain circumstances). These observations are at the root of several fundamental questions concerning neural coding, the most important one possibly being how do neurons learn to recognize multiple spatiotemporal patterns.

A very stimulating contribution in this field has been the recent introduction by Gutig and Sompolinsky [1] of a perceptron-like model neuron which is able to process spatio-temporal patterns, the

so called tempotron. In spite of its simplicity, such a device is capable of decoding the information contained in the synchrony of spike patterns through a relatively simple supervised gradient learning rule. Subsequent work [6] has analyzed by statistical physics techniques the storage capacity (i.e. the typical maximum number of distinct input-output associations which the device can in principle be trained to reproduce, assuming that the inputs and the expected responses are drawn from some probability distribution) and the geometry of the space of solutions of the tempotron for continuous synaptic weights.

In a nutshell, the tempotron is the simplest form of an integrate and fire (IF) neuron, with N input synapses of strength J_i, $i = 1, \ldots, N$ (also called synaptic weights). In the tempotron, each input corresponds to N sequences of spikes, where the set of spiking times is denoted by $\{t_i\}$. The tempotron performs a binary classification of the inputs depending on whether the membrane potential reaches or not the firing threshold θ in the given time interval. The potential at time t is given by $V(t) = \sum_i J_i \sum_{t_i<t} v(t - t_i)$, where $v(t)$ is the temporal kernel of the membrane. A standard choice for the kernel is the exponential one, namely $v(t) = v_0 \left(e^{-t/t_m} - e^{-t/t_s}\right)$, where t_m and t_s are the membrane and synaptic integration time constants. For this model the precise timings and the number of output spikes (if greater than one) play no role in the binary classification, allowing for multiple equivalent output spiking profiles for positive classifications of a given time interval.

As discussed in [6], a key parameter is the quantity $K = T/\sqrt{t_m t_s}$ where T is the duration of the input pattern. When both N and K are large, and $N \gg K$, certain time correlations can be neglected and the analysis simplifies. This has allowed the authors of [6] to estimate the storage capacity of the device for the case of continuous weights and random i.i.d. patterns. It is interesting to observe that these conditions are not far from being actually realistic [6]. The vanishing of the correlations at different times is due to the sparse regime under which the device operates, and it means that the width of the kernel $v(t)$ is much shorter than the typical interval between incoming spikes; this in turn means that, under this regime, only quasi-simultaneous input spikes actually contribute to the depolarization at any given time, which explains why in [6] the theoretical and experimental results are closely reproduced with a simplified model in which time is discretized in $K^{\text{discrete}} = K/8$ bins and the output is simply given by a perceptron rule applied on each bin (see paragraph I for additional details).

Basic devices like the temportron have the potential virtue of touching those fundamental questions in neural coding which are preserved in spite of the simplicity of the device itself. In such a framework, we have approached the problem from a different angle, namely adopting a compu-

tational scheme which that is not based on a gradient-like computation but is still fully local and distributed. We study the simplified (time-discretized) tempotron by both the replica method and the so called message-passing approach (or cavity method) which allows us to study analytically the storage capacity and at the same time to derive simple learning protocols (i.e. training rules for the modifications of the synaptic strengths, which the device applies upon receiving input patterns and being made aware of the expected response, such that at the end of the training the desired set of input-output associations is learned) which are efficient and do not rely on any continuity condition of the synaptic weights. We also show how to adapt the simplest of these learning protocols to address the original, continuous-time model. For the sake of simplicity we focus directly on the case of discrete synapses, although the results could be extended to the continuous case.

For the cases of single and multilayer perceptrons with firing rate coding and binary synapses we have shown in previous works [7–9] that the message-passing approach is indeed efficient in solving the learning problem for random patterns and that the computational scheme can be simplified to the point of providing extremely simple learning protocols. These past results together with the novel ones on spatio-temporal coding presented here should be of practical interest for large scale neuromorphic devices and hopefully for providing novel hints on aspects of synaptic plasticity.

The model We studied two tempotron scenarios, one in which synaptic conductances can take values in $\{-1,+1\}$, and one in which they can take values in $\{0,1\}$, focusing on the former case in simulations. As in the final paragraphs of [6], we worked under the simplifying assumption that input and output spike patterns can be encoded (via binning) as sparse strings of 0's and 1's, and that the relationship between the inputs and the output at any given time bin is given by a perceptron rule. As mentioned in [6] and in the Introduction, we expect that this simplification does not qualitatively alter the overall picture under the sparse regime considered, since even in the integrate-and-fire model only quasi-simultaneous input spikes affect the overall depolarization at any given time, due to the fast membrane decay constant w.r.t. the typical inter-spike interval; indeed, our numerical results (see section III C) show that it is even possible to use a time-discretized learning protocol to address the original continuous-time classification problem.

We thus consider a classification device with N binary synapses, J_i with $i \in \{1,\dots,N\}$, which has to learn to classify $M = \alpha N$ input patterns. The input patterns are $N \times K$ matrices (where K here corresponds to the K^{discrete} discussed in the Introduction) whose elements are $\xi_{it}^{\mu} \in \{0,1\}$ with $i \in \{1,\dots,N\}$, $t \in \{1,\dots,K\}$ and $\mu \in \{1,\dots,M\}$. For each pattern μ and at each time step t, the device response is given by $V_t^{\mu} = \Theta\left(\sum_{i=1}^{N} J_i \xi_{it}^{\mu} - \theta\right)$, where $\Theta(x)$ is the Heaviside step function and θ is a threshold; we call the vector $V^{\mu} = \{V_t^{\mu}\}_{t\in\{1,\dots,K\}}$ the *internal representation* for

the pattern μ. Finally, a pattern μ is classified according to $s^\mu = 1 - \prod_{t=1}^{K} (1 - V_t^\mu)$, i.e. the overall output s^μ equals 0 if the internal representation is a vector of all 0's, or it equals 1 if at least one element of V^μ is 1. Each pattern has a desired output $s^\mu_{\rm exp} \in \{0, 1\}$ which is to be compared to the actual output s^μ: the classification problem is satisfied when $s^\mu = s^\mu_{\rm exp}$ for all μ. For notational simplicity, we also define $\sigma^\mu = 2s^\mu - 1$ and $\sigma^\mu_{\rm exp} = 2s^\mu_{\rm exp} - 1$ when we need to convert the outputs so that they take values in $\{-1, +1\}$.

We studied the case in which all inputs and expected outputs are i.i.d. random variables. We call f' the output frequency (i.e. $s^\mu_{\rm exp} = 1$ with probability f') and f the input frequency (i.e. $\xi^\mu_{it} = 1$ with probability f). We will assume in the following that $f = \left(1 - (1 - f')^{\frac{1}{K}}\right)$: this ensures that the probability that a vector $\{\xi^\mu_{it}\}_{t \in \{1,\dots,K\}}$ is composed of all 0's is $(1 - f')$, i.e. has the same statistics as the internal representations which satisfy the input/output associations. Clearly, the model reduces to a standard perceptron when $K = 1$.

We assume that $N \gg 1$, $K \gg 1$ and $N \gg K$; in this case, $f \simeq -K^{-1} \log(1 - f')$, which shows that the inputs are sparse at large K with our choice for f. For simplicity, all our theoretical results and simulations will be presented for the case $f' = 0.5$, which is the value that maximizes the capacity.

II. THEORETICAL ANALYSIS

A. Replica theory

We studied the device described above within the replica theory, in a replica symmetric (RS) setting for the internal representations, in the limit of large K, both for the case in which $J_i \in \{-1, +1\}$ and $J_i \in \{0, 1\}$, and estimated the entropy, the critical capacity, the optimal value for the threshold, and studied the structure of the space of the solutions and the valid internal representations. The results are almost identical for both ± 1 and 0/1 cases, so in the following we will only specify the model when a difference arises. We confirmed the results, where possible, with the cavity method (see section II B). All details of the calculations are provided in the Appendix (section V); here we summarize the results.

The zero-temperature entropy of the device is defined as $S(\alpha) = \frac{1}{N} \langle \log(\mathcal{V}) \rangle$, where $\mathcal{V}$ is the number of solutions (valid configurations of J's) to the problem associated with some choice of the patterns ξ^μ_{it} and their expected outputs $s^\mu_{\rm exp}$ (see also eq. 15), and $\langle \cdot \rangle$ denotes the average over the patterns. $S(\alpha)$ can't be negative (since the number of valid configurations is an integer number);

the value of α at which $S(\alpha)$ vanishes is called the *critical capacity*, and represents the typical number of patterns per synapse which can be correctly classified by the device (i.e. stored) when the patterns are extracted according to the random i.i.d. distribution which we are studying.

The RS replica calculation predicts $S(\alpha) = \log(2)(1-\alpha)$, which interestingly does not depend on K (provided $K \gg 1$). This function goes to zero at $\alpha_c = 1$, which coincides with the information theoretic upper bound, i.e. the device is able to store one bit of information per synapse. This is in contrast with other related architectures, e.g. the multi-layer perceptron. After this point, the entropy is negative, and therefore the RS solution is no longer valid.

The typical value of the overlap between two different solutions to the same classification problem, defined as $q = \frac{1}{N}\left\langle \sum_{i=1}^{N} J_i^a J_i^b \right\rangle$ for two solutions $\{J_i^a\}_{i\in\{1,\dots,N\}}$ and $\left\{J_i^b\right\}_{i\in\{1,\dots,N\}}$, is constant for all values of α, and as low as possible, i.e. $q = 0$ in the ± 1 case and $q = Q^2 = 1/4$ in the 0/1 case, where Q is the typical fraction of non-null synapses (which we found to be $Q = 1/2$). In terms of the structure of the space of the solutions, this means that the clusters of solutions are isolated (point-like).

