Repeller neural networks

Andrzej Nowak
Institute for Social Studies, Warsaw University, ul. Stawki 5/7, 00–183 Warsaw, Poland

Maciej Lewenstein*
Centre d'Etudes de Saclay, Service des Photons, Atomes et Molécules, Bâtiment 22, 91191 Gif sur Yvette CEDEX, France

Wojciech Tarkowski
Centrum Fizyki Teoretycznej, Polskiej Akademii Nauk, Aleja Lotników 32/46, 02–668 Warsaw, Poland
(Received 6 May 1993)

We propose a class of network models suited for negative-choice classification. Our models change substantially the *rejected* states and do not change appreciably the *accepted* states.

PACS number(s): 87.10.+e

Most of the work on neural network modeling in recent years concentrated on the so-called attractor neural networks [1] (ANN) or on multilayered perceptrons [2,3]. In both cases the ultimate task of learning is that the network realizes a definite relation between input and output states. In the case of ANN the stored patterns are supposed to be stationary, in the case of multilayered networks a definite output pattern should correspond to a given input pattern. In many situations, ranging from biology and medicine to high-energy physics [4] one deals, however, with negative-choice classification when it is important to *reject* unwanted patterns. One can think, for instance, of the production process in which bad or failed products should be rejected. Another example is a decision process in which decisions leading to dangerous and unwanted situations should be avoided. In many cases it is also useful to have a device which transforms unacceptable inputs into more acceptable ones while preserving some of their original features. In the process of designing one often wants to avoid some unwanted features allowing at the same time for possibly large diversity of desired ones. In this paper we propose a class of networks which may perform such tasks. Repeller neural networks (RNN) can be, for instance, realized in the standard framework of ANN. Binary neurons follow then in the absence of noise a standard updating rule

$$\sigma_i(t+1) = \mathrm{sgn}\left[\sum_{j(\neq i)} J_{ij}\sigma_j(t)\right], \tag{1}$$

where $\sigma_i(t) = \pm 1$ denote their states, J_{ij}'s are connections, and $i = 1,\ldots,N$, where N is the number of network elements. We shall discuss here mainly synchronous realizations of the dynamics described by Eq. (1).

The difference between RNN and ANN lies in the learning strategy. In the standard ANN one constructs J_{ij} so that desired patterns become stationary. For Hamiltonian models, such as a Hopfield one, learning is sometimes described as "digging holes in the energy landscape." We propose an alternative approach to learning that consists in forming "hills" and "mountains" in the energy landscape. Patterns that correspond to the "tops of the hills" will become unstable repellers and will be substantially changed within one Monte Carlo (MC) step per neuron. This, however, is not sufficient, since our task is to control not only which states are *rejected*, i.e., strongly changed in the course of dynamics, but also to control which are *accepted* (i.e., not changed or weakly changed). To this aim we "fill" the energy landscape with the ground creating large regions of the flat landscape. This is achieved by introducing a term describing self-supportiveness in Eq. (1),

$$\sigma_i(t+1) = \mathrm{sgn}\left[s\sigma_i(t) + \sum_{j(\neq i)} J_{ij}\sigma_j(t)\right], \tag{2}$$

where $0 \leq s \leq 1$ is a new control parameter. Self-supportiveness has been extensively used in various models of cellular automata [5]. We have, for instance, used it in our theory of public opinion formation [6,7] to describe a natural tendency of individuals to keep their opinions.

With this new control parameter the networks will act as follows. First, they will not change or *within a finite time* change very weakly accepted states, i.e., that are sufficiently different from *rejected* configurations.

On the other hand, they will *reject*, i.e., change abruptly, states which are sufficiently similar to a given set of "unlearned" patterns. Note that we follow here the idea of Refs. [8, 9], in which an unlearning strategy was used to minimize the role of spurious memories in the network.

