IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Learning dynamics: a replica approach

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1992 J. Phys. A: Math. Gen. 25 4359

(http://iopscience.iop.org/0305-4470/25/16/013)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.58
The article was downloaded on 01/06/2010 at 16:54

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. 25 (1992) 4359–4372. Printed in the UK

Learning dynamics: a replica approach

F Pázmándi

Institute of Theoretical Physics, University of Lausanne, CH-1015 Lausanne, Switzerland

Received 2 October 1991, in final form 5 May 1992

Abstract. A continuous-time approximation for the learning process in single-layer perceptrons is presented. Both statics and dynamics are treated by the replica method within the framework of the replica symmetric theory. The Adaline rule extended by a decay factor is investigated in detail, especially for linearly-separable problems. It is shown that the poor generalization ability of this rule near $\alpha = 1$, the so called overfitting, can be cured by an appropriate decay factor or by an appropriate training time.

1. Introduction

In the last few years increasing attention has been paid to the dynamical description of the learning process in perceptrons [1–6], for a general review, see [7, 8]. A perceptron [9, 10] is a simple tool of N input units ($\xi_i = \pm 1$, $i = 1, \ldots, N$) connected to the single output unit $s = \pm 1$ via connections w_i ($i = 1, \ldots, N$), where the calculation rule is

$$s = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i - \delta\right). \tag{1}$$

The δ parameter, which is called the threshold value, will be supposed to be zero in this paper. The problem of learning in such a perceptron is that given p input patterns $\{\xi_i^\nu\}_{i=1,\ldots,N}$ with the corresponding outputs ζ^ν ($\nu = 1, \ldots, p$) we have to find the connections w_i which give 'good answers', i.e. the output $s^\nu = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_i w_i \xi_i^\nu\right)$ will be equal to the desired output ζ^ν for all ν

$$\zeta^\nu = \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} w_i \xi_i^\nu\right) \qquad \nu = 1, \ldots, p. \tag{2}$$

These equations are equivalent to the requirement that all the stabilities

$$\Delta_\nu = \frac{1}{\sqrt{N}} \zeta^\nu \sum_i w_i \xi_i^\nu \tag{3}$$

should be positive.

0305-4470/92/164359+14$04.50 © 1992 IOP Publishing Ltd

A possible solution to the learning problem is the so called Δ-rule [11]. One introduces a cost (or energy) function $E = E(\Delta_\nu)$, which measures the learning error in terms of the stabilities (equation (3)), e.g.

$$E_n = \sum_{\nu=1}^{p} (1 - \Delta_\nu)^n \Theta(1 - \Delta_\nu) \tag{4}$$

where $\Theta(x)$ is the Heaviside function, and n can be $1, 2, \ldots$. Possible good solutions can be found by minimizing E_n with respect to w_i $(i = 1, \ldots, N)$. A gradient descent search, wherein each learning step w_i is updated in proportion to $\delta w_i = -\partial E / \partial w_i$, can find some minima. A continuous-time approximation of this rule will be investigated in this paper, i.e.

$$\dot{w}_i = -\frac{\partial E}{\partial w_i} \qquad i = 1, \ldots, N. \tag{5}$$

Hertz *et al* [2] realized that if the right-hand side of equation (5) is linear in $w_j - s$, then finding the static solution $\dot{w}_i = 0$ is equivalent to solving the whole dynamical problem: both of these require determining the minima of the cost function of the same form. To see this, let us suppose an energy function of the form

$$E = \frac{1}{2} \sum_{ij} A_{ij} w_i w_j + \lambda_1 \sum_i a_i w_i + \frac{1}{2} \lambda_2 \sum_i w_i^2 \tag{6}$$

and make a Laplace transform of the dynamical equation (5) $(\mathcal{L}[f(t)] = f(s))$

$$s w_i(s) - w_i(t = 0) = -\sum_j A_{ij} w_j(s) - \frac{\lambda_1}{s} a_i - \lambda_2 w_i(s). \tag{7}$$

For simplicity we take a 'tabula rasa' initial state, i.e. $w_i(t = 0) = 0 \; \forall i$, and if we move the $s w_i(s)$ term to the right-hand side, we obtain an equation of the form $\partial E' / \partial w_i = 0$, where E' and E (equation (6)) differ only in the value of the parameters λ_1 and λ_2

$$\lambda_1 \leftarrow \lambda_1 / s \qquad \lambda_2 \leftarrow \lambda_2 + s. \tag{8}$$

Unfortunately the cost functions of the form of equation (4) have a very strong nonlinearity in the $\Theta(x)$ function, so we need a smoother one to take advantage of the transformation (8). The Adaline rule [12] is defined by the cost function

$$E = \frac{1}{2} \sum_\nu (1 - \Delta_\nu)^2 \tag{9}$$

and forces all the stabilities to be equal to 1. This rule has the desired form of equation (6) with $\lambda_1 = 1$, $\lambda_2 = 0$ and

$$A_{ij} = \frac{1}{N} \sum_\nu \xi_i^\nu \xi_j^\nu \qquad a_i = \frac{1}{\sqrt{N}} \sum_\nu \xi_i^\nu \zeta^\nu. \tag{10}$$

Non-zero λ_2 introduces a decay which penetrates the solutions with very long connection vectors ($w^2 = \sum_i w_i^2 \gg 1$). Since this decay might be useful, for other purposes, we keep the cost function in the form of equation (6) with A_{ij} and a_j given by equation (10).

