The effect of correlations in neural networks

1993 J. Phys. A: Math. Gen. 26 3165

(http://iopscience.iop.org/0305-4470/26/13/021)

J. Phys. A: Math. Gen. 26 (1993) 3165–3185. Printed in the UK

The effect of correlations in neural networks

A Wendemuth†, M Opper‡ and W Kinzel†
Institut für Theoretische Physik, Justus-Liebig Universität, Heinrich-Buff-Ring 16, W-6300 Gießen, Germany

Received 7 September 1992, in final form 4 March 1993

Abstract. The effect of correlations in neural networks is investigated by considering biased input and output patterns. Statistical mechanics is applied to study training times and internal potentials of the MINOVER and ADALINE learning algorithms. For the latter, a direct extension to generalization ability is obtained. Comparison with computer simulations shows good agreement with theoretical predictions.

With biased patterns, we find a decrease in training times and internal potentials for the MINOVER algorithm, which, however, does not lead to faster storage of a given information measure. In ADALINE training, characteristic times undergo a transition from order 1 to order N at any finite bias, for the learning of patterns as well as for the decay of the generalization error. This leads to a rescaling of the gain parameters.

1. Introduction

The methods of statistical mechanics have been extensively used in the quantitative analysis of neural networks. An interesting feature is the network's performance during the learning phase. We shall consider here two training algorithms in particular. For the ADALINE algorithm, the dynamical evolution of the training error and the generalization error have been studied [8, 10]. For the MINOVER algorithm, the distribution of learning times has been computed [11].

However, the dynamics were obtained for the simplest case of randomly chosen, uncorrelated patterns only. A question first put forward by Gardner [6] in the context of storage capacities is the effect of correlation between patterns. Gardner found that the storage capacity of the optimal network rises monotonically and continuously from $\alpha_c = 2$ for random patterns to $\alpha_c = \infty$ for fully correlated patterns.

Here, we shall investigate output and generalization errors for correlated patterns in the ADALINE algorithm. In contrast to Gardner's result for the (static) storage capacity, for the ADALINE algorithm we will find a discontinuous jump in the dynamical behaviour of the error decay for any finite correlation. This will lead us to a recalculation of typical time constants.

For the MINOVER algorithm, we shall calculate the distribution of learning times. Furthermore, we investigate the effect of redundancy of information. The information capacity of a set of patterns decreases as the patterns become more correlated; as a result, the errors decrease faster. We will show, however, that a *given* information cannot be learnt any faster by introducing redundancy, i.e. by spreading it over a larger set of correlated patterns.

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

‡ Present address: Physikalisches Institut, Julius-Maximilians Universität, Am Hubland, W-8700 Würzburg, Federal Republic of Germany.

0305-4470/93/133165+21$07.50

2. The model

We consider a single-layer perceptron with input patterns $\xi^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$ and desired outputs ('targets') $\tau^\mu = (\tau_1^\mu, \ldots, \tau_N^\mu)$. The output at site i for pattern μ is given by

$$\sigma_i^\mu(t+1) = \operatorname{sign}(h_i^\mu(t)) \tag{1}$$

where

$$h_i^\mu = \sum_{j=1}^{N} J_{ij}\xi_j^\mu \tag{2}$$

is the post-synaptic field of pattern μ. In general, this describes a feedforward network. If $\{\xi^\mu\} = \{\tau^\mu\}$, however, input and output units become identical, i.e. we then consider an autoassociative network. Our formalism can thus be applied to both network types.

An autoassociative network will be a fixed point of the dynamics (1) if

$$\xi_i^\mu h_i^\mu(t) > 0 \qquad \forall i, \mu. \tag{3}$$

It was suggested [9] that larger values of the 'internal potential' $\tau_i^\mu(t)h_i^\mu(t)$ represent stronger embedding of the pattern μ, we shall therefore also investigate the potential distribution.

Following Gardner, we impose a correlation between the patterns by choosing all of them to have a bias. To distinguish input and output, we let the input bias be m_{in}, the output bias m_{out}. The ξ_j^μ, τ_i^μ are then independent random variables with distribution

$$p(\xi_j^\mu) = \tfrac{1}{2}(1+m_{\text{in}})\delta(\xi_j^\mu - 1) + \tfrac{1}{2}(1-m_{\text{in}})\delta(\xi_j^\mu + 1) \tag{4}$$

$$p(\tau_i^\mu) = \tfrac{1}{2}(1+m_{\text{out}})\delta(\tau_i^\mu - 1) + \tfrac{1}{2}(1-m_{\text{out}})\delta(\tau_i^\mu + 1). \tag{5}$$

3. The MINOVER algorithm

To date fast algorithms have been developed which allow the learning of a set of input–output relations in perceptrons [3, 13]. Nevertheless the simplest learning strategy that can be implemented in perceptrons is given by the famous perceptron learning algorithm of Rosenblatt. As a variant of this algorithm, the MINOVER algorithm was introduced by Krauth and Mezard [9]. It optimizes the so-called stability of the patterns and has the advantage that an analytical treatment of its dynamics is possible. Numerical simulations show that the results are also good approximations to the standard perceptron learning algorithm.

The MINOVER algorithm lifts the worst internal potential higher than any required internal potential c,

$$cf_i^\mu = \tau_i^\mu h_i^\mu \geqslant c. \tag{6}$$

This equation, normalized by $|J| = \sqrt{\sum_{j=1}^{N} J_{ij}^2}$, gives the stability

$$\Delta = c/|J|. \tag{7}$$

The normalization is necessary since otherwise a global enlargement of J will have the effect of raising Δ. The MINOVER algorithm proceeds in finding, in each timestep, the pattern μ with minimal f_i^μ. The couplings are then modified according to a Hebbian rule, $\delta J_{ij} = (1/N)\tau_i^\mu \xi_j^\mu$, in parallel for all output nodes i. Starting from an empty network ($J_{ij} = 0$), it was shown to converge in a finite number of steps, maximizing for $c \to \infty$ the stability Δ, i.e. maximizing the normalized internal potential for the worst embedded pattern μ. Therefore, the MINOVER algorithm reaches the maximally possible stability limit given by Gardner [6]. This stability limit can neither be exceeded by the introduction of a threshold nor by choosing a different representation for the patterns: first, Gardner [6] has shown that, for biased patterns, the effect of any threshold is compensated for by a corresponding bias in the couplings J_{ij}. Second, for an alteration of the representation of patterns from $\xi_j^\mu = (+1, -1)$ (equations (4) and (5)) to $v_j^\mu = (1, 0)$, where $v_j^\mu = 0.5(\xi_j^\mu + 1)$ and $m(v^\mu) = 0.5(m(\xi^\mu) + 1)$ are changed accordingly, the weight vector J in the old representation can be mapped onto the weight vector in the new representation, which involves a shift in the neuron activity threshold as well. However, since we have just indicated that the Gardner limit is insensitive to the threshold, this will not affect the maximally possible stability.

For algorithms which do not reach the Gardner stability limit, the chosen representation will, however, affect the network performance [2]. For example, Amit *et al* [1] have demonstrated a stability breakdown for biased patterns in the Hopfield network, and consequent papers show [4, 14] that the combination of a modified Hopfield rule and an alternative representation of patterns will restore the Gardner limit. In this paper, we will see in section 4 that learning times in the ADALINE algorithm increase for biased patterns. The connection of this increase to possible modification of the ADALINE rule and representation of the patterns will be indicated.

For the MINOVER algorithm we shall calculate here the distribution of learning times in pattern space, following Opper [11].

Let t_ν be the number of steps pattern ν was used for in updating J. Normalizing the learning times to the threshold c, we introduce

$$x_\nu = t_\nu / c \tag{8}$$

obtaining

$$J_{ij} = \sum_{[t]} \delta J_{ij} = \frac{1}{N} \sum_{\nu=1}^{p} t_\nu \tau_i^\nu \xi_j^\nu = \frac{c}{N} \sum_{\nu=1}^{p} x_\nu \tau_i^\nu \xi_j^\nu . \tag{9}$$

Inserting (9) into (6), we obtain for all output nodes i and for all patterns ν

$$1 \leqslant \frac{1}{c} \sum_{j=1}^{N} \tau_i^\nu \xi_j^\nu J_{ij} = \sum_{\nu=1}^{p} x_i^\nu \frac{1}{N} \tau_i^\nu \tau_i^\mu \sum_{j=1}^{N} \xi_j^\nu \xi_j^\mu = f_i^\mu . \tag{10}$$

Since all output nodes i are processed in parallel, we may consider just one of them, omitting the i from now on.

