Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks

Aldo Battista, Remi Monasson

Capacity-resolution trade-off in the optimal learning of multiple low-dimensional manifolds by attractor neural networks

Aldo Battista and Rémi Monasson
Laboratory of Physics of the Ecole Normale Supérieure, CNRS UMR 8023 & PSL Research, Paris, France
(Dated: October 11, 2019)

Recurrent neural networks (RNN) are powerful tools to explain how attractors may emerge from noisy, high-dimensional dynamics. We study here how to learn the $\sim N^2$ pairwise interactions in a RNN with N neurons to embed L manifolds of dimension $D \ll N$. We show that the capacity, *i.e.* the maximal ratio L/N, decreases as $|\log \epsilon|^{-D}$, where ϵ is the error on the position encoded by the neural activity along each manifold. Hence, RNN are flexible memory devices capable of storing a large number of manifolds at high spatial resolution. Our results rely on a combination of analytical tools from statistical mechanics and random matrix theory, extending Gardner's classical theory of learning to the case of patterns with strong spatial correlations.

How sensory information is encoded and processed by neuronal circuits is a central question in computational neuroscience. In many brain areas, the activity of neurons, σ, is found to depend strongly on some continuous sensory correlate $\mathbf{r}$; examples include simple cells in the V1 area of the visual cortex coding for the orientation of a bar presented to the retina, and head direction cells in the subiculum or place cells in the hippocampus, whose activities depend, respectively, on the orientation of the head and the position of an animal in the physical space. Over the past decades, Continuous Attractor (CA) neural networks have emerged as an appealing concept to explain such findings, more precisely, how a large and noisy neural population can reliably encode 'positions' in low-dimensional sensory manifolds, $\sigma = \Phi(\mathbf{r})$, and continuously update their values over time according to input stimuli [1–5].

Models for the embedding of a CA in Recurrent Neural Network (RNN) generally assume that, after a Hebbian-like learning phase, the connection W_{ij} between the neurons i, j having their place fields centered in positions $\mathbf{r}_i$ and $\mathbf{r}_j$, takes value

$$W_{ij} = w\big(|\mathbf{r}_i - \mathbf{r}_j|\big) \ , \tag{1}$$

where $|\cdot|$ denotes the distance in the sensory space. If w is sufficiently excitatory at short distances and inhibitory at long ones, a bump state spontaneously emerges, in which active neurons tend to code for nearby positions in the sensory space. Weak external inputs suffice to move the bump and span the D-dimensional manifold of all possible positions $\mathbf{r}$ (Fig. 1(a)). This mechanism was observed in the ellipsoid body of the fly, where a bump of activity points towards the heading direction [6]. Indirect evidences for the presence of CA have been reported, *e.g.* in the grid-cell system [7] and in the prefontal cortex [8].

Hebbian connections (1) can be modified to embed in the same network of N neurons multiple, unrelated CAs (Fig. 1(a)), such as multiple hippocampal spatial maps corresponding to different environments [9] or contextual situations [10]. Assuming each one of the L maps contributes equally to the learning process, connections take the form [11]

