Storage and Learning phase transitions in the Random-Features Hopfield Model

Storage and Learning Phase Transitions in the Random-Features Hopfield Model

M. Negri,[1,2,*] C. Lauditi,[3,4] G. Perugini,[4] C. Lucibello,[4,5] and E. Malatesta[4,5]

[1]*Department of Physics, University of Rome "La Sapienza", Piazzale Aldo Moro 5, 00185 Roma, Italy*
[2]*CNR-NANOTEC, Institute of Nanotechnology, Rome Unit, Piazzale Aldo Moro, 00185 Roma, Italy*
[3]*Department of Applied Science and Technology, Politecnico di Torino, 10129 Torino, Italy*
[4]*Department of Computing Sciences, Bocconi University, 20136 Milano, Italy*
[5]*Institute for Data Science and Analytics, Bocconi University, 20136 Milan, Italy*

(Received 8 May 2023; accepted 22 November 2023; published 19 December 2023)

The Hopfield model is a paradigmatic model of neural networks that has been analyzed for many decades in the statistical physics, neuroscience, and machine learning communities. Inspired by the manifold hypothesis in machine learning, we propose and investigate a generalization of the standard setting that we name *random-features Hopfield model*. Here, P binary patterns of length N are generated by applying to Gaussian vectors sampled in a latent space of dimension D a random projection followed by a nonlinearity. Using the replica method from statistical physics, we derive the phase diagram of the model in the limit $P, N, D \to \infty$ with fixed ratios $\alpha = P/N$ and $\alpha_D = D/N$. Besides the usual retrieval phase, where the patterns can be dynamically recovered from some initial corruption, we uncover a new phase where the features characterizing the projection can be recovered instead. We call this phenomena the *learning phase transition*, as the features are not explicitly given to the model but rather are inferred from the patterns in an unsupervised fashion.

DOI: 10.1103/PhysRevLett.131.257301

The Hopfield model (HM) [1] is a paradigmatic connectionist model of associative memory with biological plausibility that allows the dynamical retrieval of stored patterns from corrupted observations. In the case of uncorrelated patterns, retrieval is possible for a number of patterns that scales linearly with the system size N, and the critical prefeature can be computed to high precision using spin-glass theory techniques [2].

Following Hopfield's seminal work, several generalizations have been investigated. A recent surge of interest involves generalizations that go beyond pairwise interactions and yield polynomial [3,4] or even exponential capacity [5–7]. Notably, the modern Hopfield network proposed in [6] is closely related to the attention mechanism that has revolutionized deep learning in the past years [8]. Other research lines preserve the pairwise structure of the standard Hopfield model (SHM) while proposing different (non-Hebb) rules for the couplings in order to address the problem of correlation among patterns decreasing the capacity [9–13]. Many sensible models of correlation in and among patterns have been proposed. For example, in [14] the authors study a biased distribution of binary patterns, that can even be generalized to a hierarchical structure of correlation as it was discussed in [15,16]. Another approach is to consider correlations in the form of Markov chains [13], which can be used to produce a correlation length both between different spins of a given pattern and between the same spin in different patterns.

Most theoretical studies of (generalized) HMs assume simple distributions for the patterns [2,3], while in practical applications the patterns are linearly or nonlinearly encoded from and decoded to a different space [17].

In this work, we addressed this limitation by proposing a generative model for the patterns where each pattern is produced by the linear combinations of a fixed vocabulary of what we call *features* weighted by pattern-specific *coefficients*, followed by an elementwise nonlinearity. We analyze the model in the high-dimensional regime using the replica method for the statistical physics of disordered systems.

This data-generating process generalizes the structure of linear superposition proposed in [18], where it was discussed in relation to the mapping between a Hopfield network and a restricted Boltzmann machine. A similar linear (but dense) mapping has been discussed in [19,20]. Our model is also deeply related to the so-called *hidden-manifold* model [21], which has been used as an analytically solvable model of feed-forward neural networks fitting data points that live on a low-dimensional submanifold of their embedding space. In fact, this low-dimensional latent structure is typical of many real-world datasets, e.g., the ones made of natural images. Here, we do not modify the Hebb rule, as we will see that it is enough to produce a new behavior of the model, in conjunction with the structure of correlation that we choose. In fact, we observe that if the correlations in the data are strong enough, the model switches from a storage phase to a learning phase, in the sense that attractors appear corresponding to the features in the data. We argue that this behavior opens up a new paradigm for this model and shows that it may have some phenomenology in common with neural networks.