2. Random Boolean functions

From now on we will take the thermodynamic limit, when both the number of input units (N) and the number of input patterns (p) go to infinity, but their ratio, $\alpha = p/N$, remains finite. We choose the input patterns randomly, i.e. $\xi_i^\nu = \pm 1$ with equal probability, and the outputs $\zeta^\nu = \pm 1$ will also be random for random-Boolean functions. We are interested in quantities which characterize the learning, such as

$$M = \left\langle\!\!\left\langle \frac{1}{p}\sum_\nu \Delta_\nu \right\rangle\!\!\right\rangle \tag{11}$$

$$q = \left\langle\!\!\left\langle \frac{1}{N}\sum_i w_i^2 \right\rangle\!\!\right\rangle \tag{12}$$

where $\langle\!\langle \cdots \rangle\!\rangle$ means an average over the random input and output. One can calculate these quantities with the help of the averaged free energy using the replica method

$$f = \lim_{N\to\infty} \left\langle\!\!\left\langle \frac{-1}{\beta N} \ln \mathrm{Tr}_w \exp -\beta E(\{w_i\}) \right\rangle\!\!\right\rangle \tag{13}$$

where $E(\{w_i\})$ is given by equation (6); in the limit $\beta \to \infty$, f describes the minima of E.

Since there is no constraint on w_i, Tr_w means an integration from $-\infty$ to $+\infty$ for all w_i. The method of calculation follows that of Amit *et al* [13], and gives

$$f = \lim_{n\to 0}\left\{ \frac{1}{n}\sum_{\alpha\leqslant\beta}^{n} r_{\alpha\beta}q_{\alpha\beta} + \frac{\alpha}{2}\left(-\lambda_1^2 + \frac{1}{\beta n}\mathrm{Tr}\ln \underline{B} + \frac{\lambda_1^2}{n}\sum_{\alpha,\beta}^{n} B^{-1}_{\alpha\beta}\right) + \Phi_n \right\} \tag{14}$$

where $\underline{B}$ is a $n \times n$ matrix, $B_{\alpha\beta} = \delta_{\alpha\beta} + \beta q_{\alpha\beta}$, and

$$\Phi_n = \frac{-1}{\beta n}\ln \mathrm{Tr}_{w^\alpha} \exp -\beta\left(\frac{\lambda_2}{2}\sum_{\alpha=1}^{n} w^{\alpha^2} - \sum_{\alpha\leqslant\beta}^{n} r_{\alpha\beta}w^\alpha w^\beta\right). \tag{15}$$

Using the replica symmetry ansatz, i.e.

$$\begin{aligned} q_{\alpha\alpha} &= q_0 \qquad & r_{\alpha\alpha} &= r_0 \\ q_{\alpha\beta} &= q_1 \qquad & r_{\alpha\beta} &= r_1 \qquad (\alpha \neq \beta) \end{aligned} \tag{16}$$

and introducing new variables

$$\begin{aligned} \varphi &= \beta(q_0 - q_1) \qquad & \rho &= r_1 - 2r_0 \\ q &= q_1 & R &= r_1/\beta \end{aligned} \tag{17}$$

one can take the limit $n \to 0$, giving

$$f = \frac{1}{2}R\varphi - \frac{\rho}{2}\left(q + \frac{\varphi}{\beta}\right) + \lambda_1^2\frac{\alpha}{2}\left(\frac{1}{1+\varphi} - 1\right) + \frac{\alpha}{2\beta}\left[\ln(1+\varphi) + \frac{\beta q}{1+\varphi}\right] + \Phi(\rho, R) \tag{18}$$

$$\Phi(\rho, R) = \Phi_0(\beta) + \frac{1}{2\beta}\ln(\lambda_2 + \rho) - \frac{1}{2}\frac{R}{\lambda_2 + \rho}.$$

The order parameters φ and q are

$$\varphi = \beta \left\langle\!\!\left\langle \frac{1}{N} \sum_i \left(\langle w_i^2 \rangle - \langle w_i \rangle^2 \right) \right\rangle\!\!\right\rangle$$
$$q = \left\langle\!\!\left\langle \frac{1}{N} \sum_i \langle w_i \rangle^2 \right\rangle\!\!\right\rangle \tag{19}$$

where $\langle \cdots \rangle$ means the temperature average. The parameters R and ρ are weight functions carrying the way of averaging. Since we need the limit as $\beta \to \infty$, we are interested in the minima of the energy, in this limit q gives the length of the connection vector. The other quantity we wanted to know was M, equation (11), which can be obtained by taking the derivative of f with respect to λ_1

$$M = -\frac{1}{\alpha} \frac{\partial f}{\partial \lambda_1} = \frac{\varphi}{1+\varphi} \lambda_1. \tag{20}$$

The values of φ, q, R and ρ are given by the saddle-point equations

$$\varphi = \frac{1}{\lambda_2 + \rho} \qquad \rho = \frac{\alpha}{1+\varphi} \tag{21}$$

$$q = \frac{R}{(\lambda_2 + \rho)^2} \qquad R = \frac{\alpha}{(1+\varphi)^2} (\lambda_1^2 + q). \tag{22}$$

The first two equations (21) are closed, giving

$$\varphi^{-1} = \lambda_2 + \frac{\alpha}{1+\varphi}. \tag{23}$$

This equation for φ is the same as that of Hertz *et al* [2] for their Green's function. Using this, q and M can be expressed by $u = \rho\varphi$

$$M = \frac{u}{\alpha} \lambda_1 \qquad q = \frac{u^2}{\alpha - u^2} \lambda_1^2 \tag{24}$$

where

$$u = \tfrac{1}{2} \left[1 + \alpha + \lambda_2 - \sqrt{(1 + \alpha + \lambda_2)^2 - 4\alpha} \right]. \tag{25}$$

Let us start with the static problem. We can recover the original Adaline rule by the substitution $\lambda_1 = 1$ and $\lambda_2 = 0$. The results are shown in figure 1. The singularity of q at $\alpha = 1$ is connected to the capacity of the Adaline rule; up to this value of α, all the stabilities are set to be 1 by the end of the learning process.

