Training noise adaptation in attractor neural networks

J. Phys. A: Math. Gen. 23 (1990) L175-L182. Printed in the UK

LETTER TO THE EDITOR

Training noise adaptation in attractor neural networks

K Y M Wong† and D Sherrington†

Department of Physics, Imperial College, London SW7 2BZ, UK

Received 23 October 1989, in final form 20 December 1989

Abstract. We consider synaptic neural networks which minimise the output error of the stored patterns when the input patterns are ensembles of their noisy versions with overlap m_t with the clean patterns. When m_t is infinitesimally less than 1, the network automatically attains maximal stability, confirming the usefulness of training noises in enhancing memory associativity. When m_t drops below 1, the field distribution has two bands for large m_t, and one continuous band for small m_t. Errorless retrieval is impossible for training noises of the order N^0. With the increase in training noise, the retrieval overlap deteriorates, although memory associativity does increase for sufficiently low storage.

Memory associativity (or content addressability) is a very important feature in attractor neural networks. This means that when a pattern $\{\xi_j = \pm 1; j = 1 \ldots N\}$ is stored in the network, the presentation of a noisy version $\{S_j\}$ of the pattern will be attracted, through the dynamics of the network, to a final configuration identical to, or at least highly correlated with, the stored pattern.

In the synaptic neural network, information about the stored patterns is encoded in the synapses J_{ij}. In the standard version of the network dynamics, the local field at a node i due to other nodes is evaluated at the tth time step, and the state at the next time step is updated according to the sign of the field, i.e.

$$S_i(t+1) = \mathrm{sgn}\left(\sum_j J_{ij} S_j(t)\right). \tag{1}$$

To monitor the associative retrieval of a stored pattern $\{\xi_j\}$, it is useful to consider the overlap $m(t)$ at each step, given by

$$m(t) = \frac{1}{N} \sum_i \xi_i S_i(t). \tag{2}$$

If $m(t)$ becomes identical to or close to 1, the pattern is successfully retrieved.

Two ways have been proposed to encode the synaptic matrix J_{ij} so that the network enhances its memory associativity. The first method [1-4] is to modify the synaptic matrix stepwise according to the perceptron learning rule, whenever the (normalised) local fields of the stored patterns at a node i do not satisfy the stability requirement

$$\xi_i^\mu \frac{1}{\sqrt{c}} \sum_j J_{ij} \xi_j^\mu > K \tag{3}$$

where $\Sigma_j J_{ij}^2 = c$, c being the connectivity of a node. At the storage ratio $\alpha = p/c$, the

† Present address: Department of Theoretical Physics, University of Oxford, 1 Keble Road, Oxford OX1 3NP, UK.

maximal stability $K(\alpha)$ is given by [1]

$$\alpha^{-1}=\int_{-\infty}^{K(\alpha)}\frac{\mathrm{d}t}{\sqrt{2\pi}}\mathrm{e}^{-t^2/2}(K(\alpha)-t)^2. \tag{4}$$

The second method, called 'training with noise' [5], is to update stepwise the synaptic matrix, again according to the perceptron learning rule, but with the input configuration $\{R_j^\mu\}$ slightly distorted by random noise. This means that

$$P(R_j^\mu)=\tfrac{1}{2}(1+m_\mathrm{t})\delta(R^\mu-\xi_j^\mu)+\tfrac{1}{2}(1-m_\mathrm{t})\delta(R^\mu+\xi_j^\mu) \tag{5}$$

where we call m_t the training overlap. Here the rationale is that if the network is adapted to associating noisy inputs of a stored pattern with its correct output during training, then it is expected to possess the same capability during retrieval. Numerical simulations have shown the success of this scheme, but no theoretical analyses have been provided.

In this letter we consider networks which minimise the averaged output error of the stored patterns $\{\xi_j^\mu\}$ when the input patterns $\{R_j^\mu\}$ are ensembles of their noisy versions according to (5). We define the cost function to be minus the output overlap. In reminiscence of Gardner and Derrida [6], we then proceed to find the corresponding free energy defined in the space of weights J_{ij}, and quench-averaged over a random distribution of p patterns. The zero temperature limit of this free energy yields the ground-state (optimal) cost function.