We expand the threshold in series of $\sqrt{N}$ and write $\theta = \theta_0 N + \theta_1 \sqrt{N}$. The optimal value of θ_0 is 0 in the ± 1 case and $\frac{f}{2}$ for the 0/1 case; the value of θ_1 does not affect the capacity, but we can set it so that synaptic values are unbiased:

$$\theta_1 = -\sqrt{2f(1-f)}\,\mathrm{erfc}^{-1}\left(2\sqrt[K]{1-f'}\right) \tag{1}$$

where erfc^{-1} is the inverse of the complementary error function.

We found that the valid internal representations follow a binomial distribution at large K, i.e. that the probability distribution for the value at each time bin is independent of the others. This fact is in agreement with the continuous model findings [6], and it is interesting for two reasons: on one hand, it confirms that different time bins are uncorrelated in the sparse limit, which is important in order to achieve efficiency in applying the cavity method. On the other hand, it means that the distributions of the input and output spike trains are identical (note that our choice of f only ensures that the all-zero string occur with the same probability in input and output, but does not imply that the non-zero strings have the same distribution of the number of spikes). In turn, this is a necessary condition for recurrent networks to be built and work under this regime, which would be an interesting direction for future research.

We also computed how the internal representations are partitioned, and found that the rescaled entropy of the dominant internal representations (i.e. the logarithm of the number of different internal representations for a pattern which are associated with the largest portions of the solution

space) is given by $\log(2)\log\left(\frac{K}{\log(2)}\right)$. This means that, as K increases, the number of valid dominant internal representations increases as $K^{M\log 2}$, while the number of synaptic states associated with each of them correspondingly shrinks, so that the overall entropy remains constant.

B. Cavity method

The cavity method has been shown [10] to provide an alternative scheme for deriving the results from replica theory in the case of the binary perceptron with binary ± 1 inputs and binary ± 1 synapses. It has also been used on single instances of the learning problem on such devices (in which case it is known as the Belief Propagation algorithm, i.e. BP) to study the space of the solutions for some particular instance and for deriving heuristic learning algorithms [7, 8]. In those studies, the problem is represented as a factor graph in which synaptic weights are represented by variable nodes, and input patterns (and their desired outputs) are represented by factor nodes; messages are exchanged between the two types of nodes along the edges of the graph, representing marginal probabilities over the states of the variables; global thermodinamic quantities such as the entropy can be computed from the messages provided they satisfy the BP equations (which is typically achieved by reaching a fixed point in an iterative algorithm). Replica theory results can then be reproduced numerically with BP by averaging the computed quantities from a large number of samples of sufficient size.

In the case of the present study, however, in which the input patterns take values in $\{0, 1\}$, the approach used in [7] can not be applied directly for the sake of reproducing replica theory results, not even in the perceptron limit $K = 1$. This is due to a violation of the underlying assumption of the cavity method, known as the clustering property, as will be explained in greater detail at the end of this section, and as a result the standard BP equations are approximate, rather than becoming asymptotically exact in the large N regime. Thus, in order to reproduce the replica theory results, the standard BP equations must be amended; however, since the heuristic algorithm described in section III A is based on the standard BP, which is simpler and more computationally efficient, and proves equally effective to the corrected-BP version, we will first derive the standard BP equations, and describe how to correct them afterwards.

Standard Belief Propagation algorithm As mentioned above, messages on the factor graph represent probabilities over the variable nodes (i.e. the synaptic weights), and therefore can be represented by a single real value: as in [7], we use average values for this purpose (also called magnetizations).

The BP equations, written in terms of the cavity magnetizations m and n, read:

$$m_{i\to\mu} \;=\; \tanh\left(\sum_{\nu\neq\mu}\tanh^{-1}\left(n_{\nu\to i}\right)\right) \tag{2}$$

$$\begin{aligned} n_{\mu\to i} \;\propto\; & P\left(\sigma^{\mu}_{\text{exp}} = 1-\prod_{t=1}^{K}\Theta\left(\theta-\sum_{j\neq i}J_j\xi^{\mu}_{jt}-\xi^{\mu}_{it}\right)\right)+ \\ & -P\left(\sigma^{\mu}_{\text{exp}} = 1-\prod_{t=1}^{K}\Theta\left(\theta-\sum_{j\neq i}J_j\xi^{\mu}_{jt}+\xi^{\mu}_{it}\right)\right) \end{aligned} \tag{3}$$

where i, j are synapse indices and μ, ν are pattern indices. The second equation is the difference between the probabilities that the pattern μ is satisfied when synapse i takes the values 1 and -1, respectively, assuming all other synaptic values are distributed according to the cavity magnetizations $m_{j\to\mu}$ (for $j \neq i$). These can be computed from the probability that the internal representation is all zero given the value of J_i:

$$B_{\mu\to i}\left(J_i\right) = \sum_{\{J_j\}_{j\neq i}}\prod_{j\neq i}\left(\frac{1}{2}+J_j\frac{\mu}{2}\right)\prod_{t=1}^{K}\Theta\left(\theta-\sum_{j\neq i}J_j\xi^{\mu}_{jt}-J_i\xi^{\mu}_{it}\right) \tag{4}$$

With this, and using the shorthand notation $B^{\pm}_{\mu\to i} = B_{\mu\to i}\left(\pm 1\right)$, we can write eq. 3 as:

$$n_{\mu\to i} = \frac{B^{+}_{\mu\to i}-B^{-}_{\mu\to i}}{B^{+}_{\mu\to i}+B^{-}_{\mu\to i}}\left(1-\frac{2s^{\mu}_{\text{exp}}}{2-B^{+}_{\mu\to i}-B^{-}_{\mu\to i}}\right) \tag{5}$$

In order to compute efficiently the function B, we use the central limit theorem, which ensures that for large N we have:

$$B_{\mu\to i}\left(J_i\right) = \int_{\mathcal{S}_{\mu\to i}(J_i)}\prod_{t=1}^{K}dy_t\;\mathcal{N}\left(\bar{y};\bar{a}_{\mu\to i},\bar{\Sigma}_{\mu\to i}\right) \tag{6}$$

where $\mathcal{N}\left(\bar{y};\bar{a},\bar{\Sigma}\right)$ is a K-dimensional multivariate Gaussian with mean $\bar{a}$ and covariance matrix $\bar{\Sigma}$, whose elements are given by:

$$\left(\bar{a}_{\mu\to i}\right)_t \;=\; \sum_{j\neq i}\xi^{\mu}_{jt}m_{j\to\mu} \tag{7}$$

$$\left(\bar{\Sigma}_{\mu\to i}\right)_{tt'} \;=\; \sum_{j\neq i}\xi^{\mu}_{jt}\xi^{\mu}_{jt'}\left(1-m^2_{j\to\mu}\right) \tag{8}$$

The region of integration is a product of semi-bounded intervals: $\mathcal{S}_{\mu\to i}\left(J_i\right) = \bigotimes_{t=1}^{K}\left(-\infty, \theta - J_i\xi^{\mu}_{it}\right]$.

Computing this integral in general is very expensive, and rapidly becomes infeasible for large K. However, the sparsity of the input patterns implies that diagonal terms are of order K^{-1}, while

off-diagonal terms are of order K^{-2} and can be neglected, simplifying the computation:

$$B_{\mu\to i}\left(J_i\right) = \prod_{t=1}^{K} \frac{1}{2}\mathrm{erfc}\left(\frac{1}{\sqrt{2}}\left(\frac{\theta - J_i\xi_{it}^{\mu} - \left(\bar{a}_{\mu\to i}\right)_t}{\left(\bar{\Sigma}_{\mu\to i}\right)_{tt}}\right)\right) \tag{9}$$

Equations 2,5,7,8 and 9 form a closed system which allows computations to be performed effectively, and which can conveniently be modified to derive a heuristic solver algorithm (see section III A). However, as stated at the beginning of this section, these equations fail to exactly reproduce the replica theory results.

Corrected Belief Propagation algorithm The reason for the failure of the standard BP equations to provide correct results (e.g. when computing the entropy) is that when the inputs are unbalanced, i.e. they don't average to 0 (as is necessarily the case when the values are in $\{0,1\}$), the clustering property, i.e. the assumption that that the messages incoming into variable nodes from different factor nodes are uncorrelated, is violated. This can be seen by considering (see [10]):

$$\begin{aligned} c_{\mu\nu\to i} &= \frac{1}{N}\left(\left\langle\left(\bar{a}_{\mu\to i}\right)_t\left(\bar{a}_{\nu\to i}\right)_t\right\rangle - \left\langle\left(\bar{a}_{\mu\to i}\right)_t\right\rangle\left\langle\left(\bar{a}_{\nu\to i}\right)_t\right\rangle\right) \\ &= \frac{1}{N}\left(\left\langle\left(\sum_{j\neq i}\xi_{jt}^{\mu}J_j\right)\left(\sum_{j\neq i}\xi_{jt}^{\nu}J_j\right)\right\rangle - \left\langle\sum_{j\neq i}\xi_{jt}^{\mu}J_j\right\rangle\left\langle\sum_{j\neq i}\xi_{jt}^{\nu}J_j\right\rangle\right) \\ &= \frac{1}{N}\sum_{j\neq i}\xi_{jt}^{\mu}\xi_{jt}^{\nu}\left(1 - m_{j\to\mu}m_{j\to\nu}\right) \end{aligned}$$

which is $\mathcal{O}\left(1\right)$ unless the average input $\bar{\xi}$ is zero, in which case it is $\mathcal{O}\left(N^{-\frac{1}{2}}\right)$ and becomes negligible. Only in that case, therefore, standard BP equations become asymptotically correct for large N; in all other circumstances, they only provide an approximation (numerical experiments show that for our model they sistematically predict a slightly lower entropy than the correct one).

