Learning grey-toned patterns in neural networks

1991 J. Phys. A: Math. Gen. 24 4941

(http://iopscience.iop.org/0305-4470/24/20/023)

J. Phys. A: Math. Gen. **24** (1991) 4941–4952. Printed in the UK

Learning grey-toned patterns in neural networks

S Mertens†, H M Köhler† and S Bos‡

† Institut für Theoretische Physik, Georg-August-Universit, D-3400 Göttingen, Federal Republic of Germany

‡ Institut für Theoretische Physik, Justus-Liebig-Universität, D-6300 Giessen, Federal Republic of Germany

Received

Abstract. The problem of learning multi-state patterns in neural networks is investigated. An analysis of the space of couplings (Gardner approach) yields the distribution of local fields, the critical storage capacity α_c and the minimum number of errors for an overloaded network. For noisy local fields the classification error is minimized if the local fields of the patterns are allowed to lie in intervals of finite width. A fast converging, adaptive learning algorithm is presented, which finds the coupling matrix of optimal stability.

1. Introduction

Guided by the analogy between networks of formal neurons and models of spin glasses, the contributions of physicists to the field of neural networks have mainly concentrated on systems of binary units (see [1] for an overview). Recently, however, several authors have considered the problem of storing multi-state patterns in attractor neural networks [2–8]. Having in mind the applications of such networks in image processing we call this kind of pattern grey-toned patterns. This technical term also indicates that our studies are mainly motivated by potential applications rather than biological relevance.

We consider a network of N formal neurons V_i which can take on Q discrete values: $V_i \in \{\sigma_k\}_{k=1}^{Q}$. The (parallel or sequential) dynamics of the system reads

$$V_i(t + \Delta t) = \mathrm{dyn}(h_i(t)) \tag{1}$$

where the local field or post-synaptic potential (PSP) is given by

$$h_i(t) = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_{ij} V_j(t). \tag{2}$$

The input/output-relation dyn maps the real axis onto the discrete set of grey levels. We think of dyn as being realized by a staircase function

$$\mathrm{dyn}(h) = \sigma_k \qquad \text{if} \qquad h \in [L(\sigma_k), U(\sigma_k)] \tag{3}$$

where the intervals $[L(\mathrm{ower}), U(\mathrm{pper})]$ form a non-overlapping partition of the real axis in Q parts. The questions we address here are:

0305-4470/91/204941+12$03.50

(i) How many random Q-state patterns $\xi_i^\nu \in \{\sigma_1, \ldots, \sigma_Q\}$, $i = 1, \ldots, N$ and $\nu = 1, \ldots, p$ can simultaneously be made fixed points of equation (1) by an appropriate choice of the couplings J_{ij}, i.e. what is the critical storage capacity of the network? What happens if we go beyond this limit?

(ii) For $Q = 2$, the perceptron with maximum separation κ of allowed PSPs is denoted 'optimal'. What is the optimal perceptron for $Q > 2$?

(iii) How can we explicitly calculate the couplings to achieve the maximum storage capacity?

To answer the first two questions we generalize the analysis of the space of synaptic couplings pioneered by Gardner [9, 10] to arbitrary Q. The third problem is solved by a generalization of the AdaTron algorithm [12] for binary-valued patterns to $Q > 2$.

2. Analytical results

A sufficient condition for the patterns to be fixed points of equation (1) is

$$h_i^\nu := \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_{ij} \xi_j^\nu \stackrel{!}{\in} [l(\xi_i^\nu), u(\xi_i^\nu)] \qquad \text{for all } i, \nu \tag{4}$$

where $[l(\xi_i^\nu), u(\xi_i^\nu)] \subset [L(\sigma_k), U(\sigma_k)]$ for $\xi_i^\nu = \sigma_k$ (see figure 1). This constraint is more severe than it is for the patterns to be fixed points (at least if $[l, u] \neq [L, U]$) but the robustness of the retrieval properties in the presence of additional noise and the desire for large basins of attraction for the patterns both require a good separation of the allowed h_i^ν [13, 14].