Defining the correlation matrix $\mathbf{C}$ by $C_{\nu\mu} = (1/N)\tau^\nu \tau^\mu \sum_{j=1}^{N} \xi_j^\nu \xi_j^\mu$, and a vector $\mathbf{1}$ with all components equal to 1, we can write in pattern space

$$\boldsymbol{f} = \mathbf{C}\boldsymbol{x} \geqslant \mathbf{1}. \tag{11}$$

From (7), we define a Hamiltonian

$$H = \frac{N}{2\Delta^2} = \frac{N}{2c^2}\sum_{j=1}^{N} J_j^2 = \frac{1}{2N}\sum_{j=1}^{N}\left[\sum_{\mu=1}^{p} x_\mu \tau_\mu \xi_j^\mu\right]^2 = \frac{1}{2}\boldsymbol{x}\mathbf{C}\boldsymbol{x} = \frac{1}{2}\boldsymbol{f}\mathbf{C}^{-1}\boldsymbol{f} \tag{12}$$

assuming $\mathbf{C}$ is invertible in the last step. If this is not the case, a corresponding condition can always be imposed by Lagrange multipliers, leaving our result unaltered. The calculation of the normalized learning times x_μ now results in the minimization of (12) under constraints (11), (4) and (5). For $f_\mu > 1$, the potential is not bound, and the minimization of H leads to $\partial H/\partial f_\mu = \sum_{\nu=1}^{p}(C^{-1})_{\mu\nu} f_\nu = x_\mu = 0$, i.e. patterns with potentials $f > 1 (= 1)$ have learning times $x = 0 (\geqslant 0)$.

3.1. Distribution of learning times

We now proceed to calculate the probability $w(x)\,\mathrm{d}x$ that for arbitrary but fixed μ, x_μ has values between x and $x + \mathrm{d}x$. Thus

$$w(x) = \langle\langle \delta(x - x_\mu)\rangle\rangle \tag{13}$$

where the average is over patterns $\{\tau_\mu\}$ and $\{\boldsymbol{\xi}^\mu\}$ with distribution (4), (5). The δ function is expressed as a Fourier transformation of the characteristic function

$$g(k) = \langle\langle \mathrm{e}^{\mathrm{i}kx_\mu}\rangle\rangle \tag{14}$$

the mean learning time being

$$\langle x\rangle = -\mathrm{i}\frac{\partial}{\partial k}\bigg|_{k=0} g(k). \tag{15}$$

The characteristic function can be written as a formal thermodynamic average [11]. We have to compute

$$g(k) = \lim_{\beta\to\infty}\left\langle\!\!\left\langle \frac{1}{Z}\int_{-\infty}^{+\infty}\prod_{\nu=1}^{p}[\mathrm{d}x_\nu\,\Theta(f_\nu - 1)]\exp(\mathrm{i}kx_{\nu=\mu} - \beta H)\right\rangle\!\!\right\rangle. \tag{16}$$

Introducing replicas, this form will be obtained in the limit $n \to 0$ if kx_μ is *not* replicated:

$$g(k) = \lim_{\substack{\beta\to\infty\\ n\to 0}}\left\langle\!\!\left\langle \int_{-\infty}^{+\infty}\prod_{\nu,a}[\mathrm{d}x_{\nu a}\,\Theta(f_\nu^a - 1)]\exp(\mathrm{i}kx_{\nu=\mu,a=1} - \beta\sum_a H_a)\right\rangle\!\!\right\rangle. \tag{17}$$

The quadratic dependence of the Hamiltonian on $\{\tau_\mu\}$, $\{\boldsymbol{\xi}^\mu\}$ must be linearized in order to perform the average. However, the term $\mathrm{i}kx_{\nu=\mu,a=1}$ in the exponential makes it impossible to introduce directly an auxiliary Gaussian field for this purpose. Instead, if we express this term as a function of the conjugate variable $\hat{h}$ of the Θ-functions, we show now that this problem is not going to arise. Rewriting the Θ-functions as exponentials, we obtain from (17)

$$g(k) = \lim_{\substack{\beta\to\infty\\ n\to 0}}\left\langle\!\!\left\langle \int_{-\infty}^{+\infty}\prod_{\nu,a}[\mathrm{d}x_{\nu a}\,\mathrm{d}\hat{h}_{\nu a}]\int_1^{\infty}\prod_{\nu,a}\mathrm{d}h_{\nu a}\right.\right.$$
$$\left.\left.\times\exp\left\{\mathrm{i}kx_{\mu 1} + \mathrm{i}\beta\sum_a \hat{\boldsymbol{h}}_a^T\boldsymbol{h}_a - \tfrac{1}{2}\beta\sum_a \boldsymbol{x}_a^T\mathbf{C}\boldsymbol{x}_a - \mathrm{i}\beta\sum_a \hat{\boldsymbol{h}}_a^T\mathbf{C}\boldsymbol{x}_a\right\}\right\rangle\!\!\right\rangle. \tag{18}$$

Transforming the $\prod_{\nu=1}^{p} \mathrm{d}x_{\nu a}\, \mathrm{d}\hat{h}_{\nu a}$ integrations into $\prod_{\nu=1}^{p} \mathrm{d}w_{\nu a}\, \mathrm{d}\hat{h}_{\nu a}$ integrations, where $w_{\nu a} = x_{\nu a} + \mathrm{i}\hat{h}_{\nu a}$, we find

$$g(k) = \lim_{\substack{\beta\to\infty \\ n\to 0}} \left\langle\!\!\left\langle \int_{-\infty}^{+\infty} \prod_{\nu,a} [\mathrm{d}w_{\nu a}\, \mathrm{d}\hat{h}_{\nu a}] \int_{1}^{\infty} \prod_{\nu,a} \mathrm{d}h_{\nu a} \right.\right.$$
$$\left.\left. \times \exp\left\{ \mathrm{i}k w_{\mu 1} + k\hat{h}_{\mu 1} - \tfrac{1}{2}\beta \sum_{a} (\mathrm{i}\hat{\boldsymbol{h}}_a^T \boldsymbol{h}_a - \boldsymbol{w}_a^T \mathbf{C} \boldsymbol{w}_a + \hat{\boldsymbol{h}}_a^T \mathbf{C} \hat{\boldsymbol{h}}_a) \right\} \right\rangle\!\!\right\rangle. \tag{19}$$

Performing $\prod_{\nu,a} \mathrm{d}w_{\nu a}$ integrations gives

$$\int_{-\infty}^{+\infty} \prod_{\nu,a} \mathrm{d}w_{\nu a} \exp\left\{ \mathrm{i}k w_{\mu 1} - \tfrac{1}{2}\beta \sum_{a} \boldsymbol{w}_a^T \mathbf{C} \boldsymbol{w}_a \right\} = \exp\left\{ -\frac{k^2}{2\beta} (\mathbf{C}^{-1})_{\mu\mu} - \frac{n}{2} \ln(\beta \det(\mathbf{C})) \right\}. \tag{20}$$

In the $(\beta \to \infty)$ limit, this becomes independent of k, which means the w terms do not contribute to $g(k)$. Now we linearize the remaining quadratic term in the exponential by auxiliary fields J_{ja}:

$$\exp\left\{ -\tfrac{1}{2}\beta \sum_{a} \hat{\boldsymbol{h}}_a^T \mathbf{C} \hat{\boldsymbol{h}}_a \right\} = \int_{-\infty}^{+\infty} \prod_{ja} \mathrm{d}J_j^a \exp -\tfrac{1}{2}\beta \sum_{ja} \left[J_{ja}^2 + \frac{\mathrm{i}}{\sqrt{N}} J_{ja} \sum_{\nu=1}^{p} \hat{h}_{\nu a} \tau_\nu \xi_j^\nu + \frac{\ln \beta}{\beta} \right].$$

The last term does not contribute to the result in the limit $\beta \to \infty$ and can therefore be omitted. Comparison with the quadratic form in the original Hamiltonian (12) shows that the Js introduced here are proportional to the components of the perceptron vector $\boldsymbol{J}$. Insertion into (18) yields

$$g(k) = \lim_{\substack{\beta\to\infty \\ n\to 0}} \left\langle\!\!\left\langle \int_{-\infty}^{+\infty} \prod_{ja} \mathrm{d}J_j^a \prod_{\nu,a} \mathrm{d}\hat{h}_{\nu a} \int_{1}^{\infty} \prod_{\nu,a} \mathrm{d}h_{\nu a} \right.\right.$$
$$\left.\left. \times \exp\left\{ k\hat{h}_{\mu 1} + \mathrm{i}\beta \sum_{\nu a} \hat{h}_{\nu a} h_{\nu a} - \sum_{ja} \left[\tfrac{1}{2}\beta J_{ja}^2 + \mathrm{i}\frac{\beta}{\sqrt{N}} J_{ja} \sum_{\nu=1}^{p} \hat{h}_{\nu a} t_\nu \xi_j^\nu \right] \right\} \right\rangle\!\!\right\rangle. \tag{21}$$

The last term in the exponential (21) is of order $1/\sqrt{N}$ and factorizes in ν and j. We may therefore perform the average with respect to $\{\boldsymbol{\xi}^\nu\}$ with distribution (4), keeping terms to second order in $1/\sqrt{N}$. Higher-order terms do not contribute to the result in the thermodynamic limit [7]. The exponential is then

$$k\hat{h}_{\mu 1} + \beta \sum_{\nu a} \hat{h}_{\nu a} h_{\nu a} - \tfrac{1}{2}\beta \sum_{j,a} J_{ja}^2$$
$$+ \sum_{\nu j} \left[-\mathrm{i} m_{\mathrm{in}} \frac{\beta}{\sqrt{N}} t_\nu \sum_{a} J_{ja} \hat{h}_{\nu a} - \frac{\beta^2}{2N} (1 - m_{\mathrm{in}}^2) \left(\sum_{a} \hat{h}_{\nu a} J_{ja} \right)^2 \right]. \tag{22}$$