We also note that, if we define $\xi_{it}^{\mu} = \bar{\xi} + \rho_{it}^{\mu}$, where $\bar{\xi} = f$ and $\rho_{it}^{\mu} \in \{-f, 1-f\}$ with average $\bar{\rho} = 0$, we can split the depolarization as such:

$$\sum_i J_i\xi_{it}^{\mu} = f\sum_i J_i + \sum_i J_i\rho_{it}^{\mu} \equiv f\sqrt{N}T + \sum_i J_i\rho_{it}^{\mu} \tag{10}$$

where we defined the overall magnetization $T = \frac{1}{\sqrt{N}}\sum_i J_i$. It becomes apparent that the depolarization distributions induced by the different patterns are all correlated via T, which is a global quantity. We can however amend the BP algorithm, and derive correct marginals and therefore correct global thermodynamic quantities, by studying a related problem, in which this contribution is removed from the factor nodes and induced by an external field instead; this suggests the following modification to the cavity equations: we start by choosing a value for the magnetization, call it T', and we consider the problem with patterns ρ instead of ξ, thereby ensuring that the

clustering property holds, and with an additional external field F applied to each variable node, thus modifying eqs. 2 and 3 as such:

$$m_{i\to\mu} \;=\; \tanh\left(\sum_{\nu\neq\mu}\tanh^{-1}\left(n_{\nu\to i}\right)+F\right) \tag{11}$$

$$\begin{aligned} n_{\mu\to i} \;\propto\; & P\left(\sigma^{\mu}_{\text{exp}}=1-\prod_{t=1}^{K}\Theta\left(\theta-\sum_{j\neq i}J_j\rho^{\mu}_{jt}-\rho^{\mu}_{it}\right)\right)+ \\ & -P\left(\sigma^{\mu}_{\text{exp}}=1-\prod_{t=1}^{K}\Theta\left(\theta-\sum_{j\neq i}J_j\rho^{\mu}_{jt}+\rho^{\mu}_{it}\right)\right) \end{aligned} \tag{12}$$

The total magnetization T can be obtained from the cavity marginals as:

$$T=\sum_{i}\tanh\left(\sum_{\mu}\tanh^{-1}\left(n_{\mu\to i}\right)+F\right) \tag{13}$$

Therefore, we can ensure that, at the fixed point, $T=T'$ by just adding an extra step to the BP iterative process in which F is modified at each iteration according to the difference $T-T'$.

After convergence, we compute the entropy $S\left(T'\right)$, and via this define $T^{\star}=\text{argmax}_{T'}S\left(T'\right)$. Then, the marginals computed for the problem defined by $T^{\star}$ are the same as those to the original problem, and are asymptotically correct (within the RS assumption), allowing us to compute all the desired properties on a given instance of the original problem via this modified cavity method. By averaging over many different samples, we can recover the results of the replica method, as shown for the entropy curves in fig. 1.

III. SOLVING SINGLE INSTANCES: LEARNING ALGORITHMS

A. Reinforced Belief Propagation

Belief propagation equations, in their message passing form over single problem instances, have repeatedly proven to provide very effective heuristics when modified in order to produce optimal configurations [8, 11, 12]. Two main ways in which this can be achieved are decimation [13, 14] and reinforcement [7], which can be seen as a "soft decimation" process. In decimation, cycles are performed which alternate message passing and fixing (or "freezing", or "decimating") the most polarized free variables, until all variables are fixed. In reinforcement, the iterative equations have an additional term which has the role of a time-dependent external field, and which is computed from the magnetizations obtained at the preceding time step, so that the system is driven towards

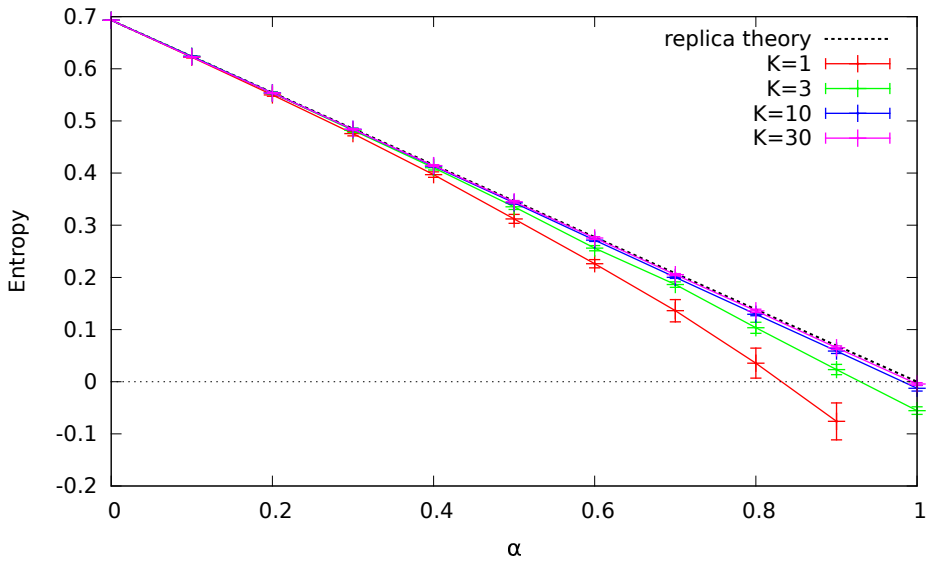

Figure 1: Entropy vs. α as computed by the cavity method at different values of K, compared to the one predicted by the replica theory for $K \gg 1$ and $N \gg K$. Each point shows the average and standard deviation over 10 random samples with $N = 1000$.

a completely polarized state. Following [7] the reinforced BP equations are the same as the normal BP equations with one single difference for eq. 2, which becomes:

$$m_{i\to\mu}^{\tau+1} = \tanh\left(\gamma\left(\tau\right)\tanh^{-1}\left(m_i^{\tau}\right) + \sum_{\nu\neq\mu}\tanh^{-1}\left(n_{\nu\to i}^{\tau+1}\right)\right) \tag{14}$$

where τ is the iteration step, $m_i^{\tau} = \tanh\left(\sum_{\nu}\tanh^{-1}\left(n_{\nu\to i}^{\tau}\right)\right)$ is the total magnetization of variable i at iteration τ, and $\gamma\left(\tau\right)$ is an iteration-dependent reinforcement parameter, which we set as $\gamma\left(\tau\right) = \gamma_0\tau$. Note that the additional term is proportional to the iteration step, and therefore dominates for large τ, ensuring that polarization towards one single configuration is eventually achieved (although in practice computational problems will arise in difficult or unsatisfiable situations, due to the limited precision of the floating point representation).

We implemented both the decimation and the reinforcement schemes, for both the standard version of BP (for which the marginals are approximate due to correlations between different messages) and the corrected version (for which marginals are exact, at the cost of increased computational complexity and running time). Since we did not find the corrected version to provide any significant advantage over the naïve version (which is not particularly surprising, considering that the approximation provided by standard BP is rather good, and that the introduction of the reinforcement term introduces spurious correlations by itself), here we will only present results for the latter case.

The value of γ_0 is a parameter of the solver algorithm; higher values of γ_0 make the algorithm

greedier, in that the messages are polarized more quickly but can get trapped into a non-zero-energy state, while reducing γ_0 improves the accuracy of the algorithm at the cost of requiring more iterations. In practical tests, we found that by using eq. 5 we were able to reach values of α as high as 0.7, but only at the cost of using extremely small values of γ_0 (of the order of 10^{-7} for $N = 1000$ and $K = 10$), and therefore of an impractically high computational time.

However, we found heuristically that, by detecting when an excessively polarized state was reached, and introducing a "depolarization event" triggered by such condition, we could achieve the same results with much higher values of γ_0, and therefore in a much shorter computational time. More in detail, we introduced, at each iteration step, a check to detect cases in which any of the terms in the denominator if eq. 5 goes to 0, indicating a numerical problem due to excessively polarized magnetizations in a state of non-zero energy. Whenever this condition is found, we divide all messages $m^{\tau}_{i\to\mu}$ and total magnetizations m^{τ}_{i} by a positive factor b (thus depolarizing all the messages), and reset $\gamma(\tau)$ to 0. In subsequent iterations, we keep increasing $\gamma(\tau)$ linearly in steps of γ_0 (until another event is detected). We obtain good results by setting the factor b to 2 initially, and increasing it by one at every invocation of this additional depolarization rule. Indeed, since $\gamma(\tau)$ does not increase monotonically any more in this scheme, this modified algorithm will not be guaranteed to polarize towards a single state, unless the state itself has zero energy and therefore represents a solution to the problem.

Fig. 2 shows the performance of this algorithm for $N = 1000$ and $K = 10$. Setting 10000 as the maximum number of iterations, a critical capacity of almost 0.8 is achieved.

