第21卷 第6期
1995 年 11 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 21, No. 6
Nov., 1995

联想记忆神经网络的训练[1)]

张承福　　赵　刚
(北京大学物理系　北京　100871)

摘　要

提出了一种联想记忆神经网络的优化训练方案，说明网络的样本吸引域可用阱深参数作一定程度的控制，使网络具有尽可能好的容错性。计算表明，训练网络可达到 $\alpha \lesssim 1(\alpha = M/N$，$N$ 是神经元数，M 是贮存样本数)，而仍有良好的容错性，明显优于外积法、正交化外积法、赝逆法等常用方案。文中还对训练网络的对称性与收敛性问题进行了讨论。

关键词：神经网络，联想记忆，容错性，吸引子，吸引域。

人工神经网络研究的生命力在于它有一定的实用价值和前景。就当前所知，它在联想记忆[1-3]、困难问题的近似优化[4]、非线性映射[5,1]等方面具有优势，已引起广泛地重视。这里主要讨论联想记忆神经网络的优化训练问题。

1　联想记忆神经网络的主要问题

联想记忆网络有否生命力取决于三个因素：(1) 贮存容量 $\alpha = M/N$；(2) 容错性；(3) 技术可行性，即硬件实施的可能性。三者是一个统一的整体，而容错性则是其核心。离开容错性谈"提高贮存容量"是毫无意义的。理由很简单：按编码方式，N 个神经元可编码 2^N 个不同的状态或模式，但网络的贮存容量一般都小于 N(Hopfield 模型 $\alpha \cong 0.15$，任何简单网络至多只能达到 $\alpha \lesssim 1$，高阶关联网络虽可有较大的容量[7]，但这是以增加联接数，即增加技术难度为代价的)。联想记忆网络的存在价值就在于它有良好的容错性。当然，资源利用效率也是重要的，若用很大网络只贮存几个样本，虽容错性很好，也难有实用价值。

联想记忆网络有两种运行模式：(1) 前传式，实现输入到输出的一步映射；(2) 演化式(即递归式)，这是一种动力学系统，输入是系统的初态，演化的终态则是输出。就容错性而言，后者明显优于前者。从动力学观点看，贮存样本集构成动力学系统的不动点吸引子，各吸引子的吸引域大小与形态决定了容错性的优劣。前传网络的吸引域明显小于异步演化网络的吸引域。这是因为：(1) 前传网络的吸引域只是同步演化网络的一部分，即其中的"一步吸引域"；(2) 异步演化网络的吸引域明显大于同步网络吸引域[8]。因

1) 得到国家非线性科学攀登项目资助的课题。
本文于 1993 年 6 月 7 日收到。

此，联想记忆网络一般都用演化模式.

网络的性能取决于联接矩阵W. 如何确定W，以使网络具有良好的性能：(1) 网络在演化时收敛，即只有不动点吸引子；(2) 有较大的贮存容量；(3) 有好的容错性，即样本吸引子有较大的吸引域，这是构造联想记忆网络的关键.

Hopfield 模型是以样本集的外积法确定W的，即

$$W = \sum_{\alpha=1}^{M} S^{\alpha}S^{\alpha} - MI. \tag{1}$$

其中 $S^{\alpha}(\alpha = 1,2,\cdots,M)$ 是M个贮存样本；I 是么矩阵；式中第二项保证 $W_{ii} = 0$，使之有好的容错性. 理论分析及数值模拟表明，当 $\alpha \lesssim 0.15$ 时它具有良好的容错性，硬件实施也是可能的，引起了广泛重视. 但低容量是该模型的严重局限，使之难以走向实用. 而且，实际样本集总有相关性，而非随机分布，其容量一般小于理论值.

用 Gram-Schmidt 正交化手续[1,2]是一种较好的改进方案. 它可解决样本的相关性问题，且可提高贮存容量 α，但仍不能达到 $\alpha = 1$. 因为此法构成的矩阵是对角项为主的，即 $W_{ii} = M$，$|W_{ij}| \ll M(i \neq j)$，$W_{ii}/ < |W_{ij}| > \gg 1$，而且当$\alpha$增加时此比值急剧增加[8]. 正的对角项虽有助于样本的贮存，但也有助于所有状态的稳定，即产生大量无用的伪吸引子，致使样本吸引域减小，容错性下降. $\alpha = 1$ 时达到极端情形，此时 $W = MI$ 成为对角矩阵，所有 2^N 状态都是不动点，吸引域为零，网络毫无意义. 若令对角项为零，可减少伪吸引子数目，在α不太大时，可改善网络性能；但当α较大时，不能保证样本的贮存. 大量计算表明，用 G-S 方案，并取零对角项，可达 $\alpha \lesssim 0.5$ 而仍有较好的容错性，但当 $\alpha > 0.5$ 后，性能明显下降[8]. 其它较流行的方案还有赝逆法[10,11]等等，亦声称可达 $\alpha = 1$，但这也是以牺牲容错性为代价的. 可以证明[8]，按文献[11]的方案给出的矩阵与 G-S 法是等价的，也只能用于 $\alpha \lesssim 0.5$ 的情形.