Considering the ν summations, we see that only terms with $\nu = \mu$ contribute to the exponent for $n \to 0$. We introduce order parameters

$$Q_a = \frac{1}{N} \sum_{j=1}^{N} J_{ja}^2 \qquad q_{ab} = \frac{1}{N} \sum_{j=1}^{N} J_{ja} J_{jb} \qquad M_a = \frac{1}{\sqrt{N}} \sum_{j=1}^{N} J_{ja} \tag{23}$$

and enforce them by δ functions. Linearizing the last squared sum by a Gauss integration over an auxiliary field z, and assuming replica symmetry, the integrand then reads

$$\exp\left\{k\hat{h}_1 + \mathrm{i}\beta\sum_a \hat{h}_a h_a - \tfrac{1}{2}\beta\sum_{ja} J_{ja}^2 - \mathrm{i}m_{\mathrm{in}}M\beta\tau\sum_a \hat{h}_a - \tfrac{1}{2}\beta^2(1-m_{\mathrm{in}}^2)(Q-q)\right.$$
$$\left.\times\sum_a \hat{h}_a^2 + \mathrm{i}\beta z\sqrt{q(1-m_{\mathrm{in}}^2)}\sum_a \hat{h}_a\right\}\prod_{a=1}^{n}[\delta\text{ functions}] \tag{24}$$

Again, the $\hat{h}_a$ integrations contribute for $a=1$ a special term $\neq 0$ to the integral. They are coupled with z and h_a integrations. The remaining variables of integration are J_{ja}, the order parameters and their conjugate fields. Since these variables are not coupled to the $\hat{h}_a$ integrations, they provide a constant factor in the exponential and can therefore be neglected. Performing the $n \to 0$ limit, we are then left with

$$g(k) = \left\langle \lim_{\beta\to\infty}\int_{-\infty}^{+\infty} \mathrm{D}z \frac{\int_{-\infty}^{+\infty}\mathrm{d}\hat{h}\int_1^{\infty}\mathrm{d}h\,\exp\{k\hat{h}+G(h,\hat{h})\}}{\int_{-\infty}^{+\infty}\mathrm{d}\hat{h}\int_1^{\infty}\mathrm{d}h\,\exp\{G(h,\hat{h})\}}\right\rangle_\tau \tag{25}$$

where Dz is the Gaussian measure $(\mathrm{d}z/\sqrt{2\pi})\mathrm{e}^{-z^2/2}$ and

$$G(h,\hat{h}) = \mathrm{i}\beta\hat{h}\left(h - m_{\mathrm{in}}M\tau + z\sqrt{q(1-m_{\mathrm{in}}^2)}\right) - \tfrac{1}{2}\beta(1-m_{\mathrm{in}}^2)(Q-q)\hat{h}^2. \tag{26}$$

Performing the $\hat{h}$ integrations, and introducing

$$H(h) = \int_h^{\infty}\mathrm{D}x \qquad \tilde{\Delta} = \frac{1}{\sqrt{q(1-m_{\mathrm{in}}^2)}} \tag{27}$$

this yields

$$g(k) = \left\langle \lim_{\beta\to\infty}\int_{-\infty}^{+\infty}\mathrm{D}z \frac{H[(z+\tilde{\Delta}(1-m_{\mathrm{in}}M\tau - \mathrm{i}k/\beta))\sqrt{q/(Q-q)}]}{H[(z+\tilde{\Delta}(1-m_{\mathrm{in}}M\tau))\sqrt{q/(Q-q)}]}\right\rangle_\tau . \tag{28}$$

The H functions will differ from each other in the $\beta\to\infty$ limit only for $z+\tilde{\Delta}(1-m_{\mathrm{in}}M\tau) > 0$. We shall therefore split the z interval:

$$g(k) = \sum_{\tau=\pm 1} p(\tau)\left(\int_{-\infty}^{-\tilde{\Delta}(1-m_{\mathrm{in}}M\tau)}\mathrm{D}z + \int_{-\tilde{\Delta}(1-m_{\mathrm{in}}M\tau)}^{\infty}\mathrm{D}z\exp\left\{\mathrm{i}\frac{\tilde{\Delta}}{\lambda}k[z+\tilde{\Delta}(1-m_{\mathrm{in}}M\tau)]\right\}\right) \tag{29}$$

with $\lambda = \beta(Q-q)/q$ and $p(\tau)$ after (5).

The parameters λ, $\tilde{\Delta}$ and $m_{\mathrm{in}}M$ have to be taken in the thermodynamic limit, i.e. at the saddle point of (17). Note that M always appears with m_{in}, reflecting that any input bias m_{in} is compensated for by a variation of the order parameter M. We will therefore write $\tilde{M}$ instead of $m_{\mathrm{in}}M$ in the following. The saddle point has to be taken in (21), omitting the term $k\hat{h}_{\mu 1}$. We can directly apply Gardner's result [6] for $\tilde{M}$ and $\tilde{\Delta}$, since in both cases we extremize the exponential under constraints given by corresponding Θ functions.

A comparison yields identical equations for $\pm vm + k \Longleftrightarrow \pm\tilde{M} + 1$; $\Delta \Longleftrightarrow \tilde{\Delta}$. The corresponding equations do *not* depend on the input bias m_{in}:

$$0 = \sum_{\tau=\pm 1} p(\tau)\tau \int_{-\tilde{\Delta}(1-\tilde{M}\tau)}^{\infty} \mathrm{D}z[z + \tilde{\Delta}(1-\tilde{M}\tau)] \tag{30}$$

$$\frac{1}{\alpha} = \sum_{\tau=\pm 1} p(\tau) \int_{-\tilde{\Delta}(1-\tilde{M}\tau)}^{\infty} \mathrm{D}z[z + \tilde{\Delta}(1-\tilde{M}\tau)]^2. \tag{31}$$

The two equations (30), (31) for the order parameters $\tilde{\Delta}$, $\tilde{M}$, can be solved numerically. The third-order parameter can also be obtained from Gardner's results [6], in one line with the total learning time, which is given by $\alpha\langle x\rangle = (1/N)(\sum_{v=1}^{p} x_v)_{x_v>0}$. For $x_v > 0$, however, the internal potential after learning is $1 = f_v = \sum_{\mu=1}^{p} C_{v\mu}x_\mu$. Multiplying yields $\alpha\langle x\rangle = (1/N)(\sum_{\mu,v} x_v C_{v\mu} x_\mu) = (2/N)\langle H\rangle$. At the saddle point, H is given by $H = \frac{1}{2}N\sum_{j=1}^{N}(J_j/c)^2 = \frac{1}{2}Nq$, thus

$$\alpha\langle x\rangle = q = \frac{1}{\tilde{\Delta}^2(1-m_{\text{in}}^2)}. \tag{32}$$

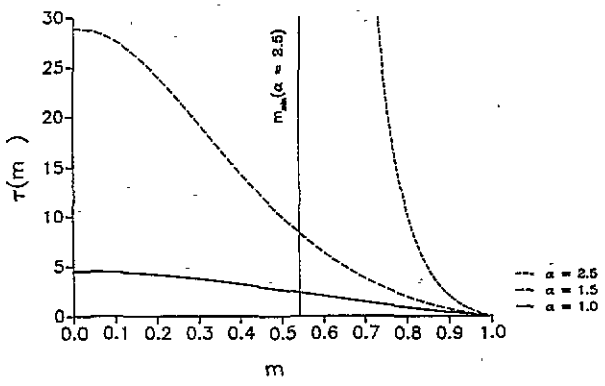

Figure 1. Total learning time τ of the MINOVER algorithm as a function of bias in the autoassociative net at fixed capacity α. Note that for $\alpha = 0.5$, a bias of 0.542 is at least necessary, as follows from the critical solution to (30), (31). (This also applies to figures 3 and 4).

Figure 1 shows the dependence of the total learning time on bias for autoassociative nets, i.e. $m_{\text{in}} = m_{\text{out}}$. Now, we may also derive the learning time from the characteristic function (29)

$$q = \alpha\langle x\rangle = -\mathrm{i}\alpha\frac{\partial}{\partial k}\bigg|_{k=0} g(k) = \sum_{\tau=\pm 1} p(\tau)\frac{\alpha\tilde{\Delta}}{\lambda}\int_{-\tilde{\Delta}(1-\tilde{M}\tau)}^{\infty} \mathrm{D}z[z + \tilde{\Delta}(1-\tilde{M}\tau)]. \tag{33}$$

Using (27) and (30), we obtain

$$\lambda(\tilde{\Delta}, \tilde{M}) = \alpha\tilde{\Delta}^3(1-m_{\text{in}}^2)(1+m_{\text{out}})\int_{-\tilde{\Delta}(1-\tilde{M})}^{\infty} \mathrm{D}z[z + \tilde{\Delta}(1-\tilde{M})]. \tag{34}$$

We now derive the distribution of learning times $w(x)$ from a Fourier transformation of $g(k)$.

$$w(x) = \sum_{\tau=\pm 1} p(\tau)\left[\delta(x)[1 - H(-\tilde{\Delta}(1-\tilde{M}\tau))] + \Theta(x)\frac{\lambda}{\tilde{\Delta}\sqrt{2\pi}}\right.$$
$$\left.\times \exp\left\{-\frac{1}{2}\left[\frac{x - \tilde{\Delta}^2(1-\tilde{M}\tau)/\lambda}{\tilde{\Delta}/\lambda}\right]^2\right\}\right]. \tag{35}$$

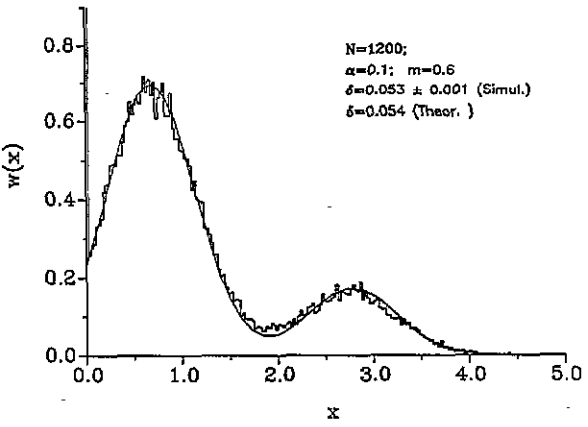

Figure 2. Distribution of normalized (to threshold c) learning times in the MINOVER algorithm.