B. Simplified BP-inspired scheme

As for the case of the simple perceptron [7–9], it is possible to drastically simplify the reinforced BP equations (in a purely heuristic way), and obtain an online algorithm which, when parameters are set to their optimal values, proves to be almost as effective at learning as reinforced BP itself, while dramatically reducing computational requirements.

This algorithm requires a hidden state h_i to be endowed with each synapse. This hidden state can only assume odd integer values, and is capped by a maximum absolute value $h^{\max}$, so that each synapse has a total of $h^{\max} + 1$ hidden states. The hidden state and the synaptic weight J_i are related by the simple expression $J_i = \mathrm{sign}(h_i)$. In all simulations, we set the initial state of the h_i states by randomly drawing values from $\{-1, 1\}$.

The learning protocol turns out to be as follows: patterns ξ^{μ} are presented in random order

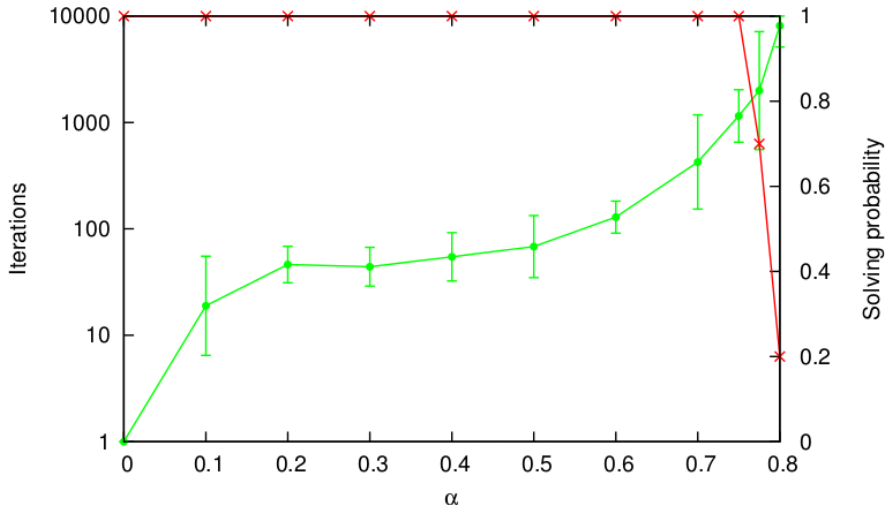

Figure 2: Number of iterations until solving (green curve) and solving probability (red curve) for different values of α, for a tempotron device with $N = 1000$ and $K = 10$, with the reinforced BP algorithm. The parameter γ was set to 0.01. For each point, 10 samples were used. The number of iterations was capped at 10000.

to the device, computing the depolarization $\Delta_t^\mu = \left(\sum_{i=1}^N J_i \xi_{it}^\mu - \theta\right)$; from this, we determine $t^\star = \text{argmax}_t \Delta_t^\mu$ and compute $\Phi^\mu = \sigma_{\text{exp}}^\mu \Delta_{t^\star}^\mu$; depending on the value of Φ^μ, we choose one of three actions:

$\Phi^\mu > 1$: do nothing

$0 < \Phi^\mu \leq 1$: with probability r, update synapses for which $\xi_{it^\star}^\mu = 1$ and $J_i = \sigma_{\text{exp}}^\mu$; with probability $(1 - r)$ do nothing

$\Phi^\mu \leq 0$: update all synapses for which $\xi_{it^\star}^\mu = 1$

The synaptic update rule is always of this form:

$$h_i \to h_i + 2\sigma_{\text{exp}}^\mu$$

which implies that synaptic values J_i are only updated in the $\Phi^\mu < 0$ case, and only if $\xi_{it^\star}^\mu = 1$ and $h_i = -\sigma_{\text{exp}}^\mu$. As stated above, we impose $-h^{\max} \leq h_i \leq h^{\max}$, so that the update rule is not applied when $h_i = h^{\max}\sigma_{\text{exp}}^\mu$.

The probability r of taking an action in the "barely correct" case $0 \leq \Phi^\mu \leq 1$ is a parameter of the algorithm, just as $h^{\max}$. We determined empirically the optimal values of r and $h^{\max}$ for different values of N and K by extensively testing the space of the parameters. Our results, shown in fig. 3, indicate that $r = 0.4$ works best for all values (we explored the values of r in steps of

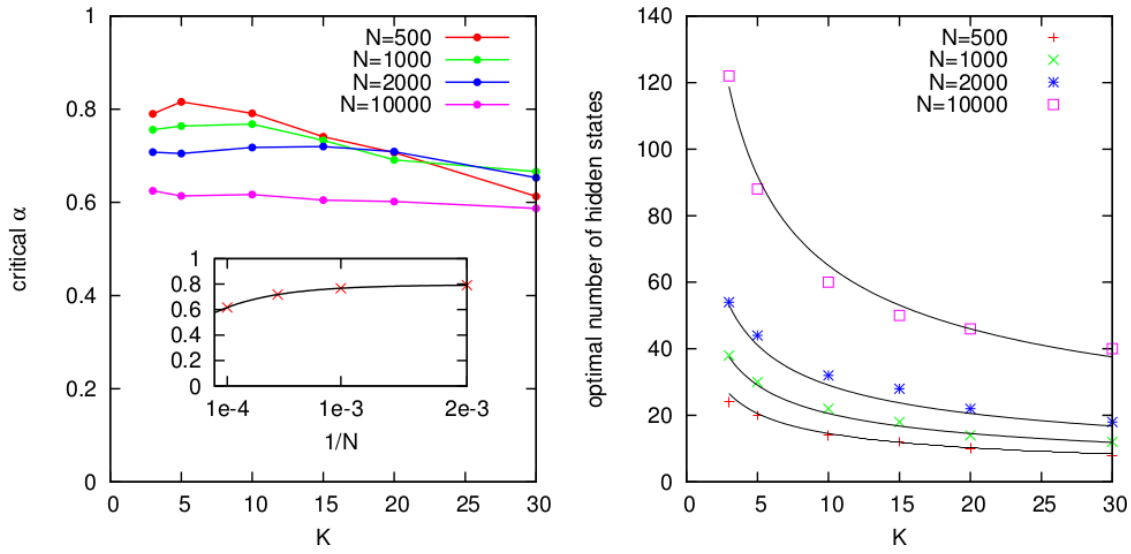

Figure 3: Optimal critical capacity (left) and optimal value of the number of hidden states $h^{\max}+1$ (right) for various values of N and K. Critical capacity is defined as the value of α for which the probability of successfully solving the problem in 10000 iterations or less is 0.5. Optimal $h^{\max}$ is defined as the value of $h^{\max}$ which yields the highest critical capacity. The parameter r is set to 0.4 in all plots shown here, since that was found to be the optimal value independently of other parameters. At least 40 random samples were generated for each combination of $(N, \alpha, K, h^{\max}, r)$ in order to determine the success probability and therefore the critical capacity. α was explored in steps of 0.05. The inset in the left panel shows the critical capacity as a function of $1/N$ for $K = 10$; the solid black line is a fit by an exponential function. The solid black lines in the right panel show the fit of $h^{\max}+1$ as a function of N and K via $\lambda\sqrt{N/K}$.

0.05), and that the optimal value of $h^{\max}$ is reasonably well fitted by a function $\lambda\sqrt{N/K}$, where $\lambda = 2.06 \pm 0.02$. The capacity decreases with N, for fixed K, but does not seem to tend to 0 asymptotically (see inset in fig. 3).

C. Generality of the discrete-time model

As a way to verify that the model and learning protocols which we studied are relevant in a more biologically realistic setting, we adapted the simplified BP-inspired scheme described in the previous section to address the continuous-time classification problem (see Introduction): for a given an instance of the problem, we discretize the time in K bins, apply the BP-inspired learning protocol (slightly modified to use continuous inputs), and test the resulting synaptic weights assignments on the continuous device (see the Appendix for details, section VI). We found that in a device with $N = 1000$ synapses, with time constants $t_m = 10ms$ and $t_s = 2.5ms$, tested on $T = 500ms$ long

input patterns discretized in 50 bins, this scheme can achieve a classification error lower than 1% up to $\alpha = 0.4$, demonstatrating that indeed under these conditions not much relevant information is typically lost in the time discretization process, and that the proposed time-discretized learning protocol can be effective even in a continuous-time setting.

IV. CONCLUSIONS

We have presented a theoretical analysis of the computational performance of the tempotron model with discretized time and discrete synaptic weights. The results show that the device is able to learn random spatio-temporal patterns at a learning rate which saturates the information theoretic bounds.

In addition to this, and possibly of more practical interest, we have been able to derive some novel learning protocols which are local and distributed and do not rely on a gradient descent process on the synaptic weights. These algorithms are based on the message-passing method and extend previous works on rate-coding networks. Specifically, we have shown that the message-passing algorithms can store spatio-temporal patterns at very high loads and that even some extremely simplified versions are still able to store an extensive number of patterns efficiently. Furthermore, we showed that the discretized-time algorithm can even be adapted to effectively address the original, continuous-time version of the problem. Our approach can be applied to both discrete and continuous synapses.

Many open problems remain to be studied, starting from how these protocols can be made even simpler in a biologically plausible modeling context. Still we believe that at least as far as artificial neural systems is concerned these results could find direct application in neuromorphic devices.

V. APPENDIX: STATISTICAL PHYSICS ANALYSIS AND REPLICA CALCULATIONS

A. Entropy

We will consider the case in which synaptic weights take values in $\{-1, 1\}$ first.