Model definition.—The Hopfield model [1] can be defined as a statistical physics model with N binary spins $s_i = \pm 1$, $i = 1, \dots, N$, and an energy function with all-to-all pairwise interactions

$$\mathcal{H}(\boldsymbol{s}) = -\frac{1}{2}\sum_{i\neq j} J_{ij} s_i s_j. \tag{1}$$

The coupling matrix J is defined through a set of P *patterns* $\{\boldsymbol{\xi}_\nu\}_{\nu=1}^P$ via the Hebbian rule

$$J_{ij} = \frac{1}{N}\sum_{\nu=1}^{P} \xi_{\nu i}\xi_{\nu j}. \tag{2}$$

In the standard statistical physics setting [2], $\xi_{\nu i}$ are independently and uniformly distributed binary spins. In this work, instead, we consider structured patterns given by a linear projection and a latent vector composed with a nonlinearity:

$$\xi_{\nu i} = \sigma\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D} c_{\nu k} f_{ki}\right), \tag{3}$$

where $\sigma(\cdot)$ is a generic nonlinear function, f_{ki} is called the matrix of *features,* and $c_{\nu k}$ is the matrix of *coefficients*; we call this the *random-features Hopfield model* (RFHM). A linear version of this structure is analyzed in Ref. [18]. The specific case we consider through the paper is the one of i.i.d. uniform binary features $f_{ki} = \pm 1$, i.i.d. standard Gaussian coefficients $c_{\nu k}$, and σ equal to the *sgn* function.

By tuning D we can switch between weakly and strongly correlated examples. In fact, in the $\alpha_D \to \infty$ we expect to recover the SHM as the examples become uncorrelated.

In this work, the numerical results and most of the analytical ones are obtained in the limit $T \to 0$. In this limit, the update rule of each spin at time t reads

$$s_i^{(t+1)} = \mathrm{sgn}\left(\sum_{j(\neq i)}^{N} J_{ij} s_j^{(t)}\right). \tag{4}$$

We use this update rule in an asynchronous way, meaning that we update one spin at the time in random order (see Sec. D1 in Supplemental Material [22] for a pseudocode). If a spin configuration $\tilde{s}_i$ satisfies the relation $\tilde{s}_i = \mathrm{sgn}(\sum_{j(\neq i)}^{N} J_{ij}\tilde{s}_i)$, then we say that $\tilde{s}_i$ is a fixed point of the dynamics. If the dynamics converges to $\tilde{s}_i$ even when a fraction of spins has been flipped, then $\tilde{s}_i$ is an attractor. The original task of the Hopfield model is to store P examples as attractors. This can also be seen as a denoising operation, since the model is capable of retrieving the stored patterns starting from noisy versions of them. In [2] the authors computed the maximum number i.i.d. patterns that can be retrieved, allowing for a small fraction of errors, in the scaling regime where $P = \alpha N$ as N grows to infinity with α fixed. They obtain a critical value $\alpha_c \simeq 0.138$ such that the model is able to retrieve all patterns if $\alpha < \alpha_c$, while above α_c the model shows a first-order phase transition referred as *catastrophic forgetting* and no storage is possible: The fixed point of the dynamics is completely uncorrelated with the patterns.

In our RFHM, the basic question that we are interested in is whether the features $\mathbf{f}_k$ can be attractors themselves, and what happens to the attractors corresponding to the patterns.

Replica analysis.—Since we are interested in the thermodynamic limit $N \to \infty$, we choose a regime where both P and D are proportional to N. At the same time, we keep the following ratios fixed:

$$\alpha = \frac{P}{N}, \qquad \alpha_D = \frac{D}{N}. \tag{5}$$

These will be the control parameters for our model. They are related via the relation $\alpha = \alpha_T\alpha_D$, where $\alpha_T = P/D$.