In figure 2, we choose parameters $m_{in} = m_{out} = 0.6$, $\alpha = 0.1$ to distinguish the two Gaussian parts from each other. Comparison with computer simulations of 500 sets of patterns at network size $N = 1200$ show excellent agreement with the theoretical results. We explore a fraction P_0 of patterns which have learning time 0, i.e. which are stored together with the other patterns *without* being learnt explicitly (figure 3):

$$P_0 = 1 - \sum_{\tau=\pm 1} p(\tau) H[-\tilde{\Delta}(1 - \tilde{M}\tau)]. \tag{36}$$

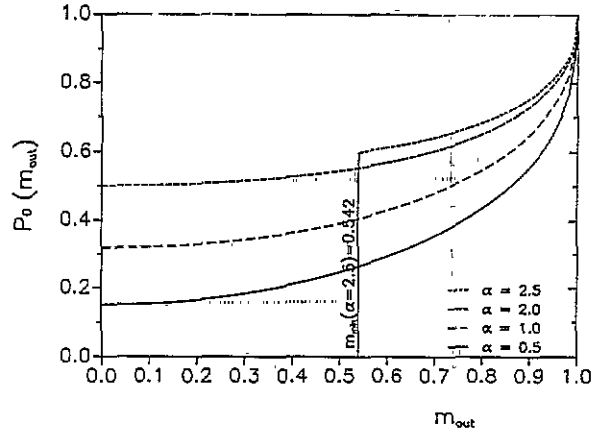

Figure 3. Fraction of patterns which are not *explicitly* learnt in the MINOVER algorithm as a function of output bias.

3.2. Internal potentials

The distribution of internal potentials $\tilde{w}(f) = \langle \delta(f - f_\mu) \rangle$ can now easily be evaluated in this framework. Again, we consider the Fourier transformation $\tilde{g}(k) = \langle e^{ikf_\mu} \rangle$. Evaluating the thermodynamic average (17), and rewriting the Θ functions as integrals over δ functions, we earlier forced $f_\mu = h_\mu$. Thus the previous calculation directly gives an analogue to (25),

$$\tilde{g}(k) = \left\langle \lim_{\beta\to\infty} \int_{-\infty}^{+\infty} Dz \frac{\int_{-\infty}^{+\infty} d\hat{h} \int_1^{\infty} dh \, \exp\{ikh + G(h, \hat{h})\}}{\int_{-\infty}^{+\infty} d\hat{h} \int_1^{\infty} dh \, \exp\{G(h, \hat{h})\}} \right\rangle_\tau \tag{37}$$

with $G(h, \hat{h})$ as in (26). Doing the $\hat{h}$ integration and splitting the z interval as before, we obtain

$$\tilde{g}(k) = \sum_{\tau=\pm 1} p(\tau)\left[\int_{-\infty}^{-\tilde{\Delta}(1-\tilde{M}\tau)} \mathrm{D}z \exp\left\{-\mathrm{i}k\left(\frac{z}{\tilde{\Delta}} - \tilde{M}\tau\right)\right\} + \mathrm{e}^{\mathrm{i}k} H[-\tilde{\Delta}(1-\tilde{M}\tau)]\right]. \quad (38)$$

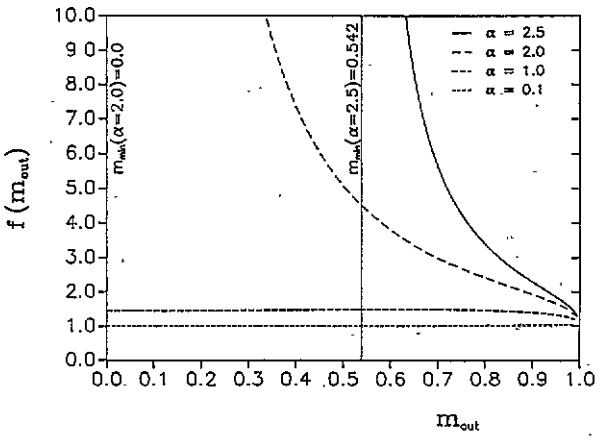

Figure 4. Average internal potential in the MINOVER algorithm as a function of output bias.

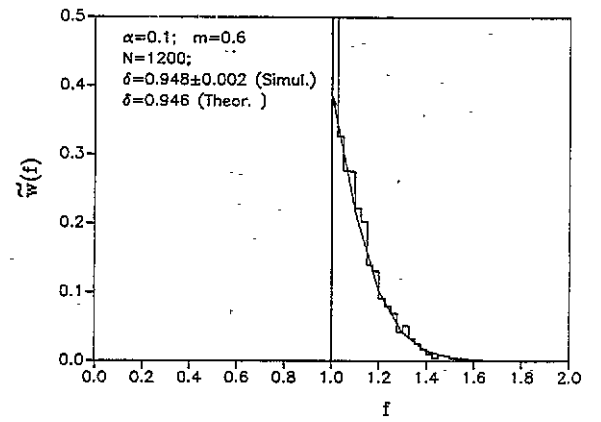

Figure 5. Distribution of internal potentials in the MINOVER algorithm.

The average internal potential (figure 4) is then given by use of (34) and (38)

$$\langle f \rangle = -\mathrm{i}\alpha \frac{\partial}{\partial k}\bigg|_{k=0} \tilde{g}(k) = m_{\mathrm{out}}\tilde{M} + \frac{\lambda}{\alpha\tilde{\Delta}^4(1-m_{\mathrm{in}}^2)}. \quad (39)$$

The Fourier transformation of $\tilde{g}(k)$ gives the potential distribution:

$$\tilde{w}(f) = \sum_{\tau=\pm 1} p(\tau)\left[\delta(f-1)H[-\tilde{\Delta}(1-\tilde{M}\tau)] + \Theta(f-1) \times \frac{\tilde{\Delta}}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\tilde{\Delta}^2(f-\tilde{M}\tau)^2\right\}\right]. \quad (40)$$

Figure 5 shows the theoretical curve and simulation data obtained with the same parameters as in figure 2. Note that the two Gaussian peaks are always centred at $f < 1$, leaving us with the truncated parts only. P_0 and the internal potentials depend on the output bias m_{out} only.

3.3. Interpretation and information storage

Comparing the distributions (35) and (40), we find that patterns with internal potentials $f \underset{(=)}{>} 1$ have learning times $x \underset{(>)}{=} 0$, as was required earlier. The two Gaussians for $\tau = \pm 1$ in the distributions show that patterns with $\tau m_{\text{out}} > 0$ are learnt more easily and are stored more stably (smaller learning times, higher internal potentials) than their negative counterparts. This seems plausible from an investigation of the high-bias limit: for $m_{\text{in}}, m_{\text{out}}$ close to 1, most of the components of J will be positive. Thus most ($\tau = 1$) patterns will satisfy the potential condition (11) automatically (high P_0), leaving the adjusting of the weights to the ($\tau = -1$) patterns. However, since their proportion is small, the total learning time decreases with the bias. Due to the easy embedding of all positive-output patterns, they will commonly satisfy the potential condition with f just slightly larger than 1. Thus $\langle f \rangle$ decreases with m_{out}, and for $m_{\text{out}} \to 1$, $\langle f \rangle \to 1$.

Finally, we investigate whether in an autoassociative network ($m_{\text{in}} = m_{\text{out}} = m$), a given information can be learnt faster by distribution into a larger number of biased patterns. The total information contained in a set of $p = \alpha N$ biased patterns with N bits is given by $I_{\text{total}} = N^2 I$, where the information capacity

$$I = \alpha \ln(0.5) \left\{ \left(\frac{1+m}{2} \right) \ln \left(\frac{1+m}{2} \right) + \left(\frac{1-m}{2} \right) \ln \left(\frac{1-m}{2} \right) \right\}. \tag{41}$$

For fixed α, I decreases monotonically from $I(m = 0) = \alpha$ to $I(|m| = 1) = 0$. However, from an information processing point of view one has to investigate the case of fixed information capacity I. Then it is known [6] that $\alpha \to \infty$ as $|m| \to 1$, i.e. sparsely coded (or biased) associative memories are able to store a diverging number of patterns.

Thus, changing free parameters from (α, m) to (I, m) (and $\alpha = \alpha(I, m)$ according to equation (41)), we solve for the total learning time $\langle x \rangle$ (equation (32)) with the conditions (30), (31), (41)). The result is given for three different information capacities in figure 6, showing that the total learning time always increases with bias m. A 'minimal' representation of information in unbiased patterns which has no redundancy is therefore favourable for fast learning. However, the embedding of information will be weaker than in the biased case.