The time needed to reach this state can be characterized by the average relaxation time of M

$$\tau_m = \lim_{s \to 0} \frac{\int_0^\infty t [M(t) - M_\infty] e^{-st} \, dt}{\int_0^\infty [M(t) - M_\infty] e^{-st} \, dt} = \lim_{s \to 0} \frac{s \partial M(s)/\partial s + M_\infty / s}{M_\infty - s M(s)} \tag{26}$$

where $M_\infty = M(t \to \infty)$ is the static value of M. $M(s)$ is given by the same equation (24), but we have to use $\lambda_1 = 1/s$ and $\lambda_2 = s$. After a little algebra, one obtains

$$\tau_m = \frac{1}{(1-\alpha)^2} \frac{2\alpha}{1 + \alpha - |1 - \alpha|} = \begin{cases} 1/(1-\alpha)^2 & \alpha < 1 \\ \alpha/(\alpha - 1)^2 & \alpha > 1. \end{cases} \tag{27}$$

Up to now we have re-derived some results of Hertz *et al* [2] and Opper [1]. The approach introduced above will be generalized to the calculation of other quantities in the next sections.

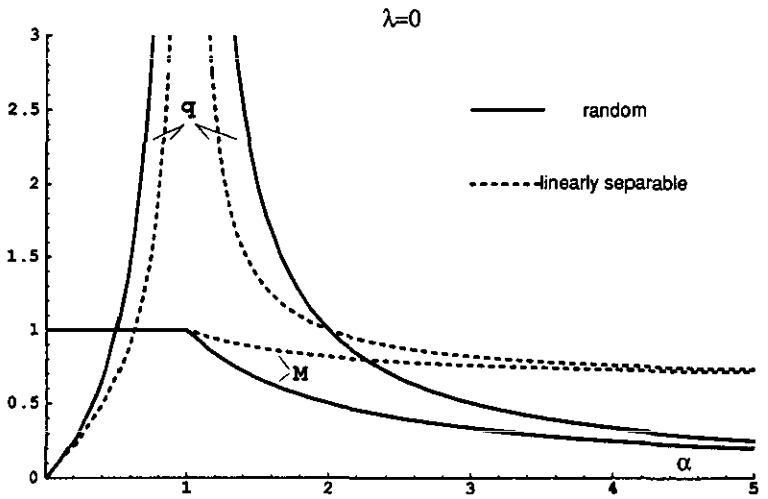

Figure 1. The length of the weight vector (q) and the average magnetisation (M) as a function of the size of the learning set ($\alpha = p/N$) for Adaline learning ($\lambda = 0$). The solid lines refer to random-Boolean functions and the dashed lines refer to linearly-separable functions.

3. The correlation function

As an example, let us calculate the following correlation function

$$C(t,t') = \left\langle\!\!\left\langle \frac{1}{N}\sum_i w_i(t)w_i(t') \right\rangle\!\!\right\rangle = \frac{1}{(2\pi\mathrm{i})^2}\int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}s \int_{-\mathrm{i}\infty}^{\mathrm{i}\infty} \mathrm{d}\hat{s}\, \mathrm{e}^{st+\hat{s}t'} C(s,\hat{s}) \tag{28}$$

where

$$C(s,\hat{s}) = \left\langle\!\!\left\langle \frac{1}{N}\sum_i w_i(s)w_i(\hat{s}) \right\rangle\!\!\right\rangle . \tag{29}$$

In particular, $C(t,t)$ will give us the $q(t)$ function. Let us denote by $C_0(t,t')$ the correlation function for random-Boolean functions. To determine $C_0(s,\hat{s})$ we introduce two systems; $\mathcal{S}$, depending on the parameters λ_1 and λ_2, and $\hat{\mathcal{S}}$, depending on $\hat{\lambda}_1$ and $\hat{\lambda}_2$. The Hamiltonian is

$$\mathcal{H}(\{w_i\},\{\hat{w}_i\};s,\hat{s}) = E(\{w_i\};\lambda_1,\lambda_2) + \hat{E}(\{\hat{w}_i\};\hat{\lambda}_1,\hat{\lambda}_2) + H\sum_i w_i\hat{w}_i \tag{30}$$

where E and $\hat{E}$ have the previous form of equation (6), λ_j and $\hat{\lambda}_j$ ($j = 1,2$) depend on s and $\hat{s}$ respectively, equation (8).

We need the external field H to extract the desired term, equation (29), by a derivation of the free energy and after that we set $H = 0$. On the other hand, when

$H = 0$, S and $\hat{S}$ are independent their minima are given by the previous equations (21) and (22).

The method of calculation of the free energy with $\mathcal{H}$ (equation (30)) is as before, but the non-zero H introduces an interaction between those replicas of S and $\hat{S}$ whose replica indices are identical, so we have

$$f = \left\langle\!\!\left\langle \frac{-1}{\beta N} \ln \mathrm{Tr}\, \mathrm{e}^{-\beta \mathcal{H}} \right\rangle\!\!\right\rangle$$

$$= f_0(\beta) + \lim_{n \to 0} \left\{ \frac{1}{n} \sum_{\alpha \leqslant \beta}^{n} (r_{\alpha\beta} q_{\alpha\beta} + \hat{r}_{\alpha\beta} \hat{q}_{\alpha\beta}) + \frac{1}{n} \sum_{\alpha\beta}^{n} a_{\alpha\beta} c_{\alpha\beta} \right.$$

$$\left. - \frac{\alpha}{2n} \underline{\lambda}_1^2 + \frac{\alpha}{2n} \underline{\lambda}_1 \underline{\underline{B}}^{-1} \underline{\lambda}_1 + \frac{\alpha}{2\beta n} \mathrm{Tr} \ln \underline{\underline{B}} + \frac{1}{2\beta n} \mathrm{Tr} \ln \underline{\underline{F}} \right\} \tag{31}$$