In order to identify the phase transitions of the RFHM, we want to compute the averaged free energy

$$\phi = \lim_{N\to\infty} -\frac{1}{\beta N}\langle \ln Z\rangle_{c,f}, \tag{6}$$

where we specified that we have two sources of disorder that must be averaged: the coefficients c and the features f. In other words, we consider the quenched average $\langle \ln Z\rangle_\xi$ where the patterns have a structured distribution. $Z = \sum_{\boldsymbol{s}} e^{-\beta\mathcal{H}(\boldsymbol{s})}$ is the partition function, where the sum is taken over the possible values of the spins $s_i = \pm 1$ for $i = 1, \dots, N$.

In order to compute the average of $\ln Z$ in Eq. (6), we use the replica method [23] that consists in writing the average of logarithm as $\langle \ln Z\rangle = \lim_{n\to 0}(\langle Z^n\rangle - 1)/n$.

The replicated partition function averaged over the disordered reads

$$\langle Z^n\rangle = e^{-(\beta/2)Pn}\sum_{\{s_i^a\}}\int\prod_{\nu a}\frac{dm_\nu^a}{\sqrt{2\pi}}e^{(\beta/2)\sum_{\nu=1}^{P}\sum_{a=1}^{n}(m_\nu^a)^2} \times\left\langle\prod\nolimits_{\nu a}\delta\left(m_\nu^a-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sigma\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D}c_{\nu k}f_{ki}\right)s_i^a\right)\right\rangle_{c,f}, \tag{7}$$

where we introduced the set of auxiliary variables

$$m_\nu^a = \frac{1}{N}\sum_i \xi_{\nu i}s_i^a, \qquad a\in[n], \qquad \nu\in[P]. \tag{8}$$

We call these *pattern magnetizations* to distinguish them from another set of order parameters, whose definition we anticipate here:

$$\mu_k^a = \frac{1}{N}\sum_i f_{ki} s_i^a, \qquad a \in [n], \qquad k \in [D]. \tag{9}$$

We call these the *feature magnetizations*. We want to see if there is a region of the α_D vs α phase diagram where $\mu_k > 0$ for some k. We also want to see what happens to the pattern magnetizations in the same phase diagram.

Similarly to [2], we make some ansatz on the structure of the solution for both these order parameters. We study two cases: the case where the model retrieves only one of the features and the case where the model retrieves only one of the examples.

Feature retrieval.—In order to analyze the retrieval of one feature only, we impose that $\mu_1 = O(1)$ and $\mu_k = O(1/\sqrt{N})$ for $k > 1$. At the same time, we impose $m_\nu = O(1/\sqrt{N})$, $\forall\ \nu$. In the thermodynamic limit, this means that we look for a solution of the form

$$\boldsymbol{\mu} = (\mu, 0, \dots, 0), \qquad \mathbf{m} = (0, \dots, 0). \tag{10}$$

In this regime, in order to compute the average over the coefficients c we must pay particular attention to the term $k = 1$ in Eq. (7), since it by itself can give a finite contribution:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \sigma\left(\frac{1}{\sqrt{D}}\sum_{k=2}^{D} c_{\nu k} f_{ki} + \frac{1}{\sqrt{D}} c_{\nu 1} f_{1i}\right) s_i^a. \tag{11}$$

We show in Sec. A in Supplemental Material [22] that the resulting distribution of m_ν^a is a Gaussian $\mathcal{N}(m_\nu^a; \bar{m}, Q)$ with mean

$$\bar{m}_\nu^a = \frac{c_{\nu 1}}{\sqrt{\alpha_D}} \mu_1 \kappa_1 \tag{12}$$

and covariance matrix

$$Q^{ab} = \kappa_*^2 q^{ab} + \kappa_1^2 p^{ab}, \tag{13}$$

where we defined the following quantities:

$$q^{ab} = \frac{1}{N}\sum_i s_i^a s_i^b, \tag{14}$$

$$p^{ab} = \frac{1}{D}\sum_{k>1} \mu_k^a \mu_k^b, \tag{15}$$

and coefficients $\kappa_0 = \int Dz\sigma(z)$, $\kappa_1 = \int Dz\, z\sigma(z)$, $\kappa_2 = \int Dz\sigma^2(z)$, and $\kappa_*^2 = \kappa_2 - \kappa_1^2 - \kappa_0^2$. This calculation goes under the name of Gaussian equivalence theorem (GET), and it has been developed in [21,24–28] and applied in cases with zero mean. The replicated partition function now reads

