Learning with Regularizers in Multi-layer Neural Networks

David Saad and Magnus Rattray

Aston University, Computer Science & Applied Mathematics, Birmingham, B4 7ET, UK.

Abstract

We study the effect of regularization in an on-line gradient-descent learning scenario for a general two-layer student network with an arbitrary number of hidden units. Training examples are randomly drawn input vectors labelled by a two-layer teacher network with an arbitrary number of hidden units which may be corrupted by Gaussian output noise. We examine the effect of weight decay regularization on the dynamical evolution of the order parameters and generalization error in various phases of the learning process, in both noiseless and noisy scenarios.

I. INTRODUCTION

One of the most powerful and commonly used methods for training large layered neural networks is that of on-line learning, whereby the internal network parameters $\{\mathbf{J}\}$ are modified after the presentation of each training example so as to minimize the corresponding error. The goal is to bring the map $f_{\mathbf{J}}$ implemented by the network as close as possible to a desired map $\tilde{f}$ that generates the examples. Here we focus on the learning of continuous maps via gradient descent on a differentiable error function.

Recent work [1]– [5] provides a powerful tool for the analysis of gradient-descent learning in a very general learning scenario [6]: that of a *student* network with N input units, K hidden units, and a single linear output unit, trained to implement a continuous map from an N-dimensional input space $\boldsymbol{\xi}$ onto a scalar ζ. Examples of the target task $\tilde{f}$ are in the

form of input-output pairs $(\boldsymbol{\xi}^\mu, \zeta^\mu)$. The output label ζ^μ for each independently drawn input $\boldsymbol{\xi}^\mu$ is provided by a *teacher* network of similar architecture, except that its number M of hidden units is not necessarily equal to K.

Here we consider the effect of regularization on the learning process in the form of weight decay, for both noiseless learning and for the case where a noise process corrupts the teacher output. Learning from corrupted examples is a realistic and frequently encountered scenario and is commonly handled by some sort of regularization. Previous analysis of noisy training scenarios and the application of regularization have been based on various approaches: Bayesian [7], equilibrium statistical physics [8] and non-equilibrium techniques for analyzing learning dynamics [9]. Here we adapt our previously formulated techniques [2] to investigate the effect of different noise mechanisms on the dynamical evolution of the learning process and the resulting generalization ability.

II. THE MODEL

We focus on a *soft committee machine* [1], for which all hidden-to-output weights are positive and of unit strength. Consider the student network: hidden unit i receives information from input unit r through the weight J_{ir}, and its activation under presentation of an input pattern $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^{\mathrm{T}}$ is $x_i = \mathbf{J}_i \cdot \boldsymbol{\xi}$, with $\mathbf{J}_i = (J_{i1}, \ldots, J_{iN})^{\mathrm{T}}$ defined as the vector of incoming weights to the i-th hidden unit. The output of the student network is $\sigma(\mathbf{J}, \boldsymbol{\xi}) = \sum_{i=1}^{K} \; g\,(\mathbf{J}_i \cdot \boldsymbol{\xi})$, where g is the activation function of the hidden units, taken here to be the error function $g(x) \equiv \mathrm{erf}(x/\sqrt{2})$, and $\mathbf{J} \equiv \{\mathbf{J}_i\}_{1 \le i \le K}$ is the set of input-to-hidden adaptive weights.

The components of the input vectors $\boldsymbol{\xi}^\mu$ are uncorrelated random variables with zero mean and unit variance. Output labels ζ^μ are provided by a teacher network of similar architecture: hidden unit n in the teacher network receives input information through the weight vector $\mathbf{B}_n = (B_{n1}, \ldots, B_{nN})^{\mathrm{T}}$, and its activation under presentation of the input pattern $\boldsymbol{\xi}^\mu$ is $y_n^\mu = \mathbf{B}_n \cdot \boldsymbol{\xi}^\mu$. In the noiseless case the teacher output is given by $\zeta_0^\mu =$

$\sum_{n=1}^{M}\ g\ (\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu)$.

The error made by a student with weights $\mathbf{J}$ on a given input $\boldsymbol{\xi}$ is given by the quadratic deviation

$$\epsilon(\mathbf{J}, \boldsymbol{\xi}, \zeta_0) \equiv \frac{1}{2}[\ \sigma(\mathbf{J}, \boldsymbol{\xi}) - \zeta_0\ ]^2 = \frac{1}{2}\left[\ \sum_{i=1}^{K} g(x_i) - \sum_{n=1}^{M} g(y_n)\ \right]^2 \quad , \tag{1}$$

measured here with respect to the noiseless teacher (we will also consider teachers corrupted by output noise, in which case deviations are with respect to the actual noisy output ζ). Performance on a typical input in the absence of noise defines the generalization error $\epsilon_g(\mathbf{J}) \equiv \ \langle \epsilon(\mathbf{J}, \boldsymbol{\xi}, \zeta_\mathbf{0}) \rangle_{\{\xi\}}$, through an average over all possible input vectors $\boldsymbol{\xi}$ to be performed implicitly through averages over the activations $\mathbf{x} = (x_1, \ldots, x_K)^{\mathrm{T}}$ and $\mathbf{y} = (y_1, \ldots, y_M)^{\mathrm{T}}$. These averages can be performed analytically [2] and result in a compact expression for ϵ_g in terms of *order parameters*: $Q_{ik} \equiv \mathbf{J}_i \cdot \mathbf{J}_k$, $R_{in} \equiv \mathbf{J}_i \cdot \mathbf{B}_n$, and $T_{nm} \equiv \mathbf{B}_n \cdot \mathbf{B}_m$, which represent student-student, student-teacher, and teacher-teacher overlaps respectively. The parameters T_{nm} are characteristic of the task to be learned and remain fixed during training, while the overlaps Q_{ik} among student hidden units and R_{in} between student and teacher hidden units are determined by the student weights $\mathbf{J}$ and evolve during training.