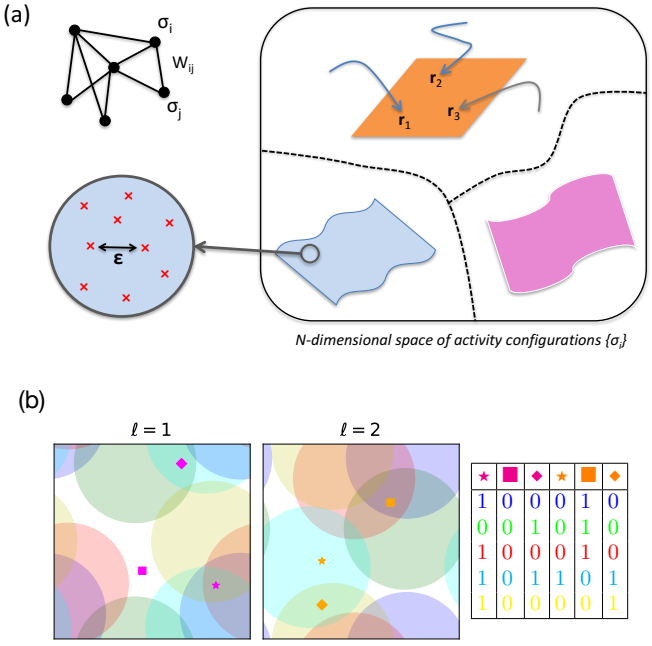

FIG. 1: (a) A recurrent network with N neurons and connectivity matrix W (top left) generates high-dimensional activity configurations attracted to multiple low-dimensional manifolds (right); on each manifold, we require to memorize p points (bottom left, red crosses), whose separation defines the spatial resolution ϵ. (b) Place fields of $N = 5$ neurons in two maps, *e.g.* pink and orange in panel (a), with periodic boundary conditions; the table lists, for each map, $p = 3$ activity patterns corresponding to the marked points.

$$W_{ij} = \sum_{\ell=1}^{L} w\big(|\mathbf{r}_i^\ell - \mathbf{r}_j^\ell|\big) \ , \tag{2}$$

where $\mathbf{r}_i^\ell$ is the center of the place field (PF) of neuron i in environment ℓ (Fig. 1(b)). Theoretical calculations show that a bump state can exist (in any map) as long

as $L < \alpha_c N$, where α_c defines the critical capacity that can be sustained by the network [12, 13].

There are, however, serious practical and conceptual issues with the current theoretical understanding of multiple CAs based on (2). First, as soon as $L \geq 2$, the activity bump gets stuck in some preferred locations in the retrieved map due to the interferences coming from the other $L-1$ non-retrieved maps [14]. In other words, rule (2) does not define truly CAs, as large barriers oppose the motion of the bump along the map [15]. The spatial error ϵ with which the environment is encoded, defined as the average discrepancy between any initial position $\mathbf{r}$ for the bump and the closest stable position in which it finally settles after neural relaxation dynamics, becomes quite large as L increases (Fig. 2(a)). The issue of spatial resolution is also unclear from a theoretical point of view. Capacity calculations [12, 13] require that a bump can form in any of the L maps, in at least one position: they offer no guarantee about the existence of other memorized positions, and, more generally, about the value of ϵ.

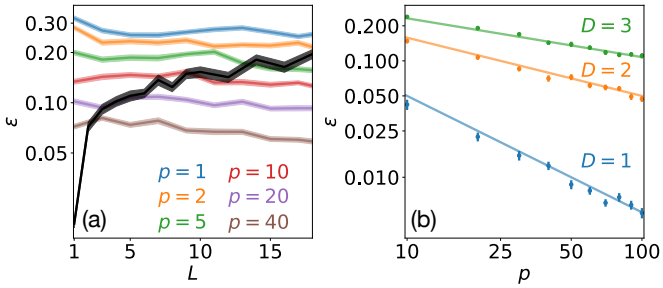

FIG. 2: (a) Spatial error ϵ vs. number L of two-dimensional maps in a network of $N = 1000$ neurons. Black: rule (2), with $w(d) = e^{-d/.01} + w_0$, where $w_0 < 0$ enforces a fraction $\phi_0 = .3$ of active cells. Colors: SVM results for different numbers p of prescribed positions. Line widths show the error bars, see SM Sec. I.E for details about the calculation of ϵ. (b) Spatial error ϵ vs. number p of positions in a network of $N = 1000$ neurons storing $L = 5$ maps, in dimensions $D = 1, 2, 3$. Lines show the expected scalings $\epsilon \sim p^{-1/D}$ in log-log scale.

Secondly, the values of the critical capacity α_c with rule (2) are generally quite low. It is reasonable to expect that the optimal storage capacity could be much higher: a ~ 15-fold increase was found from the Hebb-rule critical capacity, $\simeq 0.14$ [17], to the optimal capacity, $\alpha_c = 2$ [18] in the case of 0-dimensional attractors, corresponding to the Hopfield model [19]. Optimal learning could also provide detailed insights on the statistical structure of the neural couplings W_{ij}, which could be compared to the physiological distribution of synaptic connections [20].