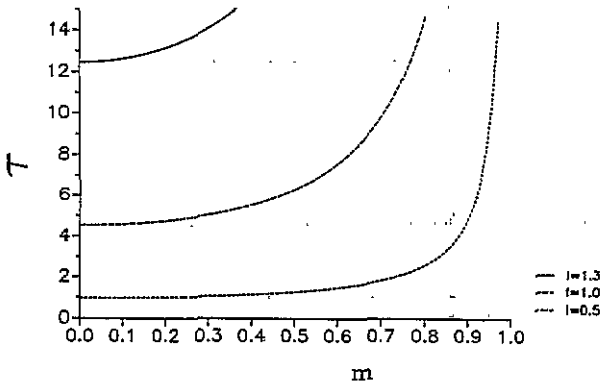

Figure 6. Total learning time τ of the MINOVER algorithm as a function of bias in the autoassociative net at fixed information capacity I.

4. The ADALINE algorithm

Another way of imposing the fixed point condition (3) is to formulate a linear algorithm which attempts to satisfy for all i, ν:

$$f_i^\nu = 1. \tag{42}$$

The ADALINE algorithm [15] proceeds by gradient descent on the error function to (42),

$$(E_i^\nu)^2 = (1 - f_i^\nu)^2 = \left(1 - \tau_i^\nu \sum_{j=1}^{N} J_{ij}\xi_j^\nu\right)^2 \tag{43}$$

altering the J_{ij} by

$$J_{ij}(t+1) = J_{ij}(t) + \delta J_{ij}(t) \qquad \forall i, j. \tag{44}$$

After parallel presentation of all patterns we obtain

$$\delta J_{ij}(t) = -\frac{\gamma}{N} \sum_{\mu=1}^{p} \left(1 - \tau_i^\mu \sum_{k=1}^{N} J_{ik}(t)\xi_k^\mu\right)\tau_i^\mu \xi_j^\mu. \tag{45}$$

We aim to compute the decay of the total learning error,

$$E(t) = \frac{1}{Np} \sum_{i\nu} E_i^\nu(t). \tag{46}$$

With (43)–(45), this obeys a recursion relation in pattern space,

$$\boldsymbol{E}_i(t+1) = (\mathbf{1} - \gamma \mathbf{C}^i)\boldsymbol{E}_i(t) \tag{47}$$

where $\mathbf{1}$ is the unit diagonal matrix and $\mathbf{C}^i$ the correlation matrix with elements

$$C_{\nu\mu}^i = \frac{1}{N}\tau_i^\nu \tau_i^\mu \sum_{j=1}^{N} \xi_j^\nu \xi_j^\mu. \tag{48}$$

Omitting the index i, and starting with an empty network ($J_{ij} = 0$)

$$E(t) = \frac{1}{p} \sum_{\mu,\nu} [\mathbf{B}^{2t}]_{\mu\nu} \qquad \text{where } \mathbf{B} = \mathbf{1} - \gamma \mathbf{C}. \tag{49}$$

We shall solve this equation first for patterns with fixed cross correlation, then for general patterns. The first problem can be treated algebraically, providing us with *exact* (i.e. non-averaged) results which will lead the calculation in the general case.

4.1. Fixed cross correlation

For the moment, we restrict ourselves to biased patterns with fixed cross correlation

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N} \xi_j^\nu \xi_j^\mu \qquad \forall \nu, \mu; \nu \neq \mu. \tag{50}$$

By complete induction, we see that all diagonal elements $b_{\mu\nu}(t)$ and all off-diagonal elements $a_{\mu\nu}(t)$ of the matrix $(\mathbf{B})^t$ can be written $b_{\mu\nu}(t) \equiv b_t$, $a_{\mu\nu}(t) \equiv a_t \cdot \tau_\mu \tau_\nu$. They obey the recursion relation

$$\begin{pmatrix} a_{t+1} \\ b_{t+1} \end{pmatrix} = \mathbf{A} \cdot \begin{pmatrix} a_t \\ b_t \end{pmatrix} \tag{51}$$

where

$$\mathbf{A} = \begin{pmatrix} (p-2)a_0 + b_0 & a_0 \\ (p-1)a_0 & b_0 \end{pmatrix} \qquad \text{and} \qquad \begin{matrix} a_0 = -\gamma\sigma^2 \\ b_0 = 1-\gamma \end{matrix}.$$

Thus we can compute the total error

$$E(t) = \frac{1}{p}\sum_{\mu,\nu} [\mathbf{B}^{2t}]_{\mu\nu} = b_{2t} + a_{2t} \cdot \tilde{\sigma}^2 \tag{52}$$

where the output correlation $\tilde{\sigma}^2 = (1/p)\sum_{\mu\nu(\neq\mu)} \tau_\mu \tau_\nu$. Applying the recursion (51) $2t$ times, we obtain

$$\begin{pmatrix} a \\ b \end{pmatrix}_{2t} = \mathbf{A}^{2t} \cdot \begin{pmatrix} a \\ b \end{pmatrix}_0 \tag{53}$$

Using the eigenvalues $\lambda_{1,2}$ of $\mathbf{A}$, we may define two coefficients β_0, β_1:

$$\lambda_{1,2}^{2t} = \beta_0 + \beta_1 \cdot \lambda_{1,2}. \tag{54}$$

By the Cayley–Hamilton theorem, the corresponding matrix equation holds with the same coefficients,

$$\mathbf{A}^{2t} = \beta_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \beta_1 \mathbf{A}. \tag{55}$$

If we draw the output from a set of τ_μ with constraint $\sum_{\mu=1}^{p} \tau_\mu = p m_{\text{out}}$, $\tilde{\sigma}^2$ is fixed, and we can compute $E(t)$ exactly. Inserting (54) and (55) into (53), we obtain the coefficients a_{2t}, b_{2t} needed for (52). The final result then is

$$E(t) = (1 - m_{\text{out}}^2)[1 - \gamma(1-\sigma^2)]^{2t} + m_{\text{out}}^2[1 - \gamma(1 + (p-1)\sigma^2)]^{2t}. \tag{56}$$

For non-vanishing input and output bias, we face the contribution of a large eigenvalue of order N to the training error, which was not present in the unbiased case [8]. While the first eigenvalue broadens into a spectrum, we will see that the large eigenvalue remains isolatedly present in the general case, forcing us to choose a gain parameter γ of order $1/N$ to ensure convergence.

4.2. General choice of patterns

We rewrite the error decay (43) with the help of two identities. First, we shall use that for a real interval of integration and small η,

$$\int_{-\infty}^{+\infty} \mathrm{d}x \, \frac{1}{x - \mathrm{i}\eta} = \mathrm{i}\pi \Longleftrightarrow \delta(x) = \frac{1}{\pi} \operatorname{Im}\left(\frac{1}{x - \mathrm{i}\eta}\right).$$

In the following, we leave the small imaginary part as understood and omit η. Then we obtain, using a method similar to that outlined in [12], for real $\boldsymbol{u}$,

$$\frac{1}{p} \boldsymbol{u}^T (\mathbf{B}^{2t}) \boldsymbol{u} = \frac{1}{p\pi} \int_{-\infty}^{+\infty} \mathrm{d}\lambda \, \lambda^{2t} \operatorname{Im}\{\boldsymbol{u}^T (\lambda \mathbf{1} - \mathbf{B})^{-1} \boldsymbol{u}\}. \tag{57}$$

Furthermore,

$$\boldsymbol{u}^T \mathbf{C}^{-1} \boldsymbol{u} = 2 \frac{\partial}{\partial(\epsilon^2)} \left[\ln \int_{-\infty}^{+\infty} \prod_{\mu=1}^{p} \mathrm{d}x_\mu \, \exp\{-\tfrac{1}{2} \boldsymbol{x}^T \mathbf{C} \boldsymbol{x} + \boldsymbol{\epsilon}^T \mathbf{U} \boldsymbol{x}\} \right] \tag{58}$$

where $\boldsymbol{\epsilon} = (\epsilon, \epsilon, \ldots, \epsilon)$ and $\mathbf{U}_{\mu\nu} = u_\mu \delta_{\mu\nu}$. For $E(t)$, $\boldsymbol{u} = \mathbf{1}$. If we use the replica trick in the form

$$\ln Z = \lim_{n \to 0} \frac{\partial}{\partial n} Z^n$$

we have to compute

$$E(t) = \frac{2}{\alpha N \pi} \int_{-\infty}^{+\infty} \mathrm{d}\lambda \, \lambda^{2t} \operatorname{Im} \frac{\partial}{\partial(\epsilon^2)} \frac{\partial}{\partial n} \bigg|_{n \to 0} Z$$

where

$$Z = \int_{-\infty}^{+\infty} \left[\prod_{\mu, a} \mathrm{d}x_\mu^a \right] \left\langle\!\!\left\langle \exp\left\{ -\tfrac{1}{2} \sum_a (\boldsymbol{x}_a^T (\lambda \mathbf{1} - \mathbf{B}) \boldsymbol{x}_a + 2 \boldsymbol{\epsilon}^T \boldsymbol{x}_a) \right\} \right\rangle\!\!\right\rangle. \tag{59}$$