A gradient descent rule for the update of student weights results in $\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^\mu + \frac{\eta}{N}\ \delta_i^\mu\ \boldsymbol{\xi}^\mu$, where the learning rate η has been scaled with the input size N, and $\delta_i^\mu \equiv g'(x_i^\mu)\left[\sum_{n=1}^{M} g(y_n^\mu) - \sum_{j=1}^{K} g(x_j^\mu)\right]$. The time evolution of the overlaps R_{in} and Q_{ik} can be written in terms of difference equations. We consider the large N limit and introduce a normalized number of examples $\alpha = \mu/N$ to be interpreted as a continuous time variable in the $N \to \infty$ limit. The time evolution of R_{in} and Q_{ik} is thus described in terms of a coupled set of first-order differential equations [2].

III. THE EFFECT OF REGULARIZERS

A common method to overcome the effects of noise and parameter redundancy, frequently used in real world training scenarios, is the use of regularizers such as weight decay [10]. The

role of regularizers has been analyzed for linear perceptrons [11–13] and optimal values for regularizers have been calculated. However, the efficacy of regularizers for on-line learning in multi-layer networks is still somewhat unclear. The effect of weight decay on the training equations is the subtraction of a term $\frac{\gamma}{N}\mathbf{J}_i^\mu$ in each weight update. The difference equation for $\mathbf{J}_i^\mu$ becomes

$$\mathbf{J}_i^{\mu+1} = \mathbf{J}_i^\mu + \frac{\eta}{N}\ \delta_i^\mu\ \boldsymbol{\xi}^\mu - \frac{\gamma}{N}\mathbf{J}_i^\mu\ , \tag{2}$$

and the resulting equations of motion for the student-teacher and student-student overlaps are given in this case by:

$$\begin{aligned} \frac{dR_{in}}{d\alpha} &= \eta\ \phi_{in} - \gamma\ R_{in}\ , \\ \frac{dQ_{ik}}{d\alpha} &= \eta\ \psi_{ik} + \eta^2\ \upsilon_{ik} - 2\gamma\ Q_{ik}\ , \end{aligned} \tag{3}$$

where $\phi_{in} \equiv \langle \delta_i\ y_n \rangle_{\{\xi\}}$, $\psi_{ik} \equiv \langle \delta_i\ x_k + \delta_k\ x_i \rangle_{\{\xi\}}$ and $\upsilon_{ik} \equiv \langle \delta_i\ \delta_k \rangle_{\{\xi\}}$. The explicit expressions [2] for ϕ_{in}, ψ_{ik}, υ_{ik} and ϵ_g depend exclusively on the overlaps Q, R and T. The only difference from the expressions in Ref. [2] is due to the presence of the weight decay terms. These equations can be solved numerically as demonstrated in Fig. 1 for the realizable training scenario of $M = K = 3$, $\eta = 0.2$ and an isotropic teacher ($T_{nm} = \delta_{nm}$). The basic features of the dynamics for both noisy and noiseless learning exist here, i.e., a short transient followed by a prolonged symmetric phase, characterized by lack of differentiation between different nodes of the student, specialization, as each student node begins to emulate a particular teacher node, and finally convergence to asymptotic values. The weight decay applied in this case, $\gamma = 0.005$ (Fig. 1(a), (c) and (d)), has a negligible effect on the location of the fixed point in the symmetric phase and on the value of the generalization error there; however, it does affect the length of the symmetric phase, the convergence phase and the asymptotic values of the order parameters and generalization error. The asymptotic values of the generalization error, the cross-correlation between vectors related to different nodes ($Q_{i\neq k}$) and overlaps between student vectors and teacher vectors imitated by different student vectors ($R_{i\neq n}$) increase with the weight decay γ, while the length of the teacher

vectors (Q_{ii}) and overlaps between student vectors and teacher vectors imitated by them (R_{ii}) decrease. Moreover, above a certain weight decay value γ_{max} the system is trapped indefinitely in the symmetric subspace as shown in Fig. 1(b), for the student overlaps, where a weight decay of $\gamma = 0.007$ is used. These effects will be analyzed in the following sections, although we are limited to the consideration of small learning rates. A different approach is introduced in section IV which allows us to determine the optimal weight decay as a function of time for arbitrary learning rates. We first attempt to derive analytical results for the dynamical behaviour during each phase of learning, with a constant weight decay parameter. For simplicity we will concentrate here on a noiseless, realizable learning scenario ($M = K$) with an isotropic teacher ($T = \delta_{nm}$).

A. The symmetric phase

Introducing weight decay modifies the fixed point during the symmetric phase. Following [2], we reduce the dimension of the system by exploiting symmetries in the dynamics which exist for realizable, isotropic learning: $Q_{ik} = Q\delta_{ik} + C(1 - \delta_{ik})$ and $R_{in} = R\delta_{in} + S(1 - \delta_{in})$ where each student node index coincides with that of the teacher node to which it will eventually specialize. One can then calculate the location of the symmetric fixed point, for small learning rates and small values of the regularization parameter, by truncating Eqs. (3) to first order in η and expanding with respect to $\tilde{\gamma} = \gamma/\eta$, regarding the solution with weight decay as a small perturbation around the $\gamma = 0$ result, $S^* = R^* = 1/\sqrt{K(2K-1)}$ and $Q^* = C^* = 1/(2K-1)$. Solving the truncated equations results in the following expressions for the new fixed point and generalization error ($S = R$ at the symmetric fixed point):

$$R^* = \frac{1}{\sqrt{K\,(2\,K-1)}} - \frac{(2\,K+1)^{3/2}\,(4\,K^2-7\,K+2)}{\pi\,K^{7/2}\,(2\,K-1)\,(2\,K-3)}\tilde{\gamma}\ ,$$
$$Q^* = \frac{1}{2\,K-1} - \frac{2\,(2\,K+1)^{3/2}\,(4\,K^2-6\,K+1)}{\pi\,(2\,K-1)^{3/2}\,(2\,K-3)\,K^2}\tilde{\gamma}\ , \qquad (4)$$

$$C^* = \frac{1}{2\,K-1} + \frac{2\,(2\,K+1)^{3/2}}{\pi\,(2\,K-1)^{3/2}\,(2\,K-3)\,K^2}\tilde{\gamma}\ ,$$
$$\epsilon_g^* = \frac{K}{\pi}\left(\frac{\pi}{6} - K\arcsin\left(\frac{1}{2K}\right)\right)\ .$$

It is interesting to note that weight decay does not modify the generalization error to first order in $\tilde{\gamma}$.