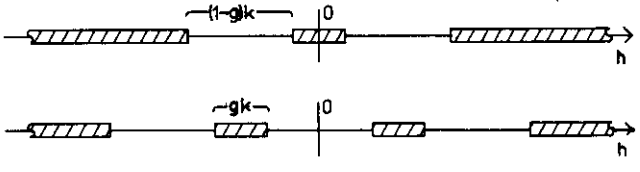

Figure 1. Intervals of desired PSPs for $Q = 3$ (upper axis) and $Q = 4$ (lower axis). To keep the requirements on J_{ij} less restrictive, the intervals for the 'black' and 'white' levels are allowed to stretch to $-\infty$ and ∞, respectively. κ denotes the distance between the midpoints of two adjacent finite size intervals, $g\kappa$ the size of these intervals. The edges of two adjacent intervals are consequently separated by $(1 - g)\kappa$. $g = 0$ forces the PSPs to discrete values (except for the saturated 'black' and 'white' levels), and for $g = 1$ the intervals touch each other, covering the whole h-axis. Note that for $Q = 2$ and $g = 0$ our κ equals twice the κ used by Gardner (cf equation (12)).

It is clear that for a growing number of patterns it will be increasingly difficult to satisfy the Np conditions (4). Following the approach of Gardner [9, 10] we can make a quantitative attack on the problem by introducing a cost function

$$H(J) = \sum_{i,\nu} (1 - \chi_{[l(\xi_i^\nu), u(\xi_i^\nu)]}(h_i^\nu)) \tag{5}$$

with

$$\chi_{[l,u]}(h) := \begin{cases} 1 & h \in [l,u] \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

H depends on the interaction matrix J_{ij} and counts the errors, i.e. the number of pattern sites that fail to obey (4). We further introduce an auxiliary parameter β and the partition function

$$Z(\beta) = \int \prod_{i,j} \mathrm{d}J_{ij}\, \rho(J) \mathrm{e}^{-\beta H(J)} \tag{7}$$

where $\rho(J)$ is a normalized *a priori* measure in J_{ij}-space. Throughout this article we will use the measure

$$\rho(J) = \frac{\prod_i \delta(\sum_j J_{ij}^2 - N)}{\int \mathrm{d}J \prod_i \delta(\sum_j J_{ij}^2 - N)} \tag{8}$$

(spherical constraint). The problem can now be cast into the language of statistical mechanics: The 'mean energy' computed from Z in the limit $\beta \to \infty$ is the minimum number of errors which can be realized by any coupling matrix J_{ij}. This quantity still depends on the specific choice of the patterns to be stored. Assuming statistically independent random patterns, meaningful general statements will be obtained by averaging over this quenched disorder. Denoting the average over the patterns by $\langle\langle\cdot\rangle\rangle$ the typical minimal number of errors is given by

$$Np f_{\min} = -\lim_{\beta\to\infty} \frac{\partial}{\partial\beta} \langle\langle \log Z(\beta) \rangle\rangle. \tag{9}$$

$f_{\min}$ is the minimum fractional part of the wrong PSPs. The calculation of $\langle\langle \log Z \rangle\rangle$ using the replica trick and the replica-symmetric ansatz has become a standard tool in neural network theory and is therefore omitted here. A detailed description of the procedure for $Q = 2$ can be found in the original work of Gardner [9, 10] or in textbooks on neural networks [1]. The extension to $Q > 2$ is straightforward.

For reasons of simplicity we assume that the ξ_i^ν are drawn independently from a distribution $w(\xi)$ with mean $\langle\xi\rangle = 0$ (unbiased patterns†) and variance $\langle\xi^2\rangle = \sigma^2$. The critical storage capacity α_c then reads

$$\alpha_c^{-1} = \left\langle \int_{-\infty}^{\hat{l}(\xi)} Dz (z - \hat{l}(\xi))^2 + \int_{\hat{u}(\xi)}^{\infty} Dz (z - \hat{u}(\xi))^2 \right\rangle \tag{10}$$

with

$$Dz := \frac{\mathrm{d}z}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \tag{11}$$

and $\hat{l} = l/\sigma$, $\hat{u} = u/\sigma$. The single angular brackets denote the average over the grey-level distribution $w(\xi)$.

† Results for biased patterns are given in the appendix.

The well known result for $Q = 2$ [9],

$$\alpha_c^{-1} = \int_{-\kappa}^{\infty} Dz(\kappa + z)^2 \tag{12}$$

is recovered from equation (10) by setting $\sigma^2 = 1$, $-l(-1) = u(1) = \infty$ and $-u(-1) = l(1) = \kappa$.

