Hindawi Publishing Corporation
Computational Intelligence and Neuroscience
Volume 2015, Article ID 191745, 7 pages
http://dx.doi.org/10.1155/2015/191745

Research Article

Memory Dynamics in Attractor Networks

Guoqi Li,[1] Kiruthika Ramanathan,[2] Ning Ning,[2] Luping Shi,[1] and Changyun Wen[3]

[1] *Centre for Brain Inspired Computing Research (CBICR), Department of Precision Instrument, Tsinghua University, Beijing 100084, China*
[2] *Department of Advanced Concepts and Nanotechnology (ACN), Data Storage Institute, A*STAR, 5 Engineer Drive 1, Singapore 117608*
[3] *School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798*

Correspondence should be addressed to Guoqi Li; liguoqi@mail.tsinghua.edu.cn

Received 25 November 2014; Revised 30 January 2015; Accepted 30 January 2015

Academic Editor: Klaus Obermayer

As can be represented by neurons and their synaptic connections, attractor networks are widely believed to underlie biological memory systems and have been used extensively in recent years to model the storage and retrieval process of memory. In this paper, we propose a new energy function, which is nonnegative and attains zero values only at the desired memory patterns. An attractor network is designed based on the proposed energy function. It is shown that the desired memory patterns are stored as the stable equilibrium points of the attractor network. To retrieve a memory pattern, an initial stimulus input is presented to the network, and its states converge to one of stable equilibrium points. Consequently, the existence of the spurious points, that is, local maxima, saddle points, or other local minima which are undesired memory patterns, can be avoided. The simulation results show the effectiveness of the proposed method.

1. Introduction

Memory is a fundamental component of our human brain; how to simulate the human memory process has attracted many scientists attention in the research of cognitive systems and architectures [1, 2]. Attractor networks [3–6] have been one of the most popular models for memory storage and retrieval in recent decades since the hypothesis of attractor dynamics is supported and observed in the neocortex and hippocanpus in various memory experiments [6–9]. In general, an attractor network is a network of recurrently connected nodes in a biological network, whose states may settle to some stable patterns. One distinguish advantage is that the network can be represented by neurons and their synaptic connections. The particular pattern of such a recurrent network, which is called its "attractor" [10–13], can be stationary, time varying, or even stochastic. In theoretical neuroscience, different kinds of attractor neural networks have been associated with different functions, such as memory, motor behavior, and classification. In this paper, we consider the patterns or the so-called attractors as the stationary memory patterns stored in the dynamic system, which allows us to employ methods in dynamical systems to quantitatively analyze the characteristics such as the stability and robustness of the network.

Usually researchers design an attractor network and then propose its energy function, often called Lyapnov function, to analyze the network [14–16], since the energy function plays a very important role in analyzing the network stability and robustness. In this paper, we design the memory dynamics in an opposite direction. We first propose an energy function and then design the attractor network. The energy function actually contains the desired memory patterns we wish to store. Different from Hopfield networks [10, 14], we introduce multiplicative algebra into the attractor network. This is biologically possible since both addition algebra and multiplication algebra are the simplest and the most widespread of all operations in the nervous system [17]. In addition, the value of our proposed energy function is nonnegative and attains zero values only at the patterns stored in the network. This makes it easy to distinguish the desired memory patterns from some other possible undesired patterns which are called spurious points [18]. It is shown that the memory patterns are stored as the stable equilibrium points of the dynamical attractor networks, which are also the local minimum points of the energy function. Compared with existing results in attractor networks [7, 10, 14], the patterns are not necessarily binary and uncorrelated in this paper. Binary patterns simplify the network design significantly as seen in [14]. Also, the uncorrelated patterns give a minimum interactions of the interactions of the network, which makes its behavior analysis much easier as seen in [7, 19].

On the other hand, when a stimulus pattern is presented as an initial input of the dynamic system, the states of the system will converge to a particular stored attractor iteratively. This process is called associative memory retrieval. The "associative" means that the memory we retrieved was by its informational content rather than by names, addresses, or relative positions. One very touchy problem in the retrieval process is how to overcome the problem that the system states may converge to spurious points; see, for example, [18, 20]. Usually, there are two kinds of equilibrium points for a dynamic system, the stable ones and unstable ones. It is shown that only the stable equilibrium points, that is, the local minimum points, of the energy function exist in the proposed designed system. We also prove that those local minima are only the memory patterns stored in the network. Thus, the spurious points, that is, local maxima, saddle points, or other local minima which are the undesired memory patterns, can be avoided.

The contributions of this paper are summarized as follows. Firstly, we have proposed a new energy function different from the energy function in Hopfield networks. This makes it easy to differentiate memory patterns from possible spurious points. Secondly, we have presented an attractor network design based on the proposed energy function. The patterns stored in the attractor network can be nonbinary and either correlated or uncorrelated. Finally, we have proven that, when an arbitrary input stimulus is presented to the designed attractor network, the states converge to one of the stored patterns. This implies that there are no spurious points in the designed dynamical systems.

The rest of the paper is organized as follows. Some background knowledge is reviewed in Section 2. Section 3 introduces the main design method of the proposed attractor network. The convergence properties are analyzed in Section 4. The simulation examples are shown in Section 5. Finally, the paper is concluded in Section 6.

2. Background Knowledge

2.1. Multiplication Algebra in Nervous Systems. Addition algebra is both the simplest and one of the most widespread of all operations in nervous systems. However, as pointed in [17], a number of biological mechanisms could, in theory, implement a multiplication algebra. Actually multiplication can be implemented based on addition. For example, when we multiply two signals x and y, we can logarithmically transform the two, add the result,and then apply an exponential:

$$xy = \exp^{(\log(x)+\log(y))}. \tag{1}$$

Thus, later it can be seen that the memory dynamics in our proposed attractor network could be implemented in a network with neurons and synapses.

