J. Phys. A: Math. Gen. 27 (1994) 7771–7784. Printed in the UK

Finite-size effects in learning and generalization in linear perceptrons

Peter Sollich†
Department of Physics, University of Edinburgh, Edinburgh EH9 3JZ, UK

Received 15 August 1994

Abstract. Most properties of learning and generalization in linear perceptrons can be derived from the average response function G. We present a method for calculating G using only simple matrix identities and partial differential equations. Using this method, we first rederive the known result for G in the thermodynamic limit of perceptrons of infinite size N, which has previously been calculated using replica and diagrammatic methods. We also show explicitly that the response function is self-averaging in the thermodynamic limit. Extensions of our method to more general learning scenarios with anisotropic teacher-space priors, input distributions, and weight-decay terms are discussed. Finally, finite-size effects are considered by calculating the $O(1/N)$ correction to G. We verify the result by computer simulations and discuss the consequences for generalization and learning dynamics in linear perceptrons of finite size.

1. Introduction

One of the main areas of research within the field of neural networks is the issue of learning and generalization. Starting from a set of training examples (normally assumed to be input–output pairs) generated by some unknown 'teacher' rule $\mathcal{V}$, one wants to find, using a suitable learning or training algorithm, a student $\mathcal{N}$ (read 'neural network') which generalizes from the training set, i.e. predicts the outputs corresponding to inputs not contained in the training set as accurately as possible. If the inputs are N-dimensional vectors $\boldsymbol{x} \in \mathcal{R}^N$ and the outputs are scalars $y \in \mathcal{R}$, then one of the simplest functional forms which can be assumed for the student $\mathcal{N}$ is the linear perceptron, which is parametrized in terms of a weight vector $\boldsymbol{w}_{\mathcal{N}}$ and implements the linear input–output mapping

$$y_{\mathcal{N}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \boldsymbol{w}_{\mathcal{N}}^T \boldsymbol{x} .$$

A commonly used learning algorithm for the linear perceptron is gradient descent on the training error E_t, i.e. the error that the student $\mathcal{N}$ makes on the training set. Using the standard squared output deviation error measure, the training error for a given set of p training examples, $\{(\boldsymbol{x}^\mu, y^\mu), \mu = 1 \dots p\}$, is

$$E_t = \sum_{\mu=1}^{p} \tfrac{1}{2}\left(y^\mu - y_{\mathcal{N}}(\boldsymbol{x}^\mu)\right)^2 = \tfrac{1}{2} \sum_{\mu=1}^{p} \left(y^\mu - \boldsymbol{w}_{\mathcal{N}}^T \boldsymbol{x}^\mu / \sqrt{N}\right)^2 .$$

To prevent the student from fitting noise in the training data, a quadratic-weight decay term $\frac{1}{2}\lambda \boldsymbol{w}_{\mathcal{N}}^2$ is normally added to the training error, with the value of the weight-decay parameter

† E-mail address: P.Sollich@ed.ac.uk

0305-4470/94/237771+14$19.50

λ determining how strongly large weight vectors are penalized. Thus, gradient descent is performed on the function $E = E_t + \frac{1}{2}\lambda w_N^2$. The corresponding learning dynamics is, in a continuous-time approximation, $dw_N/dt = -\eta \nabla E$, where the gradient is taken with respect to w_N, and η is the learning rate. The effect of η can be absorbed into a rescaling of the learning time t, and we therefore set $\eta = 1$ in the following without loss of generality. Evaluating ∇E explicitly, one obtains the learning dynamics

$$\frac{dw_N}{dt} = -\mathbf{M}_N(w_N - \mathbf{M}_N^{-1}\mathbf{a}) \tag{1.1}$$

where we have defined the input correlation matrix

$$\mathbf{A} = \frac{1}{N}\sum_\mu x^\mu (x^\mu)^T$$

and

$$\mathbf{M}_N = \lambda \mathbf{1} + \mathbf{A} \qquad \mathbf{a} = \frac{1}{\sqrt{N}}\sum_{\mu=1}^{p} y^\mu x^\mu \tag{1.2}$$

with $\mathbf{1}$ denoting the $N \times N$ identity matrix. Equation (1.1) has the unique solution

$$w_N(t) = \mathbf{M}_N^{-1}\mathbf{a} + \exp(-\mathbf{M}_N t)\left(w_N(0) - \mathbf{M}_N^{-1}\mathbf{a}\right). \tag{1.3}$$

The student weight vector w_N thus approaches its asymptotic value $\mathbf{M}_N^{-1}\mathbf{a}$ exponentially quickly with decay constants given by the eigenvalues of the matrix $\mathbf{M}_N$.

To examine what generalization performance is achieved by the above learning algorithm, one has to make an assumption about the functional form of the teacher. The simplest such assumption is that the problem is learnable, i.e. that the teacher, like the student, is a linear perceptron. A teacher $\mathcal{V}$ is then specified by a weight vector w_v and maps a given input x to the output $y_v(x) = w_v^T x/\sqrt{N}$. We assume that the test inputs for which the student is asked to predict the corresponding outputs are drawn from an isotropic Gaussian distribution $P(x) \propto \exp\left(-\frac{1}{2}x^2\right)$. The generalization error, i.e. the average error that a student $\mathcal{N}$ makes on a random input when compared to the teacher $\mathcal{V}$, is given by

$$\epsilon_g = \tfrac{1}{2}\left\langle (y_N(x) - y_v(x))^2 \right\rangle_{P(x)} = \frac{1}{2N}(w_N - w_v)^2. \tag{1.4}$$

Inserting the learning dynamics $w_N = w_N(t)$, the generalization acquires a time dependence, which, in its exact form, depends on the specific training set, teacher, and initial value of the student weight vector, $w_N(0)$. We shall confine our attention to the average of this time-dependent generalization error over all possible training sets and teachers; to avoid clutter, we write this average as simply $\epsilon_g(t)$. In order to calculate the average over training sets, we assume that the inputs x^μ in the training set are chosen independently and randomly from the same distribution as the test inputs. The corresponding training outputs are taken to be the teacher outputs corrupted by additive noise, $y^\mu = y_v(x^\mu) + \eta^\mu$, where the η^μ are independent random variables with zero mean and variance σ^2. To perform the average over teachers, we assume that the teacher weight vectors are sampled randomly from an isotropic Gaussian prior probability distribution, $P(w_v) \propto \exp\left(-\frac{1}{2}w_v^2\right)$. The resulting average generalization error in the limit of infinite learning time, $t \to \infty$, is [1]

$$\epsilon_g(t \to \infty) = \tfrac{1}{2}\left[\sigma^2 G + \lambda(\sigma^2 - \lambda)\frac{\partial G}{\partial \lambda}\right] \tag{1.5}$$

where G is the average of the *response function* over the training inputs

$$G = \langle \mathcal{G} \rangle_{P(\{x^\mu\})} \qquad \mathcal{G} = \frac{1}{N}\operatorname{tr}\mathbf{M}_N^{-1}. \tag{1.6}$$

A result of the same form also holds for learning from a perceptron teacher with a general nonlinear output function [2]. If noise is added to the learning dynamics, a term of the form $\frac{1}{2}TG$ is added to (1.5), with the 'learning temperature' T measuring the variance of the learning noise [1].