If $\alpha > \alpha_c$, i.e. if the network is overloaded, the minimum fraction of wrong PSPs is given by

$$f_{\min} = \left\langle \int_{-\infty}^{\hat{l}(\xi)-x} Dz + \int_{\hat{u}(\xi)+x}^{\infty} Dz \right\rangle \tag{13}$$

where x is the solution of

$$\alpha \left\langle \int_{\hat{l}(\xi)-x}^{\hat{l}(\xi)} Dz(z - \hat{l}(\xi))^2 + \int_{\hat{u}(\xi)}^{\hat{u}(\xi)+x} Dz(z - \hat{u}(\xi))^2 \right\rangle = 1. \tag{14}$$

Note that in the derivation of equations (10), (13) and (14) we assumed replica symmetry. This assumption is valid as long as the space of solutions J_{ij} of equation (4) is connected. This is, however, not necessarily true for $f_{\min} > 0$. An analysis of the stability of the replica symmetric solution similar to that of de Almeida and Thouless [11] for the SK model shows, that the replica symmetric solution is stable as long as

$$\frac{1}{\alpha} > \left\langle \int_{\hat{l}(\xi)-x}^{\hat{l}(\xi)} Dz + \int_{\hat{u}(\xi)}^{\hat{u}(\xi)+x} Dz \right\rangle \tag{15}$$

holds, i.e. as long as the the perceptron is not too heavily overloaded. Equation (15) especially guarantees the stability of the replica symmetric solution in the error-free regime $x \to \infty$ (see figure 5).

If for the moment we assume that the intervals $[l, u]$ are degenerated to discrete points (response levels) the equation for α_c simplifies to

$$\alpha_c = \frac{1}{1 + (\langle l^2 \rangle / \sigma^2)} \tag{16}$$

where $\langle l^2 \rangle$ is the variance of the response levels. The corresponding coupling matrix can then be expressed by means of the so-called pseudo inverse [15]:

$$J_{ij} = \frac{1}{\sqrt{N}} \sum_{\mu,\nu} l(\xi_i^\mu)(C^{-1})_{\mu\nu}\xi_j^\nu \tag{17}$$

with

$$C_{\mu\nu} = \frac{1}{N} \sum_i \xi_i^\mu \xi_i^\nu. \tag{18}$$

If κ is the distance between two adjacent response levels, $\langle l^2 \rangle$ is roughly proportional to $\kappa^2 Q^2$, i.e. α_c decreases with increasing separation of the response levels and increasing

number of grey levels. The maximum α_c is achieved for $\kappa = 0$, i.e. when the network can no longer resolve the grey levels. The $\alpha_c \propto Q^{-2}$ law has been found recently by Rieger [6] for the Hopfield matrix (see figure 4). Note, however, that α_c is independent of Q if one keeps the grey-level separation fixed, since in this case $\langle l^2 \rangle \propto \sigma^2$.

The fact that α_c decreases with increasing Q whereas the information content of a pattern vector increases with Q (like $\log Q$ for equally weighted grey levels) can be used in neural network design in order to maximize the information capacity per synapse, $I = \alpha_c \log Q$. If one allows very narrow response levels (e.g. $\kappa \propto Q^{-1}$), I can be very high (see figure 2). This unbounded growth of I with Q is based on the continuous nature of the couplings, whose representation requires an infinite amount of bits. A suitable measure of information capacity in neural networks should take this into account [8].

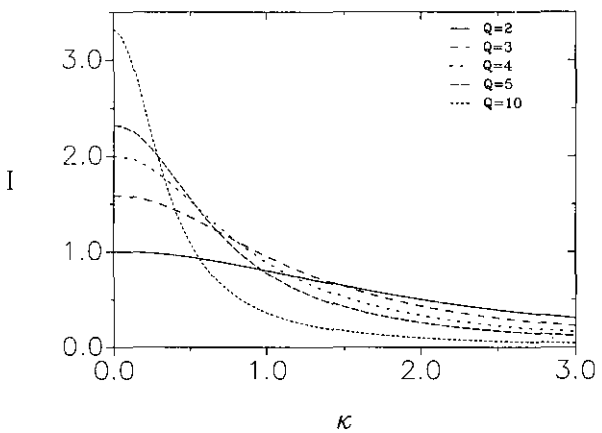

Figure 2. Information capacity $I = \alpha_c \log_2 Q$ in bits/synapse against response-level separation κ for discrete response levels (pseudo-inverse) and various values of Q and $\sigma = 1$. Curves like these can be used to choose the grey-level resolution Q that maximizes the information capacity in the network for fixed κ.