For the case shown in Fig. 1(a), (c) and (d) ($\gamma = 0.005$) we evaluated the overlaps and the generalization error to obtain: $R^* = 0.2564$, $Q^* = 0.1908$ and $C^* = 0.2005$, in close agreement with the result presented above.

A significant difference to the dynamics without weight decay is a notable reduction in the gap between the values of Q and C in the symmetric phase which may be attributed to the suppression of excessive vector length by the weight decay mechanism. This inevitably leads to higher similarity between student vectors and a delay in leaving the symmetric phase.

To investigate the effect of weight decay on the length of the symmetric phase we expanded the truncated dynamical equations, derived from Eqs.(3), around the fixed point $\{R^*, S^*, Q^*, C^*\}$ to obtain the eigenvalues which control the dynamics of the system and escape from the symmetric phase. The dynamical evolution described by the linearized equations of motion is characterized by three eigenvalues, one of which, $\lambda = \lambda_0 + \lambda_\gamma\tilde{\gamma}$, is positive and controls the escape, where

$$\lambda_\gamma = -\frac{\pi}{2}\frac{16\,K^5 - 16\,K^4 - 36\,K^3 + 22\,K^2 + 13\,K - 8}{2\,K^2\,(2\,K+1)\,(2\,K-1)\,(2\,K-3)}$$

and λ_0 is the eigenvalue obtained for the dynamics in the absence of weight decay [2]. Dependence of λ_γ on the number of hidden units K is shown in Fig. 2(a), approaching the asymptotic value of $-\pi/2$ as $K \to \infty$. The dependence on the weight decay is negative, suppressing the eigenvalue responsible for escape from the symmetric phase. The system will escape from the symmetric phase for weight decay values lower than $\tilde{\gamma}_{max}$ where

$$\tilde{\gamma}_{max} = \frac{8\,K^3\sqrt{2\,K-1}\,(2\,K-3)}{\pi^2\sqrt{2\,K+1}\,(16\,K^5 - 16\,K^4 - 36\,K^3 + 22\,K^2 + 13\,K - 8)}$$

for which $\lambda = 0$. The dependence of γ_{max} on the number of hidden units K is shown in Fig. 2(b), which decays asymptotically as $1/(K\pi^2)$. For the conditions of Fig. 1, i.e. $K = M = 3$ and $\eta = 0.2$, the maximal weight decay is $\eta\tilde{\gamma}_{max} = 0.006$ in agreement with the numerical solutions shown there.

This analysis has been carried out for the case of small learning rate which is most easily amenable to analysis. However, the more realistic case of larger η (which includes, for example, the optimal learning rate) is characterized by a different behavior with respect to γ. Analysing the large η case requires new tools and will be discussed in section IV.

B. The asymptotic regime

Asymptotically, in the realizable noiseless case with no weight decay, the secondary overlaps S and C decay to zero while R and Q approach unity, indicating full alignment for an isotropic task ($T_{nm} = \delta_{nm}$). We observe that in the presence of weight decay the student vectors converge to asymptotic values which are shorter than the teacher vectors: $Q_{ii} \to Q_\infty < 1$ and acquire a positive correlation with each other. Shorter norms for the student vectors result in a larger asymptotic generalization error.

The asymptotic phase is characterized by a fixed point solution with $R^* \neq S^*$. The coordinates of the asymptotic fixed point can also be obtained analytically in the small η approximation: $R^* = 1 + \tilde{\gamma}\ r_a$, $S^* = -\tilde{\gamma}\ s_a$, $Q^* = 1 + \tilde{\gamma}\ q_a$, and $C^* = -\tilde{\gamma}\ c_a$, with

$$r_a = \frac{12\pi\left(9+8\sqrt{3}\right)\left(3\,K-6+2\sqrt{3}\right)}{111\,K-159+56\sqrt{3}}\ ,$$
$$s_a = \frac{18\pi\left(9+8\sqrt{3}\right)}{111\,K-159+56\sqrt{3}}\ ,\quad q_a\ =\ 2r_a\ ,\quad c_a\ =\ 2s_a\ .$$

The asymptotic generalization error vanishes for the first order in γ. Expanding the asymptotic order parameters to second order in γ, one obtains for leading order in η

$$\epsilon_g^* = \frac{6\pi\left(9+8\sqrt{3}\right)\left(3\,K-6+2\sqrt{3}\right)\tilde{\gamma}^2\,K}{\left(111\,K-159+56\sqrt{3}\right)}\ . \tag{5}$$

To examine the accuracy of the results we plotted the predicted asymptotic values for the case $K = M = 3$ and $\eta = 0.2$ for various weight decay levels against the actual values obtained numerically as shown in Fig. 3: (a) shows predicted values for R and Q against actual values obtained numerically while (b) presents predicted values for S and C against actual values. The results presented in Fig. 3 show that the approximation for the asymptotic values for R, Q, S and C are very accurate for low weight decay values. Similarly, the predicted asymptotic value of the generalization error is in reasonable agreement with the numerical result; e.g., the asymptotic generalization error calculated for $K = M = 3$, $\eta = 0.2$ and $\gamma = 0.005$ shows a value of $\epsilon_g^* = 0.0193$ in comparison to the numerical result $\epsilon_g^* = 0.0169$.

C. Noisy examples and redundant parameters

From the analysis of the role played by the weight decay in the linear perceptron one would expect the weight decay to alleviate the problem of noise [11,12] and to suppress redundant parameters [13], reducing the generalization error. We therefore examined the effect of weight decay on various learning scenarios in which training examples are corrupted by noise and in the presence of redundant weights, for small and intermediate learning rates. Our numerical and analytical investigations have revealed no scenario, either when training from noisy data or in the presence of redundant parameters, where a *fixed* weight decay improves the system performance in the long run or speeds up the training process. For the asymptotic regime, especially in the case of noiseless systems with redundant units, this is probably a generic feature of on-line learning with an infinite data set, due to the absence of the numerous minima in the mean error surface which might be caused by a finite training set (i.e. the mean error is the generalization error in our case). In off-line (batch) learning, or on-line learning with recycled patterns, regularization may lead to improved performance through the modification of the error surface.