In this Letter, we present a theory of optimal storage of multiple quasi-continuous maps with prescribed spatial resolution in a RNN with N binary neurons ($\sigma_i = 0, 1$) and real-valued, oriented connections W_{ij}. A map in this context is defined through the set of the input (place) fields of the N neurons, each covering a volume fraction ϕ_0 of the D-dimensional cube (Fig. 1(b)). In practice, the centers of the fields are uniformly drawn at random in the cube, independently of each other, in all $L = \alpha N$ maps. For each map $\ell = 1...L$, we draw uniformly at random p positions $\hat{\mathbf{r}}^{\ell,\mu}$, $\mu = 1...p$, and collect the p corresponding patterns of activity: the neuron i is active ($\sigma_i^{\ell,\mu} = 1$) if the position is covered by its input field, and silent ($\sigma_i^{\ell,\mu} = 0$) otherwise (Figs. 1(a)&(b)).

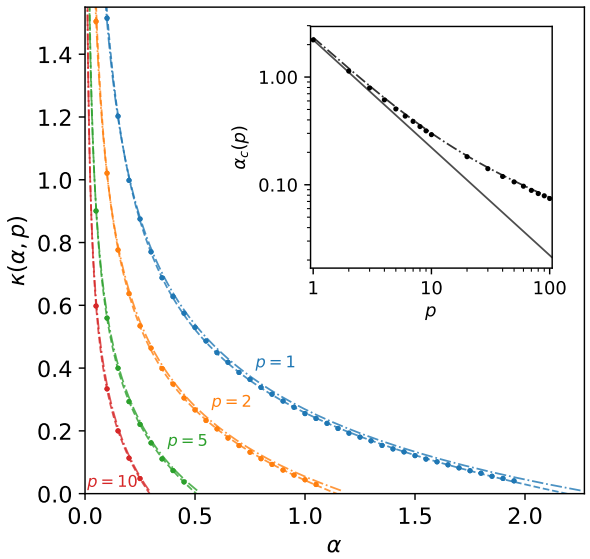

FIG. 3: Optimal stability κ as a function of the load α and the number p of positions. Dots: SVM results; Dashed lines: Gardner's theory (4); Dashed-dotted lines: quenched PF theory (7). Parameter values: $D = 2$, $\phi_0 = .3$, $N = 1000$ for SVM. Inset: $\alpha_c(p)$ decreases proportionally to $1/p$ (straight line) at low p, and much more slowly for large p. Dots indicate results from SVM ($N = 5000$), averaged over 50 samples, see SM Sec. I.D for details on the estimation of $\alpha_c(p)$; the dot size indicates the maximal error bar. The dashed-dotted line shows the predictions from the quenched PF theory.

In order to learn these patterns we use Support Vector Machines (SVM) with linear kernels and hard margin classification [21]. We train N SVM, one for every row i in the coupling matrix W_{ij}, in which we consider the neuron i as the output and the other $N-1$ neurons $j(\neq i)$ as the inputs [22]. The training set $\{\sigma_i^{\ell,\mu}\}$ is common to all SVM. Once learning is complete, we normalize each row of the coupling matrix to $\sum_{j(\neq i)} W_{ij}^2 = 1$. SVM find the coupling matrix W maximizing the stability of the stored patterns,

$$\kappa = \min_{\{i=1...N, \ell=1...L, \mu=1...p\}} \left[(2\sigma_i^{\ell,\mu} - 1) \sum_{j(\neq i)} W_{ij}\, \sigma_j^{\ell,\mu} \right] . \tag{3}$$