$$\langle Z^n \rangle = e^{-(\beta/2)Pn} \sum_{\{s_i^a\}} \int \prod_{\nu a} \frac{dm_\nu^a}{\sqrt{2\pi}} \times \exp\left\{\frac{\beta}{2}\sum_{\nu=1}^{P}\sum_{a=1}^{n} (m_\nu^a)^2\right\} \left\langle \prod\nolimits_\nu \mathcal{N}(m_\nu; \bar{m}, Q) \right\rangle_{c_1, f}, \tag{16}$$

where $\langle \cdots \rangle$ represents the average over the remaining quenched disorder f and $c_1 = \{c_{\nu 1}\}_{\nu=1}^P$.

We solve this model in the replica-symmetric (RS) ansatz. For the complete derivation, see Sec. B in Supplemental Material [22]. At the end of the long but straightforward calculation, we end up with a free energy f^{RS} that depends on eight order parameters: the feature magnetization μ, the overlap between different replicas q, the diagonal and off-diagonal parts of p^{ab}, and their four conjugate parameters $\hat{\mu}$, $\hat{q}$, $\hat{p}_{\mathrm{d}}$, and $\hat{p}$. Given the control parameters β, α, and α_D, we obtain the physical value of the order parameters by extremizing the free energy:

$$f_{\mathrm{opt}}^{\mathrm{RS}} = \underset{\mu,\hat{\mu},q,\hat{q},p_{\mathrm{d}},\hat{p}_{\mathrm{d}},p,\hat{p}}{\mathrm{extr}} f^{\mathrm{RS}}(\mu, \hat{\mu}, q, \hat{q}, p_{\mathrm{d}}, \hat{p}_{\mathrm{d}}, p, \hat{p}). \tag{17}$$

Deriving f^{RS} with respect to the order parameters, we obtain a set of eight equations that must be solved together (the so-called *saddle-point equations*). We write here only two of them, leaving the rest to Supplemental Material [22] [see Eq. (B31)]:

$$q = \mathbb{E}_{z,f} \tanh^2\left(\beta\left[z\sqrt{\alpha\hat{q}} + \hat{\mu} f\right]\right), \tag{18}$$

$$\mu = \mathbb{E}_{z,f} f \tanh\left(\beta\left[z\sqrt{\alpha\hat{q}} + \hat{\mu} f\right]\right), \tag{19}$$

where $z \sim \mathcal{N}(0, 1)$ and $f \sim \mathrm{Unif}(\{-1, +1\})$. We can observe that these equations resemble closely the ones for q and m in the SHM (see [2]): Now f has the role of the retrieved pattern and μ has the role of the magnetization. The major difference is that in our case the equation for the conjugate $\hat{q}$, reported in Supplemental Material [22] [Eq. (B31)], is more complicated and depends on the rest of the order parameters. A minor difference is that, inside the integrals of the first two equations, $\hat{\mu}$ appears instead of μ.

The solution to these equations in the limit $\beta \to \infty$ is shown in Fig. 1: For $\alpha > \alpha^{\mathrm{crit}}(\alpha_D)$, the feature magnetization becomes finite with a discontinuous jump, showing that the model is actually capable of storing the features f as attractors. This jump is a first-order phase transition similarly to the catastrophic forgetting but with the important difference that the magnetization becomes finite when α is *larger* rather than smaller than a critical value. The critical point $\alpha^{\mathrm{crit}}(\alpha_D)$ rapidly increases when α_D increases, up to the point where it diverges for $\alpha_D \simeq 0.138$. This critical value is numerically identical to the critical capacity of the SHM, and it is not a coincidence. In fact, in the limit $P \gg N, D$, we have that the coupling matrix becomes (up to a feature that can be reabsorbed in the temperature) that of a SHM where the patterns are replaced by features:

$$\frac{1}{P}\sum_{\nu=1}^{P}\xi_{\nu i}\xi_{\nu j} \overset{P\to\infty}{\simeq} \kappa_1^2 \frac{1}{D}\sum_{k=1}^{D} f_{ik} f_{jk}. \tag{20}$$