where $\underline{\underline{B}}$ and $\underline{\underline{F}}$ are $2n \times 2n$ matrices, and $\underline{\lambda}_1$ is a $2n$ vector

$$\underline{\underline{B}} = \left(\begin{array}{c|c} \underline{\underline{I}} + \beta \underline{\underline{q}} & \beta \underline{\underline{c}} \\ \hline \beta \underline{\underline{c}} & \underline{\underline{I}} + \beta \underline{\underline{q}} \end{array} \right)$$

$$\underline{\underline{F}} = \left(\begin{array}{cc|cc} \ddots & & -r_{\alpha\beta} & \\ & \lambda_2 - 2r_{\alpha\alpha} & & H\underline{\underline{I}} - \underline{\underline{a}} \\ -r_{\alpha\beta} & & \ddots & \\ \hline & & \ddots & -r_{\alpha\beta} \\ & H\underline{\underline{I}} - \underline{\underline{a}} & & \lambda_2 - 2r_{\alpha\alpha} \\ & & -r_{\alpha\beta} & \ddots \end{array} \right) \tag{32}$$

$$\underline{\lambda}_1 = (\lambda_1, \ldots, \lambda_1 \,|\, \hat{\lambda}_1, \ldots, \hat{\lambda}_1)$$

and $\underline{\underline{I}}$ is a $n \times n$ unit matrix.

The saddle points are taken with respect to $r_{\alpha\beta}$, $q_{\alpha\beta}$, $\hat{r}_{\alpha\beta}$, $\hat{q}_{\alpha\beta}$, $c_{\alpha\beta}$ and $a_{\alpha\beta}$. The replica symmetric ansatz is the same as before, equation (16), with the additional relations

$$\begin{aligned} c_{\alpha\alpha} &= c_0 \qquad a_{\alpha\alpha} = a_0 \\ c_{\alpha\beta} &= c_1 \qquad a_{\alpha\beta} = a_1 \qquad (\alpha \neq \beta). \end{aligned} \tag{33}$$

Since c_0 and c_1 characterize the interaction between the replicas of the two systems (S and $\hat{S}$) and in the $H = 0$ case we can suppose this interaction to be independent of the choice of replica indices, the difference $c_0 - c_1$ should vanish as $H \to 0$. The same is true for $a_0 - a_1$. Bearing in mind these relations, after determining the eigenvalues of $\underline{\underline{B}}$ and $\underline{\underline{F}}$, we arrive at the value of $C_0(s, \hat{s})$ (details are left to the reader)

$$C_0(s, \hat{s}) = \frac{u\hat{u}}{\alpha - u\hat{u}} \lambda_1 \hat{\lambda}_1 \tag{34}$$

where u depends on s and $\hat{u}$ depends on $\hat{s}$ but both functions are given by the equation (25).

To carry out the inverse Laplace transformation we introduce the notation

$$g = u/\sqrt{\alpha} \qquad \hat{g} = \hat{u}/\sqrt{\alpha} \tag{35}$$

and rewrite $C_0(s,\hat{s})$ as

$$\begin{aligned} C_0(s,\hat{s}) &= \frac{g\hat{g}}{1-g\hat{g}}\lambda_1\hat{\lambda}_1 \\ &= \lim_{\gamma\to 1}\frac{g\hat{g}}{1-\gamma g\hat{g}}\lambda_1\hat{\lambda}_1 \\ &= \lim_{\gamma\to 1}\sum_{n=1}^{\infty}\gamma^{n-1}g^n\lambda_1\hat{g}^n\hat{\lambda}_1. \end{aligned} \tag{36}$$

Now the inverse Laplace transforms with respect to s and $\hat{s}$ can be performed separately. With the notation

$$G_n(t) = \int_0^t \mathcal{L}^{-1}[g^n(s)](\tau)\,\mathrm{d}\tau \tag{37}$$

we have

$$C_0(t,t') = \lim_{\gamma\to 1}\frac{1}{\gamma}\sum_{n=1}^{\infty}\gamma^n G_n(t)G_n(t'). \tag{38}$$

The functions $G_n(t)$ can be calculated

$$G_n(t) = \frac{a}{\pi}\int_0^{\pi}\mathrm{d}\vartheta\,\frac{1-\mathrm{e}^{-(b-a\cos\vartheta)t}}{b-a\cos\vartheta}\sin\vartheta\,\sin(n\vartheta) \tag{39}$$

where $a = 2\sqrt{\alpha}$ and $b = 1+\alpha+\lambda$, where λ is the 'static' value of λ_2 ($\lambda_2 = \lambda + s$). First calculating the sum $\sum_{n=1}^{\infty}\gamma^n\sin(n\vartheta)\sin(n\vartheta')$ in equation (38) and then taking the limit $\gamma \to 1$ gives a Dirac's delta function times $\frac{\pi}{2}$, so finally we have

$$C_0(t,t') = \frac{a^2}{2\pi}\int_0^{\pi}\mathrm{d}\vartheta\,\sin^2\vartheta\,\frac{1-\mathrm{e}^{-(b-a\cos\vartheta)t}}{b-a\cos\vartheta}\,\frac{1-\mathrm{e}^{-(b-a\cos\vartheta)t'}}{b-a\cos\vartheta}. \tag{40}$$

For the Adaline case, figure 2 shows the $q(t) = C_0(t,t)$ function and the $M(t)$ function for $\alpha = 0.9$, where $M(t)$ is simply

$$M(t) = \frac{2}{a}G_1(t) \tag{41}$$

see equations (24,35,37).

