Learning and generalization in a linear perceptron stochastically trained with noisy data

1993 J. Phys. A: Math. Gen. 26 5767

(http://iopscience.iop.org/0305-4470/26/21/016)

J. Phys. A: Math. Gen. 26 (1993) 5767–5779. Printed in the UK

Learning and generalization in a linear perceptron stochastically trained with noisy data

A P Dunmur and D J Wallace

Department of Physics, University of Edinburgh, Kings Buildings, Mayfield Road, Edinburgh EH9 3JZ, UK

Received 8 February 1993, in final form 1 June 1993

Abstract. A linear perceptron is stochastically trained on a corrupted training data set; this enables the effect of noise on the data to be studied. The average properties of the network are calculated using the Gardner method following Seung *et al.* A weight decay term is added to the training energy and the effect on generalization studied and compared with previously known results. A prescription for setting the optimal weight decay parameter at finite temperature is presented. The results also suggest an initial temperature for an annealing schedule.

1. Introduction

Neural networks can be trained to implement a particular rule by the process of supervised learning [1]. This is done by presenting the network with a finite data set sampled from the set of all possible inputs and adjusting the network parameters (the weights) such that some cost function is minimized. The actual performance of the system is evaluated on a finite test set, again sampled from the entire set of possible inputs. However, in use, the network may be presented with any pattern selected from the set of all possible input patterns according to some distribution. The performance of the network under these conditions could be very different from that evaluated on the test set; thus it is useful to have an analytical method for estimating the generalization ability of a network over the set of all possible inputs. One possible method is to use the theory of VC dimension [2, 3] to calculate bounds on the generalization ability, another is to calculate average generalization abilities by statistical physics methods.

The statistical physics approach has been studied in some depth and numerous techniques have emerged [4]. One of these methods, presented by Krogh and Hertz [5], uses the spectrum of eigenvalue solutions of the training equation to study the network dynamics and also to estimate the asymptotic time behaviour of the network. Krogh and Hertz used this method to calculate the generalization error of a linear perceptron trained with a weight decay on a data set that has been corrupted by noise [6]. This noise is static throughout the training cycle and thus cannot be removed by simple averaging. A prescription for setting the weight decay such that the generalization error is minimized is also presented by Krogh and Hertz. The generalization error is a measure of the network's generalization ability, that is, the higher the generalization error, the worse the generalization ability.

The asymptotic time behaviour of the system may also be calculated using another statistical physics method based on the Gardner method [7, 8]. This method has been used to study the learning and generalization of simple networks where the network is trained stochastically [9–12]. Stochastic training involves adding dynamic noise to the

0305-4470/93/215767+13$07.50

weight update equation giving a Langevin equation. The asymptotic time behaviour of this equation produces a Gibbs distribution on the weight vectors. The resulting free energy may then be calculated by the replica method [13] from which the generalization abilities can be derived [11]. Training with noise has been shown to be beneficial to the generalization ability of a network [14], since it enables the network to explore more of the weight space and prevents the system from getting stuck in local minima; it plays an important role in most training procedures.

In this paper, a linear perceptron trained *with* dynamic noise *from* a noisy data set is studied within the formalism of Seung *et al*. A weight decay term is added to the training procedure in order to examine its generalization improvement capabilities. A weight decay has been shown to be a particular case of a discrete distribution of weights [12]. The use of this penalty term eliminates the need for a spherical constraint on the weight vector and thus comparisons with the results of Krogh and Hertz's method may also be drawn. The perceptron is trained on a training data set generated by adding Gaussian noise to the output of a known linear perceptron. We can therefore compare the network-generated solution (*the student*) with the known generator of the training data (*the teacher*) to calculate the generalization ability of the student.

In section 2, the model used to calculate the generalization and training errors is outlined. In section 3 the replica calculation is introduced with the associated order parameters to evaluate the average generalization and training errors. Section 4 presents results for the various limits of interest and draws some comparisons with the results of Krogh and Hertz. Section 5 uses the results derived in the previous sections to optimize the training parameters. The conclusions are discussed in section 6.

2. The model

In this section the model is presented, along with an outline of the calculation of the average generalization and training errors. The calculation follows that done by Seung *et al*, but alters the formalism to include a weight decay term and noise on the inputs.

A linear perceptron with N inputs is under study. The output, σ of the network is given as a function of the continuous valued weight vector $\boldsymbol{W}$ and input vector $\boldsymbol{s}$ by

$$\sigma = \sum_{i=1}^{N} W_i s_i .$$

The network is trained on a data set consisting of p input–output pairs $\{(s^l, \sigma^l); l = 1, \ldots, p\}$ generated by a teacher perceptron, $\boldsymbol{W}^0$, and corrupted by Gaussian noise. That is, $\sigma^l = (\boldsymbol{W}^0 \cdot s^l + \eta^l)$. The noise η is zero-mean Gaussian noise of variance γ^2 and the length of the teacher, Ω, is given by $\Omega^2 = (1/N)\boldsymbol{W}^0 \cdot \boldsymbol{W}^0$.