For the calculation of the time dependence of the average generalization error it is convenient to assume *tabula rasa* initial conditions, $w_N(0) = \mathbf{0}$. Combining (the dynamics solution) and (1.4), one obtains [1]

$$\epsilon_g(t) - \epsilon_g(t \to \infty) = \int da\, \rho(a) \frac{a}{(\lambda + a)^2} \left[(\lambda - \sigma^2) e^{-(a+\lambda)t} + \tfrac{1}{2}(\sigma^2 + a) e^{-2(a+\lambda)t} \right] \tag{1.7}$$

where $\rho(a)$ is the average eigenvalue spectrum of the input correlation matrix $\mathbf{A}$. The essence of (1.7) is that for $\lambda > 0$ the long-time behaviour of the (average) generalization error is dominated by an exponential decay with decay constant $\lambda + a_{\min}$, where $a_{\min}$ is the minimum (average) eigenvalue of $\mathbf{A}$, i.e. the smallest a for which $\rho(a)$ is non-zero. Formally, $\rho(a)$ can be defined as

$$\rho(a) = \left\langle \frac{1}{N} \sum_{i=1}^{N} \delta(a - a_i) \right\rangle_{P(\{x^\mu\})} \tag{1.8}$$

where we have denoted the eigenvalues of $\mathbf{A}$ by a_i $(i = 1 \ldots N)$. Using the identity

$$\delta(x) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} \frac{1}{x - i\epsilon}$$

one finds that $\rho(a)$ can be expressed as [3]

$$\rho(a) = \frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} G|_{\lambda = -a - i\epsilon}\,. \tag{1.9}$$

Thus, the singularities of the average response function G in the complex λ plane determine the average eigenvalue spectrum of the input correlation matrix $\mathbf{A}$.

Equations (1.5) and (1.9) show that the key quantity determining learning and generalization in the linear perceptron is the average response function G defined in (1.6). This function has previously been calculated in the 'thermodynamic limit', $N \to \infty$ at $\alpha = p/N =$ constant, using a diagrammatic expansion [4] and the replica method [5, 6]. In section 2, we present what we believe to be a much simpler method for calculating G, based on simple matrix identities. We also show explicitly that the response function $\mathcal{G}$ is self-averaging in the thermodynamic limit, which means that the fluctuations of $\mathcal{G}$ around its average G become vanishingly small as $N \to \infty$. This implies, for example, that the generalization error is also self-averaging. In section 3, we extend our method to more general cases such as anisotropic teacher-space priors and input distributions, and general quadratic penalty terms. Finally, finite-size effects are considered in section 4, where we calculate the $O(1/N)$ correction to G, verify the result by computer simulations and examine the resulting effects on generalization and learning dynamics. In section 5 we conclude with a brief summary and discussion of our results.

2. The method

Our method for calculating the average response function G is based on a recursion relation relating the values of the (unaveraged) response function $\mathcal{G}$ for p and $p+1$ training examples. Assume that we are given a set of p training examples with corresponding matrix $\mathbf{M}_N$. By

adding a new training example with input x, we obtain the matrix $\mathbf{M}_N^+ = \mathbf{M}_N + \frac{1}{N}xx^T$. It is straightforward to show that the inverse of $\mathbf{M}_N^+$ can be expressed as

$$\left(\mathbf{M}_N^+\right)^{-1} = \mathbf{M}_N^{-1} - \frac{\frac{1}{N}\mathbf{M}_N^{-1}xx^T\mathbf{M}_N^{-1}}{1+\frac{1}{N}x^T\mathbf{M}_N^{-1}x}. \tag{2.1}$$

(One way of proving this identity is to multiply both sides by $\mathbf{M}_N^+$ and exploit the fact that $\mathbf{M}_N^+\mathbf{M}_N^{-1} = \mathbf{1} + \frac{1}{N}xx^T\mathbf{M}_N^{-1}$.) Taking the trace, we obtain the following recursion relation for $\mathcal{G}$:

$$\mathcal{G}(p+1) = \mathcal{G}(p) - \frac{1}{N}\frac{\frac{1}{N}x^T\mathbf{M}_N^{-2}x}{1+\frac{1}{N}x^T\mathbf{M}_N^{-1}x}. \tag{2.2}$$

Now denote $z_i = \frac{1}{N}x^T\mathbf{M}_N^{-i}x$ $(i = 1, 2)$. With x drawn randomly from the assumed input distribution $P(x) \propto \exp\left(-\frac{1}{2}x^2\right)$, the z_i can readily be shown to be random variables with means and (co-)variances

$$\langle z_i\rangle = \frac{1}{N}\operatorname{tr}\mathbf{M}_N^{-i} \qquad \langle \Delta z_i \Delta z_j\rangle = \frac{2}{N^2}\operatorname{tr}\mathbf{M}_N^{-i-j}$$

where we have used the notation $\Delta z_i = z_i - \langle z_i\rangle$. Combining this with the fact that for $k > 0$, $\operatorname{tr}\mathbf{M}_N^{-k} \leqslant N\lambda^{-k} = \mathrm{O}(N)$, we have that the fluctuations Δz_i of the z_i around their average values are $\mathrm{O}(1/\sqrt{N})$; inserting this into (2.2), we obtain

$$\begin{aligned}\mathcal{G}(p+1) &= \mathcal{G}(p) - \frac{1}{N}\frac{\frac{1}{N}\operatorname{tr}\mathbf{M}_N^{-2}}{1+\frac{1}{N}\operatorname{tr}\mathbf{M}_N^{-1}} + \mathrm{O}\left(N^{-3/2}\right)\\ &= \mathcal{G}(p) + \frac{1}{N}\frac{\partial\mathcal{G}(p)}{\partial\lambda}\frac{1}{1+\mathcal{G}(p)} + \mathrm{O}\left(N^{-3/2}\right).\end{aligned} \tag{2.3}$$

Starting from $\mathcal{G}(0) = 1/\lambda$, we can apply this recursion p times to obtain $\mathcal{G}(p)$ up to terms which add up to at most $\mathrm{O}(pN^{-3/2})$. This shows that in the thermodynamic limit, defined by $N \to \infty$, $\alpha = p/N =$ constant, the response function $\mathcal{G}$ is self-averaging: whatever the training set, the value of $\mathcal{G}$ will always be the same up to fluctuations of $\mathrm{O}(N^{-1/2})$. In fact, in section 4 we shall show that the fluctuations of $\mathcal{G}$ are only $\mathrm{O}(1/N)$. This means that the $\mathrm{O}(N^{-3/2})$ fluctuations from each iteration of (2.3) are only weakly correlated, so that they add up like independent random variables to give a total fluctuation for $\mathcal{G}(p)$ of $\mathrm{O}((p/N^3)^{1/2}) = \mathrm{O}(1/N)$.