2.2. Notations and Definitions. Denote a nonlinear function $f(\mathbf{x})$ and a pattern $\mathbf{x} = [x_1 \ \cdots \ x_m]^T \in R^m$ where m is the dimension. For a nonlinear dynamical system $\dot{\mathbf{x}} = -f(\mathbf{x})$, we have the following definitions.

Definition 1. A pattern $\mathbf{x}^0 \in R^m$ is called an equilibrium point of $\dot{\mathbf{x}} = -f(\mathbf{x})$, if $f(\mathbf{x})$ is a zero vector at $\mathbf{x} = \mathbf{x}^0$, which is denoted as $f(\mathbf{x}^0) = \mathbf{0}$.

Definition 2. A pattern $\mathbf{x}^0 \in R^m$ is called a stable equilibrium point of $\dot{\mathbf{x}} = -f(\mathbf{x})$, if $f(\mathbf{x}) = \mathbf{0}$ at $\mathbf{x}^0$ and the Jacobian matrix at $\mathbf{x}^0$ is a positive definite matrix.

Definition 3. A pattern $\mathbf{x}^0 \in R^m$ is called an unstable equilibrium point (saddle point) of $\dot{\mathbf{x}} = -f(\mathbf{x})$, if $f(\mathbf{x}) = 0$ at $\mathbf{x}^0$ while the Jacobian matrix at $\mathbf{x}^0$ is a negative semidefinite or an indefinite matrix.

3. A New Energy Function and the Attractor Network

A classical energy based attractor network is the Hopfield network invented by Hopfield [10, 14], which serves as content addressable memory systems with binary threshold nodes. Although Hopfield networks are guaranteed to converge to a local minimum, they may converge to a spurious point (undesired memory pattern) rather than a stored pattern (desired memory pattern). To solve this problem, we propose a new energy function, and an attractor network is then designed based on the proposed energy function, in which the patterns are not necessarily binary and uncorrelated. Finally it is concluded that the memory patterns are the stable equilibrium points of the dynamic system and spurious points can be avoided.

Assume that $\mathbf{x}^1, \ldots, \mathbf{x}^k, \ldots, \mathbf{x}^n$ with $\mathbf{x}^k = [x_1^k \ \cdots \ x_m^k]^T \in R^m$ are n different stationary patterns that we wish to store. Our objective is to design an attractor network to store these patterns such that $\mathbf{x}^k$ for $k = 1, \ldots, n$ can be retrieved when an input stimulus is located around the neighborhood of $\mathbf{x}^k$. Before presenting our proposed method, we would like to point out that an ideal attractor network should preserve the following two properties.

Property 1. $\mathbf{x}^1, \ldots, \mathbf{x}^k, \ldots, \mathbf{x}^n$ are stable equilibrium points of the attractor network.

Property 2. $\mathbf{x}^1, \ldots, \mathbf{x}^k, \ldots, \mathbf{x}^n$ are the only stable equilibrium points of the attractor network.

The energy function is designed as the following form:

$$F(\mathbf{x}) = \prod_{k=1}^{k=n} d_k(\mathbf{x}) = \prod_{k=1}^{k=n} \left[\left(\mathbf{x} - \mathbf{x}^k\right)^T \left(\mathbf{x} - \mathbf{x}^k\right) \right], \tag{2}$$

where $d_k(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^k)^T(\mathbf{x} - \mathbf{x}^k)$ is an Euclidian squared distance between $\mathbf{x}$ and $\mathbf{x}^k$. The energy function is the product of the distance $d_k(\mathbf{x})$ for $k = 1, \ldots, n$. For the energy function

$F(\mathbf{x})$, it can be checked that $\forall \mathbf{x} \in R^m$, $F(\mathbf{x}) > 0$, and $F(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{x}^k$ for $k = 1, \ldots, n$.

Remark 4. As mentioned in the Introduction, compared with the energy function in the Hopfield network [10, 14], the value of the proposed energy function $F(\mathbf{x})$ is nonnegative and attains zero value only on the memory patterns stored in the network. This makes it easy to distinguish the memory patterns from spurious points [18] which may exist in an attractor network.

Note that the energy function attains its minimum only on the memory patterns $\mathbf{x}^1, \ldots, \mathbf{x}^n$. This inspires us to design a dynamical system which can make the energy function decrease iteratively. The gradient $\nabla F(\mathbf{x})$ and the Hessian matrix $\nabla^2 F(\mathbf{x})$ are given by

$$\begin{aligned}\nabla F(\mathbf{x}) &= 2\sum_{j=1}^{j=n}\left[\prod_{k=1,k\neq j}^{k=n} d_k(\mathbf{x})\right]\left(\mathbf{x}-\mathbf{x}^j\right)\\ \nabla^2 F(\mathbf{x}) &= 2\sum_{j=1}^{j=n}\left[\prod_{k=1,k\neq j}^{i=n} d_k(\mathbf{x})\right]\cdot I\\ &\quad + 4\sum_{j=1}^{j=n}\left\{\sum_{k'=1,k'\neq j}^{k'=n}\left[\prod_{k=1,k\neq j,k\neq k'}^{i=n} d_k(\mathbf{x})\right]\right.\\ &\quad \left.\cdot\left(\mathbf{x}-\mathbf{x}^{k'}\right)\left(\mathbf{x}-\mathbf{x}^{j}\right)^T\right\},\end{aligned} \tag{3}$$

where I is a m dimensional identity matrix. Let $\mathbf{v} = [v_1 \ \cdots \ v_m]^T$ be a random noise vector such that

$$\max(|\mathbf{v}|) = \max(\ |v_1| \ \cdots \ |v_m|\ \} < \mathbf{v}_{\max}, \tag{4}$$

where $\mathbf{v}_{\max}$ is a chosen small positive constant. Now an attractor network that decreases the energy function $F(\mathbf{x})$ iteratively is represented as

$$\dot{\mathbf{x}} = -\nabla F(\mathbf{x}) + \kappa \cdot F(\mathbf{x}) \cdot \mathbf{v}, \tag{5}$$

where the scalar constant κ is chosen as

$$\kappa = \begin{cases} 0 & \text{if } \max(|\nabla F(\mathbf{x})|) > 0 \\ 1 & \text{if } \max(|\nabla F(\mathbf{x})|) \le 0. \end{cases} \tag{6}$$