The training energy, E_t, is defined to be

$$E_t = \sum_{l=1}^{p} \epsilon(\boldsymbol{W}; s^l) + \tfrac{1}{2}\lambda \boldsymbol{W}^2 \tag{1}$$

where the error measure we use is

$$\epsilon(\boldsymbol{W}; s^l) = \tfrac{1}{2}\left[N^{-1/2}(\boldsymbol{W} - \boldsymbol{W}^0) \cdot s^l + \eta^l\right]^2$$

and λ is the positive weight decay parameter. The number of patterns per weight, $\alpha = p/N$.

The training process used is stochastic, that is dynamic noise with variance $2T$ is added to the weight update equation. Asymptotically, this results in a Gibbs distribution of student weights and the partition function

$$Z = \int \mathrm{d}\mu(\boldsymbol{W}) \exp\left\{ -\beta \sum_{l=1}^{p} \epsilon(\boldsymbol{W}; \boldsymbol{s}^l) \right\} \tag{2}$$

where $\beta = 1/T$. By analogy with thermodynamic terminology, T is called the temperature. The weight decay term is taken into the *a priori* distribution of student weights; $\mathrm{d}\mu(\boldsymbol{W}) = \mathrm{e}^{-\beta\lambda \boldsymbol{W}^2/2}\,\mathrm{d}\boldsymbol{W}$. The weight decay can be seen to be equivalent to taking an *a priori* Gaussian distribution of student weights with variance $1/\beta\lambda$. The Gibbs distribution can be used to calculate averages over the distribution of student weights, the so called *thermal* average, denoted by $\langle\ \rangle_T$.

This formalism involves the use of the canonical distribution for the student weights and a microcanonical distribution for the teacher weights (they have fixed length). In the thermodynamic limit, the *a priori* student distribution is equivalent to a microcanonical distribution with the weight vector constrained to be of length $\sqrt{N/\beta\lambda}$. This suggests there may be some scale invariance associated with scaling of the student weights, the temperature, β and the weight decay λ. If the error measure, $\epsilon(\boldsymbol{W}; \boldsymbol{s})$, is independent of the length of the student weight (as for a hard threshold), then a scale transformation in the student weights can be absorbed by a renormalization in the weight decay parameter λ. This is not the case for the linear perceptron; in this case, if the error measure depends only on terms quadratic in the student weights, a renormalization of β would leave the student distribution unchanged. However, the presence of the noise means that any scale transformation in the weights cannot be absorbed by a renormalization of β or λ without a corresponding transformation in the noise distribution. Thus there is no simple scaling property of the system studied in this paper.

The *a posteriori* student distribution is dependent on the instances of the training set. We can remove this dependence by averaging over the training set. This average is called a *quenched* average and is denoted by $\langle\!\langle\ \rangle\!\rangle$.

3. The replica calculation

In this section, the free energy and hence the generalization and training errors are calculated by the replica method.

3.1. Free energy

The calculation follows the formalism of Seung *et al*, in this case, there is also static noise on the training patterns which means that whenever a quenched average is performed, an average over the training data noise must also be done, since the actual manifestation of the noise affects the student weight vector and hence the average generalization and training errors.

The order parameters that evolve from the calculation are defined to be

$$R_\mu = \frac{1}{N} \boldsymbol{W}^\mu \cdot \boldsymbol{W}^0 \tag{3}$$

$$Q_{\mu\nu} = \frac{1}{N} \boldsymbol{W}^\mu \cdot \boldsymbol{W}^\nu \tag{4}$$

which represent the overlap between the teacher and solution weights, and between the solution weights respectively; μ and ν are replica indices.

The replica symmetric ansatz is assumed, so that the order parameters become

$$Q_{\mu\nu} = q_0\delta_{\mu\nu} + q_1(1-\delta_{\mu\nu})$$

$$R_\mu = R.$$

The parameters q_0, q_1 and R are defined by the equations above and $\delta_{\mu\nu}$ is the Kronecker delta. The 'conjugate' order parameters $\hat{q}_0$, $\hat{q}_1$, $\hat{R}$ arrive through introducing integral representations for the delta functions corresponding to the order parameters. The conjugate order parameters are related to the local field acting on the student weight vector [11]. Following the procedure described in Seung *et al* and after some calculation, the free energy per weight is obtained

$$\beta f = R\hat{R} + q_0\hat{q}_0 - \frac{1}{2}q_1\hat{q}_1 + \frac{1}{2}\ln(\beta\lambda - 2\hat{q}_0 + \hat{q}_1) - \frac{1}{2}\frac{\hat{R}^2\Omega^2 + \hat{q}_1}{\beta\lambda - 2\hat{q}_0 + \hat{q}_1}$$
$$+ \frac{\alpha}{2}\ln(1+\beta(q_0-q_1)) + \frac{\alpha\beta}{2}\frac{\Omega^2+\gamma^2+q_1-2R}{1+\beta(q_0-q_1)} - \frac{1}{2}\ln(2\pi). \tag{5}$$