The dependence of α_c on the variance of the response levels remains qualitatively the same if the intervals $[l, u]$ have non-zero width (figure 1). Since equation (4) is less restrictive for finite intervals, α_c is greater than for discrete response levels. Figure 3 shows α_c for different values of Q and for the interval structure depicted in figure 1. In figure 4 it can be seen, that $\alpha_c \propto Q^{-2}$ for $Q \gg 1$ (and fixed σ^2) for both discrete response-levels and intervals of non-zero width.

Figure 5 shows the minimum error $f_{\min}$ in an overloaded network as a function of α for $Q = 2, 3, 4, 5$. The slope of $f_{\min}(\alpha)$ at $\alpha = \alpha_c$ increases with the number of grey levels: For $Q \gg 1$, overloading seems to be disastrous. The broken parts of the curves indicate the region where replica symmetry has to be broken according to inequality (15). From equations (13) and (14) it follows that $f_{\min}$ is always bounded by

$$f_{\min} \leqslant 1 - \left\langle \int_{\hat{l}(\xi)}^{\hat{u}(\xi)} Dz \right\rangle \tag{19}$$

where equality holds in the limit $\alpha \to \infty$. The result $f_{\min} < 1$ even in the limit $\alpha \to \infty$ for non-zero intervals only reflects the fact that for $\alpha \to \infty$ the distribution of

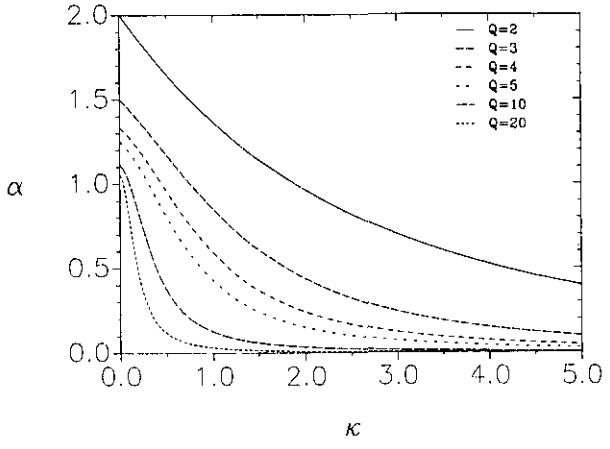

Figure 3. Critical storage capacity against κ for different values of Q ($\sigma = 1$) and $g = 0.5$. Note that $\alpha_c(\kappa = 0) = Q/(Q-1)$.

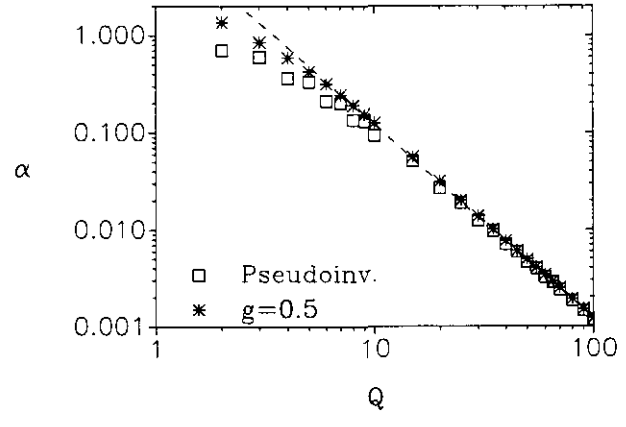

Figure 4. Critical storage capacity α_c against Q ($\sigma = 1$). The stability intervals ($g = 0.5$) respectively the discrete response levels (pseudo-inverse) were taken equidistant with $\kappa = 1$. The dashed line is given by equation (16).

the PSPs is Gaussian and there is always a non-vanishing probability of finding a PSP in any non-zero interval.