The volume of the space of the solutions for a given instantiation of the patterns can be written as:

$$\mathcal{V} = \sum_{\{J_i\}_i} \sum_{\left\{\tau_t^\mu\right\}_{\mu t}} \prod_\mu \chi\left(s_{\text{exp}}^\mu, \{\tau_t^\mu\}_t\right) \prod_{\mu t} \Theta\left(\tau_t^\mu \left(\frac{1}{\sqrt{N}} \sum_i J_i \xi_{it}^\mu - \frac{\theta}{\sqrt{N}}\right)\right) \tag{15}$$

where index $i \in \{1, \ldots, N\}$ is used for synapses, index $t \in \{1, \ldots, K\}$ is used for time bins, index $\mu \in \{1, \ldots, \alpha N\}$ is used for patterns, the auxiliary variables $\tau_t^\mu \in \{-1, 1\}$ are the internal representations (they are equal to $2V_t^\mu - 1$, see section I), and$\chi\left(s, \{\tau_t\}_t\right) = \Theta\left(s - \left(1 - \prod_{t=1}^{K}\left(1 - \frac{1}{2}\left(1 + \tau_t^\mu\right)\right)\right)\right)$ is a characteristic function ensuring that the internal representation τ_t is compatible with the output s.

From here on, for simplicity, we will omit the subscript exp from the outputs s^μ.

In order to compute the entropy, we need to compute the quenched average $\langle \log \mathcal{V} \rangle_{\xi,s}$; we do this by using the replica trick:[7, 8]

$$\langle \log \mathcal{V} \rangle_{\xi,s} = \lim_{n \to 0} \frac{\langle \mathcal{V}^n \rangle_{\xi,s} - 1}{n} \tag{16}$$

where we compute $\langle \mathcal{V}^n \rangle_{\xi,s}$ for integer values of n, and use the analytic continuation to compute the limit $n \to 0$. The average over the replicated volume is:

$$\langle \mathcal{V}^n \rangle_{\xi,s} = \left\langle \sum_{\left\{J_i^a\right\}_{ia}} \sum_{\left\{\tau_t^{\mu a}\right\}_{\mu t a}} \prod_{\mu a} \chi\left(s^\mu, \{\tau_t^{\mu a}\}_t\right) \prod_{\mu t a} \Theta\left(\tau_t^{\mu a}\left(\frac{1}{\sqrt{N}} \sum_i J_i^a \xi_{it}^\mu - \frac{\theta}{\sqrt{N}}\right)\right) \right\rangle_{\xi,s} \tag{17}$$

We used the index $a \in \{1, \ldots, n\}$ to denote the replica. We can now use the integral representation of the Θ function, $\Theta(y) = \int_{-\infty}^{\infty} \frac{dx}{2\pi} \int_0^\infty d\lambda\, e^{ix(\lambda - y)}$, and compute the average over the input patterns, using their independence and the $N \gg 1$ limit (in the following, all integrals are assumed to be on $[-\infty, \infty]$ unless otherwise specified) :

$$\begin{aligned} &\left\langle \prod_a \Theta\left(\tau_t^{\mu a}\left(\frac{1}{\sqrt{N}} \sum_i J_i^a \xi_{it}^\mu - \frac{\theta}{\sqrt{N}}\right)\right) \right\rangle_\xi = \\ &= \int \prod_a \frac{dx_t^{\mu a}}{2\pi} \int_0^\infty \prod_a d\lambda_t^{\mu a} \prod_a \exp\left(i x_t^{\mu a}\left(\lambda_t^{\mu a} - \tau_t^{\mu a}\left(\bar{\xi} \frac{1}{\sqrt{N}} \sum_i J_i^a - \frac{\theta}{\sqrt{N}}\right)\right)\right) \cdot \\ &\quad \cdot \exp\left(-\frac{v_\xi}{2N} \sum_{a,b} \tau_t^{\mu a} \tau_t^{\mu b} x_t^{\mu a} x_t^{\mu b} \sum_i J_i^a J_i^b\right) \end{aligned} \tag{18}$$

where $\bar{\xi} = f$ and $v_\xi = f(1 - f)$ are the average value and the variance of the inputs ξ_{it}^μ, respectively. We then introduce order parameters $q^{ab} = \frac{1}{N} \sum_i J_i^a J_i^b$ and $T^a = -\sqrt{N}\bar{J} + \frac{1}{\sqrt{N}} \sum_i J_i^a$ via Dirac-delta

functions, and their conjugates $\hat{q}^{ab}$,$\hat{T}^a$ via integral expansion of the deltas, and get:

$$\begin{aligned}
\langle \mathcal{V}^n \rangle_{\xi,s} \;=\; & \int \prod_a \frac{dT^a d\hat{T}^a \sqrt{N}}{2\pi} \int \prod_{a\geq b} \frac{dq^{ab} d\hat{q}^{ab} N}{2\pi} \exp\left(\sqrt{N} \sum_a \left(T^a + \sqrt{N}\bar{J} \right) \hat{T}^a - N \sum_{a\geq b} q^{ab} \hat{q}^{ab} \right) \cdot \quad (19) \\
& \cdot \left(\sum_{\left\{ J_i^a \right\}_{ia}} \prod_i \exp\left(\sum_{a\geq b} \hat{q}^{ab} J_i^a J_i^b - \sum_a \hat{T}^a J_i^a \right) \right) \cdot \\
& \cdot \Bigg\langle \sum_{\left\{ \tau_t^{\mu a} \right\}_{\mu t a}} \prod_{\mu a} \chi\left(s^\mu, \{\tau_t^{\mu a}\}_t \right) \cdot \\
& \quad \cdot \prod_{\mu t} \left(\int \prod_a \frac{dx_t^{\mu a}}{2\pi} \int_0^\infty \prod_a d\lambda_t^{\mu a} \prod_a \exp\left(i x_t^{\mu a} \left(\lambda_t^{\mu a} - \tau_t^{\mu a} \left(\bar{\xi} \left(T^a + \sqrt{N}\bar{J} \right) - \frac{\theta}{\sqrt{N}} \right) \right) \right) \cdot \right. \\
& \quad \left. \cdot \exp\left(-\frac{v_\xi}{2} \sum_{a,b} \tau_t^{\mu a} \tau_t^{\mu b} x_t^{\mu a} x_t^{\mu b} q^{ab} \right) \right) \Bigg\rangle_s \\
= \; & \int \prod_a \frac{dT^a d\hat{T}^a \sqrt{N}}{2\pi} \int \prod_{a\geq b} \frac{dq^{ab} d\hat{q}^{ab} N}{2\pi} \exp\left(\sqrt{N} \sum_a \left(T^a + \sqrt{N}\bar{J} \right) \hat{T}^a - N \sum_{a\geq b} q^{ab} \hat{q}^{ab} \right) \cdot \\
& \cdot \left(\sum_{\{J^a\}_a} \exp\left(\sum_{a\geq b} \hat{q}^{ab} J^a J^b - \sum_a \hat{T}^a J^a \right) \right)^N \cdot \\
& \cdot \Bigg\langle \sum_{\left\{ \tau_t^a \right\}_{ta}} \prod_a \chi\left(s, \{\tau_t^a\}_t \right) \cdot \\
& \quad \cdot \prod_t \left(\int \prod_a \frac{dx_t^a}{2\pi} \int_0^\infty \prod_a d\lambda_t^a \prod_a \exp\left(i x_t^a \left(\lambda_t^a - \tau_t^a \left(\bar{\xi} \left(T^a + \sqrt{N}\bar{J} \right) - \frac{\theta}{\sqrt{N}} \right) \right) \right) \cdot \right. \\
& \quad \left. \cdot \exp\left(-\frac{v_\xi}{2} \sum_{a,b} \tau_t^a \tau_t^b x_t^a x_t^b q^{ab} \right) \right) \Bigg\rangle_s^{\alpha N}
\end{aligned}$$

where in the second step we dropped indices i and μ. We expand the threshold θ in series of $\sqrt{N}$:

$$\theta = N\theta_0 + \sqrt{N}\theta_1 \quad (20)$$

from which we immediately get the relation:

$$\bar{J} = \frac{\theta_0}{\bar{\xi}} \quad (21)$$

This leaves us with:

$$\begin{aligned}\langle \mathcal{V}^n \rangle_{\xi,s} \;=\; & \int \prod_a \frac{dT^a d\hat{T}^a \sqrt{N}}{2\pi} \int \prod_{a\geq b} \frac{dq^{ab} d\hat{q}^{ab} N}{2\pi} \cdot \\ & \cdot \exp\left(\sqrt{N} \sum_a T^a \hat{T}^a + N\bar{J} \sum_a \hat{T}^a - N \sum_{a\geq b} q^{ab}\hat{q}^{ab} \right) \cdot \\ & \cdot \left(\sum_{\{J^a\}_a} \exp\left(\sum_{a\geq b} \hat{q}^{ab} J^a J^b - \sum_a \hat{T}^a J^a \right) \right)^N \cdot \\ & \cdot \left\langle \sum_{\{\tau_t^a\}_{ta}} \prod_a \chi\left(s, \{\tau_t^a\}_t\right) \prod_t \left(\int \prod_a \frac{dx_t^a}{2\pi} \int_0^\infty \prod_a d\lambda_t^a \prod_a \exp\left(i x_t^a \left(\lambda_t^a - \tau_t^a \left(\bar{\xi} T^a - \theta_1 \right) \right) \right) \right) \cdot \right. \\ & \left. \cdot \exp\left(-\frac{v_\xi}{2} \sum_{a,b} \tau_t^a \tau_t^b x_t^a x_t^b q^{ab} \right) \right) \Bigg\rangle_s^{\alpha N} \end{aligned} \tag{22}$$