Optimising the network retrieval of ensembles of noisy input patterns is equivalent to presenting similarly noisy examples to the network during training and finding the synaptic matrix that gives the least averaged output error. The previously proposed 'training with noise' algorithm can be considered as a stepwise attempt to reduce the output errors to a level which makes associative retrieval possible, but is by no means optimal. Our study of networks that minimise the output error at the input overlap m_t will therefore provide an upper bound to the network retrieval quality of the 'training with noise' procedure, and will reveal qualitative effects of training noises.

Obviously, the case $m_\mathrm{t}=1$ corresponds to the noiseless case of Gardner and Derrida [6] with *zero* stability. As we shall see, however, the network *automatically* attains *maximal* stability given by (4) when $1-m_\mathrm{t}$ is infinitesimally small. This discontinuity in network behaviour demonstrates the usefulness of training noises in enhancing the memory associativity of the system. Further increase in training noise enlarges the basins of attraction for sufficiently low storage, but causes disruption of the stored pattern. When $m_\mathrm{t}<1-\mathrm{O}((\ln\ln N)^{-1})$, errorless retrieval becomes impossible.

Let us start by minimising the output error (or equivalently, maximising the output overlap) in *one* time step when the input overlap is m_t. Since the optimisation on any one node is independent of the others, it is sufficient to consider the cost function defined on a single node i, which is

$$C_i=-\sum_\mu P(\boldsymbol{R}_i^\mu)\xi_i^\mu\,\mathrm{sgn}\left(\frac{1}{\sqrt{c}}\boldsymbol{J}_i\cdot\boldsymbol{R}_i^\mu\right) \tag{6}$$

where $\boldsymbol{R}_i^\mu$ represents the c-component input state of the noisy patterns in (5). (Subscripts i are hereafter implicit.) Performing the average over $\boldsymbol{R}_i^\mu$ explicitly, this cost function is reduced to $C=-\Sigma_\mu g(\Lambda^\mu)$, where $\Lambda^\mu=1/\sqrt{c}\xi^\mu\boldsymbol{J}\cdot\boldsymbol{\xi}^\mu$ is the (normalised) local field of the (clean) μth pattern, and

$$g(\Lambda)=-\mathrm{erf}\left(\frac{m_\mathrm{t}\Lambda}{\sqrt{2(1-m_\mathrm{t}^2)}}\right). \tag{7}$$

The corresponding free energy is now obtained by an 'annealed' average over the space of the synaptic weights J_j, treating the p patterns as 'quenched' variables. The partition function at a temperature $T=\beta^{-1}$ is then given by

$$Z=\prod_j \int \mathrm{d}J_j\,\delta\Big(\sum_j J_j^2-c\Big)\exp\Big(\beta\sum_\mu g(\Lambda^\mu)\Big). \tag{8}$$

The pattern-averaged free energy is obtained by the replica formula

$$\langle\!\langle \ln Z\rangle\!\rangle=\lim_{h\to 0}\frac{1}{n}\left(\langle\!\langle Z^n\rangle\!\rangle-1\right) \tag{9}$$

where $\langle\!\langle\ \rangle\!\rangle$ represents averaging over the p input patterns. Using the techniques of Gardner and Derrida [6], we can show that, in the replica symmetric ansatz, the optimal output overlap $f(m_t)$ (when the input overlap is m_t) is given by

$$\begin{aligned} f(m_t)&=\lim_{\beta\to\infty}\frac{z}{\beta\alpha c}\langle\!\langle \ln Z\rangle\!\rangle \\ &=\min_x\left\{\int\frac{\mathrm{d}t}{\sqrt{2\pi}}\,\mathrm{e}^{-t^2/2}\max_\theta\left(g(\theta)-\frac{2}{x^2}(\theta-t)^2\right)+\frac{2}{\alpha x^2}\right\}. \end{aligned} \tag{10}$$

Evaluating the maximum of θ explicitly, we obtain

$$f(m_t)=\int\frac{\mathrm{d}t}{\sqrt{2\pi}}\,\mathrm{e}^{-t^2/2}\operatorname{erf}\left(\frac{m_t\theta(t)}{\sqrt{2(1-m_t^2)}}\right) \tag{11}$$

where θ and t are related by

$$t(\theta)=\theta-\frac{x^2}{4}g'(\theta) \tag{12}$$

and the extremal condition for x becomes

$$\alpha^{-1}=\int\frac{\mathrm{d}t}{\sqrt{2\pi}}\,\mathrm{e}^{-t^2/2}(\theta(t)-t)^2. \tag{13}$$