In a saturated network ($\alpha = \alpha_c$), the distribution of the PSPs is more structured. It can again be calculated using the replica trick. The result is a simple generalization of the well known distribution for $Q = 2$ [14]:

$$\rho(h) = \left\langle \int_l^u \tilde{D}z\,\delta(h-z) + \delta(h-l)\int_{-\infty}^{l} \tilde{D}z + \delta(h-u)\int_u^{\infty} \tilde{D}z \right\rangle \qquad (20)$$

with

$$\tilde{D}z := \frac{\mathrm{d}z}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(z^2/2\sigma^2)}.$$

As forced by equation (4), $\rho(h) = 0$ outside the intervals $[l, u]$. In the interior of the intervals, ρ simply follows the corresponding part of a Gaussian distribution with

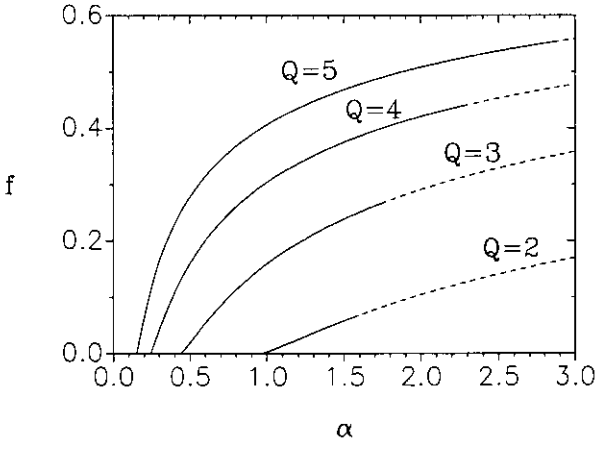

Figure 5. Minimum typical error $f_{\min}$ in an overloaded network against memory loading α ($\sigma = 1$). The curves were calculated with $g = 0.5$ and $\kappa = 2$. The dashed parts of the curves indicate the regime, where according to inequality (15) the replica symmetry has to be broken.

zero mean and variance σ^2, whereas the weight of the excluded left and right tails of the Gaussian distribution are concentrated in δ-peaks at the left and right edges respectively of the intervals (see figure 8).

For $Q = 2$ and given $\alpha \leqslant \alpha_c(\kappa = 0)$ the perceptron with maximum κ is denoted 'optimal perceptron' since the maximized separation of the response intervals improves the stability of the patterns in the presence of additional noise. For $Q > 2$, the appearance of the interior intervals of finite width makes the question for the optimal perceptron more complicated. Equation (20) shows that most of the PSPs in the saturated perceptron are concentrated at the edges of the intervals. To obtain well separated PSPs it therefore seems reasonable to choose an interval structure with maximum separation of edges of adjacent intervals. For the interval structure in figure 1 this means $g = 0$. However, a non-zero g not only leads to less separated interval edges but also to a reduced weight of δ-peaks in $\rho(h)$. The latter effect may compensate the former. To analyse this question quantitatively we add a Gaussian noise with zero mean and variance η^2 to the PSP. The resulting distribution ρ_η of the noisy PSPs reads

$$\rho_\eta(h) = \int_{-\infty}^{\infty} \frac{\mathrm{d}x}{\sqrt{2\pi\eta^2}} \exp\left(-\frac{(x-h)^2}{2\eta^2}\right) \rho(h). \tag{21}$$

Due to the noise, a unique map from h to one of the grey levels is no longer guaranteed. Figure 6 shows ρ_η for $Q = 3$. The hatched area ('crosstalk error') is the probability that a PSP is misinterpreted. An optimal perceptron should minimize this crosstalk error for given α. For the interval structure of figure 1 this means that we have to choose g in order to keep the crosstalk error low. Figure 7 shows, that the crosstalk error is minimized for non-vanishing g. Hence, for Gaussian disturbing noise and equidistant response intervals a finite interval width is superior to discrete response levels. To minimize the probability, that a PSP is mapped to the wrong grey level, the steps of the input/output relation $\mathrm{dyn}(h)$ should be placed at those values of h where the curves in figure 6 meet.

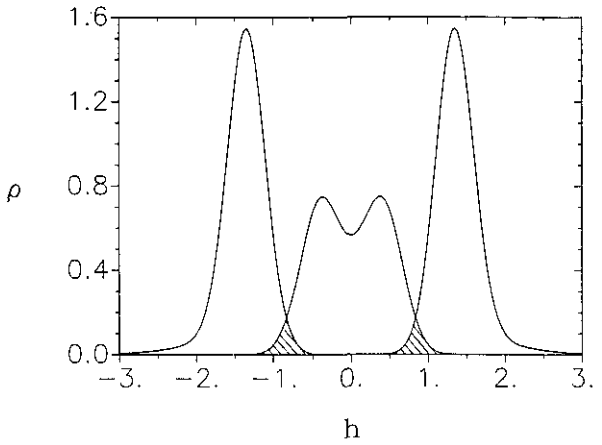

Figure 6. Distribution of the PSPs in the saturated perceptron for $Q = 3$, $\sigma = 1$, $\alpha = 0.5$ and $g = 0.5$ with added Gaussian noise of variance $\eta = 0.24$. The noise leads to an overlap of the distributions for the three grey levels (hatched area) which in turn can lead to a wrong classification ('crosstalk error').

