Associative memory networks, fault-tolerance and coding theory

Norman Hendrich
Fachbereich Informatik, Universität Hamburg,
Troplowitzstraße 7, D-2000 Hamburg 54, Germany
e-mail: hendrich@rz.informatik.uni-hamburg.dbp.de

December 5, 1990

Abstract

The binary-couplings neural network used as an associative memory is studied under the effects of synaptic damage. The model of damage is to flip synapses ($J_{ij} \to -J_{ij}$) randomly with a probability λ.

A simple model based on a network with constant stabilities is proposed that allows the calculation of the storage properties of the damaged networks. The networks prove to be fault-tolerant at low storage densities, the fraction of bit-errors in the stored patterns being finite but very small up to high concentrations of flipped synapses.

Results are compared with some standard error-correcting codes and some implications for coding-theory are discussed.

1 Introduction

The properties of neural networks used as associative memories have been studied intensively during the last few years. The basic idea is to store $P = \alpha N$ binary patterns $\xi_i^\mu = \pm 1$, $\mu = 1, \ldots, P$, $i = 1, \ldots, N$ as the fixed points of the dynamics of a neural network with N neurons S_i [Hopfield 82]. Usually a parallel dynamics is used,

$$S_i(t+1) = \operatorname{sgn}\left(||J_i||^{-1} \sum_j J_{ij} S_j(t) \right), \tag{1}$$

and therefore the patterns are stored if $\xi_i^\mu = \operatorname{sgn}(||J_i||^{-1} \sum_{j \neq i} J_{ij} \xi_j^\mu)$. To ensure finite basins of attraction a positive stability κ is introduced and the synapses are chosen to fulfil the stronger constraints

$$\kappa_{i\mu} = \xi_i^\mu \cdot h_i^\mu = \xi_i^\mu \cdot \left(||J_i||^{-1} \sum_j J_{ij} \xi_j^\mu \right) \geq \kappa > 0. \tag{2}$$

The storage capacity of such networks with real-valued couplings and a spherical normalization $||J_i|| = (\sum_j J_{ij}^2)^{1/2} = N$ has been solved recently [Gardner 87] and is known to be

$$\alpha_c(\kappa) = \left[\frac{1}{\sqrt{2\pi}} \int_{-\kappa}^{\infty} \mathrm{d}t \; \mathrm{e}^{-t^2/2} \, (t+\kappa)^2 \right]^{-1}, \tag{3}$$

so that $\alpha_c(0) = 2.0$. Information is not stored locally by this networks, but rather using a distributed representation.

The question whether the networks will gain fault-tolerance because of the distributed representation of information is therefore an interesting problem.

Biological neural networks seem to be especially robust against damaged couplings or neurons. The simple model proposed below may be used to explain this robustness and calculate the effects of damage on the storage properties of a neural network.

2 The damaged binary couplings neural network

Random dilution of synapses is a simple and biologically plausible model of damage in neural networks with real-valued couplings [Hendrich 90]. However, it is not simple to determine the information content of a real-valued coupling-matrix J_{ij}.

In order to compare the performance of damaged neural networks with some standard models of coding-theory one needs to investigate the binary-couplings ($J_{ij} = \pm 1$) network. The storage capacity of this model is known to be $\alpha_{cB} = 0.83$ [Gardner & Derrida 88] [Krauth & Mézard 89] [Krauth & Opper 89], but a storage density higher than approximately $\alpha = 0.4$ cannot be reached with the currently available learning algorithms [Koehler *et. al.* 89]. The adequate model of damage in the binary-couplings network is to flip synapses with a probability λ, that is,

$$J'_{ij} = \begin{cases} -J_{ij} & \text{with probability } \lambda, \\ J_{ij} & \text{with probability } 1-\lambda. \end{cases} \tag{4}$$

For any stability $\kappa > 0$ the patterns ξ_i^μ will be stored perfectly up to a concentration of flipped synapses of $\lambda < \kappa/(2\sqrt{N})$. However, one might expect that the patterns will remain stored — with some fraction of bit-errors — up to a much higher level of damage.

