A class of nonlinear systems with memory

A. A. Vedenov and E. B. Levchenko
I. V. Kurchatov Institute of Atomic Energy, Moscow

(Submitted 27 February 1985)
Pis'ma Zh. Eksp. Teor. Fiz. **41**, No. 8, 328–331 (25 April 1985)

Dynamic equations which follow from a simple Lagrangian describe phenomena which may be interpreted as learning, pattern recognition, bistability of perception, forgetting, and the development of a prototype.

We consider a dynamic system whose elements ("neurons") interact with each other and with a heat reservoir. The state of the i-th neuron ($i = 1, \ldots, N$) in the system is described by the variable $\sigma_i(t)$ (t is the time), which has the range

 0021-3640/85/080402-05$01.00

$-\sigma_0 \leqslant \sigma_i(t) \leqslant +\sigma_0$. The state with $\sigma_i = +\sigma_0$ corresponds to an "excited" neuron, while that with $\sigma_i = -\sigma_0$ corresponds to an "unexcited" neuron. We consider the quantity

$$L = -E = \frac{1}{2}\lambda \sum_{ij} T_{ij}\sigma_i\sigma_j \tag{1}$$

($\lambda > 0$ is a parameter) as the Lagrangian of the interaction of the system, a functional of the independent variables σ and T. Taking into account the interaction of the neurons with the heat reservoir, which gives rise to "friction forces" ($\dot{\sigma}/\gamma_1$, $\dot{T}/\gamma_2$), we find from (1) the following dynamic equations for σ and T:

$$\frac{1}{\gamma_1}\dot{\sigma}_i - f_i(\sigma) = \frac{\partial L}{\partial \sigma_i} = \lambda \sum_j T_{ij}\sigma_j, \tag{2}$$

$$\frac{1}{\gamma_2}\dot{T}_{ij} - F_{ij}(T) = \frac{\partial L}{\partial T_{ij}} = \frac{1}{2}\lambda\sigma_i\sigma_j. \tag{3}$$

The f and F terms have been added to these equations to prevent an unbounded increase in the absolute values of the variables σ and T. In the Lagrangian approach, they can be incorporated into (1) as potentials that increase rapidly near $|\sigma_i| = \sigma_0$, $|T_{ij}| = T_0$.

Two independent equations were postulated in Refs. 1–6, one describing the linear conversion of the "activity field" σ of the neurons caused by the "memory operator" T, and a second describing the change in the operator T, which is quadratic in σ. Anderson *et al.*[6] analyzed the relaxation of the initial value of σ to one of the corners (at the coordinates $\pm 1, \ldots, \pm 1$) of an N-dimensional cube. Hopfield *et al.*[1] studied this relaxation with a given matrix T and introduced the concept of an energy of the system, which decreases in the course of the relaxation.

In the present letter we take the Lagrangian approach to derive in a common way the σ relaxation equation (2) and the memory change equation (3). A further purpose is to demonstrate an analogy with various physical systems.

The dynamics of a system described by Eqs. (2) and (3) depends strongly on its history. We first consider the evolution of the variables σ and T under some special conditions: in a "learning" process. This learning can be summarized by saying that in (2) we incorporate a strong external field, which acts for a time t_1. As a result, the vector σ takes on a steady-state value φ^1, which corresponds to a "pattern" with components $\pm\sigma_0$. As a result of this learning, the matrix T_{ij} changes by an amount $\Delta T_{ij} = \gamma_2\lambda t_1 \varphi_i^1\varphi_j^1$ according to (3) [we are assuming that t_1 is considerably longer than the relaxation time of the vector $\sigma(t)$ to its steady-state value φ^1 in the external field]. This learning procedure can be repeated many times with the patterns φ^s, $s = 1, \ldots, s_0$. Assuming, for simplicity, that the condition $T_{ij} = 0$ holds before the learning begins, we find the following result after the procedure has ended:

$$T_{ij} = \sum_s \mu^s \varphi_i^s \varphi_j^s, \tag{4}$$

where the coefficients μ^s depend on the duration of the learning.

It follows from (2) and (3) [with (4)] that in the absence of an external field the scale relaxation times of the variables σ ($\tau_\sigma \sim (\gamma_1 \lambda N s_0)^{-1}$) and T ($\tau_T \sim (\gamma_2 \lambda s_0)^{-1}$ are very different: $\tau_\sigma \ll \tau_T$. Over times on the order of τ_σ, the dynamics of the system of equations (2) and (3) can be interpreted as a "pattern recognition" process; i.e., the "initial stimulus" (a vector chosen as an initial condition) relaxes in accordance with (2) to one of the "memory" vectors φ^s.