In the thermodynamic limit, only the extrema of the free energy per weight, f, contribute. Thus equation (5) must be extremized over q_0, q_1, R, $\hat{q}_0$, $\hat{q}_1$, $\hat{R}$, giving the saddle point equations

$$q_0 = q_1 + \frac{1}{\beta\lambda + \hat{R}} \tag{6}$$

$$q_1 = (\hat{R}^2\Omega^2 + \hat{q}_1)(q_0 - q_1)^2 \tag{7}$$

$$\hat{q}_0 = \tfrac{1}{2}(\hat{q}_1 - \hat{R}) \tag{8}$$

$$\hat{q}_1 = \frac{\alpha\beta^2(\Omega^2+\gamma^2+q_1-2R)}{(1+\beta(q_0-q_1))^2} \tag{9}$$

$$R = \hat{R}\Omega^2(q_0 - q_1) \tag{10}$$

$$\hat{R} = \frac{\alpha\beta}{1+\beta(q_0-q_1)} \tag{11}$$

where it is assumed that the number of training examples scales as α times the number of weights, that is $\alpha = p/N$.

It is straightforward to check that these equations reduce to those presented by Seung *et al* when $q_0 = 1$ and $\Omega = 1$ corresponding to a spherical normalization on the student and teacher weights. This result will be discussed in more detail in the next section. The form of the free energy presented in equation (5) is similar to that given by Seung *et al* for a linear perceptron learning a teacher with an unrealizable threshold, that is $\sigma^l = (\boldsymbol{W}^0 \cdot \boldsymbol{s}^l + \theta)$, where θ is the teacher threshold/bias. The free energy is identical if the spherical constraint is assumed and $\beta\lambda \to 0$, that is zero weight decay. In this case the bias is identified with the standard deviation of the noise on the teacher. Thus in this limit, an unrealizable bias on the teacher is actually equivalent to adding noise to the teacher, which is then averaged.

The saddle point equations can be solved simultaneously to give

$$q_0 = q_1 + \mathcal{Q} \tag{12}$$

$$q_1 = \frac{\alpha(\Omega^2 + \gamma^2 + \alpha\Omega^2)}{\phi^2 - \alpha} - \frac{2\alpha^2\Omega^2}{\phi(\phi^2 - \alpha)} \tag{13}$$

$$\hat{q}_0 = \tfrac{1}{2}(\hat{q}_1 - \hat{R}) \tag{14}$$

$$\hat{q}_1 = \frac{\alpha\beta^2\lambda^2(\phi^2(\Omega^2 + \gamma^2) + \alpha\Omega^2(\alpha - 2\phi))}{(\phi^2 - \alpha)(\phi - \alpha)^2} \tag{15}$$

$$R = \frac{\alpha\Omega^2}{\phi} \tag{16}$$

$$\hat{R} = \frac{\alpha\beta\lambda}{\phi - \alpha} \tag{17}$$

where

$$\mathcal{Q}(\alpha, \beta, \lambda) = \frac{1}{2\beta\lambda}(1 - \alpha - \lambda) \pm \frac{1}{2\beta\lambda}\sqrt{(1 - \alpha - \lambda)^2 + 4\lambda} \tag{18}$$

$$\phi(\alpha, \lambda) = \lambda + \lambda\beta\mathcal{Q} + \alpha. \tag{19}$$

The definition of ϕ is independent of β.

Clearly from equation (4) and the replica symmetric ansatz, q_0 is the average length squared of a student weight vector, and thus q_1/q_0 is the average normalized overlap between replicas. This varies from zero, corresponding to an infinite number of student solutions, to one, corresponding to a single student solution. $R/\Omega\sqrt{q_0}$ is the average normalized overlap between a solution and the teacher vector, which again varies between zero and one. Since $q_1/q_0 \leqslant 1$, $\mathcal{Q} = (q_0 - q_1) \geqslant 0$. Thus we need only consider the positive root of $\mathcal{Q}$.