How do we detect answers from the repeller networks? This can be done using two methods

(1) First, by looking at the overlap of the initial state $\{\sigma_i(0)\}$ with the "final" state $\{\sigma_i(\tau)\}$. Note that we con-

*Permanent address: Centrum Fizyki Teoretycznej, Polskiej Akademii Nauk, Al. Lotników 32/46, 02–668 Warsaw, Poland.

sider here on purpose finite duration τ of the dynamics. In this way we are able to distinguish patterns simply lying or slowly moving in the flat regions of the energy landscape, for which $q = \sum_i \sigma_i(0)\,\sigma_i(\tau)/N$ should be close to 1, from those lying on "tops of the hills" for which $q \ll 1$. Of course, our criterion is a convention only, but as numerical simulations show it can be used in an unambiguous way (see discussion below).

(2) An alternative method of detection that can be well applied for not too large noise levels consists in monitoring frequency of the neural flips in the initial phase of the dynamics. We have developed such a method of detection in the context of "nervous" neural networks [10, 11], and showed that it provides a very precise detection tool already within a fraction of the MC step per neuron.

Let us now turn to the theory and start with the case $s = 0$, i.e., with no self-supportiveness. Throughout this paper we shall use the Hebbian rule for constructing J_{ij}'s but any other rule or algorithm could be used equally well. The inversion of the energy landscape consists in taking a minus sign in front of the standard expression

$$J_{ij} = -\frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu, \tag{3}$$

where $\xi_i^\mu = \pm 1$, for $\mu = 1, \ldots, p$, and $p = \alpha N$, denote random rejected or unlearned patterns. For $s = 0$ the model is Hamiltonian and its statistical mechanics can be investigated along the lines of the seminal paper of Amit, Gutfreund, and Sompolinsky [12]. We use the replica method [13] and do not brake the replica symmetry. The order parameters are the same as in Ref. [12], i.e., the overlaps with ξ's, $m_\nu = \langle\langle \sum_i \xi_i^\nu \langle \sigma_i \rangle_T \rangle\rangle / N$ for $\nu = 1, \ldots, k$ and some finite k, the Edwards–Anderson order parameter $q = \langle\langle \sum_i \langle \sigma_i \rangle_T^2 \rangle\rangle / N$, and the parameter $r = \langle\langle \sum_{\mu=k+1}^{p} m_\mu^\varrho m_\mu^\sigma \rangle\rangle / N$, which measures Gaussian correlations between the noise in the local field for distinct replicas ϱ and σ. The symbol $\langle\,\rangle_T$ denotes here a thermal average, whereas $\langle\langle\,\rangle\rangle$ denotes an average over statistics of ξ's.

Equations of state have a form similar to that in Ref. [12]. We obtain thus for each $\nu = 1, \ldots, k$

$$m_\nu = \left\langle\left\langle i\xi^\nu \tanh\left[\beta\left(i\sum_{\nu'} m_{\nu'}\xi^{\nu'} + rz\right)\right]\right\rangle\right\rangle, \tag{4}$$

$$q = \left\langle\left\langle \tanh^2\left[\beta\left(i\sum_{\nu'} m_{\nu'}\xi^{\nu'} + rz\right)\right]\right\rangle\right\rangle, \tag{5}$$

$$r = \frac{\alpha q}{1 + \beta(1-q)}. \tag{6}$$

Double brackets refer now to averaging over the normally distributed Gaussian variable z, with the mean zero and variance 1, while i is the imaginary unit; $\beta = 1/T$ denotes the noise level.

It is easy to check that the above equations allow only for solutions with $m_\nu = 0$ for each ν. There are two stable phases: a paramagnetic phase and a spin glass phase. The paramagnetic phase is stable for $T \geq 0$ when $\alpha \leq 1$ and for $T \geq \sqrt{\alpha} - 1$ otherwise. This result might be interesting in itself, but unfortunately does not allow for using this Hamiltonian model as a repeller neural network. For $\alpha \leq 1$ we have no control over which states are slowly or rapidly varying. In the spin glass phase there are a lot of stationary states and slowing down of the dynamics. But, again, we do not have any control over which states are slowly varying and which are not.

