Noise and competition in neural networks

D. Sherrington, K. Wong, A. Coolen

To cite this version:


HAL Id: jpa-00246726
https://hal.archives-ouvertes.fr/jpa-00246726
Submitted on 1 Jan 1993

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Noise and competition in neural networks(*)

D. Sherrington, K.Y.M. Wong(**) and A.C.C. Coolen

Department of Physics, University of Oxford, 1 Keble Road, Oxford, OX1 3NP, G.B.

(Received 17 June 1992, accepted 3 July 1992)

Abstract. — A brief summary is given of recent results on the use of noise in the optimal training of neural networks for the retrieval of static patterns and on competition between the retrieval of static patterns and of sequences in networks with asymmetric synapses.

1. Introduction.

A neural network is an assembly of relatively simple units (neurons) which dynamically drive one another strongly and cooperatively via interactions (or local rules) which are competitive, so as to lead to a large number of possible basins of dynamic attraction of the global behaviour, but also are coded (trained) so that many of those basins are related to learned activity patterns, sequences or associations.

Noise enters into the operation and training of neural networks in several ways and can be either stochastic or quenched. It can affect the ability of a network to retrieve, the competition between different types of retrieval, and the quality of retrieval. It can also be valuable in training the system to achieve desired performance.

Retrieval (or operation) is the process of dynamical evolution of a network with a given set of interaction rules. If the system is to have many basins of attraction, corresponding to many pieces of retrievable memorized information, then there must be competition between those attractor basins. The influence of other attractors on any desired one represents a source of quenched interference noise. Damage to a network is another source of quenched noise. Stochastic noise arises in retrieval if the local rules are probabilistic, rather than deterministic. Such stochastic noise is often referred to as thermal noise. Thus, for a network of Ising neurons with a simple parallel synaptic updating rule

\[
\sigma_i(t + 1) = \text{sign}\left(\sum_j J_{ij} \sigma_j(t) + Tz_i(t) + W_i\right),
\]

(*) This paper is based on a talk given by D. Sherrington at the memorial meeting for R. Rammal held in Grenoble on May 22nd 1992.

(**) Present address: Dept. of Physics, the Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong.
where $\sigma = \pm 1$ represents the neural state (firing/non-firing) and $z_i(t)$ is chosen randomly from a Gaussian distribution of zero mean, quenched noise resides in the synapses $\{J_{ij}\}$ and weights $\{W_i\}$, while stochastic noise resides in the $\{T_z(t)\}$.

Training is the process of choosing the local rules, or $\{J_{ij}\}$ and $\{W_i\}$ in the example above, so as to achieve some operational goal. Noise enters in further aspects in this regard. The first such to be mentioned here lies in a technical procedure to determine what is achievable. If the desired aim can be formulated in terms of maximizing some performance measure, then this can be viewed as equivalent to finding the ground state energy of a ‘Hamiltonian’ in the space of rules and can be achieved (at least, in principle) by ‘simulated annealing’ in which one introduces a complementary quantity analogous to a temperature, the ‘annealing temperature’, such that the ground state is obtained by allowing this effective system to perform a stochastic dynamics which yields Gibbs measures at any annealing temperature $T_a$, and gradually cooling to $T_a = 0$ (or formally taking that limit). $T_a$ is thus another stochastic noise.

A second aspect of noise in the determination of the $\{J_{ij}\}$ lies in training using distorted, or noisy, patterns. Thus, given any training algorithm based on the presentation of clean patterns to the network, one can envisage a modification in which the presented input is a randomly distorted version of the clean data, with the degree of distortion a measure of the training noise. This procedure enables the network to ‘spread out’ its information storage and can be used to tune the subsequent retrieval performance.

In section 2 we summarize some recent progress on the analysis of training with noise in networks storing static patterns.

Networks with appropriate asymmetric synapses can also store pattern sequences in a retrievable format. One such is to take the $J_{ij}$ to be given by the modified Hebb form

$$J_{ij} = N^{-1} \sum_{\mu=1}^{p} \xi_i^{\mu+1} \xi_j^{\mu}, \mu \text{ periodic.}$$

(2)

Under parallel dynamics the network evolves, $\mu \rightarrow \mu + 1 \rightarrow \mu + 2$ etc. By contrast, the usual Hopfield-Hebb rule

$$J_{ij} = N^{-1} \sum_{\mu=1}^{p} \xi_i^{\mu} \xi_j^{\mu}$$

(3)

retrieves static patterns [1]. A combination

$$J_{ij} = N^{-1} \nu \sum_{\mu=1}^{p} \xi_i^{\mu} \xi_j^{\mu} + N^{-1} (1 - \nu) \sum_{\mu=1}^{p} \xi_i^{\mu+1} \xi_j^{\mu}$$

(4)

leads to competition between the retrieval of static patterns and sequences, depending on $\nu, T$. In section 3 we summarize recent results of such competition for intensive $p$.