To demonstrate the effect of weight decay on the evolution of the generalization error

in the case of corrupted examples and in the presence of redundant parameters, we show in Fig. 4 two typical training scenarios where weight decay has been applied. We consider additive Gaussian output noise [4] so that the teacher output is $\zeta^\mu = \rho^\mu + \sum_{n=1}^{K} g\,(\mathbf{B}_n \cdot \boldsymbol{\xi}^\mu)$, where the random variable ρ^μ is taken to be Gaussian with zero mean and variance σ^2.

The example shown in Fig. 4(a) represents a training scenario were $M = K = 3$ and examples are corrupted by Gaussian output noise with variance $\sigma^2 = 0.1$. It is clear that employing weight decay, with $\gamma = 0.001 \ldots 0.003$ in this example, has only increased the asymptotic generalization error and delayed the breaking away from the symmetric phase. The slight increase in generalization error during the symmetric phase is due to higher order effects which are not analysed in this paper. Similar results have been obtained for different types and levels of noise and weight decay, including weight decay which varies in time according to hand crafted schedules.

Figure 4(b) shows an over-realizable training scenario in which a student with five hidden nodes is trained on uncorrupted examples generated by a three node teacher. The learning rate in this case is $\eta = 0.2$. Again it is clear that optimal performance is achieved with no regularizers.

Both these simulations used a rather low value for the learning rate, significantly lower than the optimal setting. In the next section we observe how the behaviour of weight decay is significantly different during the symmetric phase for larger learning rates.

IV. GLOBALLY OPTIMAL WEIGHT DECAY

In the previous sections we have been limited to using fixed or hand-crafted weight decay terms which restrict our ability to assess the potential contribution of general weight decay terms as only a limited number of conditions can be examined. In this section we take a different approach, aiming at global optimization of a time-dependent weight decay term on the basis of previous work on globally optimal learning rates [14] and learning rules [15].

An optimal learning scenario with respect to some parameter (here γ) in a certain time

window $[\alpha_0, \alpha_1]$ corresponds to the largest decrease in generalization error between these two times; i.e., we attempt to minimize $\Delta\epsilon_g = \epsilon_g(\alpha_1) - \epsilon_g(\alpha_0)$ which may be written as an integral of the form:

$$\Delta\epsilon_g = \int_{\alpha_0}^{\alpha_1} \frac{d\epsilon_g}{d\alpha}\, d\alpha \ . \tag{6}$$

Since the generalization error depends exclusively on the overlaps Q, R and T, for which the dynamical equations are known, one can rewrite the integrand $\mathcal{L} = \frac{d\epsilon_g}{d\alpha}$ as

$$\begin{aligned}\mathcal{L} = \sum_{in} \frac{\partial\epsilon_g}{\partial R_{in}}\frac{dR_{in}}{d\alpha} + \sum_{ik} \frac{\partial\epsilon_g}{\partial Q_{ik}}\frac{dQ_{ik}}{d\alpha} - \sum_{in} \mu_{in}\left(\frac{dR_{in}}{d\alpha} - \eta\ \phi_{in} + \gamma R_{in}\right) \\ - \sum_{ik} \nu_{ik}\left(\frac{dQ_{ik}}{d\alpha} - \eta\ \psi_{ik} - \eta^2\ \upsilon_{ik} + 2\gamma Q_{ik}\right)\end{aligned} \tag{7}$$

The last two right hand terms in Eq.(7) force the correct dynamics using sets of Lagrange multipliers μ_{in} and ν_{ik} for the corresponding equations $dR_{in}/d\alpha$ and $dQ_{ik}/d\alpha$.

Using variational techniques it is straightforward to obtain a set of coupled differential equations for the Lagrange multipliers:

$$\begin{aligned}\frac{d\mu_{km}}{d\alpha} &= \gamma\mu_{km} - \eta\sum_{in}\mu_{in}\frac{\partial\phi_{in}}{\partial R_{km}} - \eta\sum_{ij}\nu_{ij}\frac{\partial\left(\psi_{ij} + \eta\ \upsilon_{ij}\right)}{\partial R_{km}} \\ \frac{d\nu_{kl}}{d\alpha} &= 2\gamma\nu_{kl} - \eta\sum_{in}\mu_{in}\frac{\partial\phi_{in}}{\partial Q_{kl}} - \eta\sum_{ij}\nu_{ij}\frac{\partial\left(\psi_{ij} + \eta\ \upsilon_{ij}\right)}{\partial Q_{kl}}\ ,\end{aligned} \tag{8}$$

as well as a set of boundary conditions

$$\mu_{in}(\alpha_1) = \left.\frac{\partial\epsilon_g}{\partial R_{in}}\right|_{\alpha_1} \quad \text{and} \quad \nu_{ik}(\alpha_1) = \left.\frac{\partial\epsilon_g}{\partial Q_{ik}}\right|_{\alpha_1} \ . \tag{9}$$

A separate equation is derived for the functional derivative of $\Delta\epsilon_g$ with respect to γ, which we use for iteratively updating γ via gradient descent:

$$\gamma(t+1) = \gamma(t) - \theta\ \delta\Delta\epsilon_g/\delta\gamma\ , \tag{10}$$

where

$$\frac{\delta\Delta\epsilon_g}{\delta\gamma} = -\sum_{in}\mu_{in}R_{in} - 2\sum_{ij}\nu_{ij}Q_{ij} \tag{11}$$

Here, t is the iteration index and θ is the learning rate for the optimization process.

All terms required for carrying out the optimization of γ using Eq.(10) can be obtained by integrating the learning dynamics in Eqs.(3) forward from some initial conditions for the overlaps, and then integrating the Lagrange multiplier dynamics backwards, using Eqs.(8) and the boundary conditions in Eq.(9). This process converges after a number of iterations and results in an exact function for the optimal weight decay over the time window.

We have employed this method to derive the optimal weight decay coefficient in several cases: structurally realizable and over-realizable noiseless scenarios with optimal and small learning rates and structurally realizable and over-realizable noisy scenarios with optimal learning rates.

For small learning rates our results support the conclusions of section III. During the symmetric phase a very small or negative value is chosen for the optimal weight decay $\gamma_{\rm opt}$, indicating that weight decay is at best useless and possibly detrimental during this phase. After the symmetric phase $\gamma_{\rm opt}$ quickly approaches zero, as required in order to achieve zero generalization error asymptotically.