See Sec. B5 in Supplemental Material [22] for the derivation. Therefore, the saddle-point equations of the RFHM must become identical to those of the SHM with μ playing the role of the magnetization and f that of the retrieved patterns (the correct scalings for this limit and the explicit calculation are shown in Sec. B4d in Supplemental Material [22]). One way to look at this behavior is to fix a value of α and to increase α_D, thus moving horizontally in the phase diagram in Fig. 1(a): When α_D is low enough, the model is able to retrieve the features; then, when they become too many, the equivalent of a catastrophic forgetting happens. This transition happens at the Hopfield critical capacity only if $\alpha = \infty$, where the matching between the two models is perfect.

The comparison between this analytical solution and numerical simulations is shown in Fig. 1(b), where we find a very good agreement for $\alpha_D = 0.03$. We test other ranges of α and α_D in Supplemental Material [22] (see Fig. D.3), and we find again good agreement.

Pattern retrieval.—For the second case, we say that $m_1 = O(1)$ and $m_\nu = O(1/\sqrt{N})$ for $\nu > 1$. At the same time, we impose that $\mu_k = O(1/\sqrt{N}),\ \forall\ k$. In the thermodynamic limit, this means that we look for a solution of the form

$$\boldsymbol{\mu} = (0, \ldots, 0), \qquad \mathbf{m} = (m, 0, \ldots, 0). \tag{21}$$

In this setting, we must be careful to apply the GET only to the vanishing pattern magnetizations, leaving the terms involving m_1 as they are. The resulting expression of the average replicated partition function reads

$$\langle Z^n\rangle = e^{-(\beta/2)Pn}\sum_{\{s_i^a\}}\int\prod_{\nu a}\frac{dm_\nu^a}{\sqrt{2\pi}}\Big\langle\prod_\nu \mathcal{N}(m_\nu;0,Q) \\ \times \exp\left\{\frac{\beta}{2}\sum_{a=1}^{n}(m_1^a)^2 + \frac{\beta}{2}\sum_{\nu>1}^{P}\sum_{a=1}^{n}(m_\nu^a)^2\right\} \\ \times \prod_a \delta\left(m_1^a - \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sigma\left(\frac{1}{\sqrt{D}}\sum_{k=1}^{D}c_{1k}f_{ki}\right)s_i^a\right)\Big\rangle_{\tilde{c}_1,f}, \tag{22}$$

where $\langle\cdots\rangle$ represents the average over the remaining quenched disorder f and $\tilde{c}_1 = \{c_{1k}\}_{k=1}^{D}$.

As we did for the feature retrieval case, we solve the model within the RS ansatz and we report the complete calculation in Supplemental Material [22] (Sec. C). This time set the order parameters do not include μ and $\hat{\mu}$, but it does include m (without the need for a conjugate variable $\hat{m}$). The order parameters also include the auxiliary variables t and $\hat{t}$ that are needed to linearize a term in