2. Training with noise.

In this section we briefly mention some recent work in Oxford on training with noise. We concentrate on Ising networks obeying local rules (1) with zero thresholds and $\{J_{ij}\}$ obeying the spherical constraints $\sum_j J_{ij}^2 = c$, where $c$ is the incoming connectivity per neuron, storing uncorrelated patterns.

One training criterion is to choose the $\{J_{ij}\}$ to maximize the average increase in overlap with a nominated pattern in one sweep of the network, starting from a state of macroscopic overlap $m_t$ with only one of the patterns; the overlap with a pattern $\mu$ is defined by
\[ m_\mu(t) = N^{-1} \sum_i \sigma_i(t) \xi_i^\mu, \]

where the \( \xi_i^\mu = \pm 1; \mu = 1, \ldots, p \) are the pattern bits. \( m_t \) is then the training overlap, with \( d_t = (1 - m_t)/2 \) the training noise. The choices of the distorted input patterns may be averaged in either an annealed or a quenched fashion; in the former one maximizes the average overlap, in the latter one averages over the maximum overlaps. As long as the number of noisy training examples is greater than the inverse training temperature \( \beta_a \) (see below), the quenched average over these examples yields the same results as the annealed average.

The annealed average output overlap is

\[ \tilde{g}_{m_t}\{\{\Lambda^\mu\}\} = (Np)^{-1} \sum_{i\mu} g_{m_t}(\Lambda_i^\mu) \]

where

\[ g_{m_t}(\Lambda) = \text{erf}\left\{ m_t\Lambda / \sqrt{2(1 - m_t^2 + T^2)} \right\} \]

and

\[ \Lambda_i^\mu = \frac{1}{\sqrt{c}} \sum_j \xi_i^\mu J_{ij} \xi_j^\mu \]

is the aligning field of pure pattern \( \xi^\mu \) at site \( i \). Maximization follows by simulated annealing; the ‘thermally’ averaged performance (output overlap) at annealing temperature \( T_a \) is given by

\[ \langle \tilde{g}_{m_t}\{\{\Lambda_i^\mu\}\}\rangle_{T_a} = \frac{\partial}{\partial \beta_a} \ln \text{Tr} \{f\} \exp(\beta_a \tilde{g}_{m_t}\{\{\Lambda\}\}) \]

where \( \beta_a = T_a^{-1} \). The maximum performance is given by the \( T_a \to 0 \) limit.

So far, this is for a particular set of patterns \( \{\xi^\mu\} \). Averaging over the specific choices can be achieved by application of the replica method, in direct analogy with the techniques developed for spin glasses [5,7], with the \( \{J_{ij}\} \) the annealed variables, analogous to the spins in the spin glass problem, and the patterns \( \{\xi^\mu\} \) the quenched random parameters, analogous to the exchange interactions and/or spin positions in the spin glass. Within a replica-symmetric ansatz and for uncorrelated patterns this gives as the averaged maximum update performance per pattern \( \int d\Lambda \rho_{m_t}(\Lambda) g_{m_t}(\Lambda) \) where \( \rho(\Lambda) \) is the aligning field distribution given by

\[ \rho_{m_t}(\Lambda) = \int Dt \delta(\Lambda - \lambda(t)) \]

where \( Dt = dt \exp(-t^2/2)/\sqrt{2\pi} \) and \( \lambda(t) \) is the value of \( \lambda \) which maximizes \( g_{m_t}(\lambda) - (\lambda - t)^2 / 2\gamma \) where \( \gamma \) is determined by \( \int Dt(\lambda(t) - t)^2 = \alpha^{-1} \) where \( \alpha = p/c \) is the storage ratio.

For \( m_t = 0^+ \) the network aligning field distribution is that of a network with Hebb synapses

\[ J_{ij} = \sqrt{p^{-1}} \sum_{\mu} \xi_i^\mu \xi_j^\mu. \]

As \( m_t \) is increased this evolves until, for \( T = 0 \) and \( m_t = 1^- \), one obtains the field distribution of a maximally stable network [8]. There are several interesting features of the evolution, discussed in detail elsewhere [4].