以上各方案都是人为设定的，因有较好性能而常被采用. 但从优化角度看，它们远非是最优的. 要人为设定一种适于各种样本集的最佳方案是困难的，应充分发挥神经网络可学习性这一优点，用训练方法使之自动找出最优或较优解. 用训练的方法确定联接矩阵的工作并不少见，但多数着眼于使样本被“记住”及提高贮存容量上，对其容错性则很少考虑. 而没有良好容错性的网络，容量再高也是没有价值的. 本文的学习方案将重点置于如何控制与改善容错性上. 计算表明，完全可以做到 $\alpha \lesssim 1$，而网络仍有良好的容错性.

2 联想记忆网络的训练方案

容错性是联想记忆网络的核心、训练的关键. 这里首先需要解决三个问题：(1) 容错性的描述与刻划；(2) 容错性的探测；(3) 容错性的控制. 下面分别予以说明.

2.1 容错性的描述与刻划. 从动力学观点看，贮存样本应成为系统的不动点吸引子，容错性的好坏则由其吸引域的大小与形态决定. 吸引域是N维空间的复杂形体（一般不是各向同性的球体），其完全描述是很困难的. 可用如下描述对其做粗略的刻划：a）样本吸引域体积，即属于该吸引域的状态总数 $B_{s\alpha}, \alpha = 1,2,\cdots,M$ 是M个样本的序号. 对于大网络，$B_{s\alpha}$ 一般是很大的数，且容易造成假象(如增加冗余神经元数目，可使 $B_{s\alpha}$ 很

快增加，但实用效率却很低）. 因此，用无量纲比值 $B_{s\alpha}/(\Omega_0/M)$ 来刻划可更确切些，其中 $\Omega_0 = 2^N$ 是状态总数，Ω_0/M 是每个样本平均可占有的最大体积. 亦可用

$$R_s = \sum_{\alpha=1}^{M} B_{s\alpha}/\Omega_0 \tag{2}$$

描述. 它反映所有样本吸引域的总比值，也代表网络资源的利用效率. 由前式可见 $0 \leqslant R_s \leqslant 1$，若 $R_s \ll 1$，则表示只有极小部分空间被利用，属于低效率. *b*）上述参数过于粗略，可用 $R_c(d)$-d 曲线对容错性做进一步描述. 其中 d 是状态与样本的哈明距离；比值 $R_c(d) \equiv B(d)/B_0(d)$，$B(d)$ 是距离为 d 并属于该吸引域的状态数；$B_0(d)$ 则是距离为 d 的这一层的状态总数，即

$$B_0(d) = C_N^d = N!/(d!(N-d)!). \tag{3}$$

$R_c(d)$ 的取值范围是 0→1，它有明确的物理意义，即输入态与样本相差 d 时，能正确识别的几率. 应该指出，上述参数并不能描述吸引域的各向异性，此任务难以用少数几个参数来完成.

2.2 吸引域的探测. 目前的理论工作远不足以对吸引域进行解析计算，主要还是靠数值计算. 网络确定之后，原则上可用搜索方法对其吸引子、吸引域做全面探测. 其中搜索方案之一是全搜索，即观察所有状态的演化及其归宿. 这里的计算工作量很大，但往往是必要的. 因为它可提供该系统的全面信息——网络的收敛性、有否周期解、有多少伪吸引子、各吸引域的大小与形态等等，便于对各种方案的优劣作出定量的比较. 这里对 $10 \leqslant N \leqslant 18$ 的较小网络作全搜索计算，此时状态总数 $\Omega_0 = 1024 - 262144$，结果有一定统计意义. 由于状态总数 Ω_0 随 N 指数地增加，对大网络再作全搜索是不可能的. 只能用抽样统计方法. 对 $N = 64$ 的较大网络的吸引域抽样统计，所得结论与小网络全搜索结果基本相符，表明该结论有一定普遍性.

2.3 吸引域的控制. 想对吸引域做全面的或“微观的”控制（即指定每一状态的归宿），显然是不可能及不必要的. 但对吸引域做粗略控制，或“宏观控制”则是可能的. 训练方案基于如下物理考虑：吸引子位于“能量”函数的局域极小处（即“坑底”），在一定限制条件下（见下），坑越深则相应面积越大，即吸引域越大. 因此，引入“阱深”参数 G^α. 训练中要求：

$$b_i^\alpha \equiv S_i^\alpha\left(\sum_{j=1}^{N} W_{ij}S_i^\alpha - \theta_i\right) > G^\alpha, \quad \begin{matrix} i = 1,2,\cdots,N, \\ \alpha = 1,2,\cdots,M. \end{matrix} \tag{4}$$

其中 $G^\alpha = A^\alpha G_0$，$A^\alpha > 0$，对不同样本可取不同值的 A^α，用以控制不同样本吸引域的相对大小（通常先取 $A^\alpha \equiv 1$，训练后若发现某些样本的吸引域太小，则增加其 A^α 值再训练）；$G_0 \geqslant 0$，是训练过程中不断增加的控制参数，开始令 $G_0 = 0$，以保证先使所有样本成为吸引子，一旦达到要求（即(3)式对所有 α, i 成立，若按通常算法，训练已结束），再使 G_0 增加一小量，继续训练. 反复此过程，直到 G_0 不再能增加为止. 这样得到的网络（在局域意义下）是使样本吸引子有最大阱深的解.