A word of caution has to be expressed about the inverse function $\theta(t)$ of $t(\theta)$. For sufficiently large m_t (and consequently x), θ is a multivalued function of t in a range of the argument, and care must be taken to choose the correct θ. Because of the maximisation requirement in (10), there exists a discontinuity of θ as a function of t at $t_0=t(\theta_>)=t(\theta_<)$, where

$$g(\theta_>)-\frac{2}{x^2}(\theta_>-t_0)^2=g(\theta_<)-\frac{2}{x^2}(\theta_<-t_0)^2. \tag{14}$$

This is equivalent to the Maxwell's construction (see figure 1)

$$\int_{\theta_<}^{\theta_>}\mathrm{d}\theta\,t(\theta)=t_0(\theta_>-\theta_<) \tag{15}$$

used in the theory of first-order thermodynamic phase transitions. Thus, by discarding the range of θ between $\theta_>$ and $\theta_<$, the function $\theta(t)$ is now single valued.

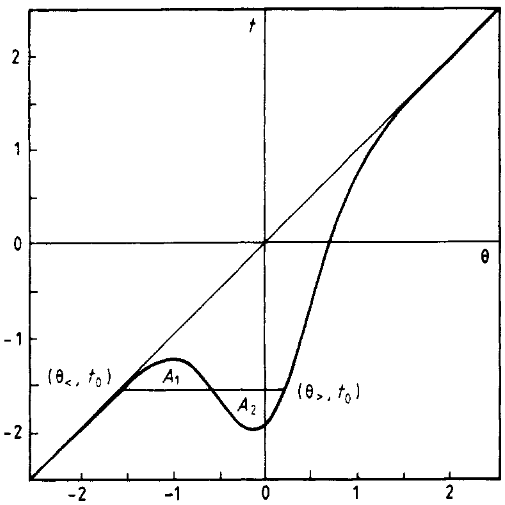

Figure 1. The Maxwell construction for $\theta(t)$. The points $(\theta_>, t_0)$ and $(\theta_<, t_0)$ are chosen such that the areas A_1 and A_2 are equal. The continuous full curve $t(\theta)$ is given by equation (12), but the physical curve for $\theta(t)$ has the discontinuity indicated. Here $\alpha = 1.5$ and $m_t = 0.9$.

We are also interested in the network retrieval. For an arbitrary input overlap m, the output overlap $\tilde{f}(m)$ is given by [7-9]

$$\tilde{f}(m) = \int \mathrm{d}\Lambda\, \rho(\Lambda)\, \mathrm{erf}\left(\frac{m\Lambda}{\sqrt{2(1-m^2)}}\right) \tag{16}$$

where $\rho(\Lambda)$ is the pattern field distribution

$$\rho(\Lambda) = \lim_{\substack{n\to 0\\ \beta\to\infty}} \left\langle\!\!\left\langle \prod_{\alpha=1}^{n} \left(\int \mathrm{d}J_j^\alpha\, \delta\left(\sum_j (J_j^\alpha)^2 - c\right)\right) \exp\left(\beta \sum_{\alpha\mu} g(\Lambda_\alpha^\mu)\right) \delta(\Lambda_a^\nu - \Lambda)\right\rangle\!\!\right\rangle. \tag{17}$$

In our noise-optimised network, this becomes

$$\rho(\Lambda) = \int \frac{\mathrm{d}t}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\, \delta(\theta(t) - \Lambda) \tag{18}$$

yielding

$$\tilde{f}(m) = \int \frac{\mathrm{d}t}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\, \mathrm{erf}\left(\frac{m\theta(t)}{\sqrt{2(1-m^2)}}\right). \tag{19}$$

Note that when $m = m_t$, the output overlap is indeed the optimised output overlap given in (11). (Hereafter we shall drop the distinction between f and $\tilde{f}$ for symbolic unity.)