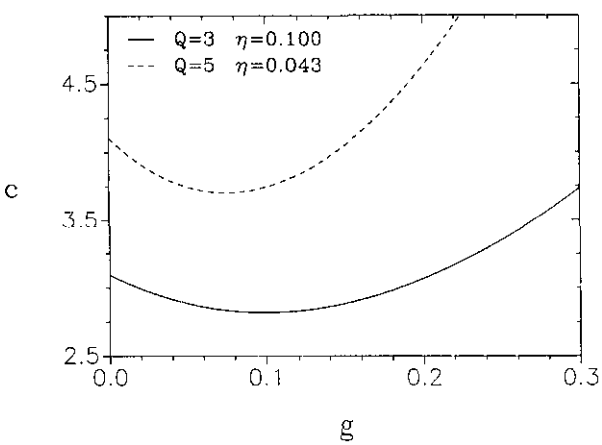

Figure 7. Crosstalk error c against g. c is the probability (in %) that a PSP is mapped to the wrong grey level. $\alpha = 1.0$ and $\sigma = 1$ for both curves.

3. Learning rule and simulations

Algorithms, which produce optimal stability and which reach the maximum possible storage capacity, are known and well understood for the two-state perceptron with continous couplings. The AdaTron algorithm as proposed by Anlauf and Biehl [12], which is a surprisingly simple application of dual quadratic programming, has proven to be fast converging. It can be generalized to grey-level patterns without making the problem much more complicated. For a given interval structure we maximize the overall scaling factor κ by minimizing a quadratic form under linear constraints. It is sufficient to regard a local learning rule, so that we can write couplings for neuron 0 as

$$J_{0j} = N^{-1/2} \sum_{\nu} x^{\nu} \xi_j^{\nu} \tag{22}$$

because components perpendicular to the linear hull of the patterns have no effect on the PSPs. We write $\mathbf{x}$ for the p component vector of x^ν. Then the quadratic norm can be written as $|J^2| = \mathbf{x}^{\mathrm{T}} C \mathbf{x}$, with C as in (18). Furthermore we require that the PSPs have values in given intervals. We denote the vector of the lower boundaries as $\mathbf{l}$ and upper boundaries as $\mathbf{u}$, so we have linear constraints

$$\mathbf{l} \leqslant C\mathbf{x} \leqslant \mathbf{u} \tag{23}$$

for each component. The difference to the original AdaTron is that we do not only have one lower bound, but several and also upper bounds. Dual transformation (see e.g. [16]) yields

$$(D) \begin{cases} & \max_{\mathbf{x}} f = -\frac{1}{2}\mathbf{x}^{\mathrm{T}} C \mathbf{x} + \mathbf{x}^{\mathrm{T}} \mathbf{b} \\ \text{subject to} & x_i \begin{cases} > 0 : \text{lower} \\ = 0 : \text{no} \\ < 0 : \text{upper} \end{cases} \Big\} \text{ boundary active} \end{cases} \tag{24}$$

where $\mathbf{b}$ is the vector of the active boundaries.

However if learning is necessary (the lower bound is active), learning intensities should be as low as possible and for the case in which unlearning is necessary (the upper bound is active) learning intensities should be as high as possible. The construction of the solution works iteratively, starting from the dual feasible point $x_0^\mu = 0$ sequentially for all patterns. The solution is found by applying

$$x_{t+1} = \begin{cases} \max(0, \frac{\gamma}{C^{\mu\mu}}(l^\mu - h^\mu) + x_t^\mu) & : h^\mu < l^\mu \\ \min(\frac{\gamma}{C^{\mu\mu}}(u^\mu - h^\mu) + x_t^\mu, \max(0, \frac{\gamma}{C^{\mu\mu}}(l^\mu - h^\mu) + x_t^\mu)) & : l^\mu \leqslant h^\mu \leqslant u^\mu \\ \min(0, \frac{\gamma}{C^{\mu\mu}}(u^\mu - h^\mu) + x_t^\mu) & : h^\mu > u^\mu. \end{cases} \tag{25}$$

Let $\delta x^\mu = x_{t+1}^\mu - x_t^\mu$. For the case when the field of an actually learned $(x^\mu, \delta x^\mu > 0)$ or unlearned pattern $(x^\mu, \delta x^\mu < 0)$ never happens to drift across the whole interval while the other patterns are learned (i.e. if g is not too small), the rule for $l^\mu \leqslant h^\mu \leqslant l^\mu$ is sufficient for all h^μ.