上述算法的基本思想与文献 [9] 的优化学习算法是一致的. 但该文并未给出具体计算结果. 本文通过大量计算表明，上述训练方案是切实可行的、有效的；所得网络的性能明显优于现有其它各种模型的性能.

关于训练方案还需说明几点:

(1) 训练中必须使 W,θ 逐行归一,即令

$$\sum_{j=1}^{N}|W_{ij}|+|\theta_i|=N+1,i=1,2,\cdots,N, \tag{5}$$

否则不能保证结果的改善. 一个极端例子是标度变换, 若令 W, $\theta\to\lambda W$, $\lambda\theta$, 其中 $\lambda>1$,则有 $b_i^a\to\lambda b_i^a$, 即 b_i^a 可随 λ 无限增大. 但这显然不会有任何实质性的改进. 在(5)式的限制条件下增加阱深,则必须对W和θ作实质性调整.

(2) 令对角项 $W_{ii}=0$, 这是改善容错性的最佳选择.

(3) 由于逐行归一,一般不能保证网络的对称性. 网络是否收敛? 这是需要回答的. 从下面可以看出,实际上网络仍是几乎收敛的,无须采取对角化措施.

3 数值模拟结果与讨论

大量计算表明,按上法训练的网络其吸引域明显大于通常的 GS 网络(Gram-Schmidt 正交化外积网络)的吸引域;而且,异步零对角 GS 网络明显优于同步非零对角 GS 网络; 同步演化 GS 网络又优于一步映射的 GS 网络. 限于篇幅,有关数据及原因分析可参阅文献[8]的图 1, 2. 关于贮存容量 α, GS 网络只能达到 $\alpha\lesssim0.5$, 当 $\alpha>0.5$ 时, 吸引域很快趋于零, 失去容错性 (因而无实用价值);而训练网络可达 $\alpha\lesssim1$ 仍有相当大的吸引域[8].