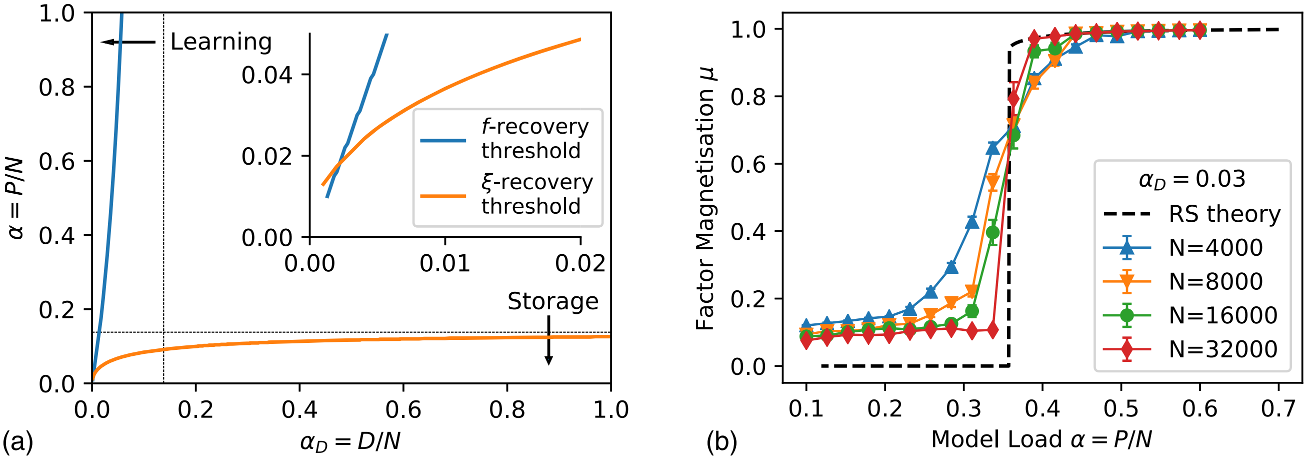

FIG. 1. Storage and learning transitions. (a) The phase diagram of the RFHM shows three regions: the *storage* phase (below the orange line), where patterns $\boldsymbol{\xi}_\nu$ are attractors; the *learning* phase (above blue line), where the features $\mathbf{f}_k$ are attractors; and the spin-glass phase (between the lines), where the attractors are uncorrelated with either $\boldsymbol{\xi}_\nu$ or $\mathbf{f}_k$. The two asymptotes are at $\alpha \simeq 0.138$ and $\alpha_D \simeq 0.138$. (b) The plot shows the feature magnetization μ along a vertical cut of the phase diagram: Increasing α, the feature magnetization μ becomes different from zero with a first-order phase transition. The dashed line is the analytical prediction of the RS theory, while the markers are numerical experiments averaged over many samples for each value of α. The simulations are performed initializing the model to a feature $\mathbf{f}_k$, running the update rule (4), and then measuring μ_k at convergence. We used 100, 50, 20, and ten samples for increasing values of N.

an intermediate integral. The definition of t is $t = (1/N)\sum_i^N \hat{v}_i s_i/$, where $\hat{v}_i$ are the conjugate variables of the auxiliary variables $v_i = (1/\sqrt{D})\sum_k^D c_{\nu k} f_{ki}$. The auxiliary variables v_i and $\hat{v}_i$ do not appear in the free energy, because they can be integrated right away. In summary, the set of nine order parameters is $m, q, \hat{q}, p_\mathrm{d}, \hat{p}_\mathrm{d}, p, \hat{p}, t, \hat{t}$.

Again, we show here only how the equation for m and q changes from the standard case in [2], and we write the rest of them in Supplemental Material [22] [Eq. (C23)].

$$q = \mathbb{E}_{x,v} \tanh^2[\beta\,(v\hat{t} + \sigma(v)m + x)] \tag{23}$$

$$m = \mathbb{E}_{x,v}\sigma(v) \tanh[\beta(v\hat{t} + \sigma(v)m + x)] \tag{24}$$

where $v \sim \mathcal{N}(0,1)$ and $x \sim \mathcal{N}(0, \alpha\hat{q} - \hat{t}^2)$. We solve the full set of saddle point equations in the limit $\beta \to \infty$ and we show the results in Fig. 1(a). A useful limit to consider is $\alpha_D \to \infty$: in this limit the equations converge to the SHM ones (see Sec. C8 in the SM), which was expected since the examples become uncorrelated. This produces an horizontal asymptote at $\alpha \simeq 0.138$ for the spinodal line of m. Decreasing α_D the example patterns become more correlated and the catastrophic forgetting happens at a lower value of α, until it happens at $\alpha = 0$ for $\alpha_D \to 0$.

The comparison between this analytical solution and numerical simulations is shown in Fig. 2: we find that, as we move from the $\alpha_D \gg 1$ regime (where we know that the simulations must match the SHM theory), the catastrophic forgetting happens at a value of α lower than the predicted one; furthermore, the mismatch increases for lower values of α_D (see also Fig. D.4 in the SM). This last fact suggests that strong correlations might be responsible of a failure of the RS ansatz. In fact, in [2], the authors found that the correct ansatz at zero temperature is indeed the full-replica-symmetry-breaking one, but the corrections to the RS calculations are small in their model. To support this hypothesis, we checked the entropy of our solution and we found that it becomes more negative the smaller the value of α_D (see Fig. D.5a in the SM). We also ruled out a possible inconsistency of the ansatz (21): in Fig. D.5b in the SM we show that both the average and the maximum of $\{m_\nu\}_{\nu>1}$ go to zero as $N \to \infty$, consistently with Eq. (21).