We have seen that, in the thermodynamic limit, $\mathcal{G}$ is identical to its average, G, because its fluctuations are vanishingly small. To calculate the value of G in the thermodynamic limit as a function of α and λ, we replace $\mathcal{G}$ by G in (2.3), insert the relation $G(p+1) - G(p) = \frac{1}{N}\partial G(\alpha)/\partial\alpha + \mathrm{O}(1/N^2)$, and neglect all finite N corrections. This yields the partial differential equation

$$\frac{\partial G}{\partial\alpha} - \frac{\partial G}{\partial\lambda}\frac{1}{1+G} = 0 \tag{2.4}$$

which can readily be solved using the method of characteristic curves. A brief account of this method can be found in the appendix. Using the initial condition $G|_{\alpha=0} = 1/\lambda$, one obtains $1/G = \lambda + \alpha/(1+G)$ which leads to the well known result (see, for example, [4])

$$G = \frac{1}{2\lambda}\left(1-\alpha-\lambda+\sqrt{(1-\alpha-\lambda)^2+4\lambda}\right). \tag{2.5}$$

In the complex λ plane, G has a pole at $\lambda = 0$ and a branch cut arising from the root; according to (1.9), these singularities determine the average eigenvalue spectrum $\rho(a)$ of $\mathbf{A}$, with the result [3]

$$\rho(a) = (1-\alpha)\Theta(1-\alpha)\delta(a) + \frac{1}{2\pi a}\sqrt{(a_+ - a)(a - a_-)} \tag{2.6}$$

where $\Theta(x)$ is the Heaviside step function, $\Theta(x) = 1$ for $x > 0$ and 0 otherwise. The root in (2.6) only contributes when its argument is non-negative, i.e. for a between the 'spectral limits' a_- and a_+, which have the values $a_\pm = (1 \pm \sqrt{\alpha})^2$. Since $\mathcal{G}$ is self-averaging, the fluctuations of the true eigenvalue spectrum of $\mathbf{A}$ around its average $\rho(a)$ are also vanishingly small in the thermodynamic limit†.

3. Extensions to more general learning scenarios

We now discuss some extensions of our method to more general learning scenarios. First, consider the case of an anisotropic teacher-space prior, $P(\boldsymbol{w}_v) \propto \exp(-\frac{1}{2}\boldsymbol{w}_v^T \boldsymbol{\Sigma}_v^{-1} \boldsymbol{w}_v)$, with symmetric positive-definite covariance matrix $\boldsymbol{\Sigma}_v$. This does not affect the definition of the response function, but (1.5) now has to be replaced by

$$\epsilon_g(t \to \infty) = \frac{1}{2}\left(\frac{1}{N}\operatorname{tr}\boldsymbol{\Sigma}_v\right)\left[\tilde{\sigma}^2 G + \lambda(\tilde{\sigma}^2 - \lambda)\frac{\partial G}{\partial \lambda}\right]$$

with a renormalized noise level $\tilde{\sigma}^2 = \sigma^2/(\frac{1}{N}\operatorname{tr}\boldsymbol{\Sigma}_v)$. The factor $\frac{1}{N}\operatorname{tr}\boldsymbol{\Sigma}_v$ determines by how much the average squared length of the teacher weight vector is now larger than for the isotropic teacher-space prior considered in the previous section. This factor also scales the size of the typical squared teacher output. Therefore, it appears as a multiplicative factor in the generalization error, and also determines the renormalized noise level (which is, effectively, a mean-square noise-to-signal ratio).

As a second extension, assume that the inputs are drawn from an anisotropic distribution, $P(\boldsymbol{x}) \propto \exp(-\frac{1}{2}\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x})$. It can then be shown that the asymptotic value of the average generalization error is still given by (1.5) if the average response function is redefined to be $G = \langle \frac{1}{N}\operatorname{tr}\boldsymbol{\Sigma}\mathbf{M}_N^{-1}\rangle$. This modified response function can be calculated as follows: first we rewrite G as $\langle \frac{1}{N}\operatorname{tr}(\lambda\boldsymbol{\Sigma}^{-1} + \tilde{\mathbf{A}})^{-1}\rangle$, where $\tilde{\mathbf{A}} = \frac{1}{N}\sum_\mu (\tilde{\boldsymbol{x}}^\mu)^T \tilde{\boldsymbol{x}}^\mu$ is the correlation matrix of the transformed input examples‡ $\tilde{\boldsymbol{x}}^\mu = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}^\mu$. Since the $\tilde{\boldsymbol{x}}^\mu$ are distributed according to $P(\tilde{\boldsymbol{x}}^\mu) \propto \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{x}}^\mu)^2\right)$, the problem is thus reduced to finding the average response function $G_L = \langle \mathcal{G}_L\rangle = \langle \frac{1}{N}\operatorname{tr}(\mathbf{L} + \mathbf{A})^{-1}\rangle$ for isotropically distributed inputs and $\mathbf{L} = \lambda\boldsymbol{\Sigma}^{-1}$. As explained in the appendix, a differential equation analogous to (2.4) holds for G_L. Together with the initial condition $G_L|_{\alpha=0} = \frac{1}{N}\operatorname{tr}\mathbf{L}^{-1}$, one obtains G_L as the solution of the implicit equation

$$G_L = \frac{1}{N}\operatorname{tr}\left(\mathbf{L} + \frac{\alpha}{1+G_L}\mathbf{1}\right)^{-1}. \tag{3.1}$$

As explained above, the modified response function $G = \langle \frac{1}{N}\operatorname{tr}\boldsymbol{\Sigma}\mathbf{M}_N^{-1}\rangle$ for the case of an anisotropic input distribution considered here is given by the value of G_L which solves (3.1) for $\mathbf{L} = \lambda\boldsymbol{\Sigma}^{-1}$. If the eigenvalue spectrum of $\boldsymbol{\Sigma}$ has a particularly simple form, then the resulting dependence of G on α and λ can be expressed analytically, but, in general, (3.1) will have to be solved numerically.

† More precisely, the fluctuations of linear functionals of the eigenvalue spectrum of $\mathbf{A}$ (which is, mathematically, a distribution) vanish as $N \to \infty$.

‡ We write $\boldsymbol{\Sigma}^{-1/2}$ for the unique positive-definite symmetric matrix which obeys $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$.

