A slightly revised version appears in Physical Rev., E-59, No. 1, 1999.

Optimal Asymptotic Learning Rate - Macroscopic vs. Microscopic Dynamics

Todd K. Leen[1], Bernhard Schottky[2] and David Saad[2]
[1] *Dept of Comp. Sci. & Engineering Oregon Graduate Institute of Science and Technology*
P.O.Box 91000, Portland, Oregon 97291-1000
[2] *Neural Computing Research Group, University of Aston, Birmingham B4 7ET, U.K.*

We investigate the asymptotic dynamics of on-line learning for neural networks and provide an exact solution to the network dynamics at late times under various annealing schedules. The dynamics is solved using two different frameworks: the master equation, and order parameter dynamics, which concentrate on microscopic and macroscopic parameters respectively. The two approaches provide complementary descriptions of the dynamics. Optimal annealing rates and the corresponding prefactors are derived for soft committee machine networks with hidden layer of arbitrary size.

I. INTRODUCTION

The asymptotic dynamics of stochastic, or on-line, learning and its dependence on the annealing schedule adopted for the learning coefficients have been studied for some time in the stochastic approximation literature [1,2] and more recently in the neural network literature [3–5]. In the latter references, the analysis is based on a master equation that describes the dynamics of the weight space probability densities.

In most cases of interest, the transition probability appearing in the master equation cannot be written in closed form (however, an integrable class of systems is discussed in [6]), so some approximate form of the dynamics is developed. Typically, as here, a small noise expansion provides a description of the dynamics in terms of suitably scaled fluctuations about a deterministic flow. This approach has been applied primarily to learning with fixed, arbitrarily small learning rate. In this realm, it provides information on the asymptotic distributions, convergence of learning with momentum, basin hopping, and learning with correlated samples (e.g., [7,8] and references within).

As discussed here, the approach can also be applied to annealed learning, where the learning rate is reduced with time to allow convergence (e.g. in mean square) of the weights. For either fixed, or annealed learning rate one can, for the equations of motion of the ensemble density, construct (ordinary differential) equations of motion for its moments, and hence evaluate the asymptotic generalization error, and its convergence rate.

Recently several authors [9,10] have developed an alternative theoretical approach based on the dynamics of order parameters for the system. While the master equation approach focuses on the stochastic dynamics of *microscopic* quantities (the weights), the order parameter approach describes deterministic dynamics of *macroscopic* quantities. The equations of motion for the order parameters can be solved numerically, enabling one to monitor the evolution of the order parameters and the system performance at all times. This approach provided insight into the dynamics at early stages of the learning process [11], the scaling of training parameters at the various stages [12], the use of regularizers in multilayer systems [13] and the optimization of learning parameters and rules [14–16].

In this paper we examine the relation between the two approaches and contrast the results obtained for different learning rate annealing schedules in the asymptotic regime. Using the master equation, we develop a perturbation approach that provides results on the asymptotic misadjustment for annealed learning rate. Although these results are known from the classic stochastic approximation literature, this particular approach has not, to our knowledge, been aired in the literature. We employ the order parameter approach to examine the dependence of the dynamics on the number of hidden nodes in a multilayer system. In addition, we report some lesser-known results on non-standard annealing schedules

II. MASTER EQUATION

Most on-line learning algorithms assume the form

$$w_{t+1} \;=\; w_t \;+\; \eta_0/t^p \, H(w_t, x_t) \tag{1}$$

where w_t is the weight at time t, x_t is the training example, and $H(w, x)$ is the weight update. The description of stochastic learning dynamics in terms of weight space probability densities starts from the master equation:

$$P(w', t+1) = \int dw \ \left\langle \delta \left(w' - w - \frac{\eta_0}{t^p} H(w,x) \right) \right\rangle_x P(w,t) \tag{2}$$

where $\langle \ldots \rangle_x$ indicates averaging with respect to the measure on x, $P(w,t)$ is the probability density on weights at time t, and $\delta(\ldots)$ is the Dirac function. A Kramers-Moyal expansion of Eq.(2), and passage to continuous time produces a partial differential equation for the weight probability density (here in one dimension for simplicity of notation) [3,4]

$$\partial_t P(w,t) \ = \ \sum_{i=1}^{\infty} \frac{(-1)^i}{i!} \left(\frac{\eta_0}{t^p} \right)^i \ \partial_w^i \left[\left\langle H^i(w,x) \right\rangle_x \ P(w,t) \right] \ . \tag{3}$$