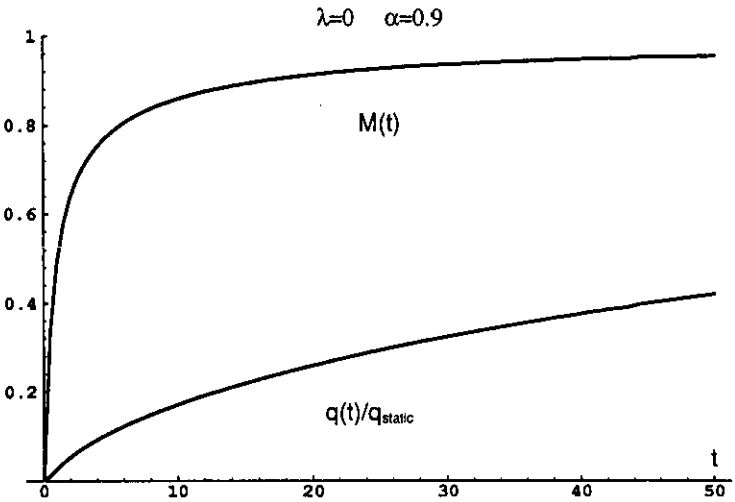

Figure 2. The time evolution of M, equation (41), and q, equation (40) at $t = t'$, for random-Boolean functions at $\alpha = 0.9$ in the Adaline-learning case. The static value of q, i.e. the limit $t \to \infty$, can be obtained from equation (24). Since $M(t \to \infty) = 1$ for $\alpha = 0.9$, see figure 1, we do not have to normalize the $M(t)$ function.

4. Linearly-separable functions

A linearly-separable function is defined through a reference (or teacher) perceptron with a weight vector $\{\mathcal{F}_i\}_{i=1,\ldots,n}$: to a given input $\{\xi_i\}$, the correct answer is the sign of $\sum_i \mathcal{F}_i \xi_i$. To fix the length of the teacher perceptron, we suppose that $N^{-1}\sum_j \mathcal{F}_{ij}^2 = 1$.

A natural quantity characterizing the learning performance can be the cosine of the angle between the weight vectors of the 'teacher' ($\underline{\mathcal{F}}$) and the 'student' ($\underline{w}$)

$$F = \left\langle\!\!\left\langle \frac{\frac{1}{N}\sum_i \mathcal{F}_i w_i}{\sqrt{\frac{1}{N}\sum_i w_i^2}} \right\rangle\!\!\right\rangle_\xi = \frac{\left\langle\!\left\langle \frac{1}{N}\sum_i \mathcal{F}_i w_i \right\rangle\!\right\rangle_\xi}{\sqrt{\left\langle\!\left\langle \frac{1}{N}\sum_i w_i^2 \right\rangle\!\right\rangle_\xi}} \tag{42}$$

where we have used the fact that $q = \frac{1}{N}\sum_i w_i^2$ is a self-averaging quantity. The so called generalization ability (P_g), which gives the probability of the correct answer from the student perceptron to a new randomly-chosen input, can be expressed by F see [14]

$$P_g = 1 - (1/\pi)\cos^{-1} F. \tag{43}$$

To describe the learning process in this case, we use the same method as before, taking the cost function of equation (6), we calculate the averaged-free energy with the help of the replica-symmetric theory. The only difference is in the way of averaging.

Since the output ζ^ν is not an independent-random variable, but is given by the relation

$$\zeta^\nu = \operatorname{sgn}\left(\frac{1}{\sqrt{N}}\sum_{j=1}^{N} \mathcal{F}_j \xi_j^\nu\right) \qquad \nu = 1, \ldots, p \tag{44}$$

we have to be more careful when we average terms like

$$\left\langle\!\!\left\langle \exp \frac{1}{\sqrt{N}} \sum_i b_{i\nu} \xi_i^\nu \zeta^\nu \right\rangle\!\!\right\rangle_{\{\xi_i^\nu\}_{i=1\ldots N}} \tag{45}$$

where $b_{i\nu}$ is a quantity of order 1. The averaging in equation (45) can be performed if N goes to infinity and $\mathcal{F}_j \propto \mathcal{O}(1)\ \forall j$. In this limit both the quantities, $z_1 = \frac{1}{\sqrt{N}} \sum_i b_{i\nu} \xi_i^\nu$ in equation (45) and $z_2 = \frac{1}{\sqrt{N}} \sum_j \mathcal{F}_j \xi_j^\nu$ in equation (44), are Gaussian variables, although they are correlated as

$$\langle\!\langle z_1 z_2 \rangle\!\rangle_\xi = \frac{1}{N} \sum_j \mathcal{F}_j b_j. \tag{46}$$

In this way, the averaging in equation (45) according to $\xi - s$ can be replaced by the averaging over z_1 and z_2, giving

$$\langle e^{z_1 \operatorname{sgn} z_2} \rangle_{z_1 z_2} = \exp\left(\frac{\sum_j b_j^2}{2N}\right)\left[1 + \operatorname{erf}\left(\frac{1}{\sqrt{2}N} \sum_j \mathcal{F}_j b_j\right)\right]. \tag{47}$$

From now on, the calculation continues in the standard way and we will have the same order parameters as before, equation (19), plus a new one coming from the argument of the $\operatorname{erf}(x)$ function in equation (47)

$$y = \left\langle\!\!\left\langle \frac{1}{N} \sum_i \mathcal{F}_i \langle w_i \rangle \right\rangle\!\!\right\rangle \tag{48}$$

with its weight function s. We can see that in the zero temperature limit y is just the numerator of equation (42).