For larger learning rates, however, we do find a positive $\gamma_{\rm opt}$ which can shorten the symmetric phase significantly for both realizable and over-realizable learning scenarios. Fig. 5(a) shows the optimal weight decay for an over-realizable example ($M = 2$, $K = 3$) and the corresponding generalization error is shown by the solid line in Fig. 5(b). The generalization error for learning in the absence of weight decay is shown as the dashed line in Fig. 5(b) and we see how the weight decay results in a shortened symmetric phase. As expected, $\gamma_{\rm opt}$ falls quickly to zero as the generalization error converges towards zero. The learning rate chosen in this example ($\eta = 0.7$) is close to the optimal value in the absence of weight decay (as determined by similar methods to those employed here for the determination of $\gamma_{\rm opt}$ [14]) and we therefore see that the inclusion of weight decay can result in an improvement on the optimal performance of standard gradient descent learning. Notice that we do not optimize η and γ simultaneously here, as we are mainly concerned with the improvements due to weight decay given a fixed learning rate schedule. Similar results are found for realizable

learning scenarios with large or near optimal learning rates.

The picture developed above is not significantly altered by the inclusion of Gaussian output noise. Fig. 6(a) shows $\gamma_{\rm opt}$ for a structurally realizable task ($M = K = 2$) with noise variance $\sigma^2 = 0.01$. The learning rate is given its optimal time-dependent value in the absence of weight decay (shown by the dotted line in Fig. 6(a)), which is initially constant at $\eta \simeq 1.6$ until a decay towards the end of the given time-window as required for the system to achieve optimal asymptotic performance [14]. As in the previous example, Fig. 6(b) shows a significant shortening of the symmetric phase when compared to learning without weight decay. However, as the system escapes the symmetric phase and the weight decay drops to zero, the generalization error approaches the same decay as in the absence of weight decay and there is no asymptotic improvement in performance.

V. CONCLUSION

In this paper we have examined the effects of a simple regularizer, weight decay, under a statistical mechanics description of the learning process which is exact in the limit of large input dimension. General results are obtained for a noiseless, isotropic and structurally matched scenario which is most amenable to analysis (a small learning rate is also assumed). In this case we find no benefit in a fixed weight decay, which results in a lengthened symmetric phase and a non-zero asymptotic generalization error. In fact, we identify a critical value for the weight decay $\gamma_{\rm max}$ above which the student will never leave the symmetric phase, resulting in very poor performance. Analytical results for both phases show this behaviour to hold for general model complexity K and we find that $\gamma_{\rm max}$ is inversely proportional to K for large K. Numerical investigations also show that weight decay is not beneficial (in terms of either transient or asymptotic performance) for small learning rates when the task being learned is over-realizable ($K > M$) or corrupted by Gaussian output noise.

In order to determine the behaviour for arbitrary learning rates we employ recent methods for determining optimal time-dependent parameters over a fixed time window [14]. For small

learning rates we find results consistent with the above discussion: the optimal weight decay parameter is very small and mostly negative during the symmetric phase, for realizable, over-realizable and noisy learning scenarios. However, for higher learning rates (we choose the optimal value in the absence of weight decay) a positive weight decay is found to be beneficial during the symmetric phase, although we never find any benefit after specialization occurs and for noisy learning the asymptotic performance is not improved upon. The shortened symmetric phase is due to non-linear effects which are not incorporated by our small η analysis.

Although we do identify a scenario in which weight decay is slightly beneficial, this is probably of little value in practice since in most situations we find fixed weight decay to be detrimental to performance, especially at late times. Other more principled, and presumably more successful, adaptations to the basic gradient descent algorithm have been suggested for reducing the length of the symmetric phase (see, for example, Ref. [16]). This is not to say that weight decay is useless in general, however, since we have only considered learning with examples drawn from an unlimited training set. One might expect some benefit during the asymptotic phase of learning in the case where training examples are drawn with replacement from a fixed sample, since one then has to deal with a fixed error surface and consequently over-fitting, resulting in a much richer optimization landscape with many local minima.

Acknowledgement

We would like to thank Sara A. Solla for useful discussions and for critical comments on the text. The work was supported by the EPSRC grant GR/L19232 and EU grant CHRX-CT92-0063.

REFERENCES

[1] M. Biehl and H. Schwarze, *J. Phys. A* **28**, 643 (1995).

[2] D. Saad and S.A. Solla, *Phys. Rev. E* **52**, 4225 (1995).

[3] D. Saad and S.A. Solla, in *Advances in Neural Information Processing Systems*, edited by D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (MIT Press, Cambridge MA, 1996) Vol. 8, p. 302.

[4] D. Saad and S.A. Solla, in *Advances in Neural Information Processing Systems*, edited by M. C. Mozer, M. I. Jordan and T. Petsche (MIT press, Cambridge MA, 1997) Vol. 9, p. 260.

[5] P. Riegler and M. Biehl, *J. Phys. A* **28**, L507 (1995).

[6] G. Cybenko, *Math. Control Signals and Systems* **2**, 303 (1989).

[7] C.M. Bishop, *Neural Networks for Pattern Recognition*, (Oxford University Press, Oxford, 1995).

[8] T.L.H. Watkin, A. Rau, and M. Biehl, *Rev. Mod. Phys.* **65**, 499 (1993).

[9] S. Amari, N. Murata, K. R. Müller, M. Finke and H. Yang, in *Advances in Neural Information Processing Systems*, edited by D. S. Touretzky, M. C. Mozer and M. E. Hasselmo (MIT Press, Cambridge MA, 1996) Vol. 8, p. 176.

[10] S. J. Hanson and L. Y. Pratt, in *Advances in Neural Information Processing Systems*, edited by D. S. Touretzky (Morgan Kaufmann Publishers, Palo-Alto, U.S.A., 1988) Vol. 1, p. 177.

[11] A. Krogh and J. A. Hertz *J Phys. A* **25**, 1135 (1992).

[12] A. D. Bruce and D. Saad, *J Phys. A* **27**, 3355 (1994).

[13] S. A. Solla, in *Neural Networks for Signal Processing II*, edited by S. Y. Kung, F. Fall-

side, J. Aa. Sorenson and C. A. Kamm (IEEE, NJ, U.S.A., 1992) p. 255.