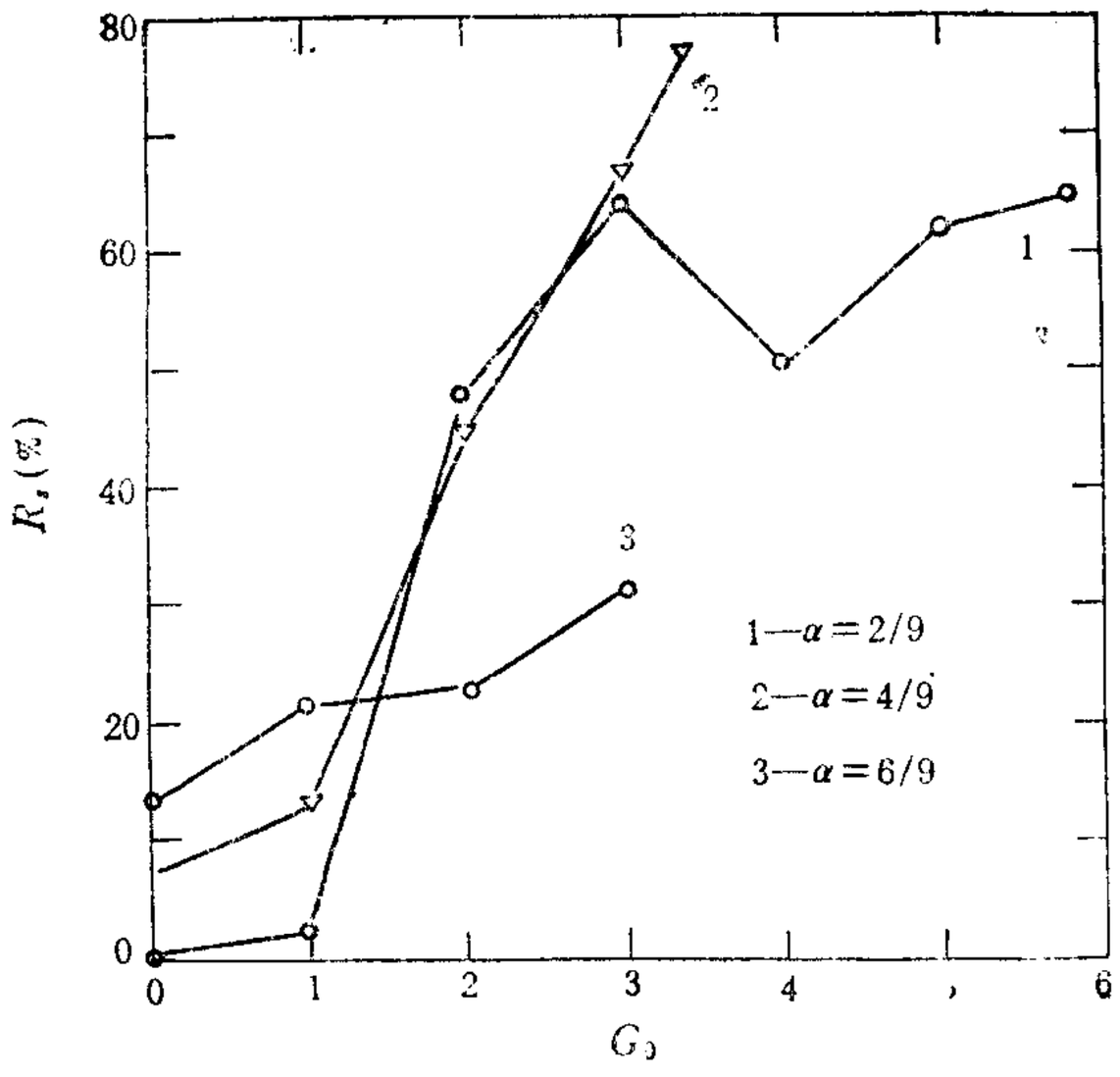


图 1 平均吸引域随阱深的变化

图 1 给出平均吸引域 R_s 随阱深参数 G_0 的变化曲线. 可见 R_s 基本上随 G_0 增加而单调上升, 表明方案是合理、可行的. 若按通常算法,一旦 $G_0\lesssim0$ 即停止训练, 其吸引域是很小的. 同时亦可看出, α 增加时, R_s 随 G_0 上升的趋势一般将减缓(当然,还与样本集的分布有关,并非严格单调).

图 2 给出 $N=14$, 贮存不同数目随机样本时, 某一样本的容错曲线 $R_c(d)$-d. 图 3 则是 $N=64$ 网络,贮存 26 个英文大写字母时, 其中四个样本的容错曲线, 此例中样本有强的相关性. 由图 2,3 可见训练网络的容错性明显优于 GS 模型.

在阱深 $G_\alpha=G_0A_s^\alpha$ 中, A_s^α 的作用通常取定 $A_s^\alpha\equiv1$, 即平等对待每一样本. 但训练出的各样本吸引域很不均匀,可有几倍之差. 这与训练初条件有一些关系,但更重要的是由样本在状态空间的分布及其相互关系所决定. 控制各吸引域相对大小的一种可行方案是: 先令所有 $A_s^\alpha\equiv1$; 训练后,探测各样本吸引域, 对于吸引域明显低于平均值的样

本，增加其 A_i^a 值，并对 A_i^a 归一化$\left(\text{即令}\sum_{a=1}^{M} A_i^a = M\right)$，再训练；如此反复几次，至满意或很难再变化为止．图 4 是一个例子，表明要使各样本吸引域完全均匀是很难做到的，但一定程度的控制则是可行的．