Following [3], we make a small noise expansion for (3) by decomposing the weight trajectory into the sum of deterministic and stochastic pieces

$$w \ \equiv \ \phi(t) \ + \ \eta_0^\gamma \ f(t) \ \xi \qquad \text{or} \qquad \xi \ = \ \left(\frac{1}{\eta_0^\gamma \ f(t)} \right) \ (w - \phi(t)) \tag{4}$$

where $\phi(t)$ is the deterministic trajectory and ξ are the fluctuations. Apart from the factor $\eta_0^\gamma \ f(t)$ that scales the fluctuations, this is identical to the formulation for constant learning in [3]. We will obtain the proper value for the unspecified exponent γ, and the form of the function $f(t)$ from homogeneity requirements.

Next, the dependence of the jump moments $\left\langle H^i(w,x) \right\rangle_x$ on η_0 is explicated by a Taylor series expansion about the deterministic path ϕ. The coefficients in this series expansion are denoted

$$\alpha_i^{(j)} \ \equiv \ \left. \frac{\partial^j \left\langle H^i(w,x) \right\rangle_x}{\partial w^j} \right|_{w=\phi}$$

for convenience, we define a new time variable

$$s \ = \ t$$

and transform the differential operators and densities in (3) as dictated by (4)

$$\begin{aligned} \partial_t &= \partial_s \ - \ \frac{1}{\eta_0^\gamma \ f(s)} \ \frac{d\phi(s)}{ds} \ \partial_\xi \ - \ \left(\frac{f'}{f} \right) \ \xi \ \partial_\xi \\ \partial_w &= \frac{1}{\eta_0^\gamma \ f(s)} \ \partial_\xi \\ P(w,t) &= (\eta_0^\gamma \ f(s))^{-1} \ \Pi(\xi, s) \ . \end{aligned} \tag{5}$$

Finally, we rewrite (3) in terms of ϕ and ξ and the expansion of the jump moments using the transformations (5), and suitably re-summing the series. These transformations leave equations of motion for the deterministic trajectory $\phi(s)$ and the density $\Pi(\xi, s)$ on the fluctuations

$$\frac{d\phi}{ds} = \left(\frac{\eta_0}{s^p} \right) \ \alpha_1^{(0)}(\phi) \ = \ \left(\frac{\eta_0}{s^p} \right) \ \langle \, H(\phi, x) \, \rangle_x \tag{6}$$

$$\begin{aligned} \partial_s \Pi = & \left(\frac{f'}{f} \right) \ \partial_\xi (\xi \, \Pi) \\ & + \ \sum_{m=2}^{\infty} \sum_{i=1}^{m} \ \frac{(-1)^i}{i!(m-i)!} \ \alpha_i^{(m-i)} \ \frac{{\eta_0}^{i(1-2\,\gamma)+m\gamma}}{s^{i\,p}} \ f(s)^{m-2i} \ \partial_\xi^i (\xi^{(m-i)} \Pi) \ . \end{aligned} \tag{7}$$

Commonly, the learning algorithm is a stochastic gradient descent, for which the weight update function $H(w,x)$ is minus the gradient of the instantaneous cost $H(w,x) = -\nabla_w E(w,x)$. Then (6) describes the evolution of ϕ as descent on the average cost. The fluctuation equation (7) requires further manipulations whose form depends on the context.