Therefore we turn to the case $s \neq 0$. That is, however, a much more difficult task, since the problem becomes non-Hamiltonian and has to be treated dynamically. We have only succeeded in solving it in two cases: for $\alpha = 0$ and arbitrary T, and in the limit of strongly diluted networks with connectivity $K \ll N$, but for $K \to \infty$. In the latter case we have used the theory of Derrida, Gardner, and Zippelius [14].

For $\alpha = 0$, i.e., for finite p, the only order parameters are m_ν's, which self–average. The dynamics takes the form

$$\sigma_i(t+1) = \operatorname{sgn}\left[s\sigma_i(t) - \sum_{\mu=1}^{p} \xi_i^\mu m_\mu(t) + r_i(t)\right], \tag{7}$$

where $r_i(t)$ denotes a white noise distributed in such a way that $\langle \operatorname{sgn}[a + r(t)] \rangle_r = \tanh(\beta a)$. Let us consider first the case when only one of the m_μ's, say m_1 (denoted below as m), is nonzero and the noise vanishes. Equation (7) then becomes

$$\sigma_i(t+1) = \sigma_i(t)\,\Theta(s - |m(t)|) - \xi_i^1 \operatorname{sgn}[m(t)]\,\Theta(|m(t)| - s), \tag{8}$$

where $\Theta(\,)$ denotes the unit step function. It indicates that σ's do not change if $|m(t)|$ is small enough, i.e., when a current configuration has sufficiently small overlap with ξ^1 (and zero overlaps with other ξ's). If $|m(t)| > s$, σ's follow periodic oscillations changing their sign in each step of the synchronous dynamics. Here we have what we wanted: all states that are sufficiently different from ξ^1 and other ξ's are stationary. All others undergo maximal possible change in each MC step.

In the presence of the noise the equation for the order parameter takes the form

$$m(t+1) = \frac{m(t)}{2}\Big(\tanh\{\beta[s - m(t)]\} + \tanh\{\beta[s + m(t)]\}\Big) + \frac{1}{2}\Big(\tanh\{\beta[s - m(t)]\} - \tanh\{\beta[s + m(t)]\}\Big). \tag{9}$$

It is easy to check that the only stable fixed point solution then corresponds to $m = 0$. One has to conclude that the only stationary state is a paramagnetic state. But, we stress that the model is not trivial, since it exhibits a *dynamical phase transition.* Although the stationary state does not change, its character changes from a stable node to a stable focus as we increase the noise level $1/\beta$. For small $m(t)$, Eq. (9) becomes $m(t+1) = \Gamma m(t)$, with

$$\Gamma = \tanh(\beta s) - \frac{\beta}{\cosh^2(\beta s)}. \tag{10}$$

We see that for large values of β, Γ differs from one only

by an exponentially small correction. The dynamics in this limit is very slow, and the reader will easily check that this takes place for finite $m(t)$ already, provided it is smaller than s. In the high-noise limit ($\beta \simeq 0$), Γ becomes negative and equal to $\simeq \beta(s-1)$. This is the regime of "repeller" paramagnetic phase for which large changes of the state occur in every step of the dynamics. This phase is of course reminiscent of the $T=0$ periodic state. The critical value of the control parameter β is determined from $\sinh(2\beta_c s) = 2\beta_c$. For s close to 1, $\beta_c \simeq \sqrt{\frac{3}{2}(1-s)}$, whereas for small s, $\beta_c \simeq -\ln(s)/2s$.

Note that when s is sufficiently small [$s \leq s_c$ with $s_c + 1 = \ln(1/s_c)$], the fixed point $m=0$ might lose its stability and the stable solution becomes a period 2 cycle, just as in the case of $T=0$. This happens when Γ becomes equal to -1, i.e., for $\beta_1 \leq \beta \leq \beta_2 \leq \beta_c$, where $2\beta_i = 1 + \exp(2\beta_i s)$ for $i = 1, 2$. Since the period 2 state is not of particular interest from a point of view of applications of RNN, we limit our discussion to the case of larger values of s only.