3.2. Generalization error

Following Seung *et al*, the average generalization error, $\epsilon_g = \langle\!\langle \langle \epsilon(\boldsymbol{W}) \rangle_T \rangle\!\rangle$, is a measure of a network's inability to solve a problem averaged over the entire data set. Assuming random, Gaussian distributed test patterns the generalization function, $\epsilon(\boldsymbol{W})$, is given by

$$\epsilon(\boldsymbol{W}) = \int \mathrm{d}\mu(s)\, \frac{1}{2N}[(\boldsymbol{W} - \boldsymbol{W}^0) \cdot s]^2$$

$$= \frac{1}{2N}\boldsymbol{W}^2 + \frac{1}{2}\Omega^2 - \frac{1}{N}\boldsymbol{W} \cdot \boldsymbol{W}^0. \tag{20}$$

This generalization function measures the network's performance at learning the uncorrupted teacher. If the student weight vector exactly equals that of the teacher $\epsilon_g = 0$; this function resembles that studied by Krogh and Hertz [6] up to a factor of two which results from a difference in the definition of the generalization function. The student could, however, be compared with the corrupted output; this would add an extra term of $\frac{1}{2}\gamma^2$ to ϵ_g to take account of the student's inability to learn the uncertainty in the teacher as studied by Seung *et al*. Hereafter, ϵ_g will refer to the network's generalization error for learning

the 'clean' teacher. When the generalization error in comparison with the corrupted teacher is referred to, the notation ϵ_g' will be used, that is $\epsilon_g' = \epsilon_g + \gamma^2/2$.

Taking the thermal average of the generalization function followed by the average over the quenched disorder gives the generalization error, again following Seung *et al*

$$\epsilon_g = \tfrac{1}{2}(\Omega^2 + q_0 - 2R). \tag{21}$$

The generalization error depends on the parameter Ω^2, representing the square length of the teacher weight vector. This is as expected, since a linear perceptron is used, the error is an absolute error and therefore the larger the weight vectors, the bigger the errors. The form of the generalization error ϵ_g' is similar to that given by Seung *et al* for an unrealizable teacher as discussed earlier.

3.3. Training error

The average training error ϵ_t is defined as

$$\epsilon_t = \langle\!\langle \langle E_t \rangle_T \rangle\!\rangle$$
$$= \frac{1}{\alpha}\frac{\partial(\beta f)}{\partial \beta}.$$

where f is the free energy per node. This is an average measure of how badly the network does on its training data set. So, differentiating equation (5) and substituting in for the saddle point equations (7)–(11) yields

$$\epsilon_t = \frac{1}{2\alpha}\left(\lambda q_0 + \frac{R}{\beta\Omega^2} + \frac{\hat{q}_1}{\beta^2}\right). \tag{22}$$

The first term in the equation above arises from the weight decay term in the training energy, since q_0 is the normalized average length of the square of a student weight vector. The temperature dependence enters the equation through the first and second terms. The final term is temperature *independent*, since from equation (15) the β^2 denominator is cancelled.

4. Network performance for various limits

Using the general expressions derived in the previous section, the system's performance in the limits of the main parameters, T, λ and α will be examined. Where the limits correspond to results from either Krogh and Hertz [6] or Seung *et al* [11] comparisons will be drawn.

4.1. Zero T and zero λ

The $T \to 0$ limit corresponds to training with no dynamic noise and the $\lambda \to 0$ limit corresponds to training with no weight decay term, and therefore no constraint on the weights. However, in this case, the integration over the weights performed in the evaluation of the partition function (2) is infinite and so $\Lambda = \beta\lambda$ is defined, with Λ finite as $T, \lambda \to 0$, giving a finite distribution of weights allowing the weight decay term to control the noise effects on weight growth. In this limit, equation (18) gives two solutions

$$Q = \begin{cases} (1-\alpha)/\Lambda & \alpha \leqslant 1 \\ 0 & \alpha > 1. \end{cases}$$

Substituting these into the solved saddle point equations, (12), (13) and (16), gives

$$q_0 = \begin{cases} \alpha\Omega^2 + (1-\alpha)/\Lambda + \alpha\gamma^2/(1-\alpha) & \alpha \leqslant 1 \\ \Omega^2 + \gamma^2/(\alpha - 1) & \alpha > 1 \end{cases} \tag{23}$$

$$q_1 = \begin{cases} \alpha\Omega^2 + \alpha\gamma^2/(1-\alpha) & \alpha \leqslant 1 \\ \Omega^2 + \gamma^2/(\alpha - 1) & \alpha > 1 \end{cases} \tag{24}$$

$$R = \begin{cases} \alpha\Omega^2 & \alpha \leqslant 1 \\ \Omega^2 & \alpha > 1. \end{cases} \tag{25}$$

Substituting these values into equation (21) yields

$$\epsilon_g = \begin{cases} \frac{1}{2}(1-\alpha)(\Omega^2 + 1/\Lambda) + \alpha\gamma^2/2(1-\alpha) & \text{for } \alpha \leqslant 1 \\ \gamma^2/2(\alpha - 1) & \text{for } \alpha > 1. \end{cases} \tag{26}$$