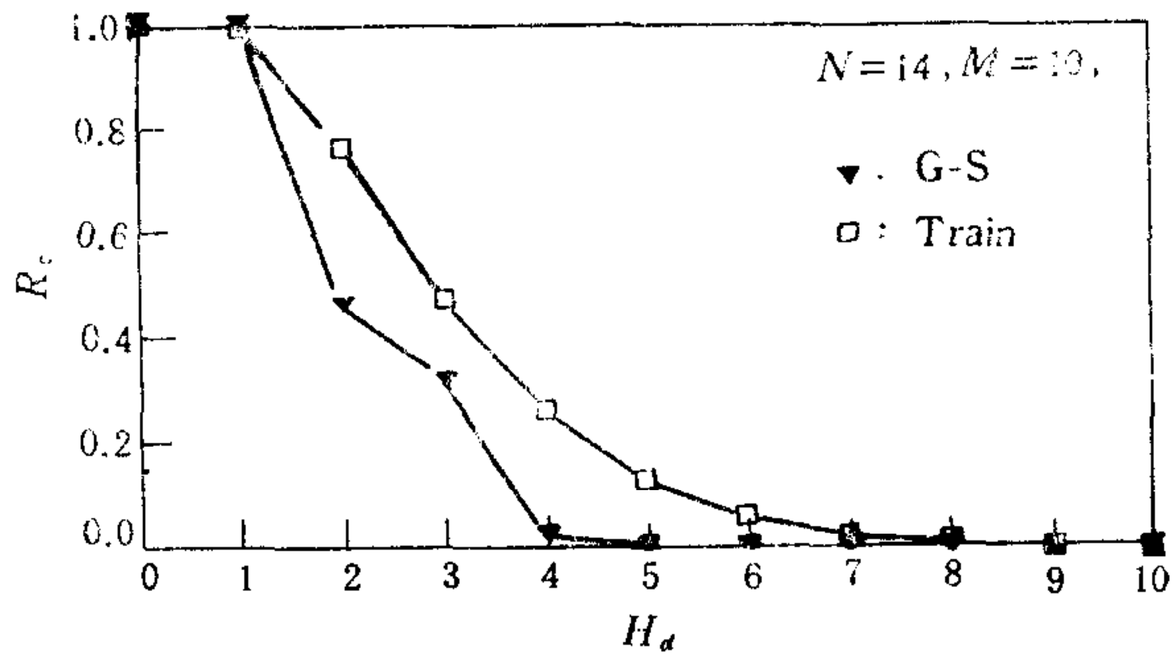


图 2　不同网络容错曲线的比较

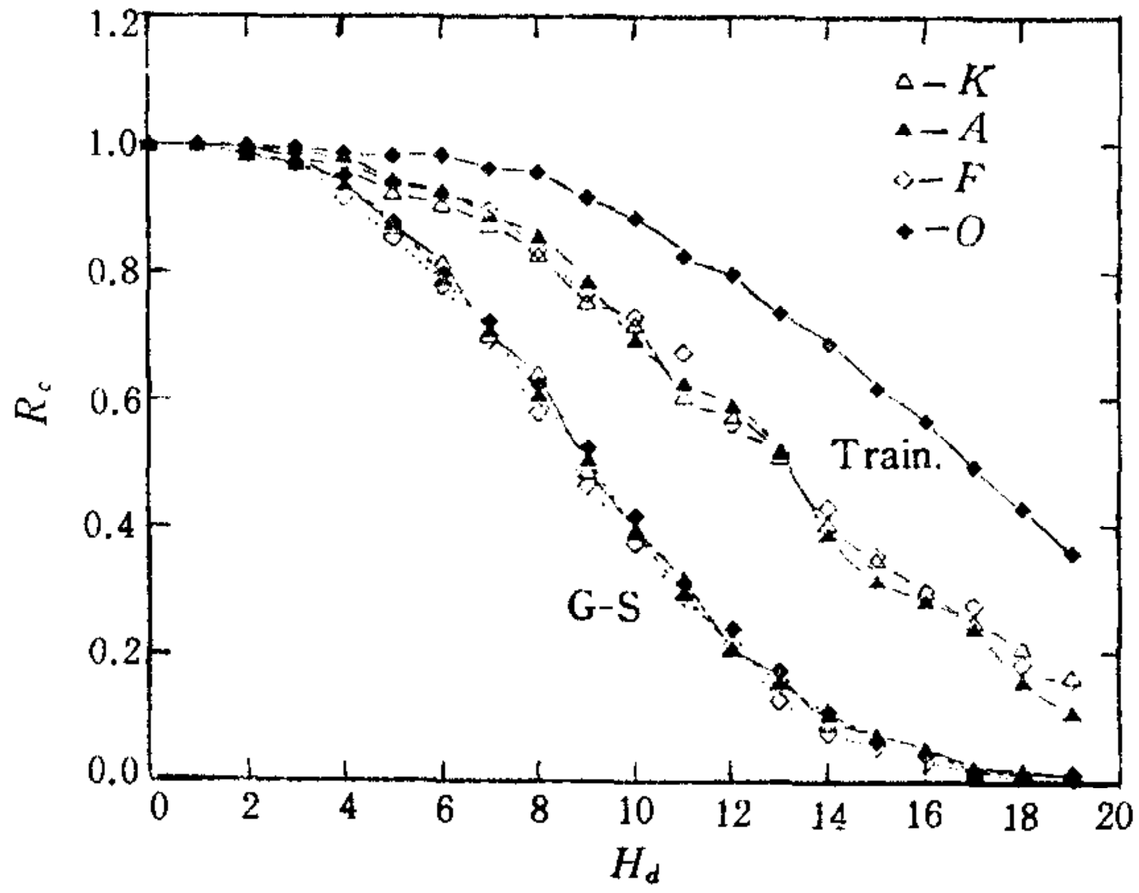