We rewrote the differential equation for each neuron x_i as

$$\dot{x}_i = -\sum_{j=1}^{n} \omega_{ij}\left(x_i - \mathbf{x}_i^j\right) + \kappa \cdot F(\mathbf{x}) \cdot v_i \tag{7}$$

for $i = 1, \ldots, m$ with

$$\omega_{ij} = 2\left[\prod_{k=1,k\neq j}^{k=n} d_k(\mathbf{x})\right]. \tag{8}$$

Equations (7) and (8) imply that ith neuron updates its state x_i according to the synaptic inputs collected from other neurons via synaptic connection strength ω_{ij}, in the presence of a random noise v_i.

Remark 5. The Hopfield model has the advantage that it can be represented by neurons and their synaptic connections. Through carefully observing (5)–(8), we note that our proposed attractor network can also be represented by neurons and their synaptic connections by introducing the multiplication algebra in Section 2.1. In addition, the value of our proposed energy function is nonnegative and attains zero values only at the patterns stored in the network. This makes it easy to distinguish the memory patterns from some other possible undesired patterns which are called spurious points [18]. It is shown that the memory patterns are stored as the stable equilibrium points of the dynamical system, which are also the local minimum points of the energy function. Compared with existing results in attractor networks [7, 10, 14], the patterns are not necessarily binary and uncorrelated in this paper. Binary patterns simplify the network design significantly as seen in [14]. Also, the uncorrelated patterns give a minimum interaction of the interactions of the network, which makes its behavior analysis much easier as seen in [7, 19].

Remark 6. As mentioned earlier, there are two kinds of equilibrium points for a dynamic system: stable ones and unstable ones. In the next section, it will be shown that only the stable equilibrium points exist in the above dynamical system. For an arbitrary initial stimulus, the states of the dynamical system converge to one of its stable equilibrium points, that is, the local minimum points of the energy function, which cannot be its local maximum points or saddle points of the energy function. Thus, Property 1 can be achieved in the design. In the next section, it will also be proven that Property 2 can also be achieved in the design; that is, $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are the only local minimum points of $F(\mathbf{x})$.

4. Convergence Analysis

Lemma 7. *For square matrix $A \in R^{m\times m}$, assume that the eigenvalues of A are not all the same. If* $\operatorname{tr}(A) \le 0$, *$A$ cannot be a positive semidefinite or positive definite matrix.*

Proof. Denote the eigenvalues of A as $\lambda_1, \ldots, \lambda_m$. Note that $\operatorname{tr}(A) = \sum_{i=1}^{i=m} \lambda_i$. If $\operatorname{tr}(A) = \sum_{i=1}^{i=m} \lambda_i < 0$, then $\exists \lambda_i$ such that $\lambda_i < 0$. If $\operatorname{tr}(A) = \sum_{i=1}^{i=m} \lambda_i = 0$, there also exists $\exists \lambda_i$ such that $\lambda_i < 0$ since $\lambda_1, \ldots, \lambda_m$ cannot be identically equal. Then this lemma holds. □

Lemma 8. *For conformable matrices A, B, and C,* $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$. *Also, if matrices A and B are addable,* $\operatorname{tr}(A + B) = \operatorname{tr}(A) + \operatorname{tr}(B)$ *[21].*

Proof. To prove $\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$, actually we only need to prove that $\operatorname{tr}(AB) = \operatorname{tr}(BA)$. As $\operatorname{tr}(AB) = \Sigma_i \Sigma_i A_{ij} B_{ji} = \operatorname{tr}(BA)$ and $\operatorname{tr}(A + B) = \Sigma_i (A_{ii} + B_{ii}) = \Sigma_i (A_{ii}) + \Sigma_i (B_{ii}) = \operatorname{tr}(A) + \operatorname{tr}(B)$, this lemma holds. □

Theorem 9. *The states of the attractor network in (5)–(8) converge to an stable equilibrium point $\mathbf{x}^* \in \{\mathbf{x}^1, \ldots, \mathbf{x}^m\}$ such* *that* $\nabla F(\mathbf{x}^*) = \mathbf{0}$, $F(\mathbf{x}^*) = 0$, *and* $\nabla^2 F(\mathbf{x}^*)$ *is positive definite which is denoted as* $\nabla^2 F(\mathbf{x}^*) > 0$.

Proof. By Definition 1, the equilibrium points of system (7) are such that

$$-\nabla F(\mathbf{x}) + \kappa \cdot F(\mathbf{x}) \cdot \mathbf{v} = \mathbf{0}. \tag{9}$$

As $\mathbf{v}$ is a random vector and κ is chosen by (6), a point is an equilibrium point of system (7) if and only if

$$\begin{gathered} \nabla F(\mathbf{x}) = \mathbf{0}, \\ F(\mathbf{x}) = 0. \end{gathered} \tag{10}$$

$F(\mathbf{x}) = 0$ gives us that all the possible equilibrium points of (7) are those points at which the values of energy function $F(\mathbf{x})$ are zero; that is, $\mathbf{x}^* \in \{\mathbf{x}^1, \ldots, \mathbf{x}^m\}$. In addition, the Jacobian matrix at a point $\mathbf{x}^i$ is the Hessian matrix $\nabla^2 F(\mathbf{x}^i)$. Then, it is easy to obtain that

$$\begin{gathered} \nabla F\left(\mathbf{x}^i\right) = 0, \\ \nabla^2 F\left(\mathbf{x}^i\right) = 2\left[\prod_{k=1,k\neq i}^{k=n} d_k\left(\mathbf{x}^i\right)\right] \cdot I > 0. \end{gathered} \tag{11}$$

So $\mathbf{x}^*$ is a stable equilibrium point of the attractor network in (5) to (8) based on Definition 2. □

Theorem 10. *For the case that the dimension* $m \leq 2$, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *are the only local minimum points of* $F(\mathbf{x})$.