Finally, one can also investigate the effect of a general quadratic weight-decay term, $\frac{1}{2}\boldsymbol{w}_N^T\boldsymbol{\Lambda}\boldsymbol{w}_N$, in the energy function E which defines the gradient descent learning dynamics. This modifies the definition (1.2) of the matrix $\mathbf{M}_N$ to $\mathbf{M}_N = \boldsymbol{\Lambda} + \mathbf{A}$, and the calculation of the average generalization error becomes more complicated in this case. In addition to the average response function $G = \langle \frac{1}{N} \operatorname{tr} \mathbf{M}_N^{-1} \rangle$, which can be obtained as the solution of (3.1) for $\mathbf{L} = \boldsymbol{\Lambda}$, one now also needs to know the modified response functions $G_{\Lambda^n} = \langle \frac{1}{N} \operatorname{tr} \boldsymbol{\Lambda}^n \mathbf{M}_N^{-1} \rangle$ for $n = 1, 2$. Fortunately, it is possible to calculate the general modified response function $G_{BL} = \langle \frac{1}{N} \operatorname{tr} \mathbf{B}(\mathbf{L} + \mathbf{A})^{-1} \rangle$ for positive-definite symmetric $\mathbf{L}$ and a general matrix $\mathbf{B}$ by extending the methods of the previous section. As outlined in the appendix, in the thermodynamic limit one obtains a differential equation for G_{BL} similar but not exactly identical to (2.4), which can be solved to give

$$G_{BL} = \frac{1}{N} \operatorname{tr} \mathbf{B} \left(\mathbf{L} + \frac{\alpha}{1 + G_L} \mathbf{1} \right)^{-1} . \tag{3.2}$$

Thus, G_{BL} can be calculated straight away once G_L is known. In the specific case of a general quadratic weight decay which we consider here, one has $\mathbf{L} = \boldsymbol{\Lambda}$ and $G_L = G$, and by setting $\mathbf{B} = \boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^2$ in (3.2), one obtains $G_\Lambda = 1 - \alpha G/(1 + G)$ and $G_{\Lambda^2} = \frac{1}{N} \operatorname{tr} \boldsymbol{\Lambda} - \alpha/(1+G) + \alpha^2 G/(1+G)^2$. Using these relations, the average generalization error can be written in terms of G alone, although the final expressions become rather more cumbersome than (1.5). We note parenthetically that expressions (3.1) and (3.2) can also be obtained using diagrammatic methods [7].

4. Finite-size effects

So far, we have focused attention on the thermodynamic limit of perceptrons of infinite size N. The results are clearly only valid approximately for real, finite systems, and it is therefore interesting to investigate corrections for finite N. This we do in the present section by calculating the $O(1/N)$ corrections to G, $\epsilon_g(t \to \infty)$, and $\rho(a)$. First note that, for $\lambda = 0$, the exact value of the average response function is [8]

$$G|_{\lambda=0} = \frac{1}{N} \langle \operatorname{tr} \mathbf{A}^{-1} \rangle = (\alpha - 1 - 1/N)^{-1} \tag{4.1}$$

for $\alpha > 1 + 1/N$. This result follows straightforwardly from the fact that the inverse input correlation matrix, $\mathbf{A}^{-1}$, obeys an 'inverted Wishart distribution' (see, for example, [9], definition 8.1 and exercise 8.7). Equation (4.1) clearly admits a series expansion in powers of $1/N$. Assuming that a similar expansion also exists for non-zero λ, we write

$$G = G_0 + G_1/N + O(1/N^2) . \tag{4.2}$$

Here G_0 is the value of G in the thermodynamic limit as given by (2.5). We calculate G_1 below, and verify the analytical result by computer simulations. Note that there is no *a priori* guarantee that an expansion of the type (4.2) exists (compare, for example, the results of [10], which suggest that for the binary perceptron, finite-size effects depend non-analytically on $1/N$). However, the simulation results presented below do provide compelling evidence for the existence of the expansion (4.2) of the average response function in powers of $1/N$.

For finite N, not only the corrections to the average response function G but also the fluctuations $\Delta\mathcal{G} = \mathcal{G} - G$ of $\mathcal{G}$ around its average value G become relevant. For $\lambda = 0$, the variance of these fluctuations is known to have a power-series expansion in $1/N$ (see, for example, [11]), and again we assume a similar expansion for finite λ,

$$\langle (\Delta\mathcal{G})^2 \rangle = \Delta^2/N + O(1/N^2) .$$

Here the first term is $O(1/N)$ and not $O(1)$ because, as discussed in section 2, the fluctuations of $\mathcal{G}$ for large N are no greater than $O(N^{-1/2})$.

To calculate G_1 and Δ^2, we start again from the recursion relation (2.2). However, now we cannot neglect terms involving Δz_i and $\Delta\mathcal{G}$, but have to expand everything up to second order in these fluctuation quantities. Averaging over the training inputs and collecting orders of $1/N$ yields, after some straightforward algebra, the known equation (2.4) for G_0 and

$$\frac{\partial G_1}{\partial \alpha} - \frac{\partial G_1}{\partial \lambda}\frac{1}{1+G_0} = \frac{\partial}{\partial \lambda}\left[\left(\frac{\partial G_0}{\partial \lambda} - \frac{\Delta^2}{2}\right)\frac{1}{(1+G_0)^2}\right] - \frac{G_1}{(1+G_0)^2}\frac{\partial G_0}{\partial \lambda} - \frac{1}{2}\frac{\partial^2 G_0}{\partial \alpha^2}. \tag{4.3}$$

By squaring the difference between (2.2) and its average over the training inputs, one can similarly derive an equation for Δ^2:

$$\frac{\partial \Delta^2}{\partial \alpha} - \frac{\partial \Delta^2}{\partial \lambda}\frac{1}{1+G_0} = -2\frac{\Delta^2}{(1+G_0)^2}\frac{\partial G_0}{\partial \lambda}. \tag{4.4}$$

Details of the solution of these two partial differential equations are again relegated to the appendix. At $\alpha = 0$, one has $\mathcal{G} = G = G_0 = 1/\lambda$ and hence $G_1 = \Delta^2 = 0$; using these initial conditions, one finds $\Delta^2 \equiv 0$ for all α and λ, and

$$G_1 = \frac{G_0^2(1-\lambda G_0)}{\left(1+\lambda G_0^2\right)^2}. \tag{4.5}$$

In the limit $\lambda \to 0$, $G_1 = 1/(\alpha-1)^2$, consistent with (4.1); likewise, the result $\Delta^2 \equiv 0$ agrees with the exact series expansion of the variance of the fluctuations of $\mathcal{G}$ for $\lambda = 0$, which begins with an $O(1/N^2)$ term [11].