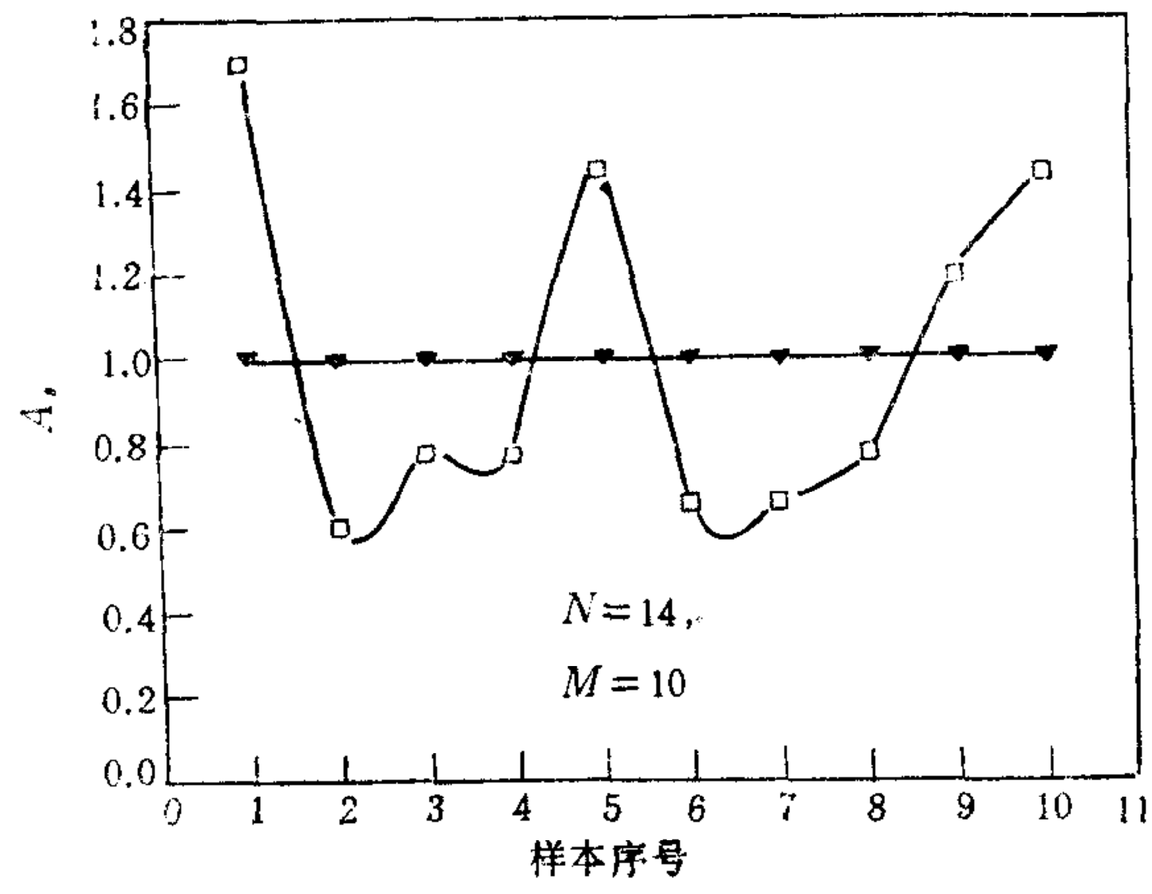
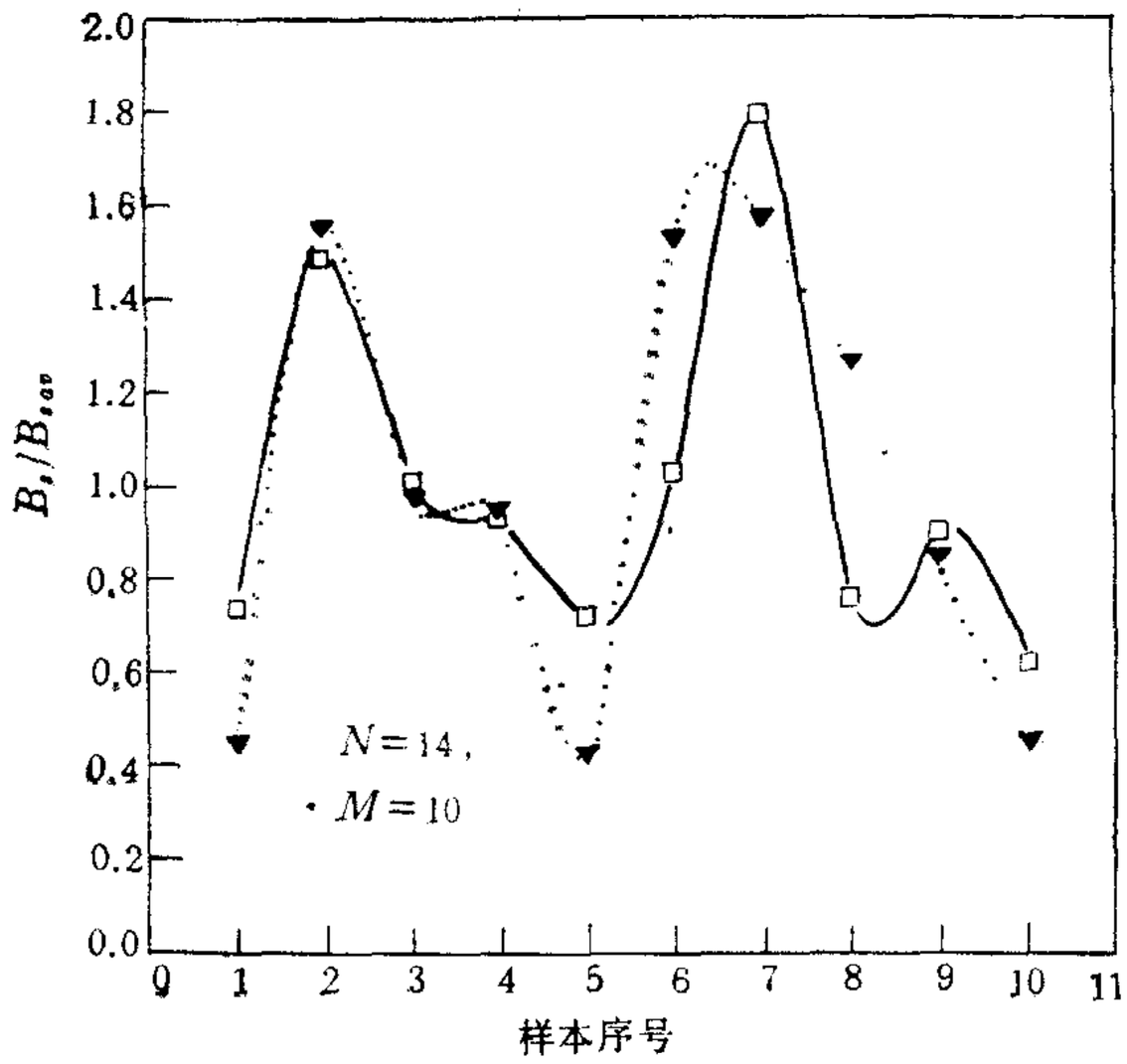