Proof. Obviously, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are local minimum points (and also global minimum points) of $F(\mathbf{x})$. We now prove that $F(\mathbf{x})$ has no other local minimum points. Let

$$\begin{aligned} A &= 2\sum_{j=1}^{j=n}\left[\prod_{k=1,k\neq j}^{k=n} d_k(\mathbf{x})\right] \cdot I, \\ B &= 4\sum_{j=1}^{j=n}\left\{\sum_{k'=1,k'\neq j}^{k'=n} \cdot \left[\prod_{k=1,k\neq j,k\neq k'}^{k=n} d_k(\mathbf{x})\right]\right. \\ &\qquad \left. \cdot \left(\mathbf{x} - \mathbf{x}^{k'}\right)\left(\mathbf{x} - \mathbf{x}^j\right)^T\right\}. \end{aligned} \tag{12}$$

It can be obtained that $\nabla^2 F(\mathbf{x}) = A + B$. It is also known that the dimension of diagonal matrix A is $m \times m$; then

$$\operatorname{tr}(A) = m \cdot 2\sum_{j=1}^{j=n}\left[\prod_{k=1,k\neq j}^{k=n} d_k(\mathbf{x})\right]. \tag{13}$$

Also, Lemma 8 gives that

$$\begin{aligned} \operatorname{tr}(B) &= 4\sum_{j=1}^{j=n}\left\{\sum_{k'=1,k'\neq j}^{k'=n}\left[\prod_{k=1,k\neq j,k\neq k'}^{k=n} d_k(\mathbf{x})\right]\right. \\ &\qquad \left. \cdot \left(\mathbf{x} - \mathbf{x}^{k'}\right)^T\left(\mathbf{x} - \mathbf{x}^j\right)\right\}. \end{aligned} \tag{14}$$

Now we assume that there is a point which is different from $\mathbf{x}_1, \ldots, \mathbf{x}_n$ but satisfies that $\nabla F(\mathbf{x}^*) = 0$. This implies that $(\nabla F(\mathbf{x}^*))^T \nabla F(\mathbf{x}^*) = 0$. This is to say that

$$G\left(\mathbf{x}^*\right) \cdot \prod_{k=1}^{k=n}\left(\mathbf{x} - \mathbf{x}^k\right)^T\left(\mathbf{x} - \mathbf{x}^k\right) = 0, \tag{15}$$

where

$$\begin{aligned} G\left(X^*\right) &= 4\sum_{j=1}^{j=n}\left[\prod_{k=1,k\neq j}^{k=n} d_k(\mathbf{x})\right] \\ &\quad + 4\sum_{j=1}^{j=n}\left\{\sum_{k'=1,k'\neq j}^{k'=n}\left[\prod_{k=1,k\neq j,k\neq k'}^{i=n} d_k\left(\mathbf{x}^*\right)\right]\right. \\ &\quad \left. \cdot \left(\mathbf{x}^* - \mathbf{x}^{k'}\right)^T\left(\mathbf{x}^* - \mathbf{x}^j\right)\right\}. \end{aligned} \tag{16}$$

Thus, we have $G(\mathbf{x}^*) = 0$ if $\mathbf{x}^*$ is different from $\mathbf{x}^1, \ldots, \mathbf{x}^n$. Let A^* and B^* be A and B at $\mathbf{x} = \mathbf{x}^*$. Combining (13)-(14) and (16), it can be obtained that

$$G\left(\mathbf{x}^*\right) = \operatorname{tr}\left(\frac{2}{m} \cdot A^* + B^*\right) = 0. \tag{17}$$

Since $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are all different from each other, the eigenvalues of $\nabla^2 F(\mathbf{x}^*)$ cannot be all the same. When $m = 1$, we have $G(\mathbf{x}^*) = \operatorname{tr}(2A^* + B^*) = 0$, which implies $\operatorname{tr}(A^* + B^*) < 0$ since A^* is a positive definite matrix. Similarly, we have $\operatorname{tr}(\nabla^2 F(\mathbf{x}^*)) = 0$ when $m = 2$. From Lemma 7, matrix $\nabla^2 F(\mathbf{x}^*)$ cannot be a positive definite matrix. It is a seminegative definite or an indefinite matrix. This implies that $\mathbf{x}^*$ cannot be a local minimum point though $\nabla F(\mathbf{x}^*) = 0$. It can be a local maximum point or a saddle point of $F(\mathbf{x})$. □

Remark 11. If the dimension $m > 2$, $\operatorname{tr}((2/m) \cdot A^* + B^*) = 0$ does not directly imply that $\nabla^2 F(\mathbf{x}^*)$ is a negative semidefinite or an indefinite matrix. Let $P_f \in R^{2\times n}$ be a full-row rank projection matrix, which projects a vector into a two-dimensional plane. The following theorem shows that $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are still the only local minimum points of $F(\mathbf{x})$.

Theorem 12. *For the case that* $m > 2$, $\mathbf{x}^1, \ldots, \mathbf{x}^n$ *are also the only local minimum points of* $F(\mathbf{x})$.

Proof. Let $\tilde{\mathbf{x}} = P_f\mathbf{x}$ and $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^n$ be the corresponding projection of the local minimum points $\mathbf{x}^1, \ldots, \mathbf{x}^n$ in the two-dimensional plane. Then, $F(\mathbf{x})$ and $\nabla F(\mathbf{x})$ become $F(\tilde{\mathbf{x}})$ and $\nabla F(\tilde{\mathbf{x}})$, respectively, on this two-dimensional plane. It can be obtained that

$$\begin{gathered} \nabla F(\tilde{\mathbf{x}}) = P_f \nabla F(\mathbf{x}) \\ \nabla^2 F(\tilde{\mathbf{x}}) = P_f \nabla^2 F(\mathbf{x}) P_f^T. \end{gathered} \tag{18}$$

If $\nabla F(\mathbf{x}) = 0$ and $\nabla^2 F(\mathbf{x}) > 0$, then $\nabla F(\tilde{\mathbf{x}}) = 0$ and $\nabla^2 F(\tilde{\mathbf{x}}) > 0$, which means that the local minimum points of $F(\mathbf{x})$ in a higher dimensional space must be also the local minimum points of $F(\widetilde{X})$ in the two-dimensional plane while the converse is not true. Obviously, $\mathbf{x}^1, \ldots, \mathbf{x}^n$ are the local minima of $F(\tilde{\mathbf{x}})$.