We need to specify γ and $f(s)$ to make further progress. We assume stochastic gradient descent in a quadratic bowl, i.e. an algorithm with a cost function whose Hessian, G is *positive-definite*. Then $\alpha_1^{(1)} = -\nabla_w^2 \left\langle E(w,x) \right\rangle_x \Big|_{w=\phi(s)} \equiv -G(\phi(s))$. To insure that for each value of m in (7) the terms in the sum over i are homogeneous in powers of η_0, we take

$$\gamma \;=\; 1/2 \;\;.$$

Similarly, to insure that for each value of m the terms in the sum over i are homogeneous in time, we take

$$f(s) \;=\; \frac{1}{s^{p/2}} \;\;.$$

For *constant* learning rate ($p = 0$), we re-scale the time as $s \to \eta_0 s$ to allow (7) to be written in a form convenient for a perturbation expansion of the solution in powers of η_0. Typically, the small learning rate limit $\eta_0 \to 0$ is invoked, and only the lowest order terms in η_0 retained (e.g. [3]). The remaining differential equation contains a simple diffusion operator, which results in a Gaussian approximation for equilibrium densities. Higher order terms have been successfully used to calculate corrections to the equilibrium moments in powers of η_0 [17].

Of primary interest here is the case of *annealed learning*, as required for convergence of the parameter estimates. Again assuming a quadratic bowl with $\gamma = 1/2$, $f(s) = 1/s^{p/2}$, the first few terms of the n-dimensional form of (7) are

$$\begin{aligned}\partial_s \Pi = \nabla_\xi \cdot \Big[\Big(\frac{\eta_0}{s^p} G(\phi(s)) - \frac{p}{2\,s} \Big)\; \xi\; \Pi \Big] \\ + \; \frac{1}{2} \frac{\eta_0}{s^p} \nabla_\xi \cdot \Big(\alpha_2^{(0)}\, \nabla_\xi\, \Pi \Big) + \; \mathcal{O}\left(\frac{\eta_0}{s^p} \right)^{3/2}\end{aligned} \tag{8}$$

where the curvature G and diffusion coefficients $\alpha_2^{(0)}$ are matrices now. As $s = t \to \infty$ the right hand side of (8) is dominated by terms explicitly written (since $0 < p \leq 1$). Precisely which terms dominate depends on p.

Classical Annealing

We first review the classical case $p = 1$ (1/t annealing). The first three leading terms on the right hand side of (8) are all of order $1/s$. For $s \to \infty$, these terms dominate, and we discard the remaining terms. The deterministic trajectory (6) is a standard gradient flow (in transformed time $\hat{s} = \ln s$), and thus $\lim_{t\to\infty} \phi(s) = w^*$, where w^* is a local minimum of the average cost $\langle E(w,x) \rangle_x$. The fluctuation dynamics carried by the first three terms of (8) are a sensible diffusion process only if the *effective* linearized drift

$$D_{\text{eff}} \;\equiv\; \eta_0\, G \;-\; 1/2$$

is a positive definite matrix. This clearly requires

$$\eta_0 > \eta_0^{\text{crit}} \;=\; 1/(2\,\lambda_{min}) \tag{9}$$

where λ_{min} is the smallest eigenvalue of $G^* \equiv G(w^*)$.

If the criticality condition in (9) is met, then the equilibrium density is a zero-mean Gaussian with covariance Σ_ξ that satisfies

$$D_{\text{eff}}\, \Sigma_\xi + \Sigma_\xi\, D_{\text{eff}} \;=\; \eta_0\, \alpha_2^{(0)} \tag{10}$$

With our choice of γ and $f(s)$, the *weight error* $v = w - w^*$ is related to the fluctuation ξ by $v = \sqrt{\eta_0/s}\,\xi$. Consequently $/sqrts\, v$ is asymptotically normal, and the expected squared weight error drops off as

$$E[|v|^2] \;=\; \text{Trace}\left(E[vv^T]\right) \;\propto\; \frac{1}{s} \tag{11}$$

which is the well-known optimal asymptotic convergence rate.