The average training error is given by equation (22)

$$\epsilon_t = \begin{cases} 0 & \alpha \leqslant 1 \\ (\gamma^2/2\alpha)(\alpha - 1) & \alpha > 1. \end{cases} \tag{27}$$

It can be seen from equation (23) that the effect of the noise is to lengthen the student vectors. This in turn means that the generalization error is expected to be higher in the presence of noise since, for a linear perceptron, the student can only generalize perfectly when it is coincident with the teacher and the student cannot coincide with the teacher if it is a different length. This is confirmed in equation (26). The effect of noise can be cancelled by presenting more patterns, that is increasing α. From equation (26), for $\alpha \leqslant 1$ the Λ parameter can be increased to reduce the generalization error. Since these equations are only valid in the zero T, λ limit, this corresponds to taking T to zero faster than λ, that is, training with an infinitesimally small but finite weight decay.

There is a discontinuity at $\alpha = 1$ in the average generalization error. Since, for this value of α there are only just enough examples to specify the N weights and the added noise in the data makes the equations unsolvable, resulting in a sudden increase in average error. Then as more patterns are presented, the student weights approach the teacher and so the errors decrease. Above $\alpha = 1$ the training error is increased from zero in the presence of noisy data, as seen in equation (27). This is because as α increases above one, the network cannot learn the random noise present on the data.

Some of the results presented by Krogh and Hertz [6] are equivalent to taking the zero T, λ limit with $\Lambda \to \infty$ and normalizing the teacher vector to be of length one, that is $\Omega^2 = 1$. Inserting these limits into equation (26), gives

$$\epsilon_g = \begin{cases} \frac{1}{2}(1-\alpha) + \alpha\gamma^2/2(1-\alpha) & \alpha \leqslant 1 \\ \gamma^2/2(\alpha - 1) & \alpha > 1 \end{cases} \tag{28}$$

which is in exact agreement with the generalization error calculated for a linear perceptron by Krogh and Hertz [6] apart from the factor of two due to the difference in definition.

4.2. Zero T and finite λ

In this section, the temperature is held at zero whilst the weight decay is allowed to increase. This corresponds to learning with a weight decay, but no dynamic noise on the weights. At zero temperature, there is only one solution to equation (18), that is, $\mathcal{Q} = 0$. This implies that there is only one solution in the student weight space. In this case, the number of replicas tends towards one.

The training error at zero temperature is given by

$$\epsilon_t = \frac{\lambda}{2}\left(\Omega^2 - \frac{(\lambda\Omega^2 - \gamma^2)}{(\phi^2 - \alpha)}\left(1 + \frac{\lambda\phi^2}{(\phi - \alpha)^2}\right)\right) \tag{29}$$

where from equation (19), $\phi = \frac{1}{2}\left(1 + \alpha + \lambda + \sqrt{(1+\alpha+\lambda)^2 - 4\alpha}\right)$.

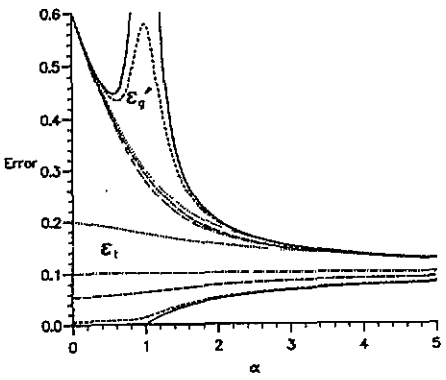

Figure 1. The training error ϵ_t (lower curves) and generalization error ϵ_g' (upper curves) against the number of patterns per weight, α for different values of the weight decay parameter, λ. The data set was corrupted by noise of variance $\gamma^2 = 0.2$. The full curve is for $\lambda = 0.0$, the broken curve is $\lambda = 0.01$, the long broken curve is $\lambda = 0.1$, the chain curve is $\lambda = 0.2$ and the dotted curve is $\lambda = 0.5$ This figure shows the asymptotic behaviour of the training and generalization errors as predicted by Seung *et al.*

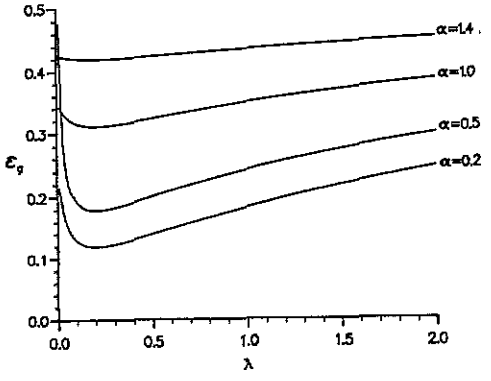

Figure 2. The generalization error ϵ_g against the weight decay parameter λ for different values of α.