Assume that there is a point $\mathbf{x}^*$ which is different from $\mathbf{x}^1, \ldots, \mathbf{x}^n$ but a local minimum of $F(\mathbf{x})$. There exists a full-row rank projection matrix $P_f^* \in R^{2\times n}$ such that $\tilde{\mathbf{x}}^* = P_f^* \mathbf{x}^*$. The full-row rank of P_f^* implies that $\tilde{\mathbf{x}}^*$ is different from $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^i, \ldots, \tilde{\mathbf{x}}^n$, where $\tilde{\mathbf{x}}_i = P_f^* \mathbf{x}_i$ for $i = 1, \ldots, n$ are the local minimum points of $F(\tilde{\mathbf{x}}) = F(P_f^* \mathbf{x})$. However, this is impossible by Theorem 10. Thus, there is no other local minimum point of $F(\mathbf{x})$. If a point $\mathbf{x}^*$ satisfies that $\nabla F(\mathbf{x}^*) = 0$ but it is different from $\mathbf{x}^1, \ldots, \mathbf{x}^n$, it can be only a local maximum point or a saddle point. So this theorem holds. □

5. Simulation Results

Example 1. Design a nonlinear dynamical system whose attractors are $\mathbf{x}^1, \ldots, \mathbf{x}^n \in R^m$ with $n = 2$, $m = 1$.

This corresponds to the one-dimensional case. The energy function is constructed as $F(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^1)^2 (\mathbf{x} - \mathbf{x}^2)^2$ from (2) with $\nabla F(X)$ and $\nabla^2 F(X)$ being given by

$$\begin{aligned}\nabla F(\mathbf{x}) &= 2\left(\mathbf{x} - \mathbf{x}^1\right)^2 \left(\mathbf{x} - \mathbf{x}^2\right) \\ &\quad + 2\left(\mathbf{x} - \mathbf{x}^2\right)^2 \left(\mathbf{x} - \mathbf{x}^1\right), \\ \nabla^2 F(\mathbf{x}) &= 2\left(\mathbf{x} - \mathbf{x}^1\right)^2 + 2\left(\mathbf{x} - \mathbf{x}^2\right)^2 \\ &\quad + 4\left(\mathbf{x} - \mathbf{x}^2\right)\left(\mathbf{x} - \mathbf{x}^1\right) \\ &\quad + 4\left(\mathbf{x} - \mathbf{x}^1\right)\left(\mathbf{x} - \mathbf{x}^2\right),\end{aligned} \tag{19}$$

respectively. The dynamic system can be then designed in (5)–(8). By solving $\nabla F(\mathbf{x}^*) = 0$, we have $\mathbf{x}^* = \mathbf{x}^1$, $\mathbf{x}^* = \mathbf{x}^2$, or $\mathbf{x}^* = (\mathbf{x}^1 + \mathbf{x}^2)/2$ as analyzed in Section 3. From

$$\left.\nabla^2 F(\mathbf{x})\right|_{\mathbf{x}=\mathbf{x}^*=(\mathbf{x}^1+\mathbf{x}^2)/2} < 0, \tag{20}$$

$\mathbf{x}^* = (\mathbf{x}^1 + \mathbf{x}^2)/2$ is a local maximum point of $F(\mathbf{x})$. So $\mathbf{x}^1$ and $\mathbf{x}^2$ are the only two local minimum points of $F(\mathbf{x})$. The energy function is shown in Figure 1 where "$*$" denotes the local minimum point and "Δ" denotes the local maximum or the saddle point. From Figure 1, when the initial stimulus $\mathbf{x} < (\mathbf{x}^1 + \mathbf{x}^2)/2$, the states converge to $\mathbf{x}^1$; when $\mathbf{x} > (\mathbf{x}^1 + \mathbf{x}^2)/2$, the states converge to $\mathbf{x}^2$. If the initial stimulus $\mathbf{x} = (\mathbf{x}^1 + \mathbf{x}^2)/2$, the states can converge to either $\mathbf{x}^1$ or $\mathbf{x}^2$, which depends on the random noise v_1 in (7).

Example 2. Example 1, analyze the dynamic system when $n = 2$, $m = 2$ and $n = 2$, $m = 3$, respectively.

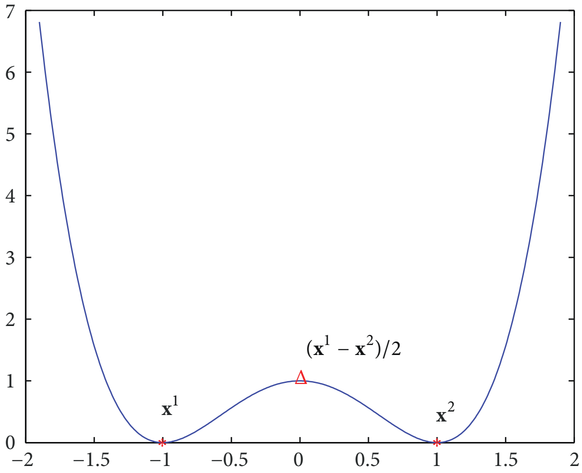

Figure 1: Energy function for one-dimensional case.