If the criticality condition is not met, the Gaussian equilibrium is not reached. The asymptotic convergence of $E[|v|^2])$ can still be calculated by developing the dynamics of the second moment $R_\xi \equiv E[\xi\xi^T]$. One obtains the differential equations of motion by multiplying (8) by $\xi_i\, \xi_j$ and integrating over $d^n\xi$ to obtain

$$\frac{d}{ds} R_\xi \;=\; -\frac{1}{s} \left(D_{\text{eff}} R_\xi + R_\xi D_{\text{eff}} \right) + \frac{\eta_0}{s} \alpha_2^{(0)} \tag{12}$$

which have the solution

$$R_\xi(s) = U(s,s_0) R_\xi(s_0) U^T(s,s_0) + \int_{s_0}^{s} d\tau \frac{\eta_0}{\tau} U(s,\tau) \alpha_2^{(0)} U^T(s,\tau) \tag{13}$$

$$U(s,s_0) = \exp\left(-\ln(s/s_0) D_{\rm eff}\right) \quad . \tag{14}$$

Transforming the result back to w coordinates, we obtain equation for the time-evolution of the weight error correlation matrix $C = E[v\, v^T]$, and hence for for the misadjustment (derived by an alternative approach in [5])

$$E[|v|^2] = \sum_{k=1}^{n} C_{kk}(s_0) \left(\frac{s_0}{s}\right)^{2\,\eta_0\,\lambda_k} + \frac{\eta_0^2\ \alpha_{2\,kk}^{(0)}}{(2\eta_0\lambda_k - 1)} \left[\frac{1}{s} - \frac{1}{s_0}\left(\frac{s_0}{s}\right)^{2\eta_0\lambda_k}\right] \tag{15}$$

where: λ_k are the eigenvalues of the curvature G^* (at the local optimum $\mathbf{w}^*$), C_{kk} and $\alpha_{2\,kk}^{(0)}$ are the diagonal components of the weight error correlation and the diffusion matrix (respectively), both in coordinates for which G^* *is diagonal*.

¿From (15), it is clear that when $\eta_0 > \eta_0^{\rm crit}$, one has $1/s$ decay of the misadjustment, while for $\eta_0 < \eta_0^{\rm crit}$ the decay progresses as $(1/s)^{2\eta_0\lambda_{\rm min}}$, i.e. *slower* than $1/s$. The above confirm the classical results [1] on asymptotic normality and convergence rate for $1/t$ annealing.

Alternative Annealing Schedules

In the case of $0 < p < 1$ the right-hand side of (8) is dominated at late times by the terms of order $1/s^p$. Now, since G^* is, by assumption, positive definite, there is no criticality or switching behaviour in the convergence. We have a Gaussian equilibrium density for ξ, with covariance that satisfies

$$G^*\,\Sigma_\xi + \Sigma_\xi\, G^* = \alpha_2^{(0)}$$

in analogy to (10). The weight error is related to the fluctuations by $v = \sqrt{\eta_0/s^p}\,\xi$ (so that $\sqrt{s^p}\,v$ is asymptotically normal) and consequently the expected squared weight error drops off asymptotically as

$$E[|v|^2] \ \propto\ \frac{1}{s^p} \quad . \tag{16}$$

Notice that i) the convergence is *slower* than $1/s$ and ii) there is *no* critical value of the learning rate to obtain a sensible equilibrium distribution. Related results are in [18].

This approach to obtain the asymptotic dynamics of the fluctuations, and hence the misadjustment $E[|v|^2]$, is quite general. The derivations assume that the minimum studied has a positive definite Hessian, and positive definite diffusion matrix $\alpha_2^{(0)}$. The latter is true for any non-realizable task, or for a realizable task with noisy cost targets. The results hold for arbitrarily large systems, though the critical learning rate for $1/t$ annealing depends on the eigenvalue spectrum of the curvature. The latter depends on the specifics of the cost function and input/target distribution.