The average one-step update from an input overlap \( m \) of networks trained with input overlap \( m_t \) is

\[ f_{m_t}(m) = \int d\Lambda \rho_{m_t}(\Lambda) g_{m}(\Lambda). \]
For a dilute asymmetric network, with connectivity $c << \ln N$, correlations are unimportant and $f_{m_t}(m)$ provides an iterative map from which both the asymptotic retrieval overlap and the basin boundary overlaps follow, respectively from the stable and unstable fixed points of the map $m \rightarrow f_{m_t}(m)$. Their appearance and disappearance as a function of $\alpha, T$ determine the phase lines in a retrieval phase diagram. Varying $m_t$ provides a means of tuning the basin structures. When the training noise is varied, the depth and width of the retrieval basins corresponding to the stored patterns change accordingly, leading to a varying degree of competition among them, and hence various retrieval and non-retrieval phases. The stable and unstable fixed points of the envelope mapping $m \rightarrow f_{m_t}(m)$ give the optimally achievable asymptotic retrieval overlaps and basin boundaries.

An analysis of the normal mode energies of replica-symmetry breaking (RSB) fluctuations in $J$-space places stability bounds on the ansatz used above. When the training noise is varied, the RSB region corresponds to a number of possible solutions competing to be the network with optimal performance. However, RSB affects only a small region of the $\alpha - m_t$ space of retrieval as determined by the dilute network maps; it excludes a larger region of $m_t - \alpha$ space for values of $\alpha$ beyond the retrieval limit.

In physics involving noisy signals, one is used to an importance lying in the signal-to-noise ratio. Another such example is found in the training with noise scenario in neural networks. Asymptotically, for small $\alpha$ at $T = 0$, the training noise region $m_t = \sqrt{\alpha}$, at which the signal strength ($m_t$) and the interference noise level due to other patterns ($\sqrt{\alpha}$) are equal, appears to mark the convergence of several novel features; among them the closing of a band gap in the $\rho(A)$ distribution, the shrinking together of upper and lower Almeida-Thouless limits on the replica-symmetric approximation's stability against small symmetry-breaking fluctuations, the minimum of a generalized $J$-susceptibility, $\chi = \lim_{T_0 \rightarrow 0} \frac{T_0}{T_0} \left( \frac{1}{J_{ij}} \right) - \langle J_{ij} \rangle_{sol}^2$, and the maximum deviation of the trajectory in weight space from the Hebbian to the maximally stable network [9]. The network behaviours are different for $m_t$ above and below $\sqrt{\alpha}$, in their way of optimizing the network performance constrained by the competing tendencies of the patterns, as illustrated by the respective presence and absence of a bandgap in the $\rho(A)$ distribution.

Figure 1 shows the loci of some of the above features for the full range of $m_t$ and for a larger range of $\alpha$. The lines of minimum $J$-susceptibility and maximum deviation of the weight space trajectory lie close to one another but above the line of band merging for larger $\alpha$ [9].

In the above, we have taken the annealing temperature $T_a$ to zero in the optimization exercise. It is also possible to extend the concept of noise training to one in which $T_a$ is kept at a finite temperature and the resulting $\{J_{ij}\}$ or $\lambda$ distribution is obtained in the thermodynamic limit. Technically, this is analogous to a spin glass problem at finite temperature. This has been considered for cost functions arising from rule learning with undistorted examples [10,11] and could be extended to cover the present case of randomly distorted patterns. In some cases, it has been argued that a non-zero annealing temperature may improve the generalization performance of the network.

For more details the reader is referred to references [2, 4].

3. Competition between pattern reconstruction and sequence processing.

In this section we report briefly on a study of the properties of recurrent neural networks with competitive synaptic efficacies as given by equation (4), with $p$ intensive (i.e. not scaling with the size of the system, which is effectively taken as infinite) and the patterns random and uncorrelated [11]. We concentrate on parallel stochastic local field dynamics in which the microstate distribution evolves according to the Glauber rule.
Fig.1. — Special lines in a noise-trained neural network; the de Almeida-Thouless lines (solid) marking the onset of instabilities against small replica-symmetry breaking fluctuations, the line marking the closing of the band-gap in the aligning field distribution (dashed), and the retrieval-nonretrieval phase boundary (dotted).

\[ P(\sigma; t + 1) = \sum_{\sigma'} W(\sigma' \rightarrow \sigma) P(\sigma'; t) \]  
(12)

where

\[ W(\sigma' \rightarrow \sigma) = \prod_j \frac{1}{2} [1 + \sigma_j \tanh (\sum_k J_{jk} \sigma'_k / T)], \]  
(13)

\[ \sigma = \sigma_1, \ldots, \sigma_N \]  
(14)

and again \( T \) is a measure of stochastic retrieval noise.

The useful macroscopic dynamical measures are the overlaps \( m^\mu(t) \) as defined in equation (6). In the thermodynamic limit they satisfy

\[ m(t + 1) = \langle \xi \tanh (\beta \xi \cdot A.m(t)) \rangle_\xi \]  
(15)

where

\[ A_{\mu\rho} = \nu \delta_{\mu\rho} + (1 - \nu)\delta_{\mu,\rho+1} \quad (\mu : \text{mod} p), \]  
(16)

\( m = (m^1, \ldots, m^p) \) and \( \langle \rangle_\xi \) denotes an average over patterns \( \xi \).