Case 1. $\mathbf{x}^1 = [-0.5 \quad -0.5]'$ and $\mathbf{x}^2 = [0.5 \quad 0.5]'$. Similar to Example 1, the energy function is constructed as $F(\mathbf{x}) = d_1(\mathbf{x}) d_2(\mathbf{x})$ with $\nabla F(\mathbf{x})$ and $\nabla^2 F(\mathbf{x})$ being given by

$$\begin{aligned}\nabla F(\mathbf{x}) &= 2d_1(\mathbf{x})\left(\mathbf{x} - \mathbf{x}^2\right) + 2d_2(\mathbf{x})\left(\mathbf{x} - \mathbf{x}^1\right), \\ \nabla^2 F(\mathbf{x}) &= 2\left(\mathbf{x} - \mathbf{x}^1\right)^T \left(\mathbf{x} - \mathbf{x}^1\right) \cdot I \\ &\quad + 2\left(\mathbf{x} - \mathbf{x}^2\right)^T \left(\mathbf{x} - \mathbf{x}^2\right) \cdot I \\ &\quad + 4\left(\mathbf{x} - \mathbf{x}^2\right)\left(\mathbf{x} - \mathbf{x}^1\right)^T \\ &\quad + 4\left(\mathbf{x} - \mathbf{x}^1\right)\left(\mathbf{x} - \mathbf{x}^2\right)^T.\end{aligned} \tag{21}$$

Solving $\nabla F(\mathbf{x}^*) = 0$ gives $\mathbf{x}^* = \mathbf{x}^1$, $\mathbf{x}^* = \mathbf{x}^2$, or $X^* = (\mathbf{x}^1 + \mathbf{x}^2)/2$. When $X^* = (\mathbf{x}^1 + \mathbf{x}^2)/2$, we have

$$\nabla^2 F(\mathbf{x}^*) = \begin{bmatrix} 0 & -2 \\ -2 & 0 \end{bmatrix}. \tag{22}$$

The two eigenvalues of $\nabla^2 F(\mathbf{x}^*)$ are $\lambda_1 = 2$ and $\lambda_2 = -2$. This is consistent with Theorem 10, which gives $\operatorname{tr}(\nabla^2 F(\mathbf{x}^*)) = 0$. Figure 2 shows the contour map of the energy function in a two-dimensional space. The contour lines around a saddle point look like a horse saddle. It can be concluded that $\mathbf{x}^1$ and $\mathbf{x}^2$ are two attractors while $(\mathbf{x}^1 + \mathbf{x}^2)/2$ is not an equilibrium point of system (5)–(8) by Theorem 9. The term $\kappa \cdot F(\mathbf{x})\mathbf{v}$ guarantees that the attractor network (5)–(8) cannot stay at the point $(\mathbf{x}^1 + \mathbf{x}^2)/2$. But from Theorem 10, we know that $(\mathbf{x}^1 + \mathbf{x}^2)/2$ is a saddle point of $F(\mathbf{x})$.

Case 2. $\mathbf{x}^1 = [-0.5 \quad -0.5 \quad -0.5]'$ and $\mathbf{x}^2 = [0.5 \quad 0.5 \quad 0.5]'$. In this case, $n = 2$, $m = 3$. When $\mathbf{x}^* = (\mathbf{x}^1 + \mathbf{x}^2)/2$, we have

$$\nabla^2 F(\mathbf{x}^*) = \begin{bmatrix} 1 & -2 & -2 \\ -2 & 1 & -2 \\ -2 & -2 & 1 \end{bmatrix}. \tag{23}$$

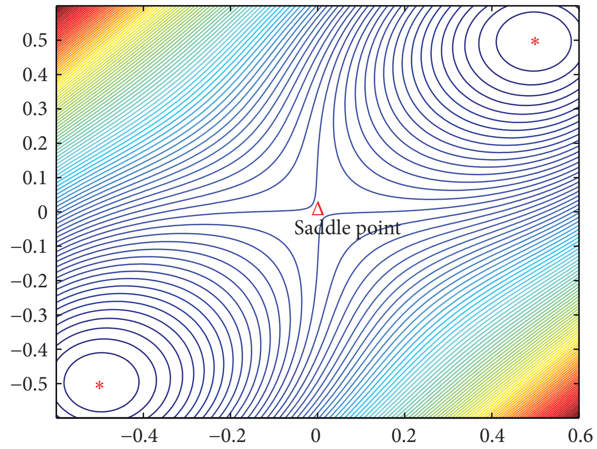

Figure 2: The contours of the energy function for Example 2.

Then $\mathbf{x}^*$ is still not an equilibrium point of system (5)–(8) but a saddle point of $F(\mathbf{x})$ as the eigenvalue of $\nabla^2 F(\mathbf{x}^*)$ is now $\lambda_1 = -3$, $\lambda_2 = 3$, and $\lambda_3 = 3$. But in this case, we have $\operatorname{tr}(\nabla^2 F(\mathbf{x}^*)) > 0$. So we cannot determine whether $\mathbf{x}^*$ is a saddle point or a local minimum point of $F(\mathbf{x})$ only based on Theorem 10. However, we can determine that it is also a saddle point by Theorem 12. This is consistent with what we observed in simulation, as the eigenvalues of above $\nabla^2 F(\mathbf{x}^*)$ is $3, 3, -3$, and thus $\mathbf{x}^*$ is a saddle point and cannot be a local minimum point of $F(\mathbf{x})$.

Example 3. Design a nonlinear dynamic system whose attractors are $\mathbf{x}^1 = [2\ \ 0]'$, $\mathbf{x}^2 = [-1\ \ \sqrt{3}]'$, and $\mathbf{x}^3 = [-\sqrt{3}\ \ -1]' \in R^2$.