It is clear from equation (29) that when the weight decay equals the variance of the noise on the data, γ^2, divided by the squared length of the teacher, the training error is constant. This can be seen in figure 1 in which the training and generalization errors are plotted for noise of variance, $\gamma^2 = 0.2$ and a teacher of normalized length one ($\Omega^2 = 1$).

The effect of a weight decay is to increase the initial training error above zero. This is because a non-zero weight decay term automatically adds a penalty term to equation (1) as well as enforcing a certain weight vector length. The result of Seung *et al* that the training and generalization error approach the same value from below and above respectively for large α can be seen to hold for values of $\lambda < \gamma^2$, but does not hold as the weight decay increases above this value since here the weight decay term eliminates part of the data, increasing both the training and generalization errors.

The optimum weight decay parameter estimated in section 5 can be picked out in figure 2, however, the minimum is not significantly lower than the surrounding generalization error for larger values of α.

4.3. Zero γ

The zero γ limit corresponds to having an uncorrupted data set. This was studied by Seung *et al* using a spherical constraint on the weights; hence we should be able to draw some comparisons between a weight decay term and a spherical constraint on the weights. Applying the same procedure using a weight decay term requires an additional order parameter, q_0, and its conjugate, $\hat{q}_0$. The parameter q_0 is associated with the mean length of a student weight vector. Using a spherical constraint, the length of the student weight vector, q_0, as well as the length of the teacher vector, Ω are constrained to be one.

In the zero T, λ limit with $\gamma^2 = 0$, from equation (23)

$$q_0 = \begin{cases} \alpha\Omega^2 + (1-\alpha)/\Lambda & \alpha \leqslant 1 \\ \Omega^2 & \alpha > 1. \end{cases}$$

Therefore having a Gaussian distribution of weights such that $\Lambda = 1/\Omega^2$ will result in $q_0 = \Omega^2$ for all α. Therefore the average length of a student weight vector is the same as the length of the teacher. This is similar to a spherical constraint though not the same, since it is the average rather than actual length of the student weight vector which lies on the sphere. To mimic the spherical constraint of Seung *et al*, Ω is set to be one. The saddle point equations of Seung *et al* for a linear perceptron with continuous weights are then homomorphic with the saddle point equations (6)–(11).

From equations (23), (24) with zero γ, T, λ and assuming the distribution of weights above, that is $\Lambda = 1/\Omega^2$

$$\frac{q_1}{q_0} = \begin{cases} \alpha & \alpha \leqslant 1 \\ 1 & \alpha > 1. \end{cases}$$

Hence for $\alpha > 1$, the average overlap between replicas is one and therefore all the replicas tend towards the same vector and so there is only one solution within the student weight space. For $\alpha \leqslant 1$, the number of possible student solutions is greater than one due to the fact that there are less than N equations specifying N unknowns, therefore the system has some freedom to find a solution. The same distribution of weights gives, from equation (25)

$$\frac{R}{\Omega\sqrt{q_0}} = \begin{cases} \alpha & \alpha \leqslant 1 \\ 1 & \alpha > 1. \end{cases}$$

The average overlap with the teacher tends towards one as α increases through one. This then makes the generalization error $\epsilon_g = 0$, as can be seen from equation (26). For $\alpha > 1$ the training set more than specifies the student and so there can only be one solution, the totally correct one. In the region, $\alpha \leqslant 1$, $\epsilon_g = (1-\alpha)\Omega^2$. The above results for $\Omega^2 = 1$ again agree with the results of Seung *et al*.

At zero temperature, zero λ and zero γ^2, ϵ_t is zero for all α. This is as expected, since in this case the student can always learn the data set exactly, that is, the problem is realizable. At finite values of λ, with zero static noise, the training and generalization errors increase from their values at $\lambda = 0$, as can be seen in figure 3. At finite temperature, the errors are increased. The presence of the weight decay in the training energy equation (1) causes the problem to be unrealizable, and therefore the generalization error and training error can never be zero.

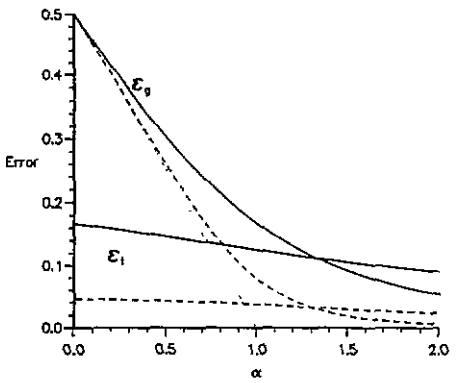

Figure 3. The generalization error ϵ_g (upper curves) and training error ϵ_t (lower curves) for zero noise on the data set ($\gamma = 0$) and zero temperature. The broken curves are for $\lambda = 0.1$ and the full curves are $\lambda = 0.5$. The dotted curve is the generalization error for zero λ.