[14] D. Saad and M. Rattray, *Phys. Rev. Lett.* **79**, 2578 (1997).

[15] M. Rattray and D. Saad, *J Phys. A* **30**, L771 (1997).

[16] A.H.L. West and D. Saad, *Phys. Rev. E* **56**, 3426 (1997).

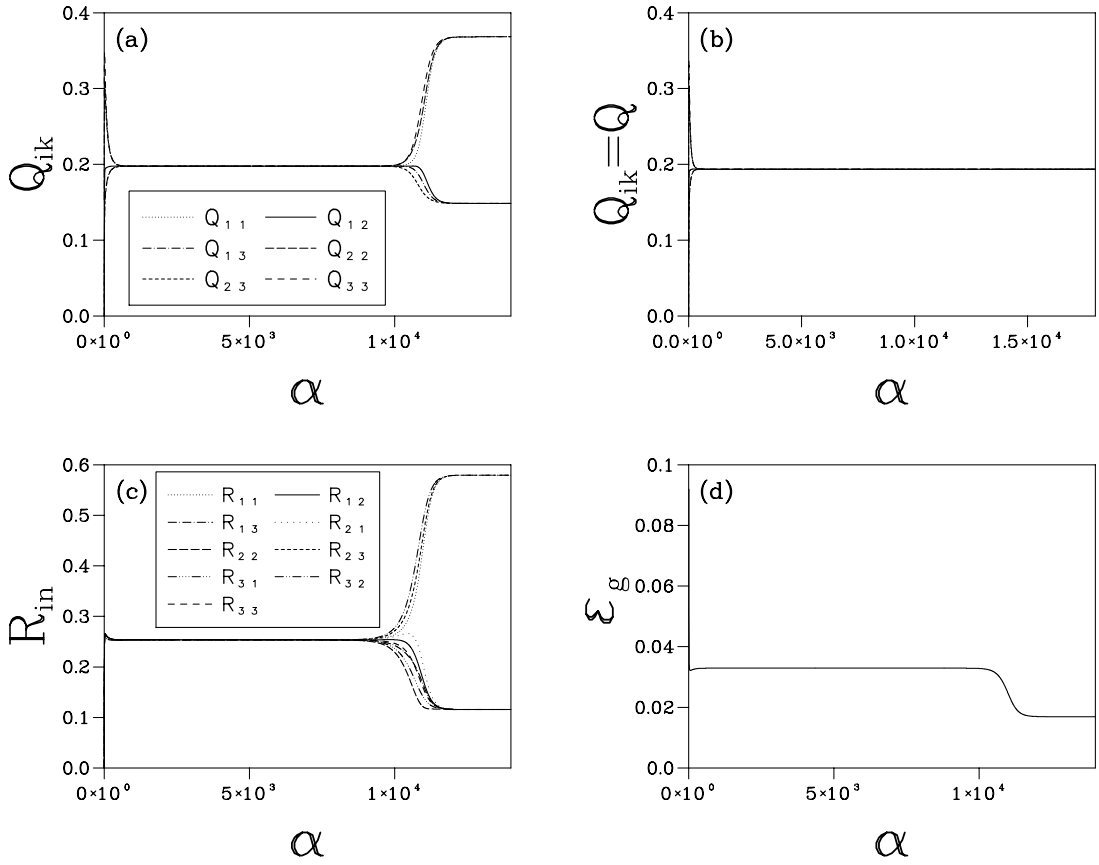

FIG. 1. The order parameters evolution for low weight decay, $\gamma = 0.005$, ($\tilde{\gamma} < \tilde{\gamma}_{max}$ - a, c and d) and high weight decay, $\gamma = 0.007$, ($\tilde{\gamma} > \tilde{\gamma}_{max}$ - b) for a noiseless scenario with $K = M = 3$ and $\eta = 0.2$. Sub-figures (a) and (b) show the evolution of student vector lengths and overlaps, (c) and (d) the overlaps between student and teacher vectors and the evolution of the generalization error respectively.

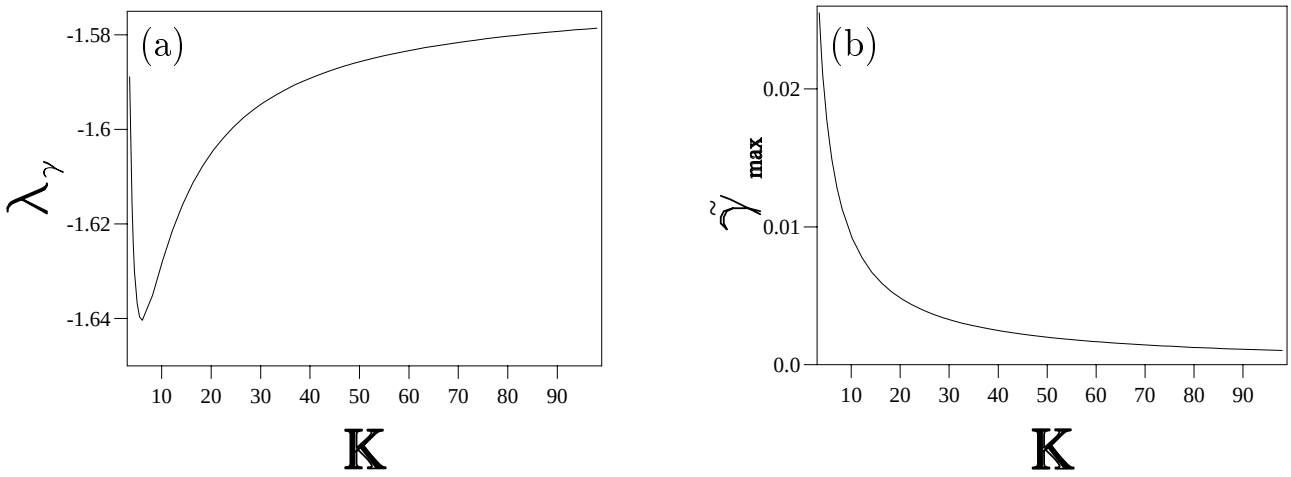

FIG. 2. Weight decay and trapping in the symmetric phase. (a) The modification of the eigenvalue controlling the escape from the symmetric phase due to weight decay, λ_γ as a function of K. (b) Maximal weight decay as a function of K.