Because of the discontinuity in θ, there exist two separate bands in the pattern field distribution $\rho(\Lambda)$ for sufficiently large m_t.

Having derived these basic results, let us consider various cases. First consider $m_t = 1$, when $g(\theta)$ in (10) becomes sgn θ. This yields a storage capacity for errorless output at $\alpha_c = 2$, where the pattern field distribution is equal to a Gaussian of unit width truncated at $\Lambda = 0$, plus a delta function of weight 1/2 at $\Lambda = 0$, agreeing with the results of Gardner and Derrida [6].

Next, consider m_t slightly less than 1. In this regime, $x \gg 1$, and we can assume that $-t_0 = -\theta_{<} \gg 1$. The Maxwell construction requires that $\theta_{>} - \theta_{<} = x$, and this in turn requires $\theta_{>} = \theta_1/\sqrt{2}$, where θ_1 is related to x via

$$\frac{x^2}{2}\exp\left(-\frac{m_t^2\theta_1^2}{\sqrt{2(1-m_t^2)}}\right)(2\pi(1-m_t^2)/m_t^2)^{-1/2} = 1. \tag{20}$$

Noting that

$$\theta(t) = \begin{cases} t & t > \theta_1 \text{ and } t < \theta_{<} \\ \theta_1 & \theta_1 > t \sim O(1) \end{cases} \tag{21}$$

then θ_1, and hence x, can be derived from the extremal condition

$$\alpha^{-1} = \int_{-\infty}^{\theta_1} \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}(\theta_1 - t)^2 \tag{22}$$

identifying θ_1 to be $K(\alpha)$, by virtue of (4). The field distribution $\rho(\Lambda)$ is then a Gaussian of width 1 truncated at $\Lambda = K$, plus a delta function at $\Lambda = K$, identical to the field distribution at maximal stability [7-9]. Since it has been proposed [5] that if a solution exists for a noisy training ensemble the 'training with noise' algorithm will converge to it, this implies that by introducing an infinitesimally small training noise, the perceptron learning algorithm automatically results in a network with maximal stability after sufficiently long training, although the stability requirement need not be imposed at each training step.

The discontinuity of network behaviour from $m_t = 1$ to $m_t = 1^-$ can be traced to a discontinuity of the training ensemble in the two cases. In the $m_t = 1$ training ensemble, all the examples of a pattern are identical, and the network is merely adapted to the retrieval of clean patterns, resulting in a memory associativity far from maximal. On the other hand, the $m_t = 1^-$ ensemble contains a full range of distinct noisy example patterns, each of whose weights decreases with its Hamming distance from the clean patterns. The network is therefore adapted to the retrieval of noisy patterns, resulting in the maximal stability.

When m_t falls further below 1, the field distribution $\rho(\Lambda)$ starts to develop two bands. The upper band is bounded below by $\theta_{>}$, and the delta function peak obtained at $\Lambda = K$ for $m_t = 1^-$ degenerates into a broader, but still sharp, peak. The lower band is bounded above by $\theta_{<}$ and its weight increases with training noise. Here

$$\rho(\Lambda) = \left(\int_{K(\alpha)}^{\infty} + \int_{-\infty}^{(K(\alpha)/\sqrt{2})-x}\right)\frac{dt}{\sqrt{2\pi}} e^{-t^2/2}\,\delta(t-\Lambda) + \int_{(K(\alpha)/\sqrt{2})-x}^{K(\alpha)} \frac{dt}{\sqrt{2\pi}} e^{-t^2/2}\,\delta(\theta(t)-\Lambda) \tag{23}$$

where x is given by (20) with $\theta_1 = K(\alpha)$, and in the second term $\theta(t) \sim K(\alpha)$ for $t \sim O(1)$. The optimised output overlap $f(m_t)$ is given by

$$\tfrac{1}{2}[1 - f(m_t)] \sim O\left[\exp\left(-\frac{m_t^2K^2(\alpha)}{2(1-m_t^2)}\right)\right] \tag{24}$$

implying that errorless retrieval at the training overlap m_t is possible only up to a training noise $d_t \equiv \frac{1}{2}(1 - m_t) = K^2(\alpha)/8 \ln N$.