The dynamical phase transition that we encounter here is very characteristic for systems with self-supportiveness [6, 7]. It does not mean much if we restrict our attention to the final states of the dynamic only. For finite times, however, it is always possible to take β and s sufficiently large that within a prescribed accuracy the desired class of states will not change appreciably, whereas the rest will be rejected. The results discussed here have been confirmed in numerical simulations, performed by us with networks consisting of 200 elements [15]. In the simulations we have introduced random repeller states with the help of the negative of the Hebbian connection matrix. As initial state we took partially deformed repeller states characterized by a macroscopic initial overlap with the ancestor state q_0 and practically vanishing overlaps with other repellers. We have performed the simulations several hundred times for a duration of ten Monte Carlo steps per neuron. Simulations were performed at $T=0$. We measured the overlap q of the final state with the initial state as a function of q_0 for several values of the self-supportiveness s (see Fig. 1). For small q_0, $q \simeq 1$ as

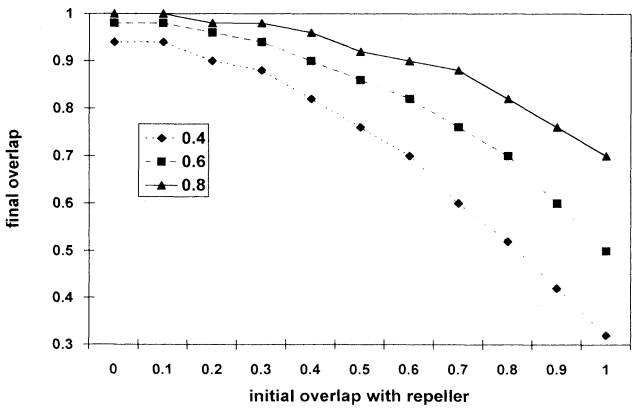

FIG. 1. The overlap q of the final state with the initial state as a function of the overlap q_0 of the initial state with the ancestor state.

we expect. For $q_0 \geq s$, q becomes gradually decreasing and becomes close to s as q_0 reaches 1. The dependence has a very well developed knick at s as it changes from the constant to monotonically decreasing. This allows us to introduce practically unambiguous criterion of distinction of accceptable and unacceptable states.

Obviously, the theory may be easily generalized to the case of two or more nonvanishing overlaps m_ν. Particularly simple is the case of two nonvanishing overlaps, which were equal initially. Such symmetric states remain symmetric in the course of dynamics. Let us introduce four dynamical order parameters:

$$m_{\pm,\pm}(t) = \frac{1}{N} \sum_{\xi_i^1 = \pm, \xi^2 = \pm} \sigma_i(t). \tag{11}$$

Obviously, $m_1 = m_{++} - m_{--} + m_{+-} - m_{-+}$, whereas $m_2 = m_{++} - m_{--} - m_{+-} + m_{-+}$. Initially $m_1 = m_2 = m(0)$, so that $m_{++}(0) - m_{--}(0) = m(0)$, and $m_{+-} = m_{-+}$. Denoting $m(t) = m_{++}(t) - m_{--}(t)$, $\tilde{m}(t) = m_{++}(t) + m_{--}(t)$, $n(t) = m_{+-}(t) - m_{-+}(t)$, $\tilde{n}(t) = m_{+-}(t) + m_{-+}(t)$, we obtain the following dynamical equations, which generalize Eq. (9) to the case of two nonvanishing symmetric overlaps:

$$\begin{aligned} m(t+1) = \frac{m(t)}{2}\big(&\tanh\{\beta[s-2m(t)]\} \\ &+ \tanh\{\beta[s+2m(t)]\}\big) \\ +\frac{1}{4}\big(&\tanh\{\beta[s-2m(t)]\} \\ &- \tanh\{\beta[s+2m(t)]\}\big), \end{aligned} \tag{12}$$