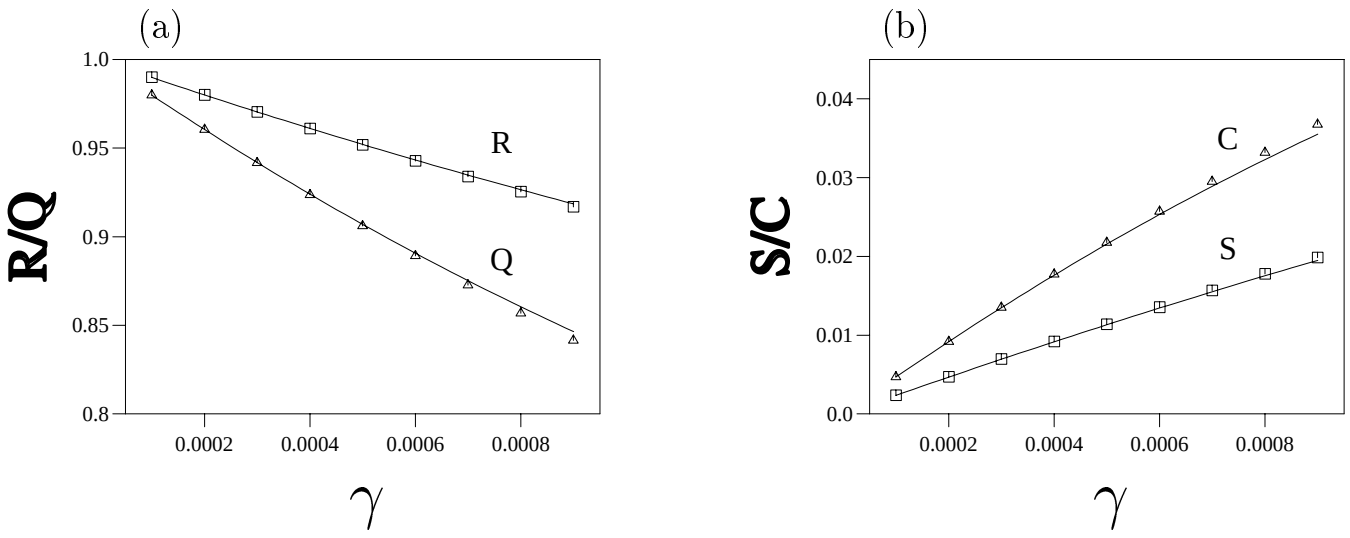

FIG. 3. Asymptotic values for the overlaps in the case $K = M = 3$ and $\eta = 0.2$ for various weight decay levels: (a) predicted values for R and Q (lines) against actual values obtained numerically (boxes and triangles). (b) predicted values for S and C (lines) against actual values (boxes and triangles).

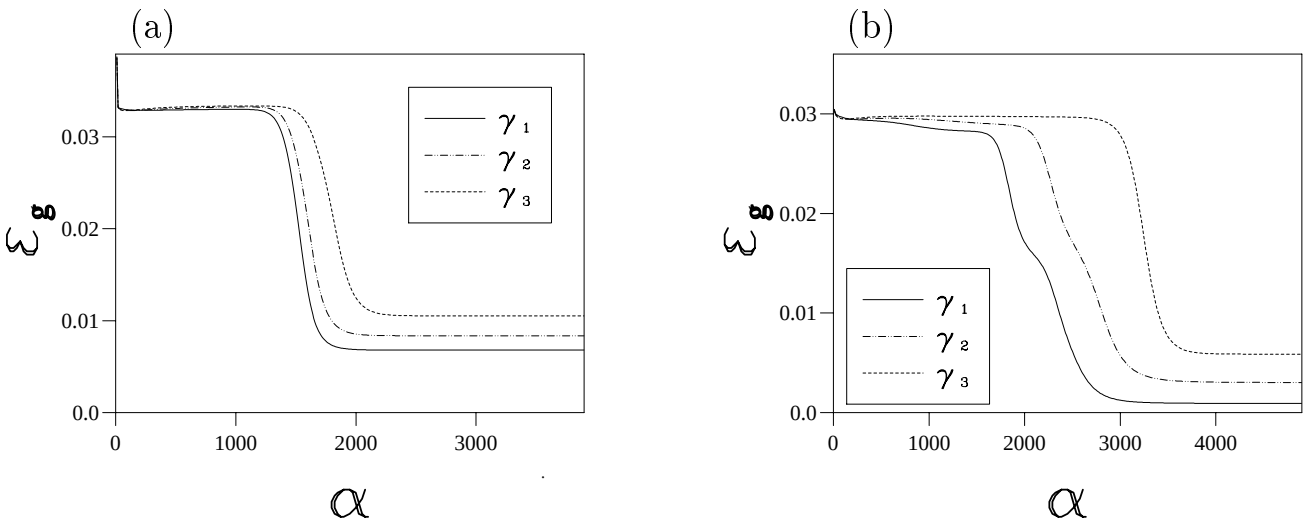

FIG. 4. The effect of weight decay on the evolution of the generalization error in two training scenarios: (a) Where examples are corrupted by output noise of zero mean and variance $\sigma^2 = 0.1$. In this case $M = K = 3$, the learning rate used is $\eta = 0.2$ and the weight decay values vary between $\gamma = 0.001 \ldots 0.003$. (b) In a highly redundant (over-realizable) training scenario with $M = 3$, $K = 5$ and $\eta = 0.2$.