Convergence to a solution with maximal κ is guaranteed for $0 < \gamma < 2$ as can be proved in a manner similar to the proof the AdaTron. It should be noted that no restrictions (like, for example, statistical independence) have to be imposed on the patterns in order to guarantee the convergence of our algorithm.

Simulations were done only for unbiased and randomly chosen patterns with equal probability for each grey level and $\gamma = 1$. The intervals were chosen as shown in figure 1. Figure 8 shows theory and simulation data κ against α for fixed g. Two curves are shown, one for three and the other for four intervals. The size of the system for the simulations was $N = 127$; the points shown were averaged over 5000 trials. The distribution of the PSPs is plotted in figure 9. Simulation data are again accumulated over 5000 systems with $N = 127$. The theoretically calculated curves are in good agreement with the numerical experiment. All δ-peaks lie within channels of high observation rates. Also the intensities of the channels with the δ-peaks are in good agreement with the theory.

These good agreements indicate that the analytical results gained in the limit $N \to \infty$ can be safely applied to networks of moderate size.

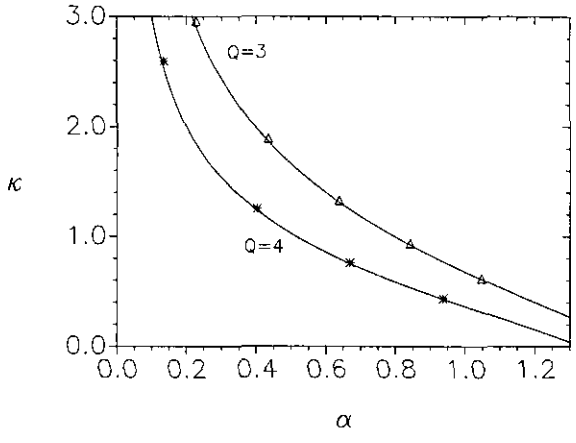

Figure 8. Maximum separation of intervals against memory loading, theory (full curve) and simulations (symbols) for $N = 127$ averaged over 5000 trials for $Q = 3$ and $Q = 4$ with $g = 0.4$ and $g = 0.25$. The error in the averaged data points is much smaller than the size of the symbols.

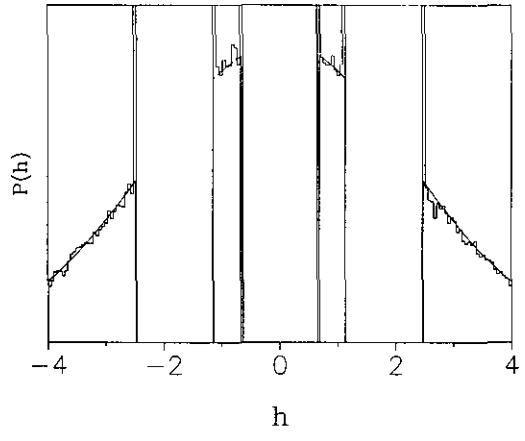

Figure 9. Distribution of PSPs in the saturated network, theory and simulation. Numerical data were accumulated over 5000 trials. $N = 127$, $Q = 4$ and $g = 0.25$. Theoretical δ-peaks are not shown; they fall into channels of high observation rates, which had to be clipped at the top of the figure.

4. Conclusions

In this contribution we have shown how the expressions for α_c, $f_{\min}$ and $\rho(h)$ have to be generalized from the case $Q = 2$ to the case $Q \geqslant 2$ and its Q response intervals $[l(\xi), u(\xi)]$. The formula for α_c (equation (10)) says that the product of α and the Gaussian integrals over the complements of the Q stability intervals must not exceed 1. This viewpoint suggests that it should be possible to arrive at equation (10) by simple geometrical arguments and the central limit theorem, avoiding the cumbersome replica calculation. For $Q = 2$ and $\kappa = 0$ this was done by Cover [17]

We have seen that the decrease in α_c with increasing Q is approximately proportional to $(\kappa Q)^{-2}$ (for fixed σ^2). For given response intervals, the number of grey levels may be chosen in order to maximize the information capacity. Further we found that

the interval structure that minimizes the 'crosstalk error' in the presence of additional noise has intervals of non-zero width.