SVM couplings share some qualitative features with their Hebbian counterparts. First, the couplings W_{ij} are

correlated with the distances $d_{ij}^{\ell} = |\mathbf{r}_i^{\ell} - \mathbf{r}_j^{\ell}|$ between the PF centers of the neurons i and j in the different maps ℓ, see SM Sec. I.C. Secondly, when simulating the trained network with simple rules for updating the neuron activities (SM, Sec. I.E), the activity bump forms and diffuses within a map, and occasionally jumps to other maps [11, 15, 16]. However, with the maximal-stability learning rule, the spatial error ϵ can be tuned at will by varying p, see Fig. 2(a). For a fixed p, ϵ remains remarkably stable as the load increases until its critical value is reached. This is in sharp contradistinction with the Hebb rule case, for which ϵ quickly increases with the number of maps. The p patterns form a discrete approximation of the map, with average spatial error scaling as $\epsilon = p^{-1/D}$, *i.e.* as the typical distance between neighboring points (Figs. 1(a)&2(b)).

The optimal stability κ (3) is shown in Fig. 3 as a function of the load α and of the number p of prescribed fixed points; it is much higher than the maximal stability achievable with rule (2) after optimization over the interaction kernel w, see SM, Sec. I.D. As expected, $\kappa(\alpha, p)$ is a decreasing function of α and p: increasing the number of maps or enforcing finer spatial resolution reduces the stability. The value of the load at which $\kappa(\alpha, p)$ vanishes defines the critical capacity $\alpha_c(p)$, that is, the maximal load sustainable by the network as a function of the required spatial resolution. Figure 3(inset) shows that $\alpha_c(p)$ decreases proportionally to $1/p$ at low p, and then much more slowly as p grows. For small p, all $L \times p$ patterns are roughly independent, and we have $\alpha_c(p) \simeq \frac{\alpha_c(1)}{p}$, where $\alpha_c(1)$ is the capacity of the perceptron with independent, biased patterns having a fraction ϕ_0 of active neurons [18]. As p gets large, substantial redundancies between the p patterns within a map appear, as nearby positions define similar patterns (Fig. 1(b)), and the capacity is expected to decrease less quickly with p. The cross-over takes place at $p_{c.o.} \sim 1/\phi_0$ (SM, Sec. I.D). The non-trivial behavior of $\alpha_c(p)$ when $p \gg p_{c.o.}$ will be characterized in the theoretical study below.

Gardner's framework [18] can, in principle, be applied to the optimal couplings corresponding to maximal stability κ (3). Following standard calculations (SM, Sec. II.A), we find that the maximal load at fixed κ and p is given by

$$\alpha_c(\kappa, p) = 1/\min_m \langle E_p(\hat{\mathcal{R}}, \mathcal{Z}, m; \kappa)\rangle_{\hat{\mathcal{R}},\mathcal{Z}} \ , \qquad (4)$$

where the minimum is taken over $m = \phi_0 \sum_{j(\neq i)} W_{ij}$. In the formula above, $\langle\cdot\rangle$ denotes the average over the vectors $\hat{\mathcal{R}} = (\hat{\mathbf{r}}_1, ..., \hat{\mathbf{r}}_p)$ of p positions $\hat{\mathbf{r}}_\mu$ drawn uniformly at random in the D-dimensional cube, and $\mathcal{Z} = (z_1, ...z_p)$ drawn from the multivariate centered Gaussian distribution with $\hat{\mathcal{R}}$–dependent covariance matrix

$$\mathbf{\Gamma}_{\mu,\nu}(\hat{\mathcal{R}}) = \Gamma\big(|\hat{\mathbf{r}}_\mu - \hat{\mathbf{r}}_\nu|\big) - \phi_0^2 \ . \qquad (5)$$

Here, $\Gamma(d)$ is the overlapping volume between two PFs, whose centers are at distance d from one another; hence, $\Gamma(0) = \phi_0$. Function E_p in (4) is defined through

$$E_p(\hat{\mathcal{R}}, \mathcal{Z}, m; \kappa) = \min_{\{t_\mu \geq \kappa + m, \mu = 1...p\}} \sum_{\mu,\nu=1}^{p} \Big(t_\mu - z_\mu - 2\,m\,\sigma(|\hat{\mathbf{r}}_\mu|)\Big)\, \mathbf{\Gamma}^{-1}_{\mu,\nu}(\hat{\mathcal{R}})\, \Big(t_\nu - z_\nu - 2\,m\,\sigma(|\hat{\mathbf{r}}_\nu|)\Big) \ , \qquad (6)$$

where $\sigma(d) = 1$ if $d < r_c$, 0 otherwise; r_c is the radius of the PF, *i.e.* the smallest number such that $\Gamma(2\,r_c) = 0$.