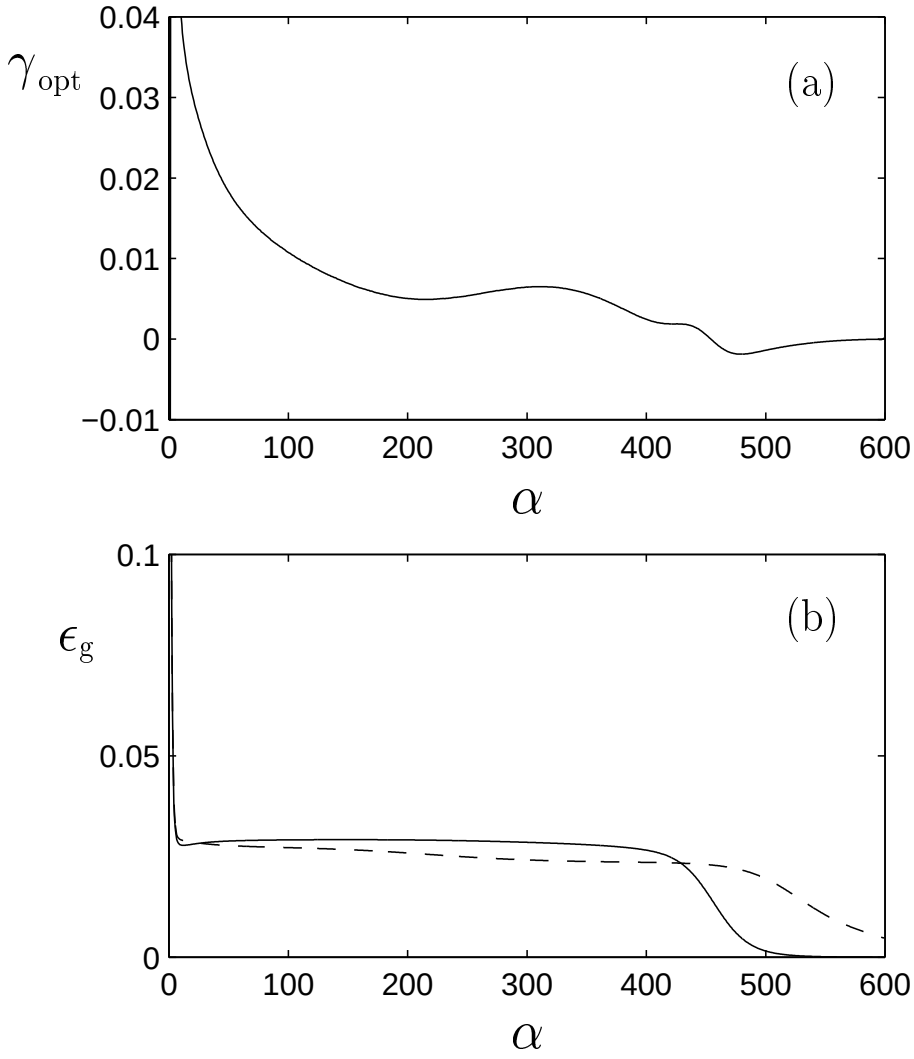

FIG. 5. The optimal time-dependent weight decay is shown in (a) for an over-realisable noiseless learning scenario with $M = 2$, $K = 3$ and $\eta = 0.7$. The corresponding generalization error is shown by the solid line in (b) where it is compared to the generalization error without weight decay (dashed line).

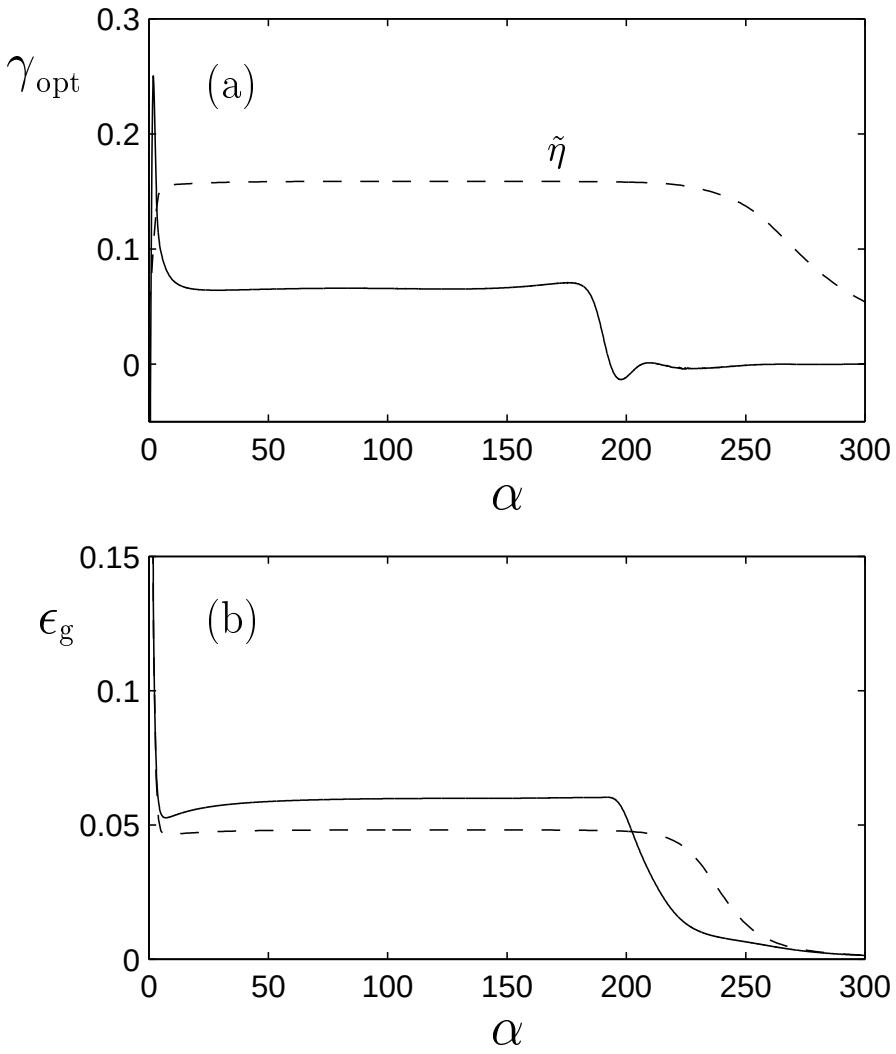

FIG. 6. The optimal time-dependent weight decay is shown by the solid line in (a) for a structurally realizable task ($M = K = 2$) with examples corrupted by Gaussian output noise of variance $\sigma^2 = 0.01$. The learning rate (dashed line) is fixed at its optimal time-dependent value in the absence of weight decay ($\tilde{\eta} = \eta/10$). The corresponding generalization error is shown by the solid line in (b) where it is compared to the generalization error without weight decay (dashed line).