We have $n = 3$ and $m = 2$ in this example. The energy function is constructed as $F(\mathbf{x}) = d_1(\mathbf{x})d_2(\mathbf{x})d_3(\mathbf{x})$ with

$$\begin{aligned}
\nabla F(\mathbf{x}) &= 2d_1(\mathbf{x})\,d_2(\mathbf{x})\left(\mathbf{x}-\mathbf{x}^3\right) \\
&\quad + 2d_1(\mathbf{x})\,d_3(\mathbf{x})\left(\mathbf{x}-\mathbf{x}^2\right) \\
&\quad + 2d_2(\mathbf{x})\,d_3(\mathbf{x})\left(\mathbf{x}-\mathbf{x}^1\right), \\
\nabla^2 F(\mathbf{x}) &= 2d_1(\mathbf{x})\,d_2(\mathbf{x})\cdot I + 2d_1(\mathbf{x})\,d_3(\mathbf{x})\cdot I \\
&\quad + 2d_2(\mathbf{x})\,d_3(\mathbf{x})\cdot I \\
&\quad + 4\left(\mathbf{x}-\mathbf{x}^3\right)\left(\mathbf{x}-\mathbf{x}^1\right)^T\left(\mathbf{x}-\mathbf{x}^2\right)^T\left(\mathbf{x}-\mathbf{x}^2\right) \\
&\quad + 4\left(\mathbf{x}-\mathbf{x}^3\right)\left(\mathbf{x}-\mathbf{x}^2\right)^T\left(\mathbf{x}-\mathbf{x}^1\right)^T\left(\mathbf{x}-\mathbf{x}^1\right) \\
&\quad + 4\left(\mathbf{x}-\mathbf{x}^2\right)\left(\mathbf{x}-\mathbf{x}^1\right)^T\left(\mathbf{x}-\mathbf{x}^3\right)^T\left(\mathbf{x}-\mathbf{x}^3\right) \\
&\quad + 4\left(\mathbf{x}-\mathbf{x}^2\right)\left(\mathbf{x}-\mathbf{x}^3\right)^T\left(\mathbf{x}-\mathbf{x}^1\right)^T\left(\mathbf{x}-\mathbf{x}^1\right) \\
&\quad + 4\left(\mathbf{x}-\mathbf{x}^1\right)\left(\mathbf{x}-\mathbf{x}^2\right)^T\left(\mathbf{x}-\mathbf{x}^3\right)^T\left(\mathbf{x}-\mathbf{x}^3\right) \\
&\quad + 4\left(\mathbf{x}-\mathbf{x}^1\right)\left(\mathbf{x}-\mathbf{x}^3\right)^T\left(\mathbf{x}-\mathbf{x}^2\right)^T\left(\mathbf{x}-\mathbf{x}^2\right)
\end{aligned} \tag{24}$$

with its dynamical property described by the attractor network in (5)–(8). Theorem 9 tells us that $\mathbf{x}^1$, $\mathbf{x}^2$, and $\mathbf{x}^3$ are the stable equilibrium points of $F(\mathbf{x})$ and the attractor network in (5)–(8). However, $F(\mathbf{x})$ has more saddle points but these saddle points are not equilibrium points (saddle points) of (5)–(8). To illustrate this, firstly, we find all the points such that $\nabla F(X^*) = 0$. If $\nabla F(X^*) = 0$, then $(\nabla F(X^*))^T \nabla F(X^*) = 0$, which means

$$G(\mathbf{x}^*)\cdot\prod_{k=1}^{k=3} d_k(\mathbf{x}^*) = 0, \tag{25}$$

where

$$\begin{aligned}
G(\mathbf{x}^*) &= 2d_1(\mathbf{x}^*)\,d_2(\mathbf{x}^*) + 2d_1(\mathbf{x}^*)\,d_3(\mathbf{x}^*) \\
&\quad + 2d_2(\mathbf{x}^*)\,d_3(\mathbf{x}^*) \\
&\quad + 4\left(\mathbf{x}^*-\mathbf{x}^3\right)^T\left(\mathbf{x}^*-\mathbf{x}^1\right)d_2(\mathbf{x}^*) \\
&\quad + 4\left(\mathbf{x}^*-\mathbf{x}^3\right)^T\left(X^*-\mathbf{x}^2\right)d_1(\mathbf{x}^*) \\
&\quad + 4\left(\mathbf{x}^*-\mathbf{x}^2\right)^T\left(\mathbf{x}^*-\mathbf{x}^1\right)d_3(\mathbf{x}^*) \\
&\quad + 4\left(\mathbf{x}^*-\mathbf{x}^2\right)^T\left(\mathbf{x}^*-\mathbf{x}^3\right)d_1(\mathbf{x}^*) \\
&\quad + 4\left(\mathbf{x}^*-\mathbf{x}^1\right)^T\left(\mathbf{x}^*-\mathbf{x}^2\right)d_3(\mathbf{x}^*) \\
&\quad + 4\left(\mathbf{x}^*-\mathbf{x}^1\right)^T\left(\mathbf{x}^*-\mathbf{x}^3\right)d_2(\mathbf{x}^*).
\end{aligned} \tag{26}$$

We know that $(\nabla F(\mathbf{x}^*))^T \nabla F(\mathbf{x}^*) = 0$ gives $\mathbf{x}^* = \mathbf{x}^1$, $\mathbf{x}^* = \mathbf{x}^2$, $\mathbf{x}^* = \mathbf{x}^3$ or $G(\mathbf{x}^*) = 0$. But $G(\mathbf{x}^*) = 0$ implies that

$$\operatorname{tr}\left(\nabla^2 F(\mathbf{x}^*)\right) = 0. \tag{27}$$

Thus, $\mathbf{x}^*$ will be a saddle point of $F(\mathbf{x})$ but not the attractor network in (5)–(8) if $\nabla F(\mathbf{x}^*) = 0$ while $\mathbf{x}^*$ is different from $\mathbf{x}^1, \ldots, \mathbf{x}^n$. As the attractor network in (5)–(8) does not have any saddle points from Theorem 9, usually, a saddle point of $F(\mathbf{x})$ is located in between two local minimum points. As seen from Figure 3, two saddle points of $F(\mathbf{x})$ are in between the three local minimum points on the plane. One is located about $(-1.05, 0.22)$ and the other is about $(0.58, 0.31)$.