Before exploring the consequences of the result (4.5) for finite-size effects in generalization and learning dynamics, we present the results of computer simulations, performed to test our analytic predictions. For perceptron sizes between $N = 4$ and 80, we calculated the response function by direct matrix inversion, averaging over between 1200 (for $N = 80$) and 200 000 (for $N = 4$) randomly sampled sets of training inputs to obtain an 'experimental' value of the average response function. In figure 1, we plot the results

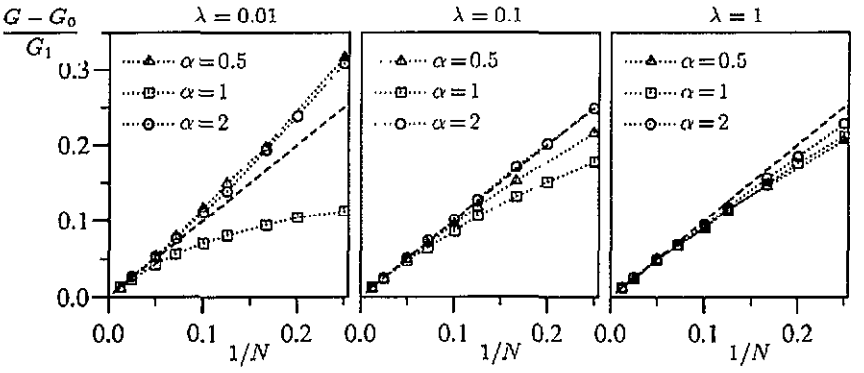

Figure 1. Simulation results for the average response function G at finite perceptron size N, for different values of the weight decay parameter, λ, and the number of training examples (normalized by N), α. The plots of $(G - G_0)/G_1$ versus $1/N$ show that as $1/N$ approaches zero, the results (symbols connected by dotted curves as a visual aid) are well approximated by $(G - G_0)/G_1 = 1/N$ (broken curve), in agreement with (4.2). Statistical errors due to the finite numbers of simulation samples are smaller than the symbol size.

in the form $(G - G_0)/G_1$ versus $1/N$ for $\alpha = 0.5$, 1 and 2, and $\lambda = 0.01$, 0.1 and 1, using the results for G_0 and G_1 from (2.5) and (4.5). The simulation results are seen to agree well with the theoretical prediction from (4.2), namely, $(G - G_0)/G_1 = 1/N + O(1/N^2)$. The $O(1/N^2)$ terms, which correspond to corrections to G of second and higher order in $1/N$, appear as deviations from the straight line $(G - G_0)/G_1 = 1/N$ in figure 1 for larger values of $1/N$. These higher-order corrections are expected to be negligible as long as $1/N \ll G_0/G_1$, because this entails that the first-order correction G_1/N is already small compared to the zeroth-order contribution G_0. Correspondingly, the strongest higher-order corrections in the plots in figure 1 are seen to occur for $\lambda = 0.01, \alpha = 1$, which can be checked to see that it has smallest value of G_0/G_1 amongst the plots.

We now turn to the finite-size corrections to the generalization error. From the $1/N$ expansion (4.2) of G we obtain a corresponding expansion of the asymptotic value of the average generalization error, which we write as

$$\epsilon_g(t \to \infty) = \epsilon_{g,0} + \epsilon_{g,1}/N + O(1/N^2).$$

From (1.5), the explicit expressions for $\epsilon_{g,0}$ and $\epsilon_{g,1}$ are

$$\epsilon_{g,0} = \frac{1}{2}\left[\sigma^2 G_0 + \lambda(\sigma^2 - \lambda)\frac{\partial G_0}{\partial \lambda}\right] \qquad \epsilon_{g,1} = \frac{1}{2}\left[\sigma^2 G_1 + \lambda(\sigma^2 - \lambda)\frac{\partial G_1}{\partial \lambda}\right].$$

For given $\epsilon_{g,0}$ and $\epsilon_{g,1}$, we can distinguish three regimes for the size of the perceptron, N. For $N \gg N_c = |\epsilon_{g,1}/\epsilon_{g,0}|$, the result $\epsilon_g(t \to \infty) = \epsilon_{g,0}$ obtained in the thermodynamic limit is a good approximation. For smaller values of N, the first-order correction $\epsilon_{g,1}/N$ has to be taken into account. Finally, for $N \approx N_c$ we expect the series expansion of $\epsilon_g(t \to \infty)$ in powers of $1/N$ to break down altogether, since all terms in the series become of comparable order of magnitude. In figure 2, we plot $\epsilon_{g,0}$ and $\epsilon_{g,1}$ for several values of λ and σ^2. The graphs suggest that the relative correction $|\epsilon_{g,1}/(N\epsilon_{g,0})|$—and hence $N_c = |\epsilon_{g,1}/\epsilon_{g,0}|$—is largest when λ is small and α is close to 1. The exact dependence of N_c on α, λ and σ^2, however, is rather complicated. We confine ourselves to bounding N_c in the form

$$N_c \leqslant \max_{\sigma^2} N_c(\alpha, \lambda, \sigma^2) = N_c(\alpha, \lambda)$$

by maximizing over the noise parameter σ^2 (which, in an experimental setting, is beyond our control anyway). This bounding operation is easily performed since $N_c = |\epsilon_{g,1}/\epsilon_{g,0}|$ attains its maximum WRT σ^2 either at $\sigma^2 = 0$ or for $\sigma^2 \to \infty$, due to the monotonicity of $\epsilon_{g,1}/\epsilon_{g,0}$ as a function of σ^2. We plot the resulting $N_c(\alpha, \lambda)$ in figure 3 for several values of λ. $N_c(\alpha, \lambda)$ is maximal for $\lambda \to 0$; evaluating this limit, we obtain†

$$N_c \leqslant \max_{\lambda} N_c(\alpha, \lambda) = N_c(\alpha) = \begin{cases} 1/(1-\alpha) & \text{for} \quad 0 < \alpha < 1 \\ (3\alpha + 1)/[\alpha(\alpha - 1)] & \text{for} \quad \alpha > 1. \end{cases} \tag{4.6}$$

Therefore, results for $\epsilon_g(t \to \infty)$ derived in the thermodynamic limit will be valid for any λ and σ^2 provided that $N \gg N_c(\alpha)$. For large α, $N_c(\alpha) = 3/\alpha + O(1/\alpha^2)$, and the condition $N \gg N_c(\alpha)$ will easily be fulfilled. For finite λ and near $\alpha = 1$, the bound (4.6) is unnecessarily pessimistic, as figure 3 shows. To remedy this, we have verified numerically that for $\lambda > 2$, $N_c(\alpha, \lambda)$ attains its maximum WRT α for $\alpha \to 0$. From this one can derive an alternative bound for N_c,

$$N_c \leqslant \max_{\alpha} N_c(\alpha, \lambda) = \frac{2\lambda - 1}{(\lambda + 1)^2} \qquad \text{for} \quad \lambda > 2$$

† Note that both $N_c(\alpha, \lambda)$ and $N_c(\alpha)$ are discontinuous at $\alpha = 0$: their limits as $\alpha \to 0$ are, in general, non-zero, whereas at $\alpha = 0$, where $\epsilon_{g,1}$ vanishes, they are both zero.