We must perform the average over the term

$$\exp\left\{ -\frac{\gamma}{2N} \sum_{\mu, \nu} \sum_a x_\mu^a x_\nu^a \tau_\mu t_\nu \sum_{j=1}^{N} \xi_j^\mu \xi_j^\nu \right\}. \tag{60}$$

Linearizing the exponential with auxiliary Gaussian fields z_{ja} yields

$$\int_{-\infty}^{+\infty} \prod_{ja} \mathrm{D}z_{ja} \prod_\mu \left\langle \prod_j \left\langle \exp\left\{ \tau_\mu \xi_j^\mu \sqrt{\frac{-\gamma}{N}} \sum_a z_{ja} x_\mu^a \right\} \right\rangle_{\xi_j^\mu} \right\rangle_{\tau_\mu}. \tag{61}$$

As in (21) ff, we expand the exponential to order $1/N$, perform the average and introduce order parameters in z space

$$Q_a = \frac{1}{N} \sum_{j=1}^{N} z_{ja}^2 \qquad q_{ab} = \frac{1}{N} \sum_{j=1}^{N} z_{ja} z_{jb} \tag{62}$$

enforcing them by δ functions with conjugate parameters $\hat{Q}_a, \hat{q}_{ab}$. Assuming replica symmetry, we obtain in the $n \to 0$, $N \to \infty$ limit, after standard integrations,

$$Z = \int_{-\infty}^{+\infty} [\mathrm{D}]\{\exp\{\tfrac{1}{2}nN(G_1 + \alpha \cdot G_2)\}\}_{\{\tau_\mu\}}$$

where

$$G_1 = \left(R\sum_{\mu=1}^{p}\tau_\mu\right)^2 - \ln(\hat{r}) - \frac{\mathrm{i}\hat{q}(1+q\hat{r})}{\hat{r}} - \frac{1}{N}\ln\left(\frac{NS^2}{\hat{r}}\right) + (\hat{r} + \mathrm{i}\hat{q} - 1)(r+q) \tag{64}$$

$$G_2 = \frac{\epsilon^2 + \tilde{\gamma}q}{\tilde{\lambda} - \tilde{\gamma}r} - \ln(\tilde{\lambda} - \tilde{\gamma}r) \tag{65}$$

$$R = \sqrt{\frac{-\gamma}{N}}\frac{\epsilon m_{\mathrm{in}}}{\tilde{\lambda} - \tilde{\gamma}r}\frac{1}{\sqrt{\hat{r}(1+NS^2)}} \tag{66}$$

$$S^2 = \frac{\alpha m_{\mathrm{in}}^2 \gamma}{\tilde{\lambda} - \tilde{\gamma}r} \qquad \hat{r} = 2\mathrm{i}\hat{Q} - \mathrm{i}\hat{q} \qquad \tilde{\lambda} = \lambda - 1$$
$$r = Q - q \qquad \tilde{\gamma} = \gamma(1 - m_{\mathrm{in}}^2)$$

[D] denotes integrations over $q, \hat{q}, r, \hat{r}$. Averaging over the output represented in the first term of (64) yields, for $n \to 0$,

$$\exp\{\tfrac{1}{2}n\alpha N R^2(1 - m_{\mathrm{out}}^2 + \alpha N m_{\mathrm{out}}^2)\}. \tag{67}$$

Note for later reference that this result holds only for distribution (5) with the τ_μ drawn *independently* of each other, giving

$$\langle \tau_\mu t_\nu \rangle_{\mu \neq \nu} = \langle \tau_\mu \rangle \langle t_\nu \rangle = m_{\mathrm{out}}^2. \tag{68}$$

If we just imposed the condition

$$\sum_{\mu=1}^{p}\tau_\mu = \alpha N m_{\mathrm{out}} \tag{69}$$

we would obtain

$$\langle \tau_\mu t_\nu \rangle = m_{\mathrm{out}}^2 \pm \mathcal{O}\left(\frac{1 - m_{\mathrm{out}}^2}{\sqrt{N}}\right). \tag{70}$$

This will be relevant later to study the ($m_{\mathrm{out}} = 0$) behaviour of $E(t)$, for which we also keep all 'small' factors of $\mathcal{O}(1/N)$ for the moment.

After computing the saddle point (SP) with respect to the set of parameters $\{q, \hat{q}, r, \hat{r}\}$, we derive the argument of the integral (59):

$$(1+\tilde{\lambda})^{2t}\frac{2}{\pi\alpha N}\,\mathrm{Im}\,\frac{\partial}{\partial(\epsilon^2)}\frac{\partial Z}{\partial n}\bigg|_{n\to 0} = -(1+\tilde{\lambda})^{2t}\frac{1}{\pi\alpha}\,\mathrm{Im}\,\frac{\partial(G_1 + \alpha G_2)}{\partial(\epsilon^2)}\bigg|_{\mathrm{SP}}$$
$$= (1+\tilde{\lambda})^{2t}\frac{1}{\pi}\,\mathrm{Im}\left\{\frac{1 - (m_{\mathrm{out}}^2 + (1 - m_{\mathrm{out}}^2)/\alpha N)(1 + ((\tilde{\lambda} - \tilde{\gamma}r_{\mathrm{SP}})/\gamma m_{\mathrm{in}}^2 \alpha N r_{\mathrm{SP}}))^{-1}}{\tilde{\lambda} - \tilde{\gamma}r_{\mathrm{SP}}}\right\} \tag{71}$$

with r_{SP} given by

$$\tilde{\lambda} = \tilde{\gamma} r_{SP}\left(1 + \frac{\alpha}{r_{SP} - 1}\right). \tag{72}$$

After insertion of r_{SP} from (72) into (71), we obtain imaginary parts (i), denoted I_1, due to roots of negative numbers; and (ii), denoted by I_2, via δ functions, due to zero denominators $[\tilde{\lambda} - \tilde{\lambda}_0]^{-1}$.

We find a spectrum of I_1 terms in the interval $\tilde{\lambda}_1 < \tilde{\lambda} < \tilde{\lambda}_2$, where

$$\tilde{\lambda}_{1,2} = \tilde{\gamma}(1 \mp \sqrt{\alpha})^2. \tag{73}$$

In this region, for $m_{in} = 0$,

$$I_1\Big|_{m_{in}=0} = \frac{(1+\tilde{\lambda})^{2t}}{2\pi\alpha\tilde{\gamma}\tilde{\lambda}}\sqrt{(\tilde{\lambda} - \tilde{\lambda}_1)(\tilde{\lambda}_2 - \tilde{\lambda})}. \tag{74}$$

At $m_{in} \neq 0$, we expand for

$$N\gamma\alpha m_{in}^2 \gg |\tilde{\lambda}_2| \Longleftrightarrow |m_{in}| \gg (1/\sqrt{N})(1 + \sqrt{\alpha}) \tag{75}$$

$$I_1\Big|_{m_{in}\neq 0} = I_1\Big|_{m_{in}=0}\left(1 - m_{out}^2 - \frac{1 - m_{out}^2}{\alpha N}\right). \tag{76}$$

Since for $\alpha \neq 1$, $\tilde{\lambda}_{1,2} > 0$, the denominator of I_1 remains non-zero. We may thus compute I_2 more easily by changing the variable of integration to $r_{SP} = r(\tilde{\lambda})$, given by (72), excluding the I_1 interval by this transformation:

$$I_2 = \frac{-1}{\alpha\pi}\int_{\substack{1-\sqrt{\alpha}\\ r\neq 1}}^{1+\sqrt{\alpha}} \mathrm{d}r\,(1+\tilde{\lambda}(r))^{2t}\frac{\mathrm{d}\tilde{\lambda}}{\mathrm{d}r}\,\mathrm{Im}\left(\frac{\partial}{\partial(\epsilon^2)}(G_1 + \alpha G_2)[\tilde{\lambda}(r)]\right). \tag{77}$$

We obtain poles of the integrand at

$$r_1 = 0 \qquad r_2 = 1 + \frac{1 - m_{in}^2}{N m_{in}^2}. \tag{78}$$

The transformation mapped the large eigenvalue to r_2. It now becomes clear that we had to keep terms of $\mathcal{O}(1/N)$ in the calculation, otherwise we would have lost r_2 in the contour of integration.