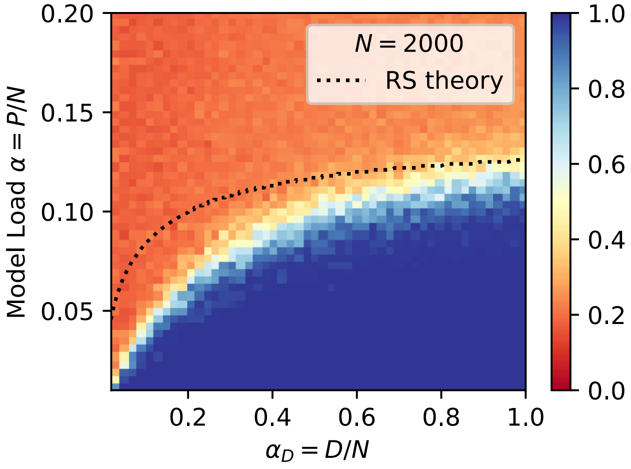

FIG. 2. Comparison with numerical results for the retrieval of one pattern. Each pixel represents the mean pattern magnetization for given values of α and α_D, averaged over 25 samples of size $N = 2000$. The simulations are performed initializing the model to a pattern $\boldsymbol{\xi}_\nu$, running the update rule (4), then measuring m_ν at convergence.

Learning transition.—Summing up the results, we have a phase diagram with two transition lines demarking three regions [see Fig. 1(a)]: the feature retrieval region, for which we obtain a nonzero feature magnetization solution to saddle point (10); the pattern retrieval region, for which we have nonzero pattern magnetization solutions for Eq. (21); and a spin-glass region between the two. The behavior that we call *learning transition* can be observed following a vertical line in the phase diagram, namely, fixing a value of α_D and increasing α. Starting from small α, we obtain a model of storage of correlated patterns: The capacity is smaller than the uncorrelated case, but the phenomenology is similar, since there is a maximum number of patterns that can be stored, and attempting to store a larger number results in catastrophic forgetting. The surprising result is that, when we have $\alpha_D \leq 0.138$ (i.e., when the correlations are strong enough), if we keep increasing the number of patterns, we find another phase beyond the spin-glass one. In this new phase, attractors corresponding to the features $\mathbf{f}_k$ appear. If we interpret the patterns $\boldsymbol{\xi}_\mu$ as an unsupervised training dataset, we see that, if the dataset is big enough, the model is capable of inferring the features hidden in the data. This behavior resembles the feature extraction that deep neural networks and some shallow generative models perform [4,29–31]. Our model represents an extension to the classical Hopfield settings that, while being amenable to theoretical analysis, can potentially capture the phenomenology of much more complex architectures, similarly to what the hidden-manifold model does for the supervised learning phenomenology [21]. An additional step in this direction would be to consider the features $\mathbf{f}_k$ as part of the machine instead of the data and the external stimuli as entirely provided by c. In this setting, one could think of a learning rule beyond the Hebb one, where the features are chosen to maximize the likelihood of a training set. It could be also interesting to extend the analysis proposed in this work to modern versions of the Hopfield model, such as the superlinear capacity ones analyzed in Refs. [4,6,7].

The code used in this work is available [32].

We thank Marc Mézard for many useful comments and discussions. M. N. acknowledges the support of LazioInnova—Regione Lazio under the program Gruppi di ricerca 2020—POR FESR Lazio 2014–2020, Project NanoProbe (Application Code No. A0375-2020-36761), and PNRR MUR Project No. PE0000013-FAIR.

*Corresponding author: matteo.negri@uniroma1.it

[1] J. J. Hopfield, Neural networks and physical systems with emergent collective computational abilities, Proc. Natl. Acad. Sci. U.S.A. **79**, 2554 (1982).

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Statistical mechanics of neural networks near saturation, Ann. Phys. (N.Y.) **173**, 30 (1987).

[3] E. Gardner, Multiconnected neural network models, J. Phys. A **20**, 3453 (1987).