To calculate the effects of this damage the distribution of stabilities in the network is studied. Assume a neuron with constant stabilities $\kappa_{i\mu} = \kappa_0$ for all patterns ξ_i^μ, $\mu = 1 \ldots P$ before any damage is introduced to the network, that is, the initial distribution of stabilities is a δ-function,

$$\rho^{(0)}(\kappa_{i\mu}) = \delta(\kappa_{i\mu} - \kappa_0). \tag{5}$$

After damaging (flipping a finite fraction of the couplings) the distribution of stabilities $\rho^{(\lambda)}(\kappa_{i\mu})$ will be approximately gausssian, with some mean value κ_λ and width σ_λ. Because of condition (5) the terms $J_{ij}\xi_j^\mu\xi_i^\mu$ in the calculation of the stabilites are correlated, $\langle J_{ij}\xi_j^\mu\xi_i^\mu \rangle \approx \kappa_0/\sqrt{N}$, and the mean-stability will be $\kappa_\lambda = (1-2\lambda)\kappa_0$. The calculation of the width of the distribution $\rho^{(\lambda)}(\kappa)$ is more complicated. Numerical evidence is that σ_λ as a function of λ is given by $\sigma_\lambda = 2\sqrt{\lambda(1-\lambda)}$. (Note that the flipping of all couplings $J_{ij} \to -J_{ij}$ implies $\kappa_{i\mu} \to -\kappa_{i\mu}$. That is, κ_λ and σ_λ as as a function of λ must obey $\kappa_\lambda = -\kappa_{(1-\lambda)}$ and $\sigma_\lambda = \sigma_{(1-\lambda)}$.) Therefore the distribution of stabilities in the damaged network will be

$$\rho^{(\lambda)}(\kappa) = \frac{1}{\sqrt{2\pi}\,\sigma_\lambda} \exp\left[-\frac{1}{2}\frac{(\kappa-\kappa_\lambda)^2}{\sigma_\lambda^2}\right] \tag{6}$$

with $\kappa_\lambda = (1-2\lambda)\kappa_0$ and $\sigma_\lambda = 2\sqrt{\lambda(1-\lambda)}$ as above.

Now the fraction $\Gamma(\kappa_{\min}, \lambda)$ of patterns that are stored with stabilities $\kappa_{i\mu} < \kappa_{\min}$ per neuron can be calculated as a function of initial stability κ_0 and damage λ. One has

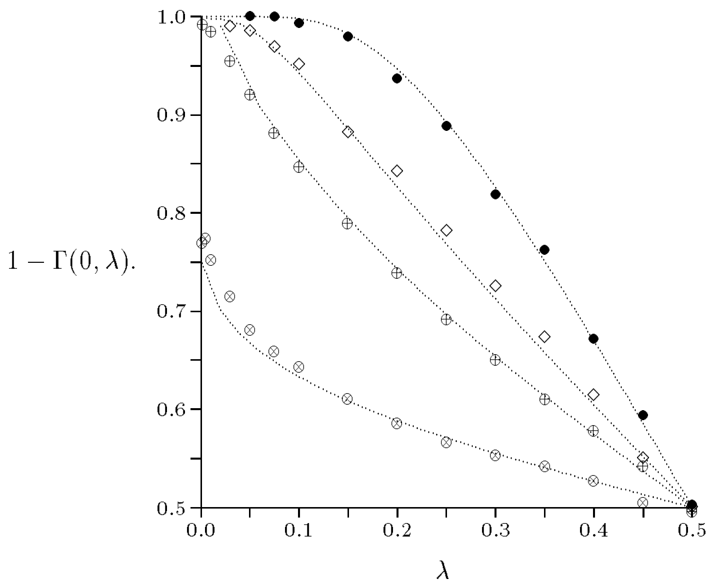
$$\Gamma(\kappa_{\min}, \lambda) = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\kappa_{\min} - \kappa_\lambda}{\sigma_\lambda\sqrt{2}}\right)\right), \tag{7}$$

and therefore the fraction of errors in the stored patterns is given by $\Gamma(0, \lambda)$.

If the initial network does not have constant stabilities, but its initial distribution $\rho^{(0)}(\kappa)$ of stabilities is known, the corresponding fraction $\Gamma(\kappa_{\min}, \lambda)$ of stabilities less than $\kappa_{\min}$ as a function of λ can be found by a superposition of constant-stability terms,

$$\Gamma(\kappa_{\min}, \lambda) = \frac{1}{2}\left(1 + \int \mathrm{d}\kappa_0 \;\; \rho^{(0)}(\kappa_0) \;\; \operatorname{erf}\left(\frac{\kappa_{\min} - \kappa_\lambda(\kappa_0)}{\sigma_\lambda\sqrt{2}}\right)\right). \tag{8}$$

The prediction (8) is plotted in figure 1 as a function of network damage λ, where the initial distribution of stabilities $\rho^{(0)}(\kappa)$ was taken from the simulations. Obviously the model is a rather accurate description of the storage properties of the damaged network at all storage densities. Note that the used learning algorithm of [Koehler *et. al.* 89] did not succeed to store the patterns perfectly at $\alpha = 0.4$.



Fraction of patterns stored perfectly per neuron $1-\Gamma(0,\lambda)$ as a function of concentration of damaged ("flipped") synapses. ($N = 256$, $\alpha = 0.05$ $\bullet$, 0.15 $\diamond$, 0.30 $\oplus$, 0.40 $\otimes$).
.