In practice, computing $\alpha_c(\kappa, p)$ from (4) is quite involved from a numerical point of view, as it requires to solve the p-dimensional semi-definite quadratic optimization problem in (6), as well as to average over the random vectors $\hat{\mathcal{R}}$ and $\mathcal{Z}$. This can be accurately done for small enough p, with results in excellent agreement with the SVM simulations, see Fig. 3. Notice that, for $p = 1$, our calculation reproduces Gardner's critical capacity $\alpha_c(1)$ for independent and biased patterns (SM, Sec. II.B). This is expected as spatial correlations between patterns within a map appear when $p \geq 2$.

Formula (4) seems, unfortunately, intractable for large p. The intricate dependence on p, *e.g.* showing up through the Gaussian correlations between the p random fields z_μ in (6), stems from the average (in each map ℓ) over the N PF centers, $\{\mathbf{r}_i^{\ell}\}$, at fixed positions $\{\mathbf{r}^{\ell,\mu}\}$. To avoid introducing these correlations and have an explicit dependence on the parameter p, we consider an alternative calculation scheme, where the p positions in each map are averaged out, while keeping the $L \times N$ centers quenched. To further simplify the calculation we neglect in the effective action all terms of order ≥ 3 in the couplings W_{ij} [23]; this Gaussian approximation is expected to be exact in the large-p limit. Details about the calculation can be found in SM, Sec. II.C. Within our quenched PF theory the optimal load $\alpha_c(\kappa, p)$ is the root

of F defined through

$$F(\alpha; m, q, U, V, T) = V\left(q + U - \frac{m^2}{1 - \frac{4}{g(U)} + 4U}\right) \quad (7)$$

$$+T\left(1 + \frac{U\, g(U) - 1}{V}\right) - \alpha p(q - m^2) \int_x^\infty dz\, \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} (z - x)^2$$

with $x = \frac{m-\kappa}{\sqrt{q-m^2}}$. In (7), $m = \sum_{j(\neq i)} (2\boldsymbol{\mathcal{C}}_{ij} - \phi_0) W_{ij}$, $q = \sum_{j,k(\neq i)} W_{ij} \boldsymbol{\mathcal{C}}_{jk} W_{ik}$, and the Lagrange multipliers U, V, T enforcing, respectively, the normalization of W and the definition of the order parameters, are all chosen to optimize F. $\boldsymbol{\mathcal{C}}$ denotes the $N \times N$ multi-space Euclidean Random Matrix (ERM)

$$\boldsymbol{\mathcal{C}}_{jk}\big(\{\mathbf{r}_i^\ell\}\big) = \frac{1}{L} \sum_{\ell=1}^{L} \Gamma\big(|\mathbf{r}_j^\ell - \mathbf{r}_k^\ell|\big) \ , \quad (8)$$

with resolvent $g(U) = \frac{1}{N} \mathrm{Trace}\, (U\, \mathbf{Id} + \boldsymbol{\mathcal{C}})^{-1}$. While ERM have been intensively studied in the literature [24], superimpositions of ERM mixing up different spaces have not been considered so far to our knowledge. The resolvent $g(U)$ can nevertheless be computed using tools from Random Matrix Theory [25], and shown to be solution of the implicit equation

$$U = \frac{1}{g(U)} - \sum_{\mathbf{k} \neq \mathbf{0}} \frac{\alpha\, \hat{\Gamma}(\mathbf{k})}{\alpha + g(U)\, \hat{\Gamma}(\mathbf{k})} \ , \quad (9)$$

where the $\hat{\Gamma}(\mathbf{k})$'s are the components of the Fourier transform of Γ on the D-dimensional infinite reciprocal cube.