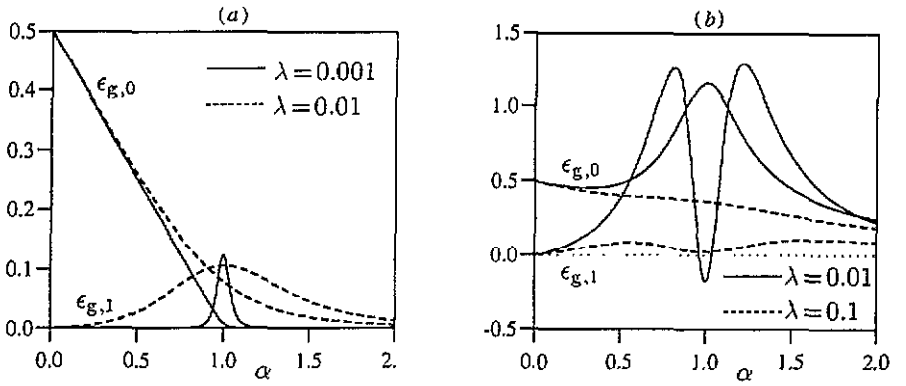

Figure 2. Average generalization error: result for $N \to \infty$, $\epsilon_{g,0}$, and $O(1/N)$ correction, $\epsilon_{g,1}$. Curves are labelled by the value of the weight decay parameter λ. (*a*) Noise-free teacher, $\sigma^2 = 0$. (*b*) Noisy teacher, $\sigma^2 = 0.5$.

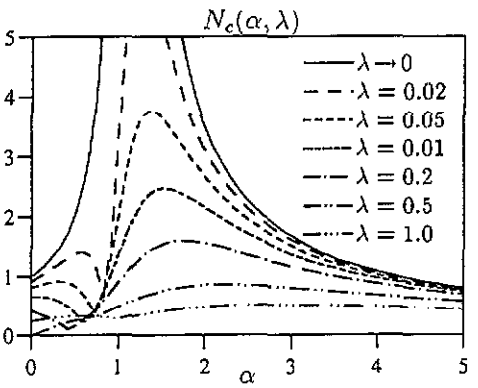

Figure 3. Critical perceptron size $N_c(\alpha, \lambda)$: for $N \gg N_c(\alpha, \lambda)$, the results for the average generalization error $\epsilon_g(t \to \infty)$ obtained in the thermodynamic limit are valid to a good approximation, for any noise level σ^2. Note that the maximum of $N_c(\alpha, \lambda)$ WRT λ is obtained for $\lambda \to 0$.

which is independent of α and σ^2 and will be tighter than (4.6) near $\alpha = 1$ and for sufficiently large λ.

We now turn to the $O(1/N)$ correction to the average eigenvalue spectrum $\rho(a)$ of the input correlation matrix **A**. We set

$$\rho(a) = \rho_0(a) + \rho_1(a)/N + O(1/N^2) \tag{4.7}$$

where $\rho_0(a)$ is the $N \to \infty$ result given by (2.6). From (1.9) and (4.5) one then derives

$$\rho_1(a) = \tfrac{1}{4}\delta(a - a_+) + \tfrac{1}{4}\delta(a - a_-) - \frac{1}{2\pi}\frac{1}{\sqrt{(a_+ - a)(a - a_-)}} . \tag{4.8}$$

Figure 4 shows sketches of $\rho_0(a)$ and $\rho_1(a)$. Note that $\int da\, \rho_1(a) = 0$, as expected since from the definition (1.8) the normalization of $\rho(a)$, $\int da\, \rho(a) = 1$, is independent of N. Furthermore, there is no $O(1/N)$ correction to the δ-peak in $\rho_0(a)$ at $a = 0$, since this peak arises from the $N - p$ zero eigenvalues of **A** for $\alpha = p/N < 1$ and therefore has an exact height of $1 - \alpha$ for any N. The δ-peaks in $\rho_1(a)$ at the spectral limits a_+ and a_- *are an artefact of the truncated* $1/N$ *expansion:* $\rho(a)$ *is determined by the singularities of* G as a function of λ, and the location of these singularities is only obtained correctly by resumming the full $1/N$ expansion. The δ-peaks in $\rho_1(a)$ can be interpreted as 'precursors' of a broadening of the eigenvalue spectrum of **A** to values which, when the whole $1/N$

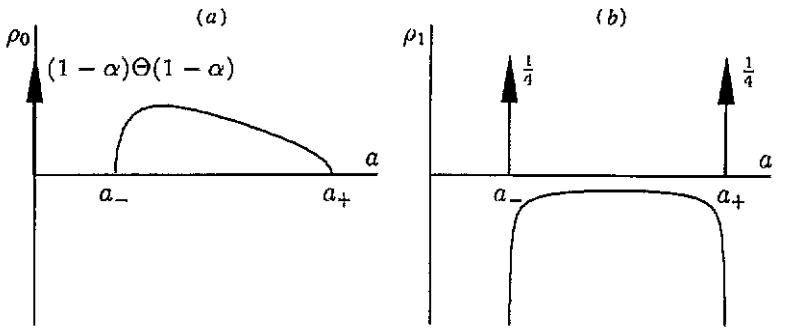

Figure 4. Schematic plot of the average eigenvalue spectrum $\rho(a)$ of the input correlation matrix **A**. (*a*) Result for $N \to \infty$, $\rho_0(a)$. (*b*) $O(1/N)$ correction, $\rho_1(a)$. Arrows indicate δ-peaks and are labelled by the corresponding heights.

series is resummed, will lie outside the $N \to \infty$ spectral range $[a_-, a_+]$. The negative term in $\rho_1(a)$ represents the corresponding 'flattening' of the eigenvalue spectrum between a_- and a_+. We can thus conclude that the average eigenvalue spectrum of **A** for finite N will be broader than for $N \to \infty$. This implies that the learning dynamics given by (1.7) will be slower for finite N than predicted from the thermodynamic-limit results, since the smallest eigenvalue $a_{\min}$ of **A** will be smaller than a_-, the lower spectral limit for $N \to \infty$.

Note that our prediction of a broadening of $\rho(a)$ for finite N can also be confirmed by considering the extreme case $N = 1$: in this case, the matrix **A** becomes the scalar sum of p Gaussian random variables with zero mean and unit variance. Hence, $\rho(a)$ is just the probability density of a χ^2-distribution with p degrees of freedom, which is non-zero for all $a > 0$, i.e. over a much broader range than the spectrum $[a_-, a_+]$ predicted for $N \to \infty$.

From our result for $\rho_1(a)$ we can also deduce when the $N \to \infty$ result $\rho_0(a)$ is valid for finite N; the condition turns out to be $N \gg a/[(a_+ - a)(a - a_-)]$. Consistent with our discussion of the broadening of the eigenvalue spectrum of **A**, N has to be larger for a near the spectral limits a_- and a_+ if $\rho_0(a)$ is to be a good approximation to the finite-N average eigenvalue spectrum of **A**.