In the $N \gg 1$ limit, this integral can be computed by the saddle point method: we introduce the RS Ansatz for the solution: $T^a = T\,\forall a$, $q^{ab} = q\,\forall a, b : a \neq b$, $q^{aa} = Q\,\forall a$, and analogous expressions

for the conjugate parameters. Therefore:

$$
\begin{aligned}
\langle \mathcal{V}^n \rangle_{\xi,s} \;=\; & \exp\left(N\bar{J}\hat{T} + N\frac{n}{2}\hat{q}q - Nn\hat{Q}Q\right)\cdot \\
& \cdot\left(\sum_{\{J^a\}_a}\exp\left(\frac{\hat{q}}{2}\left(\sum_a J^a\right)^2 - \frac{1}{2}\left(\hat{q}-2\hat{Q}\right)\sum_a (J^a)^2 - \hat{T}\sum_a J^a\right)\right)^N \cdot \\
& \cdot\left\langle \sum_{\{\tau_t^a\}_{ta}}\prod_a \chi\left(s,\{\tau_t^a\}_t\right)\prod_t\left(\int\prod_a\frac{dx_t^a}{2\pi}\int_0^\infty\prod_a d\lambda_t^a\prod_a\exp\left(ix_t^a\left(\lambda_t^a-\tau_t^a\left(\bar{\xi}T-\theta_1\right)\right)\right)\cdot\right.\right. \\
& \left.\left.\cdot\exp\left(-\frac{v_\xi}{2}\left(q\left(\sum_a\tau_t^a x_t^a\right)^2+(Q-q)\sum_a(x_t^a)^2\right)\right)\right)\right\rangle_s^{\alpha N} \\
=\; & \exp\left(N\bar{J}\hat{T} + N\frac{n}{2}\hat{q}q - Nn\hat{Q}Q\right)\cdot \\
& \cdot\left(\int Du\left(\sum_{\{J\}}\exp\left(-\frac{1}{2}\left(\hat{q}-2\hat{Q}\right)J^2+\left(\sqrt{\hat{q}}u-\hat{T}\right)J\right)\right)^n\right)^N\cdot \\
& \cdot\left(\int\prod_t Du_t\left\langle\left(\sum_{\{\tau_t\}_t}\chi\left(s,\{\tau_t\}_t\right)\prod_t\left(\int\frac{dx_t}{2\pi}\int_0^\infty d\lambda_t\,\exp\left(-\frac{v_\xi}{2}(Q-q)(x_t)^2\right)\cdot\right.\right.\right. \\
& \left.\left.\left.\cdot\exp\left(ix_t\left(\lambda_t-\tau_t\left(\bar{\xi}T-\theta_1-\sqrt{v_\xi q}u_t\right)\right)\right)\right)\right)^n\right\rangle_s\Bigg)^{\alpha N} \\
=\; & \exp\, Nn\left(\bar{J}\hat{T}+\frac{1}{2}\hat{q}q-\hat{Q}Q+\right. \\
& \quad+\int Du\,\log\left(\sum_{\{J\}}\exp\left(-\frac{1}{2}\left(\hat{q}-2\hat{Q}\right)J^2+\left(\sqrt{\hat{q}}u-\hat{T}\right)J\right)\right) \\
& \quad\left.+\,\alpha\int\prod_t Du_t\left\langle\log\left(\sum_{\{\tau_t\}_t}\chi\left(s,\{\tau_t\}_t\right)\prod_t H\left(-\tau_t\frac{\bar{\xi}T-\theta_1-\sqrt{v_\xi q}u_t}{\sqrt{v_\xi(Q-q)}}\right)\right)\right\rangle_s\right)
\end{aligned}
\tag{23}
$$

where in the second step we introduced auxiliary Gaussian integrals (we use the shorthand notation $Du = du\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$ and define $H(x) = \int_x^\infty Dy$), which allows to drop the replica index a, and in the last step we used the $n \to 0$ limit. Finally, we obtain the expression for the entropy:

$$
\mathcal{S} = \frac{1}{N}\langle \log \mathcal{V}\rangle_{\xi,s} \;=\; \bar{J}\hat{T}+\frac{1}{2}\hat{q}q-\hat{Q}Q+\mathcal{Z}_J\left(\hat{Q},\hat{q},\hat{T}\right)+\mathcal{Z}_S\left(Q,q,T\right) \tag{24}
$$

$$
\mathcal{Z}_J\left(\hat{Q},\hat{q},\hat{T}\right) \;=\; \int Du\,\log\left(\sum_{\{J\}}\exp\left(-\frac{1}{2}\left(\hat{q}-2\hat{Q}\right)J^2+\left(\sqrt{\hat{q}}u-\hat{T}\right)J\right)\right) \tag{25}
$$

$$
\mathcal{Z}_S\left(Q,q,T\right) \;=\; \alpha\int\prod_t Du_t\left\langle\log\left(\sum_{\{\tau_t\}_t}\chi\left(s,\{\tau_t\}_t\right)\prod_t H\left(-\tau_t\,\eta\left(u_t,Q,q,T\right)\right)\right)\right\rangle \tag{26}
$$

$$
\eta\left(u,Q,q,T\right) \;=\; \frac{\bar{\xi}T-\theta_1-\sqrt{v_\xi q}u}{\sqrt{v_\xi(Q-q)}} \tag{27}
$$

The expression for $\mathcal{Z}_J$ is the familiar expression for perceptron models, and it can be written more

explicitly for the two cases $J \in \{-1, 1\}$ and $J \in \{0, 1\}$:

$$\mathcal{Z}_J^{\pm}\left(\hat{Q}, \hat{q}, \hat{T}\right) = -\frac{1}{2}\left(\hat{q} - 2\hat{Q}\right) + \int Du \, \log\left(2\cosh\left(\sqrt{\hat{q}}u - \hat{T}\right)\right) \tag{28}$$

$$\mathcal{Z}_J^{01}\left(\hat{Q}, \hat{q}, \hat{T}\right) = \int Du \, \log\left(1 + \exp\left(-\frac{1}{2}\left(\hat{q} - 2\hat{Q}\right) + \sqrt{\hat{q}}u - \hat{T}\right)\right) \tag{29}$$

The expression for $\mathcal{Z}_S$ can be manipulated further:

$$\mathcal{Z}_S\left(Q, q, T\right) = \alpha\left(1 - f'\right)K \int Du \log H\left(\eta\left(u, Q, q, T\right)\right) + \tag{30}$$

$$+\alpha f' \int \prod_t Du_t \log\left(1 - \prod_t H\left(\eta\left(u_t, Q, q, T\right)\right)\right) \tag{31}$$

In the limit of $K \gg 1$, we can use the central limit theorem and keep only the higher order terms in K, and obtain:

$$\mathcal{Z}_S\left(Q, q, T\right) = \alpha\left(\left(1 - f'\right)K\Lambda\left(Q, q, T\right) + f' \log\left(1 - \exp\left(K\Lambda\left(Q, q, T\right)\right)\right)\right) \tag{32}$$

where we defined:

$$\Lambda\left(Q, q, T\right) = \int Du \log H\left(\eta\left(u, Q, q, T\right)\right) \tag{33}$$

The saddle point equation for T gives:

$$0 = \frac{\partial \mathcal{Z}_S}{\partial T} = \alpha K\left(1 - f'\frac{1}{1 - e^{K\Lambda}}\right)\frac{\partial \Lambda\left(Q, q, T\right)}{\partial T}$$

which implies:

$$\Lambda\left(Q, q, T\right) = \frac{1}{K}\log\left(1 - f'\right) \tag{34}$$

This in turn puts to zero $\hat{q}$ and $\hat{Q}$:

$$\hat{q} = -2\frac{\partial \mathcal{Z}_S}{\partial q} = -2\alpha K\left(1 - f'\frac{1}{1 - e^{K\Lambda}}\right)\frac{\partial \Lambda\left(Q, q, T\right)}{\partial q} = 0$$

$$\hat{Q} = \frac{\partial \mathcal{Z}_S}{\partial Q} = \alpha K\left(1 - f'\frac{1}{1 - e^{K\Lambda}}\right)\frac{\partial \Lambda\left(Q, q, T\right)}{\partial Q} = 0$$

Optimizing with respect to θ_0, i.e. imposing $\frac{\partial S}{\partial \bar{J}} = 0$, we also get $\hat{T} = 0$.

The remaining equations are different for the cases ± 1 and 01. For the ± 1 case:

$$q = -2\frac{\partial \mathcal{Z}_J}{\partial \hat{q}} = 1 - \frac{1}{\sqrt{\hat{q}}}\int Du \, u \, \tanh\left(\sqrt{\hat{q}}u - \hat{T}\right) \tag{35}$$

$$Q = \frac{\partial \mathcal{Z}_J}{\partial \hat{Q}} = 1 \tag{36}$$

$$\bar{J} = -\frac{\partial \mathcal{Z}_J}{\partial \hat{T}} = \int Du \, \tanh\left(\sqrt{\hat{q}}u - \hat{T}\right) \tag{37}$$

The result $Q = 1$ is obvious. From $\hat{T} = 0$ and $\hat{q} = 0$, and since $\bar{\xi} \neq 0$, we get $q = 0$ and $\theta_0 = 0$.