图 3　不同网络容错曲线的比较



(a)



(b)

图 4 各样本相对吸引域的控制

4 训练网络的对称性与收敛性问题

通常认为演化网络必须对称，否则不能保证网络收敛(即只有不动点吸引子)．上述训练方案不保证网络对称，也未采取对称化措施，自然会产生网络是否收敛的疑问．大量计算表明，除个别情况有极少数状态可能进入短周期解外，训练网络几乎都收敛．其原因分析如下．

首先讨论网络对称度的描述．对称性条件 $W_{ij}=W_{ji}(i,j=1,2,\cdots,N)$ 涉及 $N(N-1)/2$ 个条件，这是一个很严的要求．能否用少数(例如一个)参数刻划网络对称程度呢？分析表明，一个合适的参数是[6]

$$S_{\mathrm{sym,max}}=\max\{S_{\mathrm{sym}}(LW);\forall L\}. \tag{6}$$

其中

$$S_{\mathrm{sym}}(W)=\mathrm{Trace}(WW)/\mathrm{Trace}(WW^T); \tag{7}$$

$\mathrm{Trace}(\cdot)$ 是矩阵求迹；L 是任意正对角矩阵(即 $L_{ii}>0$; $L_{ij}=0$, $i\neq j$)．大量计算表明[6]，网络收敛条件比严格对称要求(即 $S_{\mathrm{sym,max}}=1$) 宽得多，通常当 $S_{\mathrm{sym,max}}\lesssim 0.6$ 时网络就几乎收敛(即仅个别情况有极少数状态可能进入短周期解)．此类网络可称为本质上准对称的．

图 5 给出了从不同初条件出发，网络对称性参数随训练时间的变化过程及最终的 $S_{\mathrm{sym,max}}$ 值．可见：(1) 训练网络虽非严格对称，但本质上是准对称的，因而是几乎收敛的；(2) 若从不对称的随机初条件出发，S_{sym} 值随训练时间单调上升，最后才达到准对称状态．所以，充分训练不仅可扩大吸引域，且有助于网络收敛．

最后指出，准对称网络的收敛性还与其演化模式有关．表 1 给出两种模式下收敛性的比较．其中 SDAN 是“最陡下降异步网络”[6]，即每次改变使“能量”下降最快的神经元状态；SQAN 是“顺序异步网络”，即按顺序改变神经元状态．表中的伪吸引子是指样