Figure 1: Fraction of patterns with $\kappa_{i\mu} > 0$

The calculation of the basins of attraction of a neural network model is still an unsolved problem. However, a few models relying on the distribution of stabilities [Kepler & Abbott 88] have been proposed. Therefore these models can be used to approximately calculate the basins of attraction in the damaged networks using the distribution (6).

3 Comparison with coding theory

The function of an associative memory is very similar to the function of an error-correcting code: The stored memory patterns correspond to the code vectors, and the task of the code is to recognize noisy input vectors as code vectors.

This is done (for the symmetric binary channel) by choosing the code vector which has the largest overlap with the input pattern. A linear code which stores k patterns of length N bit is called a (N, k) code. The code may correctly restore any input pattern with less than t errors (wrong bits) if it has minimum weight $d = 2t + 1$, that is, if any two code vectors differ in at least $2t + 1$ bits.

While a classification of neural networks using the methods of coding theory is far beyond the scope of this paper, it is interesting to compare the performance of the networks with some standard error-correcting codes. For sake of simplicity we choose Reed-Muller codes which are characterized by a parameter r, so that

$$k = 1 + \binom{m}{1} + \binom{m}{2} + \cdots + \binom{m}{r}. \qquad (9)$$

with $N = 2^m$. The resulting (N, k) code has a minimum weight $d = 2^{m-r}$. For instance, the *repetition codes* use $r = 0$ and may correct up to $t < N/2$ wrong bits. The $r = 1$ Reed-Muller code is a $(N, \ln_2 N)$ code that may correct up to $t < N/4$ errors.

A neural network which stores $P = \alpha N$ patterns is equivalent to a $(N, \ln_2 P)$ code, because $\ln_2 P$ bits suffice to index the patterns. As noted above, the accurate calculation of the basins of attraction of neural networks is still an open problem. Simulations indicate that neural networks at moderate storage densities $\alpha \leq 0.2$ have large basins of attraction and may correct up to $t \approx N/4 \ldots N/2$ errors.

That is, neural networks may be viewed as a new class of low rate random error-correcting codes. While the rate of the networks is less than the rate of $r = 1$ Reed-Muller codes they have greater error-correcting capabilities.

However, there is also a neural network impact on coding theory philosophy. Usually the encoder and decoder of a code are assumed to be fault-free, the transmission channel being the only source of noise. While this assumption is a good description of some tasks, for example satellite communications, it is questionable when dealing with neural systems: Any neural network should tolerate damage of some neurons and therefore the decoder itself has to be fault-tolerant.

Neural networks used as an associative memory are an elegant solution to the problem of designing a fault-tolerant decoder for an error-correcting code.

4 Conclusion

A simple model for the effects of random damage on the binary-couplings neural network has been proposed. Using the distribution of stabilities, the storage properties of the damaged networks can be calculated as a function of the concentration of damaged couplings. The predictions agree very well with simulations of damaged networks.

The networks prove to be quite fault-tolerant at low storage densities (which correspond to high values of the stabilities). However, the fraction of errors in the patterns as a function of the concentration of damaged synapses increases quite fast at higher storage densities.

The function of an associative memory is equivalent to the decoding of an error-correcting code. The binary-couplings associative memory neural networks have been shown to be decoders of low-rate random codes with capabilities comparable to standard error-correcting codes.

References

[Gardner 87] E. Gardner, Maximum Storage Capacity in Neural Networks, *Europhysics Letters* 4, 481–485 (1987).

[Gardner & Derrida 88] E. Gardner, B. Derrida, Optimal storage properties of neural network models, *Journal of Physics* A 21, 271–284 (1988).

[Hopfield 82] J. J. Hopfield, Neural networks and physical systems with emergent collective computational capabilities, *Proc. Nat. Acad. Sci.* 79, 2554–2558 (1982).

[Hendrich 90] N. Hendrich, Associative memory in damaged neural networks, submitted to *Journal of Physics* A (1990).

[Kepler & Abbott 88] T. B. Kepler and L. F. Abbott, Domains of attraction in neural networks, *Journal de Physique (France)* 49, 1657–1662 (1988).

[Koehler *et. al.* 89] H. Koehler, S. Diederich, W. Kinzel and M. Opper, Learning Algorithm for a Neural Network with Binary Synapses, *private communication* (Universität Giessen preprint).

[Krauth & Opper 89] W. Krauth and M. Opper, Critical storage capacity of the $J = \pm 1$ neural network, *Journal of Physics* A 22, L519–L523 (1989).

[Krauth & Mézard 89] W. Krauth and M. Mézard, Storage capacity of memory networks with binary couplings, *Journal de Physique (France)* 50, 3057–3066 (1989).