Resolution of these equations gives access to $\kappa(\alpha, p)$, in very good agreement with the numerical results obtained with SVM (Fig. 3). Small deviations can, however, be noticed and diminish with increasing p as expected. The order parameters q and m are shown as functions of p in Fig. 4, in good agreement with SVM results for large p ($\gg p_{c.o.}$). The value of p at which the confluence between the results from the quenched theory and SVM takes place is a decreasing function of the PF size ϕ_0 (SM, Sec. II.D) and of the map dimension D (SM, Sec. I.D).

Due to the explicit dependence of F on p in (7) the asymptotic behaviour of the critical capacity can be analytically determined in the large–p limit:

$$\alpha_c(p) \sim A(D)\, \frac{\phi_0^{-(D-1)}}{(\log p)^D} \qquad (p \to \infty) \ , \quad (10)$$

where the constant A is made explicit in SM, Sec. II.C. Equation (10) is our main result. Informally speaking, the very slow decay of the critical capacity with p (Fig. 3, inset) means that recurrent neural nets can efficiently store multiple spatial maps, even at high spatial resolution. More precisely, enforcing a strong reduction of the spatial error, such as $\epsilon \to \epsilon^2$, results in a moderate drop of the maximal sustainable load, $\alpha_c \to \alpha_c / 2^D$. In addition, the capacity is predicted to be a decreasing function of the PF size in dimensions $D = 2, 3$, but not in dimension $D = 1$. This asymptotic statement is qualitatively corroborated by SVM results, even for moderate values of p (SM, Sec. I.D).

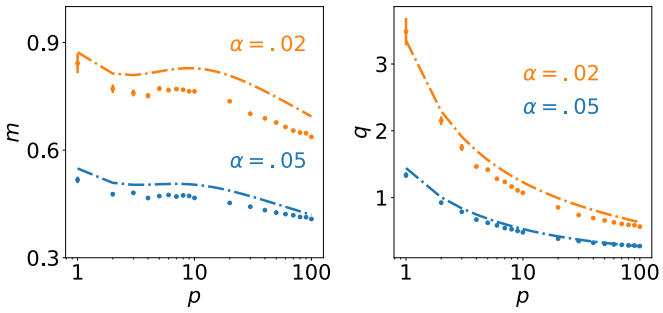

FIG. 4: Order parameters m (left) and q (right) vs. p. Dots: SVM results ($N = 2500$), averaged over 50 samples; Dashed-dotted lines: quenched PF theory (7). Parameters: $D = 2$, $\phi_0 = .3$, $\alpha = .02$ (top) and .05 (bottom), for which up to, respectively, $p_c \simeq 2500$ and $p_c \simeq 250$ points can be memorized.

Many extensions of the current work can be contemplated. First, our theory can be easily generalized to the case of spatial resolutions varying from map to map, by substituting p with its average value over the maps in (10). This suggests that the fraction of maps with finest spatial resolution ϵ should not exceed $\sim \epsilon^D$ when $\epsilon \to 0$, in order not to affect too much the critical capacity.

Secondly, while we have assumed here for the sake of simplicity that the spatial resolution was statistically uniform across space, this need not be the case in practice. Experiments have shown that spatial representations of environments are enriched in place fields close to spots of interests (such as water pots [26] or objects [27]) with respect to void regions. Numerical simulations reported in SM, Sec. I.D&E show that increasing the density of prescribed positions in regions of the physical space allows us to carve specific attractors in the neural activity space, representing preferentially those regions. This result is compatible with recent studies establishing the link between PF distribution and behavioral place preference [28]. Interestingly, our quenched PF theory can be applied to any set of PF centers and sizes, not necessarily homogeneously distributed over space; knowledge of the PF characteristics, *e.g.* from experimental measurements, allows us to determine the multispace correlation matrix $\boldsymbol{\mathcal{C}}$ in (8) and to make specific predictions. A proof of principle of this approach is shown in SM, Sec. II.D, where we compare the couplings found with SVM and with our quenched PF theory on synthetic data.