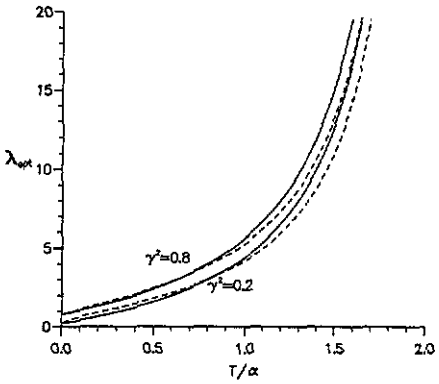

Figure 4. The optimum weight decay parameter, λ plotted against the temperature divided by the number of patterns per weight, T/α. The top two curves are for $\gamma^2 = 0.8$ and the bottom two $\gamma^2 = 0.2$. The full curves are for normalized number of patterns $\alpha = 1.0$ and the dotted curves are for $\alpha = 0.2$.

4.4. Large α

The large α limit is of interest, since it displays the behaviour of the network as the number of training patterns is increased significantly. The approximations for various order parameters are to second order in $1/\alpha$

$$q_0 = \Omega^2\left\{1 + \frac{1}{\alpha}\left(\frac{T}{\Omega^2} + \frac{\gamma^2}{\Omega^2} - 2\lambda\right) + O(\alpha^{-2})\right\} \tag{30}$$

$$\frac{q_1}{q_0} = 1 - \frac{T}{\alpha\Omega^2} + O(\alpha^{-2}) \tag{31}$$

$$\frac{R}{\sqrt{q_0}\Omega} = 1 - \frac{1}{\alpha}\left\{\frac{T}{\Omega^2} + \frac{\gamma^2}{\Omega^2}\right\} + O(\alpha^{-2}) \tag{32}$$

$$\epsilon_g = \frac{1}{2\alpha}(T + \gamma^2) + O(\alpha^{-2}) \tag{33}$$

$$\epsilon_t = \frac{1}{2}\left\{\gamma^2 + \frac{1}{\alpha}(T - \gamma^2 + \lambda\Omega^2)\right\} + O(\alpha^{-2}). \tag{34}$$

These results are in agreement with those presented by Seung *et al* for large α. The large α behaviour of the perceptron is similar to that predicted using a spherical constraint, except, the training error is increased by the presence of a weight decay parameter. The weight decay has no influence on the generalization error or the average overlap with the teacher at large enough α, thus training with a weight decay cannot degrade generalization for a large training set. The average number of solutions is unaffected by either the weight decay or the noise on the data set from equation (31).

4.5. Large λ

The large λ limit corresponds to narrowing the distribution from which the student weight vectors are drawn. The generalization error in this limit is

$$\epsilon_g = \frac{1}{2}\left\{\Omega^2 + \frac{(T-2\alpha)\Omega^2}{\lambda} - \frac{\alpha}{\lambda^2}(3\Omega^2(\alpha+1) - \gamma^2 + T) + O(\lambda^{-3})\right\}. \tag{35}$$

Thus for large weight decays, the amount of noise in the data set has little effect on the generalization error since it appears at $O(\lambda^{-2})$. The order λ^{-1} term increases the error for $T > 2\alpha$ which suggests that with large weight decays, better results can be obtained by training at temperatures less than 2α; further comments are made in the next section

5. Optimal weight decay parameter

It has been shown by Krogh and Hertz that there exists a λ_{opt} which minimizes the generalization energy. This can be found by differentiating ϵ_g at finite λ.

Given the generalization error from equation (21)

$$\frac{\partial \epsilon_g}{\partial \lambda} = \frac{1}{2}\frac{\partial q_0}{\partial \lambda} - \frac{\partial R}{\partial \lambda}$$

$$= \left[\frac{\alpha\Omega^2\phi(\lambda - \gamma^2/\Omega^2)}{(\phi^2 - \alpha)^2}\right]\frac{\partial \phi}{\partial \lambda} + \frac{\partial \mathcal{Q}}{\partial \lambda}.$$

For zero T, $\partial\mathcal{Q}/\partial\lambda = 0$ and putting $\partial\epsilon_g/\partial\lambda = 0$ yields

$$\lambda_{opt} = \gamma^2/\Omega^2 \tag{36}$$

which for $\Omega^2 = 1$ agrees with the result of Krogh and Hertz.

Now consider the general case of finite T. At finite temperature the condition, $\partial\epsilon_g/\partial\lambda = 0$ gives

$$4\alpha\lambda^2\Omega^2\left(\lambda - \frac{\gamma^2}{\Omega^2}\right) + T\psi^{3/2}(a-1) - T\psi(\psi - \lambda(1+\alpha+\lambda)) = 0$$

where $\psi = ((1+\alpha+\lambda)^2 - 4\alpha)$.