[4] D. Krotov and J. J. Hopfield, Dense associative memory for pattern recognition, Adv. Neural Inf. Process. Syst. **29**, 1172 (2016).

[5] M. Demircigil, J. Heusel, M. Löwe, S. Upgang, and F. Vermet, On a model of associative memory with huge storage capacity, J. Stat. Phys. **168**, 288 (2017).

[6] H. Ramsauer, B. Schäfl, J. Lehner, P. Seidl, M. Widrich, T. Adler, L. Gruber, M. Holzleitner, M. Pavlović, G. K. Sandve *et al.*, Hopfield networks is all you need, arXiv:2008.02217.

[7] C. Lucibello and M. Mézard, The exponential capacity of dense associative memories, arXiv:2304.14964.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. **30**, 5998 (2017).

[9] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Information storage in neural networks with low levels of activity, Phys. Rev. A **35**, 2293 (1987).

[10] J. F. Fontanari and W. Theumann, On the storage of correlated patterns in Hopfield's model, J. Phys. (Les Ulis, Fr.) **51**, 375 (1990).

[11] R. Der, V. Dotsenko, and B. Tirozzi, Modified pseudo-inverse neural networks storing correlated patterns, J. Phys. A **25**, 2843 (1992).

[12] J. Van Hemmen, Hebbian learning, its correlation catastrophe, and unlearning, Network **8**, V1 (1997).

[13] M. Löwe, On the storage capacity of Hopfield models with correlated patterns, Ann. Appl. Probab. **8**, 1216 (1998).

[14] H. Gutfreund, Neural networks with hierarchically correlated patterns, Phys. Rev. A **37**, 570 (1988).

[15] C. Cortes, A. Krogh, and J. Hertz, Hierarchical associative networks, J. Phys. A **20**, 4449 (1987).

[16] A. Krogh and J. Hertz, Mean-field analysis of hierarchical associative networks with 'magnetisation', J. Phys. A **21**, 2211 (1988).

[17] J. Steinberg and H. Sompolinsky, Associative memory of structured knowledge, Sci. Rep. **12**, 21808 (2022).

[18] M. Mézard, Mean-field message-passing equations in the hopfield model and its generalizations, Phys. Rev. E **95**, 022117 (2017).

[19] E. Agliari, A. Barra, A. De Antoni, and A. Galluzzi, Parallel retrieval of correlated patterns: From Hopfield networks to Boltzmann machines, Neural Netw. **38**, 52 (2013).

[20] M. Smart and A. Zilman, On the mapping between hopfield networks and restricted Boltzmann machines, arXiv:2101.11744.

[21] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the influence of data structure on learning in neural networks: The hidden manifold model, Phys. Rev. X **10**, 041044 (2020).

[22] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.131.257301 for a complete derivation of the calculations and for additional figures.

[23] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications* (World Scientific, Singapore, 1987), Vol. 9.

[24] S. Mei and A. Montanari, The generalization error of random features regression: Precise asymptotics and the double descent curve, Commun. Pure Appl. Math. **75**, 667 (2022).

[25] F. Gerace, B. Loureiro, F. Krzakala, M. Mézard, and L. Zdeborová, Generalisation error in learning with random features and the hidden manifold model, in *Proceedings of the International Conference on Machine Learning* (IOP Publishing, 2020), pp. 3452–3462.

[26] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, The gaussian equivalence of generative models for learning with shallow neural networks, in *Mathematical and Scientific Machine Learning* (MIT Press, 2022), pp. 426–471.

[27] H. Hu and Y. M. Lu, Universality laws for high-dimensional learning with random features, IEEE Trans. Inf. Theory **69**, 1932 (2022).

[28] C. Baldassi, C. Lauditi, E. M. Malatesta, R. Pacelli, G. Perugini, and R. Zecchina, Learning through atypical phase transitions in overparameterized neural networks, Phys. Rev. E **106**, 014116 (2022).

[29] G. E. Hinton, S. Osindero, and Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. **18**, 1527 (2006).

[30] G. E. Hinton and R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science **313**, 504 (2006).

[31] J. Tubiana and R. Monasson, Emergence of compositional representations in restricted Boltzmann machines, Phys. Rev. Lett. **118**, 138301 (2017).

[32] https://github.com/ArtLabBocconi/random_feature_hopfield.