6. Conclusion

The contributions of this paper are summarized as follows.

(1) We have proposed a new energy function which includes the information of the stored patterns, and it is different from the energy function in Hopfield network. The proposed energy function makes it easy to differentiate memory patterns from possible spurious points.

(2) We have presented an attractor network design based on the proposed energy function. The patterns stored in the attractor network can be nonbinary and either correlated or uncorrelated. The memory patterns are

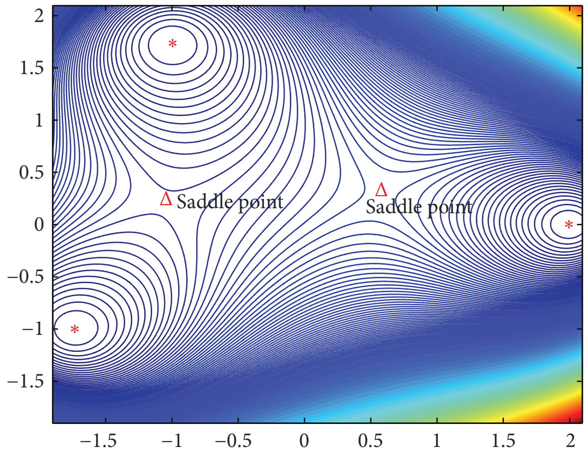

Figure 3: The contours of the energy function for Example 3.

the attractors of the network and the equilibrium points of the dynamic system.

(3) When an arbitrary input stimulus is presented to the designed attractor network, it has been proved that the states converge to one of the stored patterns. There are no spurious states in the designed dynamic systems.

Our future work is to construct a biological plausible dynamical system in hardware (neuromorphic chip) which can stimulate the behavior of the designed network. This sheds new lights on the research towards the realization of artificial cognitive memory.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work is funded by Brain Inspired Computing Research, Tsinghua University (20141080934).

References

[1] L. P. Shi, K. J. Yi, K. Ramanathan et al., "Artificial cognitive memory-changing from density driven to functionality driven," *Applied Physics A: Materials Science and Processing*, vol. 102, no. 4, pp. 865–875, 2011.

[2] G. Li, N. Ning, K. Ramanathan, W. He, L. Pan, and L. Shi, "Behind the magical numbers: hierarchical chunking and the human working memory capacity," *International Journal of Neural Systems*, vol. 23, no. 4, Article ID 1350019, 2013.

[3] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Spin-glass models of neural networks," *Physical Review, A: Third Series*, vol. 32, no. 2, pp. 1007–1018, 1985.

[4] D. J. Amit and G. Mongillo, "Spike-driven synaptic dynamics generating working memory states," *Neural Computation*, vol. 15, no. 3, pp. 565–596, 2003.

[5] B. Poucet and E. Save, "Attractors in memory," *Science*, vol. 308, no. 5723, pp. 799–800, 2005.

[6] M. Tsodyks, "Attractor neural networks and spatial maps in hippocampus," *Neuron*, vol. 48, no. 2, pp. 168–169, 2005.

[7] H. Tang, H. Li, and R. Yan, "Memory dynamics in attractor networks with saliency weights," *Neural Computation*, vol. 22, no. 7, pp. 1899–1926, 2010.

[8] Y. Miyashita, "Neuronal correlate of visual associative long-term memory in the primate temporal cortex," *Nature*, vol. 335, no. 6193, pp. 817–820, 1988.

[9] A. Bakker, C. B. Kirwan, M. Miller, and C. E. L. Stark, "Pattern separation in the human hippocampal CA$_3$ and dentate gyrus," *Science*, vol. 319, no. 5870, pp. 1640–1642, 1999.

[10] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 79, no. 8, pp. 2554–2558, 1982.

[11] J. Conklin and C. Eliasmith, "A controlled attractor network model of path integration in the rat," *Journal of Computational Neuroscience*, vol. 18, no. 2, pp. 183–203, 2005.

[12] H. Sompolinsky, A. Crisanti, and H.-J. Sommers, "Chaos in random neural networks," *Physical Review Letters*, vol. 61, no. 3, pp. 259–262, 1988.

[13] T. J. Wills, C. Lever, F. Cacucci, N. Burgess, and J. O'Keefe, "Attractor dynamics in the hippocampal representation of the local environment," *Science*, vol. 308, no. 5723, pp. 873–876, 2005.

[14] M. K. Müezzinoglu, C. Güzeliş, and J. M. Zurada, "An energy function-based design method for discrete Hopfield associative memory with attractive fixed points," *IEEE Transactions on Neural Networks*, vol. 16, no. 2, pp. 370–378, 2005.

[15] M. Hurley, "Lyapunov functions and attractors in arbitrary metric spaces," *Proceedings of the American Mathematical Society*, vol. 126, no. 1, pp. 245–256, 1998.

[16] S. Hélie, "Energy minimization in the nonlinear dynamic recurrent associative memory," *Neural Networks*, vol. 21, no. 7, pp. 1041–1044, 2008.

[17] C. Koch and I. Segev, "The role of single neurons in information processing," *Nature Neuroscience*, vol. 3, pp. 1171–1177, 2000.

[18] A. V. Robins and S. J. R. McCallum, "A robust method for distinguishing between learned and spurious attractors," *Neural Networks*, vol. 17, no. 3, pp. 313–326, 2004.

[19] W. Senn and S. Fusi, "Learning only when necessary: better memories of correlated patterns in networks with bounded synapses," *Neural Computation*, vol. 17, no. 10, pp. 2106–2138, 2005.

[20] R. S. Zemel and M. C. Mozer, "Localist attractor networks," *Neural Computation*, vol. 13, no. 5, pp. 1045–1064, 2001.

[21] G. Li, C. Wen, W. X. Zheng, and Y. Chen, "Identification of a class of nonlinear autoregressive models with exogenous inputs based on kernel machines," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2146–2159, 2011.