Finally, in figure 5 we present exemplary results from computer simulations of the average eigenvalue spectrum of the input correlation matrix **A**. We show the results for $\alpha = 10$ and $N = 4$, $N = 8$, which are based on 10^7 and 2×10^6 randomly sampled sets of training inputs, respectively. The average eigenvalue spectrum $\rho(a)$ was found by sorting the numerically determined eigenvalues of **A** into 100 histogram slots, evenly spaced across the spectral range shown in figure 5, and then applying a suitable normalization. Instead of displaying the resulting $\rho(a)$ directly, in figure 5 we plot the quantity $N(\rho(a) - \rho_0(a))$, which should approach $\rho_1(a)$ for large N from (4.7). This approach can already clearly be seen for the relatively small values of N used in our simulations; consistent with our discussion above, the deviations of $N(\rho(a) - \rho_0(a))$ from $\rho_1(a)$ are largest near the $N \to \infty$ spectral limits a_- and a_+, where the finite-size corrections to $\rho(a)$ of first, and hence also of higher, order in $1/N$ are largest. We note parenthetically that the results of more extensive computer simulations for perceptron sizes $N = 2 \ldots 6$ suggest that for any α, $N(\rho(a) - \rho_0(a))$ as a function of a has $2N$ turning points between a_- and a_+ (compare with figure 5(*a*)); it remains an open question whether this remains true for large N and if so, whether there exists a proof for this property.

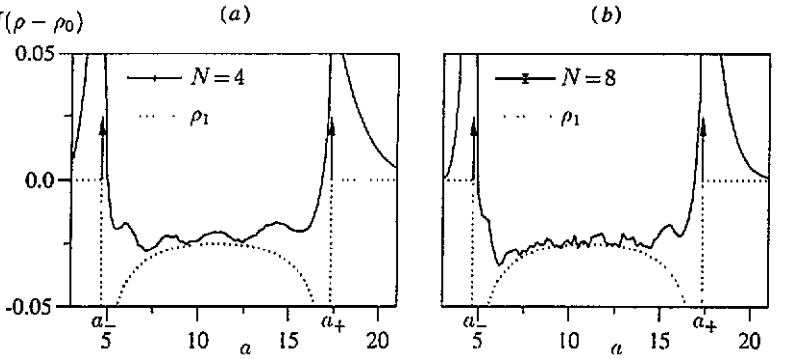

Figure 5. Simulation results for the average eigenvalue spectrum, $\rho(a)$, of the input correlation matrix **A**, for $\alpha = 10$ and (*a*) $N = 4$, (*b*) $N = 8$. Shown is the scaled difference $N(\rho(a)-\rho_0(a))$ (full curve), which should approach $\rho_1(a)$ (dotted curve) for large N from (4.7). The arrows indicate the delta-peak contributions to ρ_1 at the $N \to \infty$ spectral limits $a_\pm = (1 \pm \sqrt{\alpha})^2$ (compare (4.8) and figure 4); typical error bars for the simulation results are shown in the legend.

5. Summary and discussion

We have presented a new method, based on simple matrix identities, for calculating the response function $\mathcal{G}$ and its average G which determine most of the properties of learning and generalization in linear perceptrons. In the thermodynamic limit, $N \to \infty$, we have recovered the known result for G and have also shown explicitly that $\mathcal{G}$ is self-averaging. We have then demonstrated the versatility of our method by extending it to more general learning scenarios. Finally, we have calculated the $O(1/N)$ correction to G, which was found to agree well with the results of computer simulations. The corresponding correction to the average generalization error has been obtained, and explicit conditions have been derived on how large N has to be for the results obtained in the thermodynamic limit to be valid. We have also calculated the $O(1/N)$ correction to the average eigenvalue spectrum of the input correlation matrix **A** and interpreted it in terms of a broadening of the spectrum for finite N, which will cause a slowing down of the learning dynamics.

We remark that the $O(1/N)$ corrections which we have obtained can also be used in different contexts. For example, the generalization error can be estimated by the test error, obtained by comparing the outputs of student and teacher on a finite number of randomly chosen test inputs. Using our results, test error fluctuations can be analysed, and an optimal test-set size can be derived for the case where the total number of training and test examples is limited [11]. Another application is in an analysis of the evidence procedure in Bayesian inference for finite N, where optimal values of 'hyperparameters' like the weight decay parameter λ are determined on the basis of the training data [12]. We hope, therefore, that our results will provide the basis for a systematic investigation of finite-size effects in learning and generalization.

Acknowledgments

The author wishes to thank David Saad and David Barber for helpful comments and careful reading of the manuscript, as well as one of the reviewers for pointing out reference [10].

Appendix.

In this appendix, we briefly describe the method of characteristic curves for the solution of partial differential equations, following the exposition in [13]. We then apply the method to obtain the solutions of the differential equations for the various response functions introduced in the paper.

Consider the following quasi-linear first-order partial differential equation for $f(x, y)$,

$$a\,\frac{\partial f}{\partial x} + b\,\frac{\partial f}{\partial y} - c = 0 \tag{A.1}$$

where a, b and c are functions of x, y and f. The solution $f = f(x, y)$ can be thought of as a surface in (x, y, f) space, which has normal vectors proportional to $(\partial f/\partial x, \partial f/\partial y, -1)$. Equation (A.1) can then be interpreted as defining a vector field (a, b, c) of 'characteristic directions', which are orthogonal to the normal vectors of the solution surface. This suggests that any curve starting at a point within the solution surface remains within that surface if it follows the characteristic direction at every point. Formally, such 'characteristic curves' are defined by the requirement $\mathrm{d}(x, y, f)/\mathrm{d}t = (a, b, c)$, where t parametrizes the points along the curve. *It can be shown rigorously [13] that the solution surface is indeed given by* the union of all characteristic curves which pass through a one-parameter family of points defining the initial conditions for $f(x, y)$.

Now consider equation (2.4) for the average response function G in the thermodynamic limit. The characteristic curves are the solutions of

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = 1 \qquad \frac{\mathrm{d}\lambda}{\mathrm{d}t} = -\frac{1}{1+G} \qquad \frac{\mathrm{d}G}{\mathrm{d}t} = 0$$

which are given by

$$\alpha = \alpha_0 + t \qquad \lambda = \lambda_0 - \frac{t}{1+G_0} \qquad G = G_0\,. \tag{A.2}$$

The initial condition $G|_{\alpha=0} = 1/\lambda$ selects the characteristic curves with $\alpha_0 = 0$, λ_0 arbitrary, $G_0 = 1/\lambda_0$. Inserting this into (A.2), one can eliminate α_0, λ_0, G_0 and t to obtain $1/G = \lambda + \alpha/(1+G)$. This yields the solution (2.5) for $G(\alpha, \lambda)$.