If, however, we are interested in errorless retrieval for clean input, the restriction on training noise is less stringent. The storage of clean patterns (i.e. input overlap being 1) is determined by $f(1)$, which we shall call the storage overlap. Their outputs are errorless provided

$$f(1)=\operatorname{erf}\left(\frac{x-K/\sqrt{2}}{\sqrt{2}}\right)\sim 1-\mathrm{O}(N^{-1}) \tag{25}$$

or $d_t = K^2(\alpha)/8 \ln\ln N$. This shows that any training noise of the order N^0 results in the disruption of the stored patterns, revealing a disadvantage of the 'training with noise' scheme. Because of the double logarithmic dependence on N this restriction, however, is not too stringent.

A further reduction in m_t results in a further deterioration of both the optimised overlap $f(m_t)$ and the storage overlap $f(1)$, a narrowing of the band gap and a smoothening of the peak in the field distribution. To study the effects on the final overlap and the basins of attraction, we restrict our discussion to the case of large but dilute connectivity, namely $1 \ll c \ll \ln N$, when the one-step relation between input and output overlaps can be extended to successive retrieval steps, since time correlations of the configurations may be neglected [10]. Starting from an initial configuration close to a stored pattern, the final overlap after iteration is then given by a stable fixed point $m^* = f(m^*)$, and the basin boundary of attraction is defined by an unstable fixed point m_B. We find that training noise indeed enlarges the basin of attraction (i.e. reduces m_B) for sufficiently low storage, although the fixed point overlap m^* inevitably deteriorates because of training noise disruption. (For high storage levels, training noise may shrink the basin of attraction near the first-order-retrieval-non-retrieval transition. The reader is referred to [12] for details.)

At $m_t = m_c$ where $t'(\theta_c) = t''(\theta_c) = 0$, the two bands merge. In the extremely noisy limit $m_t \to 0$, the field distribution $\rho(\Lambda)$ becomes a Gaussian of width 1, and mean $1/\sqrt{\alpha}$. This means that the system becomes a Hebb-rule network with $J_{ij} = \Sigma_\mu \xi_i^\mu \xi_j^\mu / \sqrt{\alpha} c$. Our recent studies [11] showed that the Hebb-rule net minimises the output error among all Boolean networks in the high training noise limit. Since synaptic networks can be emulated by Boolean networks, the Hebb-rule must also minimise the output error among all synaptic networks in that limit, which is indeed the present result.

Figure 2 shows the training noise dependence of the optimised overlap $f(m_t)$, the storage overlap $f(1)$, the fixed point overlap $m^* = f(m^*)$ and the boundary overlap m_B. Figure 3 shows the training noise dependence of the field distribution $\rho(\Lambda)$ from the maximally stable limit ($m_t = 1^-$) to the Hebb-rule limit ($m_t \to 0$).

We now have a clearer picture of the differences between the maximally stable perceptron network and the Hebb-rule network. Since the maximally stable perceptron is associated with low training noise, it has a higher storage capacity, less retrieval error, but weaker memory associativity (i.e. smaller basins of attraction). On the other hand, the Hebb-rule net is associated with high training noise, giving it lower storage capacity, larger retrieval error, but stronger memory associativity. In fact, the basins of attraction are so large that the transition from retrieval to non-retrieval is second order on increasing α.

Finally, we discuss the noise-optimal networks in terms of the universality classes of Abbott and Kepler [13]. Identifying the Boltzmann factor $e^{-\beta g(\Lambda^\mu)}$ as their *a priori* weight, we see that for m_t below m_c, where $1-(x^2/4)g'' > 0$ holds, we have the one-band field distribution belonging to the 'singular' universality class of Abbott and Kepler. In particular, as $m_t \to 0$, $g(\theta) \sim \theta$ gives their result for a Hebb matrix. On the other hand, for m_t above m_c, the two-band field distribution belongs to a new universality class. As $m_t \to 1^-$, however, this class approaches the Gardner case.