The free energy is

$$\begin{aligned} f &= \frac{1}{2} R\varphi - \frac{\rho}{2}\left(q + \frac{\varphi}{\beta}\right) + ys + \Phi(\rho, R, s) + \Psi(\varphi, q, y) - \frac{\alpha}{2}\lambda_1^2 + f_0(\beta) \\ \Phi &= \frac{\ln(\lambda_2 + \rho)}{2\beta} - \frac{s^2 + R}{2(\lambda_2 + \rho)} \\ \Psi &= \frac{\alpha}{2\beta} \ln(1 + \varphi) + \frac{\alpha}{2(1 + \varphi)}\left(\lambda_1^2 + q - 2\sqrt{\frac{2}{\pi}} y\lambda_1\right). \end{aligned} \tag{49}$$

In the $\beta \to \infty$ limit, the saddle-point equations are

$$\begin{aligned} \varphi &= \frac{1}{\lambda_2 + \rho} & \rho &= \frac{\alpha}{1 + \varphi} \\ y &= \frac{1}{\lambda_2 + \rho} s & s &= \sqrt{\frac{2}{\pi}} \frac{\alpha}{1 + \varphi} \lambda_1 \\ q &= \frac{R + s^2}{(\lambda_2 + \rho)^2} & R &= \frac{\alpha}{(1 + \varphi)^2}\left(\lambda_1^2 + q - 2\sqrt{\frac{2}{\pi}} y\lambda_1\right) \end{aligned} \tag{50}$$

where the first two equations are the same as in the random case, equation (21), giving the same expression for the $u(\lambda_2)$ function, equation (25).

The other quantities of interest, such as y, q and M, can be expressed by this $u(\lambda_2)$ function

$$y = \sqrt{\frac{2}{\pi}}\, u\lambda_1 \tag{51}$$

$$q = \frac{u^2}{\alpha - u^2}\left(1 + \frac{2\alpha}{\pi} - \frac{4}{\pi}u\right)\lambda_1^2 \tag{52}$$

$$M = \frac{u}{\alpha}\left[1 + \frac{2}{\pi}(\alpha - u)\right]\lambda_1. \tag{53}$$

For the static Adaline problem, i.e. $\lambda_1 = 1$ and $\lambda_2 = 0$, the length of the weight vector w, i.e. q, equation (52), and the average stability, i.e. M, equation (53), behave qualitatively as in the random case, equation (24), for $\alpha \propto O(1)$; only a small correction is due to the linear separability of the input patterns. For $\alpha \gg 1$, this property of inputs becomes important, and both q and M tend to finite values as $\alpha \to \infty$, rather than of tending to zero, showing that the network was able to learn something, figure 1. The average relaxation time, defined by means of M, equation (26), is similar to the case of random functions, equation (27). The small correction disappears for $\alpha \to \infty$

$$\tau_m = \begin{cases} \dfrac{1}{(1-\alpha)^2}\left(1 - \dfrac{2\alpha}{\pi}\,\dfrac{1-\alpha}{1-2\alpha/\pi}\right) & \alpha < 1 \\ \dfrac{\alpha}{(1-\alpha)^2}\left(1 - \dfrac{2}{\pi\alpha}\,\dfrac{\alpha-1}{1+(2/\pi)(\alpha-2)}\right) & \alpha > 1. \end{cases} \tag{54}$$

5. The generalization ability

Although the above quantities, equations (51–54), characterize the learning performance, the most important quantity is the generalization ability of the network, equation (43), which will be measured by F, equation (42), in the rest of the paper.

In the static case we can use equations (51) and (52) directly to obtain the $F(\lambda_2, \alpha)$ function

$$F = \sqrt{\frac{\alpha - u^2}{\frac{1}{2}\pi + \alpha - 2u}} \tag{55}$$

where u is the previous expression of equation (25). Since F depends on λ_2, it may be worth asking whether F is maximal at a given value of λ_2. Solving the equation $\partial F/\partial\lambda_2 = 0$ gives

$$\lambda_2^{\text{opt}} = \tfrac{1}{2}\pi - 1 \tag{56}$$

for any value of $\alpha = p/N$, and at this point

$$F_{\text{opt}} = \sqrt{\tfrac{1}{2}\left[\tfrac{1}{2}\pi + \alpha - \sqrt{(\tfrac{1}{2}\pi + \alpha)^2 - 4\alpha}\right]}. \tag{57}$$

This generalization ability should be compared to the Adaline case [14], where $\lambda_2 = 0$, giving

$$F_{\mathrm{Ad}} = \begin{cases} \sqrt{\dfrac{\alpha(1-\alpha)}{\frac{1}{2}\pi - \alpha}} & \alpha < 1 \\ \sqrt{\dfrac{\alpha - 1}{\frac{1}{2}\pi + \alpha - 2}} & \alpha > 1 \end{cases} \tag{58}$$

and to the Hebb case [14, 15], which can be recovered in the limit $\lambda_2 \to \infty$ where the learning process stops in a very short time, making w_i proportional to a_i, equation (10), which is nothing but the Hebb rule

$$F_{\mathrm{Hebb}} = \sqrt{\frac{\alpha}{\frac{1}{2}\pi + \alpha}} \tag{59}$$

(see figure 3).

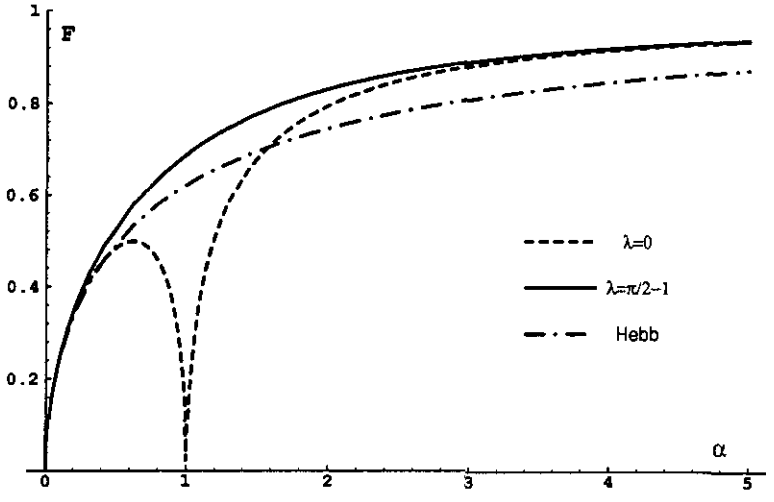

Figure 3. The cosine of the angle between the weight vectors of the teacher perceptron ($\underline{\mathcal{T}}$) and of the student perceptron ($\underline{w}$) as a function of $\alpha = p/n$ for linearly-separable functions equation (42). The solid line denotes the network with optimal weight decay. The dashed line and the dash-dotted line refer to the Adaline learning ($\lambda = 0$) and to the Hebb rule ($\lambda \to \infty$) respectively.