We now turn to (3.1) for the modified response function $G_L = \langle \mathcal{G}_L \rangle = \langle \frac{1}{N} \mathrm{tr}(\mathbf{L}+\mathbf{A})^{-1} \rangle$. To obtain this result, one first replaces the matrix $\mathbf{L}$ by $\mathbf{L}+\lambda\mathbf{1}$. The recursion relation (2.2) between $\mathcal{G}(p+1)$ and $\mathcal{G}(p)$ remains valid for $\mathcal{G}_L$, and results, in the thermodynamic limit, in a differential equation for G_L exactly analogous to (2.4), with G replaced by G_L. The corresponding characteristic curves are the same as in (A.2), but the initial condition $G_L|_{\alpha=0} = \frac{1}{N}\mathrm{tr}(\mathbf{L}+\lambda\mathbf{1})^{-1}$ now selects a different set of characteristic curves. This leads to the equation $G_L = \frac{1}{N}\mathrm{tr}[\mathbf{L}+(\lambda+\alpha/(1+G_L))\mathbf{1}]^{-1}$, from which (3.1) is obtained by setting $\lambda = 0$.

The solution for the general modified response function $G_{BL} = \langle \frac{1}{N}\mathrm{tr}\,\mathbf{B}(\mathbf{L}+\mathbf{A})^{-1} \rangle$ given in (3.2) is obtained as follows: first, one again replaces the matrix $\mathbf{L}$ by $\mathbf{L}+\lambda\mathbf{1}$. Multiplying the matrix equation (2.1) by $\mathbf{B}$ and taking the trace, one can follow the procedure described in section 2 to obtain, in the thermodynamic limit, the differential equation

$$\frac{\partial G_{BL}}{\partial \alpha} - \frac{\partial G_{BL}}{\partial \lambda}\frac{1}{1+G_L} = 0\,.$$

Since G_L is a fairly complicated function of α and λ, the corresponding characteristic equations

$$\frac{\mathrm{d}\alpha}{\mathrm{d}t} = 1 \qquad \frac{\mathrm{d}\lambda}{\mathrm{d}t} = -\frac{1}{1+G_L(\alpha,\lambda)} \qquad \frac{\mathrm{d}G_{BL}}{\mathrm{d}t} = 0$$

might seem hard to solve. However, G_L is, in fact, constant along the characteristic curves: as pointed out above, G_L obeys (2.4) (with G replaced by G_L), and hence

$$\frac{dG_L}{dt} = \frac{d\alpha}{dt}\frac{\partial G_L}{\partial \alpha} + \frac{d\lambda}{dt}\frac{\partial G_L}{\partial \lambda} = \frac{\partial G_L}{\partial \alpha} - \frac{1}{1+G_L}\frac{\partial G_L}{\partial \lambda} = 0.$$

Therefore, the characteristic curves are

$$\alpha = \alpha_0 + t \qquad \lambda = \lambda_0 - \frac{t}{1+G_L} \qquad G_{BL} = \text{constant}.$$

Together with the initial condition $G_{BL}|_{\alpha=0} = \frac{1}{N}\operatorname{tr}\mathbf{B}(\mathbf{L}+\lambda\mathbf{1})^{-1}$, this yields $G_{BL}|_{\alpha=0} = \frac{1}{N}\operatorname{tr}\mathbf{B}[\mathbf{L}+(\lambda+\alpha/(1+G_L))\mathbf{1}]^{-1}$. Equation (3.2) is recovered for $\lambda = 0$.

Finally, for the solution of (4.3) and (4.4), one first verifies that $\Delta^2 \equiv 0$ satisfies (4.4) and the corresponding initial conditions; of course, this solution can also be obtained using the method of characteristic curves. One can then simplify (4.3) by inserting $\Delta^2 = 0$ and by using the fact that G_0, the value of G in the thermodynamic limit, obeys (2.4) (with G replaced by G_0). After some algebra, one obtains

$$\frac{\partial G_1}{\partial \alpha} - \frac{\partial G_1}{\partial \lambda}\frac{1}{1+G_0} = \tfrac{1}{2}G_0'' - G_1\frac{G_0'}{1+G_0}. \tag{A.3}$$

Here we have introduced the abbreviations $G_0' \equiv \partial G_0/\partial\alpha$ and $G_0'' \equiv \partial^2 G_0/\partial\alpha^2$. By the same reasoning as above, one can show that G_0 is constant along the characteristic curves of (A.3). The characteristic curves obeying the initial condition $G_1|_{\alpha=0} = 0$ are therefore given by

$$\alpha = t \qquad \lambda = \lambda_0 - \frac{t}{1+G_0} \qquad \frac{dG_1}{dt} = \tfrac{1}{2}G_0'' - G_1\frac{G_0'}{1+G_0}$$

with $G_1(t=0) = 0$. The constant value of G_0 along a characteristic curve is related to λ_0 by $G_0 = G_0(t=0) = G_0|_{\lambda=\lambda_0,\alpha=0} = 1/\lambda_0$. Using the explicit form of $G_0(\alpha,\lambda)$, both G_0' and G_0'' can be expressed as functions of G_0 and λ alone as follows:

$$G_0' = -\frac{1}{\lambda + 1/G_0^2} \qquad G_0'' = \frac{2/G_0^3}{(\lambda + 1/G_0^2)^3}.$$

This finally leads to the following linear differential equation for G_1 as a function of λ along a characteristic curve with a given value of G_0:

$$\frac{dG_1}{d\lambda} = -(1+G_0)\frac{dG_1}{dt} = -\frac{(1+G_0)/G_0^3}{(\lambda+1/G_0^2)^3} - \frac{G_1}{\lambda+1/G_0^2}. \tag{A.4}$$

Since $\lambda = \lambda_0 = 1/G_0$ at $t = 0$, the initial condition is $G_1(\lambda = 1/G_0) = 0$. The integration of (A.4) is straightforward and yields directly the solution (4.5) given in the text.

References

[1] Krogh A and Hertz J A 1992 *J. Phys. A: Math. Gen.* **25** 1135–47

[2] Sollich P 1994 Minimum entropy queries for linear students learning nonlinear rules (in preparation)

[3] Krogh A 1992 *J. Phys. A: Math. Gen.* **25** 1119–33

[4] Hertz J A, Krogh A and Thorbergsson G I 1989 *J. Phys. A: Math. Gen.* **22** 2133–50

[5] Opper M 1989 *Europhys. Lett.* **8** 389–92

[6] Kinzel W and Opper M 1991 *Models of Neural Networks* ed E Domany, J L van Hemmen and K Schulten (Berlin: Springer) pp 149–71

[7] Saad D 1994 General Gaussian priors for improved generalization *Neural Comput.* submitted

[8] Hansen L K 1993 *Neural Networks* **6** 393–6

[9] Eaton M L 1983 *Multivariate Statistics—A Vector Space Approach* (New York: John Wiley)
[10] Derrida D, Griffiths R B and Prügel-Bennett A 1991 *J. Phys. A: Math. Gen.* **24** 4907–40
[11] Barber D, Saad D and Sollich P 1994 Finite size effects and optimal test set size in linear perceptrons. *J. Phys. A: Math. Gen.* submitted
[12] Marion G 1994 Data dependent hyperparameters and evidence in linear hypothesis evaluation (in preparation)
[13] John F 1978 *Partial Differential Equations* 3rd edn (New York: Springer)