Dynamic system (2), (3) has been studied by the method of discrete stochastic simulation,[1)] specifically, by the procedure of Metropolis *et al.*[7] (the model of a system in a heat reservoir) and the algorithm proposed by Hopfield.[1] The continuous variables σ_i are replaced by discrete variables which take on the values $\sigma_{i0} = \pm 1$, and the dynamics of the system is specified in one of several ways: by means of a "spin-flip" probability, in accordance with the sign of the "molecular field"[1] $T\sigma$, or in accordance with the change in the energy of system (1) at a given temperature.[7]

In the discrete model, the state with $\sigma = \varphi^s$ corresponds to a local minimum of the energy of the system (this energy decreases as σ evolves), and it is stable if the number of patterns, s_0, is small. We set $\sigma^s = \varphi^s + \delta^s$, and we consider a variation of the energy (1) ($\sigma_0 \equiv 1$, $\mu^r \equiv 1$):

$$\frac{1}{\lambda}\delta E = -N\varphi^s\delta^s - \sum_{r \neq s}(\varphi^r\varphi^s)(\varphi^r\delta^s) - \frac{1}{2}\sum_r(\varphi^r\delta^s)^2, \tag{5}$$

where $\varphi\psi = \Sigma\varphi_i\psi_i$. If the different φ^s are approximately orthogonal, the second term in (5) is small in comparison with the first, and for a variation of δ^s, which corresponds to a reversal of one of the symbols ± 1 in pattern φ^s, the first term will be $2N$, and the third will be $2s_0$. We thus find $\delta E > 0$—the pattern is stable—if $s_0 \lesssim N$. This estimate agrees qualitatively with the results of an analysis[1] of the errors of recognition by means of matrix (4).

In a discrete simulation, the stable states of system (2), (3) may be vectors which are not the same as φ^s (Refs. 1 and 8). For example, if a group of patterns obtained upon slight random distortions of some vector φ_0 (but which does not contain φ_0) is written in matrix T in (4), then an analysis similar to that of (5) shows that φ_0 may be a stable state of the system, i.e., may be a "prototype"[2)] (Ref. 9). In a simulation by the algorithm of Ref. 1, "false patterns" φ^* may also turn out to be stable (with $s_0 = 3$, for example, φ^* is stable if the quantities $p_i = \varphi^*\varphi^i > 0$ satisfy the triangle inequalities $p_1 \leqslant p_2 + p_3$, $p_2 \leqslant p_1 + p_3$, $p_3 \leqslant p_1 + p_2$), which vanish with increasing temperature in the method of Metropolis *et al.*[7]

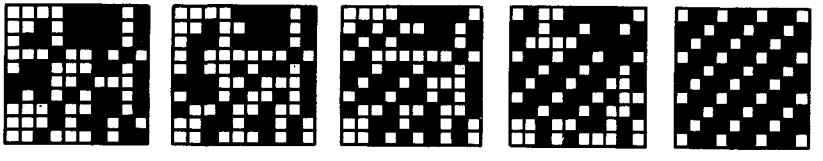

FIG. 1. Sequence of states of a system of neurons which arises during the relaxation of the "initial stimulus" (the picture at the left) to a "pattern" (that at the right) according to a discrete simulation of Eq. (2) by the algorithm of Ref. 1.

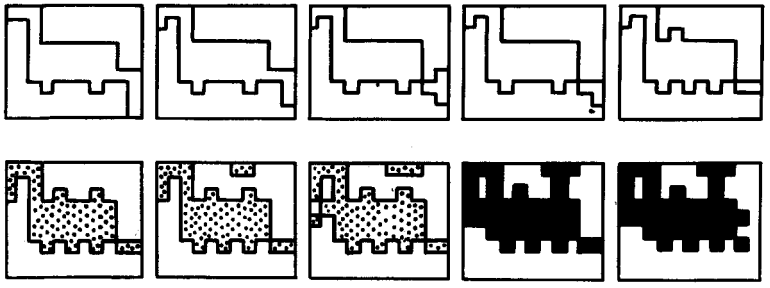

FIG. 2. Sequence of "stimuli" presented as an initial condition in the discrete simulation of (2) and (3). In the absence of "learning" ($\gamma_2 = \epsilon = 0$), the stable state changes after the fifth picture is presented. A sufficiently large modification of the matrix T ($\epsilon = 0.025$) results in a hysteresis: The change in patterns as we go from the "pangolin" to the "steam engine" occurs in the eighth picture, while it occurs in the third picture when we go in the opposite direction.