This equation may be solved numerically and the results for $\Omega = 1$ are presented in figure 4. The solutions tend to infinity as $\alpha\beta \to 0.5$ from below, above this temperature, the optimum λ is infinite. This value of the weight decay parameter corresponds to having a prior weight distribution of zero weights, i.e. a random guess gives the best generalization. Equation (35) agrees with this result since above $\alpha\beta = 0.5$ the generalization error gets worse. The result can be explained in terms of the signal-to-noise ratio and is therefore natural. Thus an initial temperature for an annealing schedule may be postulated.

For the larger values of T, the optimum generalization error is not significantly less than the surrounding values and therefore training at λ_{opt} is not strictly necessary. The effect of the noise on the training set is to increase the optimal λ at low values of T/α.

6. Conclusion

A linear perceptron was studied using a statistical mechanics formalism based on the replica method as presented by Seung *et al* [11]. The network's training data set was generated by a known teacher perceptron and then corrupted by additive Gaussian noise. The perceptron was then trained stochastically to learn the hidden rule by minimizing an energy function (the training energy) with respect to the network's weights. A weight decay term was introduced into the calculation eliminating the need for a spherical constraint as used in previous calculations. The training energy was used to generate a Gibbs distribution on the possible student weight vectors for asymptotic times and from this distribution the average properties of the network were calculated.

The relationship between the weight decay and a spherical constraint was investigated; the spherical constraint was interpreted as a special case of the weight decay in the zero temperature limit. The zero weight decay limit was shown to make sense only in terms of taking the zero temperature limit at the same time, this limit agreed with the results of Krogh and Hertz [6]. The form of the free energy was seen to be similar to that given by Seung *et al* for the case of an unrealizable threshold on the teacher in the limit of zero weight decay. It is natural therefore to expect that the effect of noise on the training set can be reduced by training with a threshold.

The weight decay was shown to be useful in decreasing the generalization error when the data set contained noise and the best (minimum generalization error) weight decay parameter was calculated at finite training temperature. Adding noise to the data set produces a corrupted energy surface. Training a network stochastically smooths the energy surface to within limits prescribed by the value of the training temperature. However, training with a weight decay term smooths the energy surface by eliminating those weights with small eigenvalues and therefore only minor contributions to the underlying solution. This means that when training a network stochastically as well as with a weight decay term, the optimum generalization error is found at zero temperature and the value of the weight decay parameter is as described before.

There is a temperature ($T = 2\alpha$) above which the optimum weight decay is infinite. This can be interpreted as the value of the temperature where the noise swamps the data. This could suggest an initial temperature for an annealing schedule. When the weight decay is increased above the level of the noise on the training data set the training error is higher for low numbers of patterns (small α) and decreases as the number of patterns increases. This can be understood because for higher values of the weight decay parameter, the penalty term is removing some of the information contained in the data as well as the spurious information in the noise on the data.

Acknowledgments

We are very grateful to David Saad for many helpful discussions. APD is supported by an SERC CASE studentship in collaboration with British Gas Plc.

References

[1] McClelland J L and Rumelhart D E 1986 *Parallel Distributed Processing* vol 1 (Massachusetts, MA: MIT Bradford)

[2] Vapnik V N and Chervonenkis A Y 1971 *Theory Prob. Appl.* **16** 264–80

[3] Hertz J A, Krogh A and Palmer R G 1991 *Introduction to the Theory of Neural Computation* (Redwood City, CA: Addison-Wesley)
[4] Watkin T H L, Rau A and Biehl M 1992 The statistical mechanics of learning a rule *Preprint* Oxford University
[5] Hertz J A, Krogh A and Thorbergsson G I 1989 *J. Phys. A: Math. Gen.* **22** 2133–50
[6] Krogh A and Hertz J A 1992 *J. Phys. A: Math. Gen.* **25** 1135–47
[7] Gardner E 1988 *J. Phys. A: Math. Gen.* **21** 257
[8] Gardner E and Derrida B 1988 *J. Phys. A: Math. Gen.* **21** 271
[9] Györgyi G and Tishby N 1990 Statistical theory of learning a rule *Neural Networks and Spin Glasses* ed W K Theumann and R Köberle (Singapore: World Scientific) pp 3–36
[10] Bös S, Kinzel W and Opper M 1992 *Phys. Rev.* E **47** 1384–91
[11] Seung H S, Sompolinsky H and Tishby N 1992 *Phys. Rev.* A **45** 6056–91
[12] de Menezes F S and Wallace D J 1993 Learning and generalization in neural network models: the effect of dynamic range of weights, in preparation
[13] Edwards S F and Anderson P W 1975 *J. Phys. F: Met. Phys.* **5** 965
[14] Gardner E, Stroud N and Wallace D J 1987 Training with noise: application to word and text storage *Neural Computers: From Computational Neuroscience to Computer Design* ed R Eckmiller (Berlin: Springer) pp 251–60