Thirdly, several improvements could be brought in terms of biological plausibility. In particular one should study the case of continuous rather than binary neurons, explicitly distinguish excitatory and inhibitory neurons and impose Dale's law on the associated synapses, and take into account the sparse nature of synapses [29] and

of place-cell activity [9] observed in CA3. Border effects, known to be important for hippocampal maps [30], should also be considered instead of the simple periodic boundary conditions assumed here. Finally, it would be extremely interesting to study the dynamics of learning, in particular how the network progressively maturates to account for more and more fixed points and eventually defines a quasi-continuous attractor (SM, Sec. I.B), as seems to be the case during the first weeks of development in rodents [31].

Acknowledgements. We are grateful to A. Treves for useful discussions. This work was funded by the HFSP RGP0057/2016 project.

[1] S. Amari. *Bio. Cyber.* **27**, 77-87 (1977)
[2] M. Tsodyks, T. Sejnowski, *Int. J. Neur. Syst.* **6**, 81-86 (1995)
[3] B. Ben-Yishai, R. Bar-Or, H. Sompolinsky. *Proc. Natl. Acad. Sci.* **92**, 3844-48 (1995)
[4] C.C.A. Fung, K.Y.M. Wong, S. Wu, *Neural Comp.* **22**, 752-92 (2010)
[5] W. Zhong *et al.*, arXiv:1809.11167 (2018)
[6] S.S. Kim *et al*, *Science* **356**, 849-853 (2017)
[7] K. Yoon *et al*, *Nature Neurosci.* **16**, 1077 (2013)
[8] K. Wimmer *et al*, *Nature Neurosci.* **17**, 431 (2014)
[9] C.B. Alme *et al*, *Proc. Natl. Acad. Sci.* **111**, 18428-35 (2014)
[10] K. Jezek *et al*, *Nature* **478**, 246 (2011)
[11] A. Samsonovich, B.L. McNaughton, *J. Neurosci.* **17**, 5900 (1997)
[12] F.P. Battaglia, A. Treves. *Phys. Rev. E* **58**, 7738 (1998)
[13] R. Monasson, S. Rosay. *Phys. Rev. E* **87**, 062813 (2013)
[14] E. Cerasti, A. Treves. *Front. Cell. Neurosci.* **7**, 112 (2013)
[15] R. Monasson, S. Rosay. *Phys. Rev. E* **89**, 032803 (2014)
[16] R. Monasson, S. Rosay. *Phys. Rev. Lett.* **115**, 098101 (2015)
[17] D. Amit, H. Gutfreund, H. Sompolinsky. *Phys. Rev. Lett.* **55**, 1530 (1985)
[18] E. Gardner. *J. Phys. A* **21**, 257 (1988).
[19] J.J. Hopfield, *Proc. Nat. Acad. Sci.* **79**, 2554-58 (1982)
[20] N. Brunel. *Nature Neurosci.* **19**, 749 (2016)
[21] B. Scholkopf, A.J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press (2001)
[22] The case of hetero-associative classification of manifolds was recently studied by S.Y. Chung, D.D. Lee, H. Sompolinsky, *Phys. Rev. X* **8**, 031003 (2018)
[23] R. Monasson, *J. Physique I* **3**, 1141-52 (1993)
[24] A. Goetschy, S.E. Skipetrov, arXiv:1303.2880 (2013)
[25] A. Battista, R. Monasson, *in preparation* (2019)
[26] S.A. Hollup et al. *J. Neurosci.* **21**, 1635-44 (2001)
[27] R. Bourboulou et al., *eLife* **8**:e44487 (2019)
[28] O. Mamad *et al*, *PLoS Biology* **15**:e2002365 (2017)
[29] S.J. Guzman *et al. Science* **353**, 1117-23 (2016)
[30] C. Barry *et al. Reviews in the Neurosciences* **17**, 71-97 (2006)
[31] U. Farooq, G. Dragoi, *Science* **363**, 168-173 (2019)