$$\begin{aligned} \tilde{m}(t+1) = \frac{\tilde{m}(t)}{2}\big(&\tanh\{\beta[s-2m(t)]\} \\ &+ \tanh\{\beta[s+2m(t)]\}\big), \end{aligned} \tag{13}$$

$$\begin{aligned} n(t+1) = \frac{n(t)}{2}\big(&\tanh\{\beta[s-2n(t)]\} \\ &+ \tanh\{\beta[s+2n(t)]\}\big) \\ +\frac{1}{4}\big(&\tanh\{\beta[s-2n(t)]\} \\ &- \tanh\{\beta[s+2n(t)]\}\big), \end{aligned} \tag{14}$$

$$\begin{aligned} \tilde{n}(t+1) = \frac{\tilde{n}(t)}{2}\big(&\tanh\{\beta[s-2n(t)]\} \\ &+ \tanh\{\beta[s+2n(t)]\}\big). \end{aligned} \tag{15}$$

One easily sees that $n(t) = \tilde{n}(t) = 0$. $\tilde{m}(t)$ decays also to zero. For large values of β that decay is relatively fast for $m(t) \geq s/2$. As soon as $m(t)$ decreases below $s/2$, the rate of decay of $\tilde{m}(t)$ becomes exponentially small. $m(t)$ tends also toward zero, but its decay will become exponentially slow for $m(t) \leq s/2$. The rate in the limit of $m(t) \simeq 0$ is given by the same expression as in Eq. (10). We encounter again the same kind of dynamical phase transition as in the previously discussed case that occurs at the same critical value β_c. The only difference is that in the case of single overlap the "accepted," i.e., slowly varying states should have $m(t) \leq s$, whereas now

"accepted" states have to be more strongly repelled from the two repellers ξ^1 and ξ^2 and have $m(t) = m_1(t) = m_2(t) \leq s/2$.

Let us now turn to the other limit—the strong dilution of the network. Generalization of the theory of Derrida, Gardner, and Zippelius [14] to the case of networks with self-supportiveness has been recently done by us [6, 7]. We consider here a network with connectivity $K \ll N$, but with $K \to \infty$. We restrict our discussion to the case of $T = 0$, since extension of this theory to the case of finite T is straightforward. We shall also restrict our discussion to configurations of the network that have a nonvanishing overlap with one and only one of the αK *unlearned* patterns, say ξ. We assume therefore that $\sigma_i(t)$ are uncorrelated for different i and that $\langle\langle \xi\sigma_i(t) \rangle\rangle = m(t)$. The local field for $K \to \infty$ becomes a Gaussian process with the mean $m(t)$ and the dispersion α. The dynamics becomes thus

$$m(t+1) = m(t)\langle\langle\theta(s - |m(t) + \sqrt{\alpha}z|)\rangle\rangle - \langle\langle \mathrm{sgn}\left[m(t) + \sqrt{\alpha}z\right]\theta(|m(t) + \sqrt{\alpha}z| - s)\rangle\rangle, \tag{16}$$

where $\langle\langle\ \rangle\rangle$ denotes average over normally distributed variable z. Again, the only stable solution of Eq. (16) is $m = 0$. But we encounter once more the similar dynamical phase transition, since for small m, $m(t+1) = \Gamma m(t)$, with

$$\Gamma = \mathrm{erf}\left(\frac{s}{\sqrt{2\alpha}}\right) - \left(\frac{2}{\pi\alpha}\right)^{1/2} \exp\left(\frac{-s^2}{2\alpha}\right). \tag{17}$$