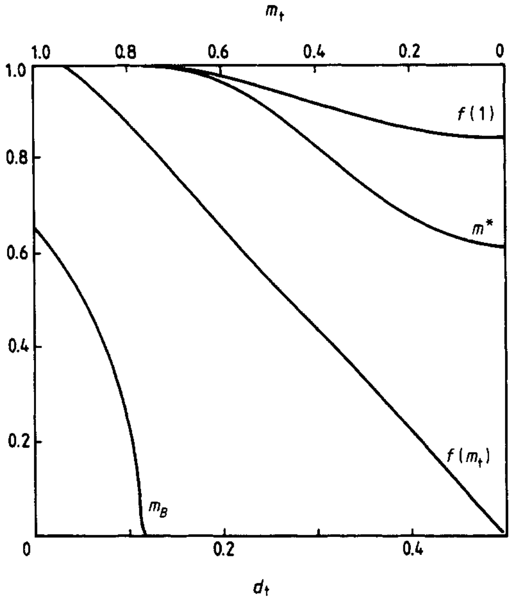

Figure 2. The dependence on the training noise $d_t \equiv \frac{1}{2}(1-m_t)$ of the optimised overlap $f(m_t)$, storage overlap $f(1)$, fixed point overlap m^* and boundary overlap m_B for $\alpha = 0.5$. Here $m_B = 0$ for $d_t > 0.11$.

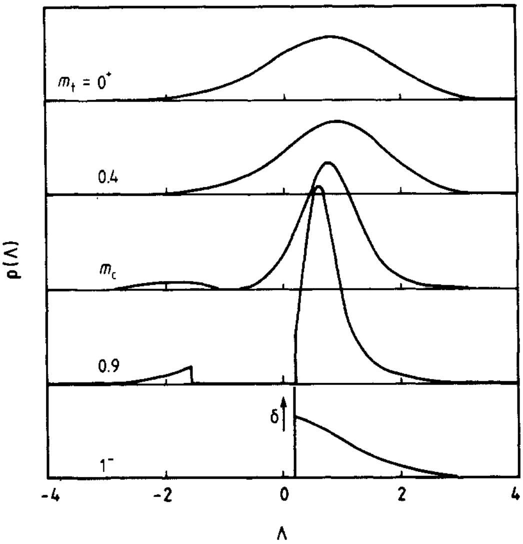

Figure 3. The field distribution $\rho(\Lambda)$ for different training overlaps from the maximally stable limit to the Hebb-rule limit, at $\alpha = 1.5$. Starting from the bottom, the curves correspond to $m_t = 1^-$, 0.9, m_c, 0.4 and 0^+. Here $m_c = 0.7798$. The vertical scale has units such that the separation between horizontal axes is 0.6, but the zero is reset for each curve in the sequence; the asymptotic values of $\rho(\Lambda)$ at large Λ are zero. The curve for $m_t = 1^-$ has a delta function peak at $K(\alpha) = 0.1861$.

The idea of using noisy example patterns for training has also been applied recently to the feedforward network [14-16].

We thank C Campbell, D Amit and M Evans for meaningful discussions. This work was supported partially by SERC.

References

[1] Gardner E 1987 *Europhys. Lett.* **4** 481; 1988 *J. Phys. A: Math. Gen.* **21** 257
[2] Diederich S and Opper M 1987 *Phys. Rev. Lett.* **58** 949
[3] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
[4] Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
[5] Gardner E, Stroud N and Wallace D J 1989 *J. Phys. A: Math. Gen.* **22** 2019
[6] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[7] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1657
[8] Krauth W, Nadal J-P and Mézard M 1988 *J. Phys. A: Math. Gen.* **21** 2995
[9] Gardner E 1989 *J. Phys. A: Math. Gen.* **22** 1969
[10] Derrida B, Gardner E and Zippelius A 1987 *Europhys. Lett.* **4** 167
[11] Wong K Y M and Sherrington D 1989 *Europhys. Lett.* **10** 419
[12] Wong K Y M and Sherrington D 1990 Retrieval properties of noise-optimal attractor neural networks, to be published
[13] Abbott L F and Kepler T B 1989 *J. Phys. A: Math. Gen.* **22** 2031
[14] Refregier Ph and Vignolle J-M 1989 *Europhys. Lett.* **10** 387
[15] Hansel D and Sompolinsky H 1989 Learning from examples in a single layer neural network *preprint*
[16] Gyorgyi G and Tishby N 1989 Statistical theory of a learning rule *Proc. STATPHYS 17 Workshop on Neural Networks and Spin Glasses* (Singapore: World Scientific) in press