For $\alpha > 1$, r_1 lies within the interval of integration. For r_2, we must have

$$r_2 < 1 + \sqrt{\alpha} \Longleftrightarrow |m_{in}| > 1/\sqrt{N\sqrt{\alpha}}. \tag{79}$$

Using $\mathrm{Im}[r - r_i]^{-1} = \pi\delta(r - r_i)$, the integral (77) can be performed. For $m_{in} = 0$, r_2 is not inside the integration area, and

$$I_2\Big|_{m_{in}=0} = \frac{\alpha - 1}{\alpha}\Theta(\alpha - 1). \tag{80}$$

For m_{in} obeying (79), we obtain exactly the large eigenvalue given in (56). This follows since correlations between patterns are of order $\mathcal{O}(1/\sqrt{N})$, leaving the large eigenvalue correct to $\mathcal{O}(1)$†.

$$I_2\Big|_{m_{\text{in}}\neq 0} = (1-m_{\text{out}}^2)\frac{\alpha-1}{\alpha}\Theta(\alpha-1)+(m_{\text{out}}^2)[1-\gamma(1+(N\alpha-1)m_{\text{in}}^2)]^{2t}\Theta\left(|m_{\text{in}}|-\frac{1}{\sqrt{N\sqrt{\alpha}}}\right). \tag{81}$$

Before omitting $\Theta(|m_{\text{in}}|-1/\sqrt{N\sqrt{\alpha}})$ and expressions of order $\mathcal{O}(1/N)$ relative to the leading terms, we shall have a closer look at the distribution of outputs. With condition (69), we had to replace m_{out}^2 by (70) in our results, yielding a prefactor to the large eigenvalue term

$$m_{\text{out}}^2+\mathcal{O}((1-m_{\text{out}}^2)/\sqrt{N}). \tag{82}$$

If we choose a gain parameter $\gamma=\mathcal{O}(1)$, for $m_{\text{out}}=0$, $m_{\text{in}}\neq 0$, we would have a contribution

$$\mathcal{O}(1/\sqrt{N})(1-N\alpha\gamma m_{\text{in}}^2)^{2t}=\mathcal{O}(N^{2t-1/2}) \tag{83}$$

which diverges immediately. Computer simulations confirm this divergent behaviour. Regardless of m_{out}, we shall therefore *always* choose $\gamma=\mathcal{O}(1/N)$ for $m_{\text{in}}\neq 0$. For $m_{\text{in}}=0$, however, the Θ function guarantees that the large eigenvalue will vanish. For small N simulations with low bias, though, the Θ function indicates that the large eigenvalue may not yet be encountered, e.g. for $m_{\text{in}}=0.1$, $\alpha=0.3$, $N=100$ (figure 7). This argument is very loose, since the derivations are in the $(N\to\infty)$ limit (equation (75)), but might be a useful estimate for practical purposes.

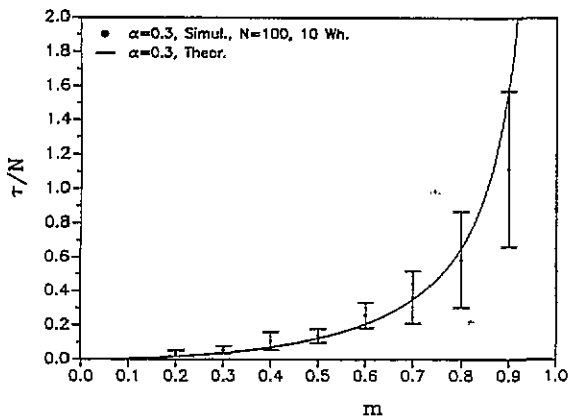

Figure 7. Characteristic training error decay time in the ADALINE algorithm as a function of input bias.

With this result in mind, we finally obtain the decay of the training error, restoring the original variables of integration:

$$E(t)\Big|_{m_{\text{in}}=0} = \frac{\alpha-1}{\alpha}\Theta(\alpha-1)+\int_{\lambda_1}^{\lambda_2}\mathrm{d}\lambda\,(1-\gamma\lambda)^{2t}\frac{\sqrt{(\lambda-\lambda_1)(\lambda_2-\lambda)}}{2\pi\alpha\lambda} \tag{84}$$

$$E(t)\Big|_{m_{\text{in}}\neq 0} = [1-m_{\text{out}}^2]\left\{\frac{\alpha-1}{\alpha}\Theta(\alpha-1)+\int_{\lambda_1}^{\lambda_2}\mathrm{d}\lambda\,(1-\gamma(1-m_{\text{in}}^2)\lambda)^{2t}\frac{\sqrt{(\lambda-\lambda_1)(\lambda_2-\lambda)}}{2\pi\alpha\lambda}\right\} + \left[m_{\text{out}}^2+\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)\right](1-N\alpha\gamma m_{\text{in}}^2)^{2t} \tag{85}$$

† For a detailed treatment, see [5].

with $\lambda_{1,2} = (1 \mp \sqrt{\alpha})^2$.

The $m_{\text{in}} = 0$ result is identical to [8]. Using the methods outlined there, we can minimize the characteristic training error decay time τ with respect to γ,

$$E(t) = E_\infty + E_0 e^{-t/\tau}. \tag{86}$$

For $m_{\text{in}} = 0$, $\gamma_{\text{opt}} = (1+\alpha)^{-1}$ yields

$$\tau_{\min} = \frac{1}{2\ln((1+\alpha)/2\sqrt{\alpha})}. \tag{87}$$

For $m_{\text{in}} \neq 0$,

$$\gamma_{\text{opt}} = 2[N\alpha m_{\text{in}}^2 + (1-m_{\text{in}}^2)(1+(1-\sqrt{\alpha})^2)]^{-1}$$

yields

$$\tau_{\min} = \frac{N\alpha m_{\text{in}}^2}{4(1-m_{\text{in}}^2)(1-\sqrt{\alpha})^2}. \tag{88}$$

For any bias, τ increases (decreases) with α for $\alpha < 1 (> 1)$, diverging at $\alpha = 1$ according to $\tau \sim (1-\sqrt{\alpha})^{-2}$. In figure 7, we compare simulations at $\alpha = 0.3$, $N = 100$ with theoretical predictions. As mentioned earlier, for $m_{\text{in}} = 0.1$ the large eigenvalue could not be encountered. Since we are evaluating decay times of $\mathcal{O}(1/N)$, simulations cannot be expected to be in excellent agreement with theory, but they prove to be within the error bounds.

In both $m_{\text{in}} = 0$ and $m_{\text{in}} \neq 0$ cases, $E_\infty \sim \Theta(\alpha - 1)$ resembles the fact that, for $E_\infty = 0$ to be obtained, conditions (43) are αN linear equations for N variables J_j, which can be satisfied only for $\alpha \leqslant 1$.

4.3. Internal potentials

Finally, we shall calculate the distribution of internal potentials after learning for $\alpha > 1$. Choosing the Hamiltonian (43), we evaluate $w(f) = \langle \delta(f - f_\mu) \rangle$ by the characteristic function

$$\tilde{g}(k) = \left\langle\!\!\left\langle \int_{-\infty}^{+\infty} \prod_{ja} \mathrm{d}J_{ja} \exp\{-\tfrac{1}{2}\beta \sum_{a\nu} (1 - f_{\nu a})^2 + \mathrm{i}k f_{\mu 1}\} \right\rangle\!\!\right\rangle. \tag{89}$$

We proceed exactly as in section 3, deriving order parameters from the partition function

$$Z = \left\langle\!\!\left\langle \int_{-\infty}^{+\infty} \prod_{ja} \mathrm{d}J_{ja} \exp\left\{ -\frac{\beta}{2} \sum_{a\nu} \left(1 - \frac{1}{\sqrt{N}} t_\nu \sum_{j=1}^{N} J_j^a \xi_j^\nu\right)^2 \right\} \right\rangle\!\!\right\rangle. \tag{90}$$

The result is

$$\langle f \rangle = \frac{1 + (\alpha - 1) m_{\text{out}}^2}{\alpha} \tag{91}$$

$$w(f) = \frac{\alpha}{\sqrt{2\pi(1-m_{\text{out}}^2)(\alpha-1)}} \sum_{\tau=\pm 1} p(\tau) \exp\left\{ -\frac{(f - (1 + (\alpha-1) m_{\text{out}} \tau)/\alpha)^2}{2(1-m_{\text{out}}^2)(\alpha-1)/\alpha^2} \right\}. \tag{92}$$

In figure 8, we choose parameters $\alpha = 8$, $m_{\text{out}} = 0.6$ to distinguish the two Gaussian peaks. Simulation and theory are in good agreement, especially considering that with four runs at $N = 50$, the sample data are small. Again, the internal potentials are independent of the input bias. For $\alpha \to 1$, $w(f) \to \delta(f-1)$, showing that the N equations (43) are satisfied exactly. For $2 > \alpha > 1$, the width of the Gaussians increases. For $\alpha > 2$, the Gaussians are again sharper, approaching $w(f) \to \sum_{\tau=\pm 1} p(\tau)\delta(f - m_{\text{out}}\tau)$ for $\alpha \to \infty$. This shows that a particular pattern's output 'sits' in the mean field created by all the other patterns' outputs. Since we are considering the distribution after learning, all dependence on the input bias has been removed. With increasing α, the stabilization of patterns with output $\tau = -\operatorname{sign}(m_{\text{out}})$ worsens, see figure 8. This leads to $\langle f \rangle \to m_{\text{out}}^2$ for $\alpha \to \infty$.

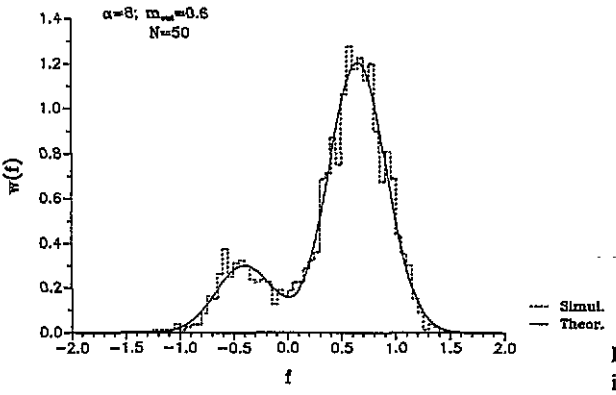

Figure 8. Distribution of internal potentials in the ADALINE algorithm after learning.