In the limit $\alpha \gg 1$, both F_{opt} and F_{Ad} behave like $\sqrt{1-(\pi/2\alpha)(1-2/\pi)}$, so using λ_2^{opt} instead of $\lambda_2 = 0$ makes no difference in this limit. The region where non-zero λ_2 can be useful may be estimated by the region where the inequality $F_{\mathrm{Ad}} < F_{\mathrm{Hebb}}$ holds, i.e. when the Adaline rule is even worse than the Hebb rule. This happens if

$$1 - \tfrac{1}{2}\pi < (\alpha - 1) < \tfrac{1}{2}\pi - 1. \tag{60}$$

To describe the time evolution of F, we need the functions $y(t)$ and $q(t)$. The former can be obtained by the inverse Laplace transformation of equation (51) with

the substitution of equation (8), giving $y(t) = \sqrt{2\alpha/\pi}\, G_1(t)$. The latter requires the calculation of the correlation function of equation (28). The Laplace transform of $C(t,t')$ can be calculated in the same way as in the random case, giving

$$C(s,\hat{s}) = \frac{u\hat{u}}{\alpha - u\hat{u}}\left[1 + \frac{2}{\pi}(\alpha - u - \hat{u})\right]\lambda_1\hat{\lambda}_1 \tag{61}$$

where we have used the notation of the previous section. The inverse Laplace transformation is made with the help of the same trick as before equation (36), so we have

$$C(t,t') = \left(1 + \frac{a^2}{2\pi}\right)C_0(t,t') - \frac{2a}{\pi}C_1(t,t') \tag{62}$$

where $C_0(t,t')$ is given by equation (40) and

$$\begin{aligned} C_1(t,t') = {} & \frac{a^2}{2\pi}\int_0^\pi \mathrm{d}\vartheta\, \sin^2\vartheta\, \cos\vartheta \frac{\left(1 - \mathrm{e}^{-(b - a\cos\vartheta)t}\right)\left(1 - \mathrm{e}^{-(b-a\cos\vartheta)t'}\right)}{(b - a\cos\vartheta)^2} \\ & + \frac{a^2}{2\pi^2}\int_0^\pi \frac{\mathrm{d}\vartheta\mathrm{d}\vartheta'}{\cos\vartheta' - \cos\vartheta}\sin^2\vartheta\, \sin^2\vartheta' \\ & \times \frac{1 - \mathrm{e}^{-(b-a\cos\vartheta)t}}{(b - a\cos\vartheta)}\frac{1 - \mathrm{e}^{-(b-a\cos\vartheta')t'}}{(b - a\cos\vartheta')} \end{aligned} \tag{63}$$

where $a = 2\sqrt{\alpha}$, $b = 1 + \alpha + \lambda$ as before, and $\int$ means principal value integration. Fortunately, if $t = t'$, the second term in $C_1(t,t')$ disappears (because it changes sign for $\vartheta \leftrightarrow \vartheta'$), so $q(t) = C(t,t)$ is given by a one-dimensional integral. The $F(t) = y(t)/\sqrt{q(t)}$ function for $\alpha = 0.9$ and for various values of λ_2 is shown in figure 4.

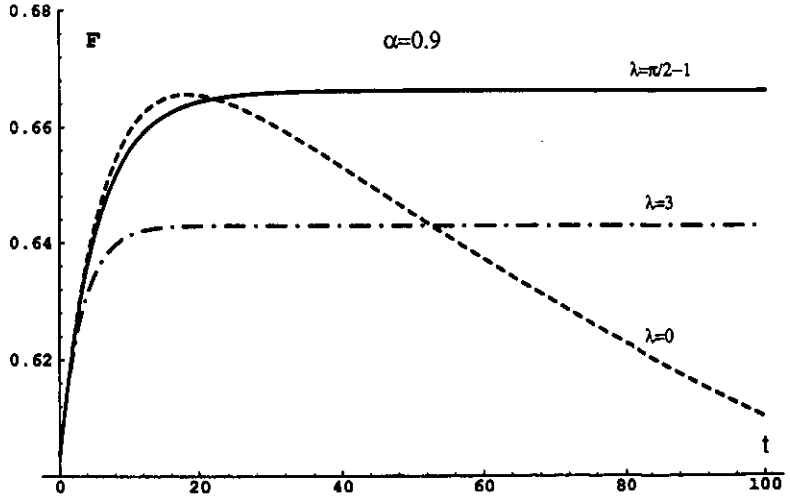

Figure 4. The generalization ability, measured by F, equation (42), as a function of the training time for $\alpha = 0.9$. For the Adaline learning ($\lambda = 0$, dashed line), there is an overfitting effect in time: at $t \approx 18$ the generalization ability is at a maximum. The optimal decay ensures a smooth approach to the optimal generalization ($\lambda = \frac{\pi}{2} - 1$, solid line); higher decays give worse generalization (e.g. $\lambda = 3$, dash-dotted line).