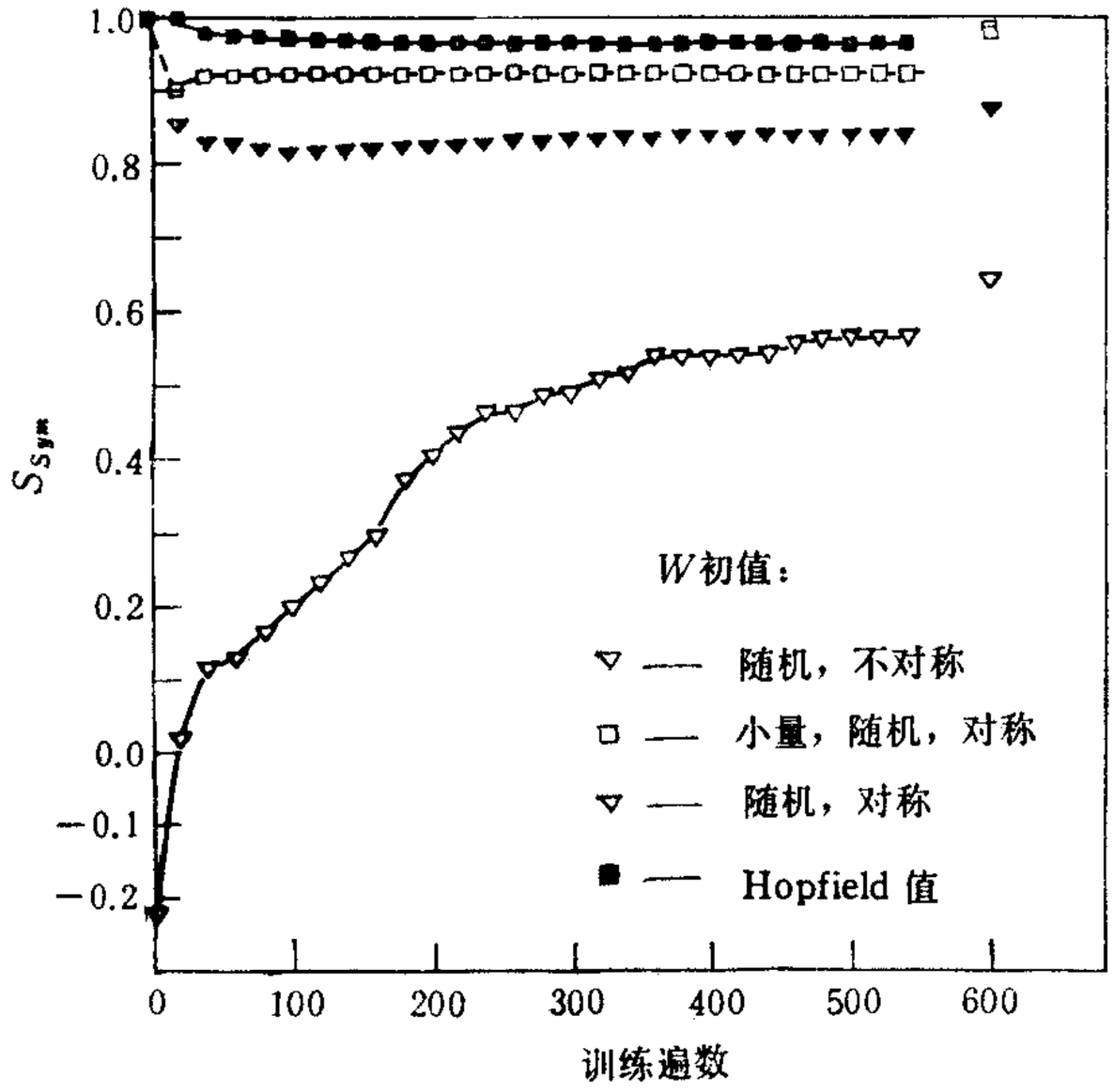


图 5　网络对称性参数随训练时间的变化

本集之外的吸引子，以 $nC_m(x)$ 表示有 n 个 m 周期的伪吸引子，其吸引域总体积为 x. 由表可见，SDAN 没有 $m>1$ 的吸引子. 即没有周期解，网络均收敛；而 SQAN 则可有少量（与总状态数 $\Omega_0 = 2^{14} = 16384$ 相比）态进入短周期解. 说明 SDAN 具有更强的抗不对称能力，且更易收敛.

表 1　训练网络的对称性与收敛性 $N=14$, $M=7$

初 条 件	最终 $S_{sym,max}$	演化模式	伪吸引子
随机不对称	0.64	SDAN	$6c_1(2015)$
		SQAN	$6c_1(2212), 4c_4(474), 5c_5(172), 1c_8(40)$
随机对称	0.87	SDAN	$19c_1(8417)$
		SQAN	$19c_1(9568), 2c_4(462)$
随机小量	0.94	SDAN	$17c_1(5315)$
		SQAN	$17c_1(6292), 1c_4(16)$
Hopfield 模　型	0.98	SDAN	$16c_1(8374)$
		SQAN	$16c_1(7580)$

5　结论

通过阱深参数对联想记忆神经网络的样本吸引域进行宏观控制是可行的. 训练网络可以达到贮存容量 $\alpha \gtrsim 1$ 而仍有良好的容错性，明显优于 Hopfield 的外积法、Gram-

Schmidt 正交化外积法、赝逆法等常用方案. 训练网络虽非严格对称,但是准对称的, 网络演化时是几乎全收敛的.

参 考 文 献

[1] Kohonon T. Self-organization and associative memory. Berlin: Springer-Verlag, 1984.

[2] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities. Proc. Nati. Acad. Sci. U. S. A., 1982, **79**: 2554.

[3] Hopfield J J. Neurons with graded response have colletive computational properties. Proc. Nat. Acad. Sci. U. S. A., 1984, **81**: 3088.

[4] Hopfield J J, Tank D W. *Biol. Cybern.*, 1985, **52**: 141.

[5] Rumelhart D E et al. Parallel distributed processing: Exploration in the microstructure of cognition. Cambridge MA, MIT Press, 1986.

[6] Zhang C F, Wang X J, Zhao K H. On the dynamical properties of evolutionary neural network. *Commun. Theor. Phys.*, 1992, **18**: 233.

[7] Lee Y C et al. *Physica*. 1986, **D22**: 276.

[8] 张承福,赵 刚. 关于联想记忆神经网络的若干问题. 自动化学报,1994,**20**(5): 513—521.

[9] Krauth W et al. *J. Phys.*, 1987, **A20**: L745.

[10] Personnaz L et al. *J. Phys. (Paris) Lett.*, 1985, **46**: L359.

[11] Kanter I, Sompolinsky H. *Phys. Rev.*, 1987, **A35**: 380.

[12] Denker J S. *Physica*, 1986, **D22**.

ON THE TRAINING OF NEURAL NETWORK FOR ASSOCIATIVE MEMORY

ZHANG CHENGFU ZHAO GANG

(*Department of Physics, Beijing University, Beijing* 100083)

ABSTRACT

In this paper, an optimized training scheme of neural network for associative memory is proposed. We show that the basins of attraction for samples attractors can be controlled in some extent by a pitfall depth parameter, therefore, the fault-tolerance of network can be made as good as possible. Numerical simulations show that with this scheme, the capacity of network can reach $\alpha \gtrsim 1$ ($\alpha = M/N$, here N is the number of neurons and M is the number of stored samples) and still with good fault-tolerance. The results are much better than the popular schemes such as outer-product scheme, orthogonalized outer-product scheme, pseudo-inverse matrix scheme and etc.. The problems on symmetry and convergence of trained networks are discussed too.

Key words: Neural network, associative memory, fault-tolerance, attractor, basin of attraction.

张承福,赵 刚 简介及照片见本刊第20卷第 6 期.