4.4. Generalization ability

Recently, Krogh and Hertz [10] have studied the generalization ability of the ADALINE algorithm in a continuous-time limit. They obtained for the generalization error

$$F(t) = \frac{1}{N}\sum_{i=1}^{N} v_i^2 \tag{93}$$

where $v_i = u_i - J_i$, u being the weight vector of the 'teacher'. The time development of v_i was determined according to

$$\frac{\mathrm{d}v_i}{\mathrm{d}t} = -\sum_{j=1}^{N} A_{ij}v_j \qquad \text{with} \qquad A_{ij} = \frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu. \tag{94}$$

Returning to discrete time steps, we write this as a vector in sites i:

$$v(t+1) = v(t) - \gamma \mathbf{A}v(t) = (\mathbf{1} - \gamma\mathbf{A})^t u \tag{95}$$

where we started with an empty network $J(t=0) = \mathbf{0}$. Using equations (57) and (58), we write the generalization error

$$F(t) = \frac{1}{N}u^{\mathrm{T}}(\mathbf{1} - \gamma\mathbf{A})^{2t}u = \frac{2}{N\pi}\int_{-\infty}^{+\infty}\mathrm{d}\lambda\,\lambda^{2t}\,\mathrm{Im}\,\frac{\partial}{\partial(\epsilon^2)}\frac{\partial}{\partial n}\bigg|_{n\to 0}\tilde{Z} \tag{96}$$

where

$$\tilde{Z} = \left\langle \int_{-\infty}^{+\infty} \prod_{aj} \mathrm{d}x_{aj} \, \exp\left\{ -\tfrac{1}{2} \sum_{a} (\boldsymbol{x}_a^T \mathbf{B} \boldsymbol{x}_a + 2\epsilon^T \mathbf{U} \boldsymbol{x}_a) \right\} \right\rangle_{\{\xi_j^\mu\}} \tag{97}$$

with $\mathbf{B} = \mathbf{1} - \gamma \mathbf{A}$ and $(\mathbf{U})_{ij} = u_i \delta_{ij}$. Note that the $\{\boldsymbol{x}\}$ span the space of sites now, in contrast to the pattern space earlier. We may therefore identify (97) with (59) if we formally let $\tau_\mu \equiv 1$ (as $\mathbf{A}$ contains no output), write ϵu_μ instead of ϵ , and exchange labels i and μ in the end, which will transform α into α^{-1}. Due to the distribution (4), there is no problem associated with the interchange $\xi_j^\mu \to \xi_\mu^j$. Finally, the leading factor α^{-1} which scaled $E(t)$ has to be removed. This procedure leads us to replace in (65)

$$\epsilon^2 \to \epsilon^2 \frac{1}{p} \sum_{\mu=1}^{p} u_\mu^2 = \epsilon^2 \tag{98}$$

if we scale $|\boldsymbol{u}|^2 = N$ after $\mu \leftrightarrow j$ exchange. In (64), however, the replacement leads to $\epsilon \sum_{\mu=1}^{p} \tau_\mu \to \epsilon \sum_{\mu=1}^{p} u_\mu \tau_\mu \equiv \epsilon \sum_{\mu=1}^{p} u_\mu$, and we can formally reinterpret m_{out} as the *bias* of $\boldsymbol{u}$, i.e. after $\mu \leftrightarrow j$ exchange, we define

$$m_{\text{out}} \hat{=} \frac{1}{N} \sum_{j=1}^{N} u_j \equiv \tilde{u} \qquad |\tilde{u}| < 1. \tag{99}$$

As in (70), the typical product $\langle u_\mu u_\nu \rangle$ will again attain a variance of $\mathcal{O}(1/\sqrt{N})$, leading to a non-vanishing, large eigenvalue term even for $\tilde{u} = 0$. The resulting generalization error now follows directly from (84), (85):

$$F(t)\Big|_{m=0} = (1-\alpha)\Theta(1-\alpha) + \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \, (1-\gamma\lambda)^{2t} \frac{\sqrt{(\lambda-\lambda_1)(\lambda_2-\lambda)}}{2\pi\lambda} \tag{100}$$

$$F(t)\Big|_{m\neq 0} = [1-\tilde{u}^2]\left\{ (1-\alpha)\cdot\Theta(1-\alpha) + \int_{\lambda_1}^{\lambda_2} \mathrm{d}\lambda \, (1-\gamma(1-m^2)\lambda)^{2t} \frac{\sqrt{(\lambda-\lambda_1)(\lambda_2-\lambda)}}{2\pi\lambda} \right\}$$
$$+ \left[\tilde{u}^2 + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) \right] (1 - N\gamma m^2)^{2t} \tag{101}$$

with $\lambda_{1,2} = (1 \mp \sqrt{\alpha})^2$.

In the continuous time limit, $(1-\gamma\lambda)^{2t} \to \mathrm{e}^{-2\lambda t}$, where t was rescaled to $t\gamma$. Then (100) is identical to the solution given by Krogh and Hertz [10]. Again, for $m \neq 0$ one has to operate in very small time steps, $\gamma_{\text{opt}} \sim \mathcal{O}(1/N)$, and the resulting generalization error will be $F(\infty) = [1-\tilde{u}^2](1-\alpha)\Theta(1-\alpha)$ which is less than in the case of unbiased patterns. From (99), one concludes that generalization becomes best for those 'teachers' pointing in an all-positive (or all-negative) direction of $\boldsymbol{u}$ space. Note that in previous works on generalization, the input was unbiased and therefore the results were independent of a specific teacher due to spherical symmetry. Here, with this symmetry broken, the result depends on the task to be learnt. If the teacher *and* the examples are biased, the learning updates add coherently in the teacher's direction. Note that this results in a smaller generalization error than in the unbiased case, even though for fixed α the information content of the biased patterns is smaller than in the unbiased case (equation (41)). At first sight, this results seems counter-intuitive since generalization improves upon the presentation of *less* information in the

patterns. However, the fact that the presented patterns are *biased* can already be regarded as additional information in itself, since it evokes a biased perceptron vector J constructed in the learning process. Any pattern bias will produce the mentioned learning updates which are coherent with the teacher, hence the difference in generalization errors after learning (equations (99) and (100)). This clearly shows the influence of the teacher in the learning problem.

5. Outlook

We discuss whether and how the effects of correlation presented in this paper can be taken into account by possible modifications to the learning algorithms.

For the MINOVER algorithm, we have seen in section 3 that the effect of correlation is marginal. Therefore, there is no need to alter the learning procedure.

For the ADALINE algorithm, we have found the learning times to be proportional to N for biased patterns, regardless of whether the patterns are classified according to a given output, or whether the classification is done by a rule. The investigation of fixed cross correlation had already indicated that this behaviour is due to the appearance of a large eigenvalue of order $\mathcal{O}(N)$ in the spectrum of the correlation matrix. This effect has been observed in the calculations for a general choice of patterns as well.

Thus the aim of possible modifications to the ADALINE algorithm must be to remove this large eigenvalue. We expect that by constructing weights (in the spirit of Amit *et al* [1]) from patterns which are shifted by their input bias, the resulting modification of equation (45) would yield this removal and lead to reduced learning times. One then has to investigate whether a different representation of the patterns, as indicated in section 3, can enhance their stability. Furthermore, note that the storage capacity realized by ADALINE-type algorithms is always $\alpha_c = 1$ since a linear system of equations (42) has to be solved, therefore no increase in the storage capacity is possible by a modification of the algorithm or the pattern representation.

Clearly, such investigations are beyond the scope of this paper. We hope that the results obtained and the ideas expressed will, in the future, generate suitably constructed modifications to gradient descent algorithms like ADALINE, which are adapted to sets of correlated patterns.

Acknowledgments

We thank Michael Biehl for the simulation data used in figures 2 and 5, and Timothy Watkin for useful comments on the manuscript. One of us (AW) would like to acknowledge support by the Science and Engineering Research Council of Great Britain and the Friedrich–Naumann–Stiftung.

References

[1] Amit D J, Gutfreund H and Sompolinsky H 1987 *Phys. Rev.* A **35** 2293
[2] Amit D J 1989 *Modeling Brain Function* (Cambridge: Cambridge University Press)
[3] Anlauf J K and Biehl M 1989 *Europhys. Lett.* **10** 687
[4] Buhmann J, Divko R and Schulten K *Phys. Rev.* A **39** 2689
[5] Edwards S F and Jones R C 1976 *J. Phys. A: Math. Gen.* **9** 1595

[6] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[7] van Hemmen J L and Kühn R 1991 *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)
[8] Kinzel W and Opper M 1991 *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer)
[9] Krauth W and Mezard M *J. Phys. A: Math. Gen.* **20** L745
[10] Krogh A and Hertz J A 1991 *Advances in Neural Information Processing Systems* vol III (San Mateo, CA: Morgan Kaufmann)
[11] Opper M 1988 *Phys. Rev.* A **38** 3824; 1989 *Europhys. Lett.* **8** 389
[12] Opper M, Diederich S and Anlauf J 1988 *Neural Networks from Models to Applications* ed L Personnaz and J Dreyfus (Paris: IDSET)
[13] Rujan P 1991 *Preprint* University of Oldenburg
[14] Tsodyks M V and Feigel'man M V 1988 *Europhys. Lett.* **6** 101
[15] Widrow B and Hoff M E 1960 *IRE WESCON Convention Report* **4** 4–96