We calculated the minimum error that can be achieved in an overloaded network and found that the perceptron reacts more sensitive on overloading for larger values of Q.

We have seen how the AdaTron algorithm has to be generalized to cope with the multi-interval restrictions (equation (4)). It should be possible to generalize other learning algorithms for $Q = 2$ in a similar way.

Acknowledgments

We are indebted to A Engel, G A Kohring, R Kree, W Kinzel and A Zippelius for useful hints. SM acknowledges financial support from the Studienstiftung des deutschen Volkes, SB from the Stiftung Volkswagenwerk. The computer time for the simulations was granted by the Forschungszentrum Jülich, Federal Republic of Germany. This work was supported by the BRAIN initiative of the commission of European Communities.

Appendix

The analysis of the space of the couplings is not restricted to unbiased patterns. For the sake of completeness we give the results for patterns which are drawn from a distribution with

$$a := \langle \xi \rangle \qquad \sigma^2 := \langle \xi^2 \rangle - \langle \xi \rangle^2. \tag{26}$$

The minimum error now reads

$$f_{\min} = \left\langle \int_{-\infty}^{(l(\xi)-aM-x)/\sigma} Dz + \int_{(u(\xi)-aM+x)/\sigma}^{\infty} Dz \right\rangle \tag{27}$$

where M and x have to be determined from

$$\frac{\sigma^2}{\alpha} = \left\langle \int_{(u-aM)/\sigma}^{(u-aM+x)/\sigma} Dz(u - aM - \sigma z)^2 + \int_{(l-aM-x)/\sigma}^{(l-aM)/\sigma} Dz(l - aM - \sigma z)^2 \right\rangle \tag{28}$$

and

$$0 = \left\langle \int_{(u-aM)/\sigma}^{(u-aM+x)/\sigma} Dz \int_{(l-aM-x)/\sigma}^{(l-aM)/\sigma} Dz \right\rangle. \tag{29}$$

The critical storage capacity is obtained for $x \to \infty$. The additional parameter M can be interpreted as the bias of the couplings (see [9]).

References

[1] Hertz J, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Reading, MA: Addison-Wesley)

Müller B and Reinhardt J 1991 *Neural Networks. An Introduction* (Berlin: Springer)
[2] Kanter I 1988 *Phys. Rev.* A **37** 2739
[3] Cook J 1989 *J. Phys. A: Math. Gen.* **22** 2057
[4] Yedidia J S 1989 *J. Phys. A: Math. Gen.* **22** 2265
[5] Treves A 1990 *Phys. Rev.* A **42** 2418
[6] Rieger H 1990 *J. Phys. A: Math. Gen.* **23** L1273
[7] Mertens S 1991 *J. Phys. A: Math. Gen.* **24** 337
[8] Kohring G A 1991 *J. Stat. Phys.* **62** 563
[9] Gardner E 1987 *Europhys. Lett.* **4** 481; 1988 *J. Phys. A: Math. Gen.* **21** 257
[10] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[11] de Almeida J R L and Thouless D J 1978 *J. Phys. A: Math. Gen.* **11** 983
[12] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687; 1990 *Europhys. Lett.* **11** 387
Biehl M, Anlauf J K and Kinzel W 1991 Perceptron learning by constrained optimization: the adatron algorithm *Proc. Ninth ASI Summer Workshop on Mathematical Physics (Clausthal)* (Berlin: Springer) to be published
[13] Krauth W and Mézard M 1987 *J. Phys. A: Math. Gen.* **20** L745
Forrest B M 1988 *J. Phys. A: Math. Gen.* **21** 245
[14] Kepler T B and Abbott L F 1988 *J. Physique* **49** 1656
[15] Kohonen T 1984 *Self-Organization and Associative Memory* (Berlin: Springer)
Personnaz L, Guyon I and Dreyfus G 1986 *Phys. Rev.* A **34** 4217
Kanter I and Sompolinsky H 1987 *Phys. Rev.* A **35** 380
[16] Fletcher R 1987 *Practical Methods of Optimization* (Chichester: Wiley)
[17] Cover T M 1965 *IEEE Trans. on Electron. Comput.* **EC-14** 326