Once more, for α small (more precisely for $s/\sqrt{2\alpha} \gg 1$) Γ differs from 1 only exponentially. In this regime we deal with quasistable states. In the limit of large α we enter the phase of the stable focus which is reminiscent of the $\alpha = 0$ periodic orbit. Critical values of α are $\alpha_c \simeq 1/[6(1-s)]$ for s close to 1, and $\alpha_c \simeq -s^2/\ln s$ for s close to zero. The limit $s \to 1$ is especially interesting since we may then formally "unlearn" an unlimited number of patterns in the repeller network. However, in order to make the repeller network work [i.e., make $m(t+1)/m(t) \simeq 1$] we have to take $s/\sqrt{2\alpha} \leq 1$, i.e., for all practical purposes the critical value of α_c for $s \to 1$ is of the order of 1. Nevertheless, for small enough α and finite duration time of the dynamics we may always be sure that states with $m \leq O(s)$ will not change appreciably, i.e., will be detected as *accepted* states. Again, for small s the fixed point $m = 0$ might become unstable and the stationary state becomes periodic, but we skip the discussion of this regime for the same reasons as in the case $\alpha = 0$.

In concluding we would like to point out that with our model we have formulated a paradigm of network modeling, which we hope will turn out very useful for negative-classification tasks. The major innovation of repeller neural networks consists in simultaneous introduction of unlearning and self-supportiveness. In the standard attractor networks it is important to control the size of the basins of attraction of the remembered patterns. In the present case of RNN it is equally important to be able to control the size of basins of repulsion of unlearned patterns, i.e., to control the size of the regions of the configuration space that are unacceptable and should be rejected by the networks. Self-supportiveness allows for such control in a very efficient way.

This paper has been financed by the KBN Grant Nos. 2-2417-92-03 and 1-1113-92-02. M.L. thanks Centre d'Etudes de Saclay (SPAM) for hospitality and financial support.

* Permanent address: Centrum Fizyki Teoretycznej, Polskiej Akademii Nauk, Al. Lotników 32/46, 02-668 Warsaw, Poland.

[1] D. J. Amit, *Modeling Brain Function: The World of Attractor Neural Networks* (Cambridge University Press, Cambridge, England, 1989).

[2] M. Minsky and S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969). The new edition is dated 1989.

[3] *Parallel Distributed Processing: Explorations in Microstructure of Cognition*, edited by D. E. Rumelhart and J. L. McClelland (MIT Press, Cambridge, Mass., 1986) Vols. I and II.

[4] *Proceedings of the Workshop on "Neural Networks: From Biology to High Energy Physics,"* edited by O. Benhar, C. Bosio, P. Del Guidice and E. Tabet (Editrice, Pisa, 1992).

[5] *Theory and Applications of Cellular Automata*, edited by S. Wolfram (World Scientific, Singapore, 1986).

[6] A. Nowak, J. Szamrej, and B. Latané, Psych. Rev. **97**, 362 (1990).

[7] M. Lewenstein, A. Nowak, and B. Latané, Phys. Rev. A **45**, 763 (1992).

[8] J. J. Hopfield, D. I. Feinstein, and R. G. Palmer, Nature (London) **305**, 159 1983.

[9] J. L. van Hemmen and R. Kühn, in *Models of Neural Networks*, edited by E. Domany, J. L. van Hemmen, and K. Schulten (Springer-Verlag, Berlin, 1992).

[10] M. Lewenstein and A. Nowak, Phys. Rev. Lett. **62**, 225 (1989).

[11] M. Lewenstein and A. Nowak, Phys. Rev. A **40**, 4652 (1989).

[12] D. J. Amit, H. Gutfreund, and H. Sompolinsky, Phys. Rev. A **32**, 1007 (1985); Ann. Phys. (N.Y.) **173**, 30 (1987).

[13] *Spin Glass Theory and Beyond*, edited by M. Mezard, G. Parisi, and M. Virasoro (World Scientific, Singapore, 1987).

[14] B. Derrida, E. Gardner, and A. Zippelius, Europhys. Lett. **4**, 167 (1987).

[15] A. Nowak, M. Lewenstein, and W. Tarkowski (unpublished).