Manipulations of (15) lead to useful equalities and inequalities from which follow important consequences for the possible fixed point solutions. In particular, for \( T > 1 \) the only fixed point is \( m = 0 \) and for \( 1 > T > 2\nu - 1 \) the only fixed point is uniform, \( m = m(1,1,..,1) \). Thus, in neither of these regions is there a retrieval fixed point. In fact, for \( p > 2 \) any such Hopfield-like retrieval solutions are restricted by first-order transitions to \( \nu > \nu_c(T) \) which is given to a (numerically) good approximation by
\[ \nu_c(t) = \frac{1}{2} + T - T^2 + \frac{3}{4} T^3 - \frac{1}{4} T^4. \]  

(17)

Fixed points are not necessarily stable. Stability requirements restrict further the stability of the uniform fixed point to regions close to \( T = 1 \) and, for \( p \) odd, \( T = 0 \), which shrink rapidly as \( p \) is increased. Without stable fixed points the only alternatives are periodic sequential solutions.

One such set of sequential solutions follows from a general symmetry mapping. This mapping relates solutions of (15) for \( \nu \) to corresponding solutions for \( \nu \rightarrow (1 - \nu) \). For any solution \( M = m(1), m(2), \ldots \) there exists a solution with identical stability given by \( M' = D M \) where \([D M](n) \equiv D(n)m(n)\) and \( D(n) = S^n K \) with \( S_{\mu\rho} = \delta_{\mu,\rho+1} \) and \( K_{\mu\rho} = \delta_{\mu,p+1-\rho} \).

In particular, therefore any Hopfield solution for \( \nu > \nu_c \) has as its counterpart a \( p \)-periodic sequential solution for \( \nu \rightarrow 1 - \nu \), i.e. in the region \( \nu < (1 - \nu_c) \), with first-order transitions at the limits.

More interesting sequential solutions lie between these first order boundaries \((1 - \nu_c(T))\) and \( \nu_c(T) \). These evolve sequentially and periodically through the \( p \) patterns more slowly than those discussed in the last paragraph. In fact, in the limit of large \( p \), and numerically to a good approximation even for relatively small \( p \), the relative sequence speed \( \omega \equiv p/\Omega \), where \( \Omega \) is the period, is given by

\[ \lim_{p \rightarrow \infty} \omega_p = 1 - \nu \]  

(18)

This is illustrated in Figure 2 which shows values of \((1 - \omega)\) obtained asymptotically for systems with various \( \nu \) in simulations started in pure Hopfield states, with superimposed the first order boundaries of the regions of Hopfield and \( p \)-periodic cyclical solutions and also the stability boundary of the uniform state. \( p \approx 10 \) in the case exhibited.

Thus, by varying the values of \( \nu, T \) a system described by equations (4) and (12) can be made to retrieve either static patterns or sequences, with the transition speed of the sequence tunable.

Further details may be found in reference [11].

Since the above discussion is restricted to \( p \) intensive, there is no significant effect of interference noise between the patterns. It would be interesting to investigate how the above could be extended to an extensive number of sets of (finite) pattern sequences, when such interference could be expected to play a role. Similarly, it would be interesting to consider the effect of replacing the Hebbian synapses of eqn. (4) by corresponding noise-trained forms analogous to those considered in section 2, with the added feature that different training noises could be used in generating the pattern retrieval and sequence generating synapses, using chemical modulatres to switch on and off different modes of training [13].

Acknowledgement.

The work reported above has benefitted from discussions with several other members of the condensed matter theory group at Oxford and has been performed within the context of an EC SCIENCE twinning programme, "Statistical Mechanics and its Applications to Complex Problems in Physics, Engineering and Biology". Until his untimely death, Rammal Rammal was the coordinator of the Grenoble branch of that twinning and all members of the Oxford group honour his memory.
Fig. 2. — Superposition of the phase diagram and the measure \( (1-\omega) \) of the relative sequence processing speeds of a competitive network described by equations (4) and (12) for \( p = 10 \). \( H, M, C \) denote regions in \( \nu, T \) space, respectively for Hopfield-like pattern retrieval, the uniform mixture state, and \( p \)-periodic cycles; in the remaining region \( C' \) only sequences with periods different from \( p \) are possible. The phase boundaries between \( H, C \) and \( C' \) are first order, between \( M \) and \( C' \) they are second order. The 'points' show the values of \( (1-\omega) \) reached in the asymptotic behaviour of systems started in pure pattern states for \( \nu = 0.1, 0.2, 0.9 \) (from bottom to top) as a function of the retrieval temperature.

References