In the limit $t \to 0$, we get back the generalization ability of the Hebb rule (equation (59)), as we should, and for $t \to \infty$, we arrive at the static value of F (equation (55)). If the latter is the smaller, we can expect an overfitting phenomenon in time. The generalization ability first increases and later decreases to its static value. For the Adaline learning ($\lambda_2 = 0$), this 'overfitting region' of α is given by equation (60). Unfortunately, there is no simple analytical formula for the optimal training time, so this quantity can only be obtained numerically.

6. Conclusion and possible extensions

We have shown that the continuous-time approximation of the learning process in single-layer perceptrons can be treated by the replica method. The crucial point of the approximation is that if the dynamical equations (equation 5) are linear in the dynamical variables w_i, we can use the replica method on the Laplace-transformed equations. As a demonstration, we calculated the dynamical properties of the Adaline rule extended by a decay factor for the case of random and linearly-separable-Boolean functions. For the latter, an optimal decay, $\lambda_2^{\text{opt}} = \frac{1}{2}\pi - 1$, was found in the sense that this decay optimizes the generalization ability of the network.

For the original Adaline rule, the poor generalization properties near $\alpha = 1$ are due to the long training time, and an appropriate training time can give as good a generalization as we have for the optimal decay (see figure 4). We have found that these overfitting effects occur mainly in the region $\alpha \in (\alpha_-, \alpha_+)$, where $\alpha_\pm = 1 \pm (\frac{1}{2}\pi - 1)$.

A straightforward extension of the present model could be the case of the 'unreliable teacher'. There are several ways of defining an 'unreliable teacher' [5]; here we take only the simplest case, where instead of equation (44) the outputs in the training set are

$$\zeta^\nu = \epsilon_\nu \operatorname{sgn}\left(\frac{1}{\sqrt{N}} \sum_j \mathcal{F}_j \xi_j^\nu\right) \qquad \nu = 1, \ldots, p \tag{64}$$

where $\epsilon_\nu = \pm 1$ with probability $(1 \pm \epsilon)/2$ and constant during the learning process. Equation (64) means that, for some inputs, the teacher does not know the correct answer but he is faithful to the incorrect ones. In this case, the average free energy differs from that of equation (49) only in the expression of Ψ, instead of the term $-2\sqrt{2/\pi}\, y \lambda_1$ we have ϵ times the same expression. This means that in the equations (51)–(55) we have to replace $\sqrt{2/\pi}$ by $\sqrt{2/\pi}\,\epsilon$ and $2/\pi$ by $2\epsilon^2/\pi$.

The most important changes due to $\epsilon < 1$ are those in $\lambda_2^{\text{opt}}(\epsilon) = \pi/2\epsilon^2 - 1$ and in the width of the 'overfitting region' $\alpha_\pm(\epsilon) = 1 \pm (\pi/2\epsilon^2 - 1)$: we need bigger decay and the overfitting region becomes wider. Other types of noises in the learning process can be treated in a similar fashion.

It is a more difficult matter to apply this method to the cost functions of the type of equation (4). For example, if we choose $E_2 = \sum_\nu (1 - \Delta_\nu)^2 \Theta(1 - \Delta_\nu)$ (the so called Adatron rule), and introduce new dynamical variables $V_\nu := (1 - \Delta_\nu)\Theta(1 - \Delta_\nu)$, the equations of motion will be

$$\frac{\partial V_\rho}{\partial t} = -\Theta(V_\rho) \frac{\partial}{\partial V_\rho} \left[\frac{1}{2} \sum_{\mu\nu} Q_{\mu\nu} V_\mu V_\nu + \frac{1}{2} \lambda_2 \sum_\mu (V_\mu - 1)^2\right] \tag{65}$$

where $Q_{\mu\nu} = N^{-1}\sum_i \xi_i^\nu \xi_i^\mu$ is the overlap matrix of the input patterns. If we are interested only in the static properties, the expression $[\cdots]$ in equation (65) is a good Lyapunov function; the minima of this expression with the constraint $V_\mu > 0$ ($\mu = 1, \ldots, p$) give the stationary solutions of equation (65). This problem can be solved by the replica method. On the other hand, the constraint of $V_\mu(t) > 0$ makes the Laplace transformation useless, we cannot use equation (65) to describe the dynamics.

Acknowledgments

The support of the Swiss National Science Foundation through grant 20-28846.90 is gratefully acknowledged. The author thanks P Erdős and T Geszti for stimulating discussions.

References

[1] Opper M 1989 *Europhys. Lett.* **8** 389
[2] Hertz J A, Krogh A and Thorbergsson G I 1989 *J. Phys. A: Math. Gen.* **22** 2133
[3] Kinzel W and M. Opper *Physics of Neural Networks* ed J L van Hemmen *et al* (Berlin: Springer)
[4] Dev R 1990 *J. Phys. A: Math. Gen.* **23** 763
[5] Krogh A and Hertz J A 1992 *J. Phys. A: Math. Gen.* **25** 1135
[6] Krogh A 1992 *J. Phys. A: Math. Gen.* **25** 1119
[7] Hertz J A, Krogh A and Palmer RG 1991 *Introduction to the Theory of Neural Computation* (New York: Addison-Wesley)
[8] Györgyi G and Tishby N 1990 *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific)
[9] Minsky M and Papert S 1988 *Perceptrons* (Cambridge, MA: MIT Press)
[10] Rosenblatt F 1958 *Psychoanalytic Review* **65** 386
[11] Rumelhart D E, McClelland J L and the PDP research group 1960 *Parallel Distributed Processing* (Cambridge, MA: MIT Press)
[12] Widrow B and Hoff M E 1960 *WESCON Convention, Report IV*
[13] Amit D G, Gulfreund G and Sompolinsky H 1987 *Ann. Phys.* **173** 30
[14] Opper M, Kinzel W, Kleinz J and Nehl R 1990 *J. Phys. A: Math. Gen.* **23** 581
[15] Vallet F 1989 *Europhys. Lett.* **8** 747