For the 01 case:[7, 8]

$$q = -2\frac{\partial \mathcal{Z}_J}{\partial \hat{q}} = \int Du \frac{1}{1 + \exp\left(\frac{1}{2}\left(\hat{q} - 2\hat{Q}\right) - \sqrt{\hat{q}}u + \hat{T}\right)} \left(1 - \frac{u}{\sqrt{\hat{q}}}\right) \tag{38}$$

$$Q = \frac{\partial \mathcal{Z}_J}{\partial \hat{Q}} = \int Du \frac{1}{1 + \exp\left(\frac{1}{2}\left(\hat{q} - 2\hat{Q}\right) - \sqrt{\hat{q}}u + \hat{T}\right)} \tag{39}$$

$$\bar{J} = -\frac{\partial \mathcal{Z}_J}{\partial \hat{T}} = Q \tag{40}$$

From $\hat{q} = 0$ and $\hat{Q} = 0$ these simplify to:

$$\begin{aligned} Q &= \bar{J} = \frac{1}{1 + e^{\hat{T}}} = \frac{1}{2} \\ q &= \frac{1}{\left(1 + e^{\hat{T}}\right)^2} = Q^2 = \frac{1}{4} \\ \theta_0 &= \frac{f}{2} \end{aligned}$$

From $q = Q^2$ we see that the cross-overlap is as low as possible, like in the ± 1 case: the physical interpretation is that clusters of solution are isolated, i.e. point-like.

The only remaining order parameters are T and θ_1, which are related by eq. 34 and give:

$$T = \frac{1}{f}\left(\theta_1 + \sqrt{2f\left(1 - f\right)}\mathrm{erfc}^{-1}\left(2\sqrt[K]{1 - f'}\right)\right) \tag{41}$$

Therefore, in order to have unbiased synapses, i.e. $T = 0$, we may set θ_1 to:

$$\theta_1 = -\sqrt{2f\left(1 - f\right)}\mathrm{erfc}^{-1}\left(2\sqrt[K]{1 - f'}\right) \tag{42}$$

With our choice for the distribution of the inputs, $f = 1 - \sqrt[K]{1 - f}$, this formula starts from 0 at $K = 1$, has a maximum for $K = 6$ and slowly decreases (as $\sqrt{\frac{\log(K)}{K}}$) to 0 as K diverges; the reason for this behaviour is that there are two competing tendencies at work as K increases: on one hand, the increase in the length of the internal representation while f' is kept constant requires that more and more individual bins fall below threshold; on the other hand, the sparsification of the inputs reduces the fluctuations in the depolarization; this second contribution dominates for large K and so the threshold goes to 0, but for practical purposes (i.e. for biologically relevant values of K) it does not become negligible.

From the above results, we can determine the entropy:

$$\mathcal{S} = \log\left(2\right) - \alpha\left(\left(1 - f'\right)\log\left(1 - f'\right) + f'\log\left(f'\right)\right) \tag{43}$$

which goes to 0 at:

$$\alpha_c = \left(1-f'\right)\log_2\left(1-f'\right) + f'\log_2\left(f'\right) \tag{44}$$

which coincides with the information theoretic upper bound.

B. Distribution of output spikes

The probability distribution of the number of output spikes (i.e. 1's in the internal representation) can be obtained by taking the ratio between the volume of the solution space in which one pattern is restricted to produce Y spikes and the total volume:

$$\begin{aligned} P\left(Y\right) \;=\; & \frac{1}{\mathcal{V}}\left\langle \sum_{\left\{\tau_t^\mu\right\}_{\mu l}} \delta_k\left(\sum_t\left(\frac{1+\tau_t^1}{2}\right), Y\right)\cdot \right. \\ & \left.\cdot \sum_{\{J_i\}_i}\prod_\mu \chi\left(s^\mu,\{\tau_t^\mu\}_t\right)\prod_{\mu t}\Theta\left(\tau_t^\mu\left(\frac{1}{\sqrt{N}}\sum_i J_i\xi_{it}^\mu - \frac{\theta}{\sqrt{N}}\right)\right)\right\rangle_{\xi,s} \end{aligned} \tag{45}$$

where $\delta_k\left(x,y\right)$ is the Kronecker delta function:

$$\delta_k\left(x,y\right) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

We can write $\mathcal{V}^{-1} = \lim_{n\to 0}\mathcal{V}^{n-1}$, restrict ourselves to integer n and obtain an expression almost identical to eq. 17, except for the Kronecker delta affecting pattern $\mu = 1$:

$$\begin{aligned} P\left(Y\right) \;=\; & \left\langle \sum_{\left\{\tau_t^{\mu a}\right\}_{\mu t a}} \delta_k\left(\sum_t\left(\frac{1+\tau_t^{1a}}{2}\right) - Y\right)\cdot \right. \\ & \left.\cdot \sum_{\left\{J_i^a\right\}_{ia}}\prod_{\mu a} \chi\left(s^\mu,\{\tau_t^{\mu a}\}_t\right)\prod_{\mu t a}\Theta\left(\tau_t^{\mu a}\left(\frac{1}{\sqrt{N}}\sum_i J_i^a\xi_{it}^\mu - \frac{\theta}{\sqrt{N}}\right)\right)\right\rangle_{\xi,s} \end{aligned} \tag{46}$$

The computation follows the one for the entropy; the only affected term is $\mathcal{Z}_S$, and only one term survives the limit $n\to 0$, giving:

$$P\left(Y\right) = \delta_k\left(Y,0\right)\left(1-f'\right) + \left(1-\delta_k\left(Y,0\right)\right)f'\int\prod_t Du_t \frac{\binom{K}{Y}\prod_{t=1}^{Y}H^+\left(u_t\right)\prod_{t=Y+1}^{K}H^-\left(u_t\right)}{1-\prod_{t=1}^{K}H^-\left(u_t\right)} \tag{47}$$

where we wrote $H^\pm\left(u\right) = H\left(\mp\eta\left(u,Q,q,t\right)\right)$ (see eq. 27) for short. For large K and finite Y, this approximates to:

$$\begin{aligned} P\left(Y\right) \;&=\; \delta_k\left(Y,0\right)\left(1-f'\right) + \left(1-\delta_k\left(Y,0\right)\right)f'\binom{K}{Y}\frac{e^{(K-Y)\Lambda}}{1-e^{K\Lambda}}\left(1-e^\Lambda\right)^Y \\ &=\; \binom{K}{Y}\left(e^\Lambda\right)^{K-Y}\left(1-e^\Lambda\right)^Y \end{aligned} \tag{48}$$

where $\Lambda = \Lambda(Q, q, T)$ (see eq. 33), and in the second step we used eq. 34 from the saddle point solution. The result is a binomial distribution, in which the probability of producing a spike is $1 - e^{\Lambda} = 1 - \sqrt[K]{1 - f'}$, which is our choice for the input frequency f.

C. Structure of the internal representations

We can study the structure of the space of the internal representations by following [15, 16]: we consider the volume of each internal representation $\mathcal{V}_{\mathcal{T}}$, where $\mathcal{T} = \{\tau_t^{\mu}\}_{\mu t}$ is an internal representation, such that the overall volume can be written as $\mathcal{V} = \sum_{\mathcal{T}} \mathcal{V}_{\mathcal{T}}$; then we define:

$$\mathcal{V}(r) = \sum_{\mathcal{T}} (\mathcal{V}_{\mathcal{T}})^r \tag{49}$$

and study the free energy defined by:

$$g(r) = -\frac{\langle \log(\mathcal{V}(r)) \rangle}{Nr} \tag{50}$$

which, once known, allows to derive the size of the internal representations from the quantity

$$w(r) = \frac{\partial}{\partial r}(-rg(r)) \tag{51}$$

(for $r = 1$, $w(1) = \frac{1}{N} \log \mathcal{V}_{\mathcal{T}}^{\star}$ where $\mathcal{V}_{\mathcal{T}}^{\star}$ is the typical volume of the dominant internal representations), and their number from the micro-canonical entropy

$$\mathcal{N}(r) = -\frac{\partial}{\partial 1/r} g(r) \tag{52}$$

(for $r = 1$, $\mathcal{N}(1)$ is the logarithm of the typical number of internal representation of size $\mathcal{V}_{\mathcal{T}}^{\star}$, divided by N).