Let us assume that two orthogonal patterns separated by a Hamming distance (the number of symbols $+1$ or -1 which distinguish one pattern from the other) of m are written in T, and let us choose as an initial condition for system (2), (3) the elements of a sequence of vectors (Fig. 2), whose first term coincides with the first ("left-handed") pattern and whose last term coincides with the second ("right-handed") pattern. The numerical simulation of the recognition process with learning [after each event in which a stimulus is recognized as a distorted pattern φ^s, a term $\epsilon\varphi_i^s\varphi_j^s$ is added to the matrix T in accordance with (3)] shows that the intermediate stimuli of the sequence may be perceived as either a left-handed or right-handed pattern, depending on the order in which the stimuli are presented; i.e., a hysteresis occurs. There is an analogous phenomenon in psychology: a hysteresis of visual perception.[10] We can estimate the width (x) of the hysteresis region by assuming that at the boundary of this region the energy in (1) corresponding to the interaction of the stimulus with a left-handed pattern is increased by a factor of $[1 + \epsilon(m/2 + x/2)]$, which is equal to the initial energy corresponding to the interaction of the stimulus with the right-handed pattern:

$$\left[1 + \epsilon\left(\frac{m}{2} + \frac{x}{2}\right)\right]\left[N - 2\left(\frac{m}{2} + \frac{x}{2}\right)\right]^2 = \left[N - 2\left(\frac{m}{2} - \frac{x}{2}\right)\right]^2 .$$

For $x \ll N_1 \equiv N - m$ we then find the half-width of the hysteresis region to be $x = m/[(8/\epsilon N_1) - 1]$. The hysteresis spans the entire sequence of stimuli ($x = m$) in the case $\epsilon \geqslant 4/N_1$, or it contracts to a single figure (with decreasing N_1 at constant values of ϵ and m) at $N_1 = 8/\epsilon m$.

The restriction on T_{ij} which follows from the nonlinear term in (3) has the consequence that newly written patterns reduce the relative weight of the old patterns and also cause a partial distortion of these old patterns (the old patterns are "forgotten," by analogy with the psychological phenomenon[9]). Treating the increase in x as a widening of a "funnel" which is being entered by stimuli close to some pattern (e.g., a left-handed pattern), we can write an approximate equation to estimate the width of the funnel for a left-handed pattern upon the presentation of random stimuli at a frequen-

cy Ω: $\dot{x} = \epsilon\Omega(2x/m)$, where m is the Hamming distance to the nearest pattern. The funnel thus widens exponentially over time, causing a displacement of one pattern by another.

There is a qualitative analogy between the phenomena described here and the modulational instability (or collapse) of a turbulent plasma[11] and also with certain phenomena in nonlinear optics, which are described by a system of equations similar to (2) and (3).

[1)]The numerical calculations were carried out by A. A. Ezhov, L. A. Knizhnikova, and Yu. G. Chernov, to whom the author expresses his deep gratitude.

[2)]The possible appearance of a "prototype" during a discrete simulation by the algorithm of Ref. 1 has been demonstrated by A. A. Ezhov and L. A. Knizhnikova.

[1]J. J. Hopfield, Proc. Nat. Acad. Sci. USA **79**, 2554 (1982); **81**, 3088 (1984); J. J. Hopfield, D. J. Feinstein, and R. G. Palmer, Nature **304**, 158 (1983).

[2]L. Ingber, Physica (Utrecht) **5D**, 83 (1982); Phys. Rev. **8A**, 395 (1983); **9A**, 3346 (1984).

[3]W. A. Little, Math. Biosci. **19**, 101 (1974); W. A. Little and G. L. Show, Math. Biosci. **39**, 281 (1978); D. S. Levine, Math. Biosci. **66**, 1 (1983).

[4]P. Peretto, Biol. Cybern. **50**, 51 (1984).

[5]J. W. Clark, J. V. Winston, and J. Rafelski, Phys. Lett. **102A**, 207 (1984).

[6]J. A. Anderson, J. W. Silverstein, S. A. Ritz, and R. S. Jones, Psych. Rev. **84**, 413 (1977); G. E. Hinton and J. A. Anderson (editors), Parallel Models of Associative Memory, New York, 1981.

[7]N. Metropolis *et al.*, J. Chem. Phys. **21**, 1087 (1953); S. Kirkpatrick, C. D. Gellat, and M. P. Vecchi, Science **220**, 671 (1983).

[8]F. C. Crick and G. Mitchison, Nature **304**, 111 (1983).

[9]R. Klacky, Human Memory: Processes and Structures (Russ. transl. Mir, Moscow, 1978).

[10]T. Poston and I. Stewart, Catastrophe Theory and Its Applications, Fearon Pitman, Belmont, Calif., 1978 (Russ. Transl. Mir, Moscow, 1980).

[11]M. V. Goldman, Rev. Mod. Phys. **56**, 709 (1984).

Translated by Dave Parsons