The computation is performed by using the replica trick for r integer and then performing an analytic continuation:

$$\langle \mathcal{V}(r)^n \rangle = \left\langle \sum_{\{\tau_t^{\mu a}\}_{\mu t a}} \sum_{\{J_i^{a\nu}\}_{ia\nu}} \prod_{\mu a} \chi\left(s^{\mu}, \{\tau_t^{\mu a}\}_t\right) \prod_{\mu t a \nu} \Theta\left(\tau_t^{\mu a}\left(\frac{1}{\sqrt{N}} \sum_i J_i^{a\nu} \xi_{it}^{\mu} - \frac{\theta}{\sqrt{N}}\right)\right) \right\rangle_{\xi, s} \tag{53}$$

where we introduced the new internal representation replica index $\nu \in \{1, \ldots, r\}$. The computation follows the steps of the entropy computation of section V A, but requires the introduction of order parameters with 2 replica indices; in particular $q^{a\nu, b\phi} = \frac{1}{N} \sum_i J_i^{a\nu} J_i^{b\phi}$, which in the RS Anzatz can

take 3 values:

$$q^{a\nu,b\phi} = \begin{cases} Q & \text{if } a = b, \nu = \phi \\ q_1 & \text{if } a = b, \nu \neq \phi \\ q_0 & \text{if } a \neq b \end{cases}$$

We obtain, for large K:

$$\begin{aligned}
g(r) &= -\bar{J}\hat{T} - \frac{1}{2} r q_0 \hat{q}_0 + \frac{r-1}{2} q_1 \hat{q}_1 + \hat{Q} Q - \frac{1}{r}\mathcal{Z}_J + \frac{1}{r}\mathcal{Z}_S && (54)\\
\mathcal{Z}_J &= \int Du \, \log\left(\int Dz \left(\sum_{\{J\}} e^{\frac{1}{2}\left(2\hat{Q}-\hat{q}_1\right)J^2 + \left(\sqrt{\hat{q}_0}u + \sqrt{\hat{q}_1 - \hat{q}_0}z - \hat{T}\right)J}\right)^r\right) && (55)\\
\mathcal{Z}_S &= \alpha\left(\left(1-f'\right) K \Lambda\left(Q, q_1, q_0, T\right) + f' \log\left(e^{K\Phi(Q,q_1,q_0,T)} - e^{K\Lambda(Q,q_1,q_0,T)}\right)\right) && (56)\\
\Lambda\left(Q, q_1, q_0, T\right) &= \int Du \, \log\left(\int Dz \, H\left(\eta\left(u, z, Q, q_1, q_0, T\right)\right)^r\right) && (57)\\
\Phi\left(Q, q_1, q_0, T\right) &= \int Du \, \log\left(\int Dz \, H\left(-\eta\left(u, z, Q, q_1, q_0, T\right)\right)^r + H\left(\eta\left(u, z, Q, q_1, q_0, T\right)\right)^r\right) && (58)\\
\eta\left(u, z, Q, q_1, q_0, T\right) &= -\frac{u\sqrt{v_\xi q_0} - z\sqrt{v_\xi\left(q_1 - q_0\right)} + \theta_1 - T\xi}{\sqrt{v_\xi\left(Q - q_1\right)}} && (59)
\end{aligned}$$

The saddle point equations for $r = 1$ give the same results as before, as expected; in particular, we find $q_0 = q_1 = 0$ in the ± 1 case and $q_0 = q_1 = 1/4$ in the 01 case, and $g(1) = -\mathcal{S}$ as expected. Furthermore, we have:

$$\left.\frac{\partial \mathcal{Z}_{\mathcal{S}}}{\partial r}\right|_{r=1} = \alpha K \int Du \int Dz \left(H^+(u,z) \log H^+(u,z) + H^-(u,z) \log H^-(u,z)\right) \tag{60}$$

where we used the shorthand notation $H^{\pm}(u,z) = H\left(\mp\eta\left(u, z, Q, q_1, q_0, T\right)\right)$. From this and from the saddle point equations at $r = 1$, in particular from eq. 34, we obtain the weight and the entropy of the dominant internal representations for $f' = 1/2$:

$$\begin{aligned}
w(1) &= \left.\frac{\partial}{\partial r}\left(-r g(r)\right)\right|_{r=1} = \log 2\left(-1 + \alpha + \alpha \log K - \alpha \log\log 2\right) && (61)\\
\mathcal{N}(1) &= -\left.\frac{\partial}{\partial 1/r} g(r)\right|_{r=1} = -\alpha\left(\log\log 2 - \log K\right)\log 2 && (62)
\end{aligned}$$

From these, we can find the leading terms of the number of different dominant internal representations:

$$e^{N\mathcal{N}(1)} = \left(\frac{K}{\log 2}\right)^{N\alpha \log 2} \tag{63}$$

and their volume:

$$e^{-Nw(1)} = 2^{N(1-\alpha)} \left(\frac{\log 2}{K}\right)^{N\alpha \log 2} \tag{64}$$

VI. APPENDIX: TIME DISCRETIZATION

A. Modified BP-inspired learning scheme for continuous inputs

The learning protocol presented in section III B can be easily generalized to the case in which the input patterns ξ_{it}^{μ} are not binary, but positive and continuous: the only required change is that the update rules, rather then being applied only to those synapses for which $\xi_{it^\star}^{\mu} = 1$, are applied to all synapses with probability $p_i^{\mu} = \min\left(\xi_{it^\star}^{\mu}, 1\right)$. Therefore, the actions taken upon determining $t^\star$ and the value Φ^{μ} are:

$\Phi^{\mu} > 1$: do nothing

$0 < \Phi^{\mu} \leq 1$: with probability r, update synapses for which $J_i = \sigma_{\text{exp}}^{\mu}$, each with probability p_i^{μ}; with probability $(1 - r)$ do nothing

$\Phi^{\mu} \leq 0$: update all synapses, each with probability p_i^{μ}

In order for this generalization to be effective without furher modifications of the algorithm, it is crucial that a normalization step is applied to the inputs (see next section).

As an additional generalization, we also introduce a robustness parameter ρ and re-define $\Phi^{\mu} = \sigma_{\text{exp}}^{\mu}\Delta_{t^\star}^{\mu} - \rho\theta$, where θ is the firing threshold: this forces the learning algorithm to seek solutions in which the depolarization is far from the threshold. In the numerical experiments described in section III C we used the value $\rho = 0.2$, increasing it from 0 in steps of 0.01 for 1000 iterations at each step; the other parameters of the model used in those tests were $N = 1000$, $K = 50$, $h_{\max} = 25$ and $r = 0.3$.

B. Pattern time-discretization

In this section we describe the time-discretization process mentioned in section III C: we consider a continuous-time model as described in the Introduction; then, for any given input spike train, we compute the post-synaptic-potential trace $R_i^{\mu}(t) = \sum_{t_i^{\mu} < t} v\left(t - t_i^{\mu}\right)$, where $v(t)$ is the temporal kernel of the membrane. We divide the time window T in K equal bins, and for each bin k we compute the input ξ_{ik}^{μ} as the fraction of the membrane kernel in that bin:

$$\xi_{ik}^{\mu} = \frac{1}{t_m - t_s} \int_{k-\text{th bin}} dt \; R_i^{\mu}(t) \tag{65}$$

where we used the fact that $\int_0^{\infty} dt \; v(t) = t_m - t_s$ with out choice of v.

Note that the resulting ξ_{ik}^{μ} can be greater than 1, but this is rare under the sparsity regime which we considered.

The time-discretized patterns can then be passed to the discrete algorithm for deriving a vector of synaptic weights, which in turn can be tested on the original model. In our numerical experiments, we generated input spike trains by a Poisson process with a rate chosen as to obtain the correct value of the input frequency f (see section I) after the discretization in K time bins. When testing the solution, we used the value of the firing threshold for the continuous unit which gave the lowest number of errors.

---

[1] Gütig R and Sompolinsky H 2006 *Nature Neuroscience* **9** 420–428 ISSN 1097-6256 URL `http://www.nature.com/neuro/journal/v9/n3/abs/nn1643.html`

[2] Johansson R S and Birznieks I 2004 *Nature Neuroscience* **7** 170–177 ISSN 1097-6256 URL `http://www.nature.com/neuro/journal/v7/n2/full/nn1177.html`

[3] deCharms R C and Merzenich M M 1996 *Nature* **381** 610–613 ISSN 0028-0836

[4] Meister M, Lagnado L and Baylor D A 1995 *Science* **270** 1207–1210 ISSN 0036-8075

[5] Wehr M and Laurent G 1996 *Nature* **384** 162–166 ISSN 0028-0836

[6] Rubin R, Monasson R and Sompolinsky H 2010 *Physical Review Letters* **105** 218102 URL `http://link.aps.org/doi/10.1103/PhysRevLett.105.218102`

[7] Braunstein A and Zecchina R 2006 *Physical Review Letters* **96** 030201 URL `http://prl.aps.org/abstract/PRL/v96/i3/e030201`

[8] Baldassi C, Braunstein A, Brunel N and Zecchina R 2007 *Proceedings of the National Academy of Sciences* **104** 11079–11084 ISSN 0027-8424, 1091-6490 PMID: 17581884 URL `http://www.pnas.org/content/104/26/11079`

[9] Baldassi C 2009 *Journal of Statistical Physics* **136** ISSN 0022-4715 (Print) 1572-9613 (Online) URL `http://www.springerlink.com/content/r077721167526045/`

[10] Mézard, M 1989 *Journal of Physics A: Mathematical and General* **22** 2181–2190 ISSN 0305-4470, 1361-6447

[11] Braunstein A, Kayhan F, Montorsi G and Zecchina R 2007 Encoding for the blackwell channel with reinforced belief propagation *IEEE International Symposium on Information Theory (ISIT07)* pp 1891–1895

[12] Bailly-Bechet M, Borgs C, Braunstein A, Chayes J, Dagkessamanskaia A, François J and Zecchina R 2010 *Proceedings of the National Academy of Sciences* **108** 882–887 ISSN 0027-8424 URL `http://www.pnas.org/content/108/2/882.short`

[13] Mézard M, Parisi G and Zecchina R 2002 *Science* **297** 812 – 815 ISSN 10959203

[14] Braunstein A, Mézard M and Zecchina R 2005 *Random Structures and Algorithms* **27** 201–226

[15] Monasson R and Zecchina R 1996 *Physical Review Letters* **76** 2205–2205 URL `http://link.aps.org/doi/10.1103/PhysRevLett.76.2205.3`

[16] Cocco S, Monasson R and Zecchina R 1996 *Physical Review E* **54** 717–736 URL `http://link.aps.org/doi/10.1103/PhysRevE.54.717`