Connection to neural networks

In the context of neural networks, the cost function $E(w,x)$ measures the deviation between the trained networks output (termed here students) and the output of the underlying process represented here by a teacher network, specified by some weight vector. The performance measure of interest is then not the misadjustment but the generalization error, defined for a given (student-)weight distribution $p(w)$ as

$$\epsilon_g = \langle\langle E(w,x)\rangle_x\rangle_w \,. \tag{17}$$

error $\epsilon_g - \epsilon_{min}$ (where ϵ_{min} is the least generalization error achievable in the area of the (possibly local) minimum considered) follows the same decay rate as the misadjustment. Using the Taylor series expansion of (17), to lowest order in the weight error, one has

$$\epsilon_g - \epsilon_{min} \ =\ \frac{1}{2} E[v^T G^* v] = \frac{1}{2} {\rm Trace}\,(G^*\,C)\,.$$

Thus the excess generalization error is bounded above (below) by the maximum (minimum) eigenvalue of G^* times the misadjustment $E[|v|^2]$. These eigenvalues and eigenvectors depend on the actual architecture of the neural network considered.

In the next section we will use a different approach, relying on the order parameter ansatz, to derive the learning behaviour for a neural network of a concrete architecture, namely the soft committee machine.

III. ORDER PARAMETERS

In the master equation approach, one focuses attention on the weight space distribution $P(w,t)$ and calculates quantities of interest by averaging over this density. An alternative approach is to choose a smaller set of *macroscopic* variables, called order parameters, that are sufficient for describing principal properties of the system such as the generalization error (in contrast to the evolution of the weights w which are *microscopic*).

Formally, one can replace the parameter dynamics presented in Eq.(2) by the corresponding equation for macroscopic observables which can be easily derived from the corresponding expressions for w and the formal definition of macroscopic observables [19]. By choosing an appropriate set of macroscopic variables and invoking the thermodynamic limit (i.e., looking at systems where the number of parameters is infinite), one obtains a closed set of equations of point distributions for the order parameters, rendering the dynamics deterministic. Note that in contrast to the master equation approach, which provides an approximation to the weight space distribution, the order parameter approach provides an exact closed set of deterministic equations which fully describe their dynamics and can be employed for calculating other observables which are of interest. The disadvantages of this approach is that we are restricted to calculating observables which can be formulated in terms of the order parameters and it is exact only in the thermodynamic limit (although finite-size analysis shows quite good agreement between theory and simulations also for small systems [20]).

Practically, the formal replacement of the microscopic parameters by the order parameters in (2) is usually unnecessary and it is possible to set up the equations for the order parameters straight away once the appropriate order parameters have been identified. We use this approach now to investigate the asymptotic behaviour of the training dynamics with annealed learning rate of a soft committee machine (SCM), which is a generic two-layer neural network [9], extending the results obtained in the master equation approach. The SCM maps inputs $\boldsymbol{x} \in \Re^N$ to a scalar, realized through a model $\rho(\mathbf{w}, \boldsymbol{x}) = \sum_{i=1}^{K} g(\mathbf{w}_i \cdot \boldsymbol{x})$. The activation function of the hidden units is $g(u) \equiv \mathrm{erf}(u/\sqrt{2})$ and $\mathbf{w}_i$ is the set of input-to-hidden adaptive weights for the $i = 1 \ldots K$ hidden nodes. The hidden-to-output weights are set to 1. The learning dynamics and generalization error evolution in this architecture are similar to that of a general two-layer network. Several researchers [9,10] have already employed the order parameter approach for calculating the training dynamics of a SCM and the formalism can be easily extended to accommodate adaptive hidden-to-output weights [21].

The training examples $(\boldsymbol{x}, y)$ are independently drawn input vectors with zero mean, and unit variance and the corresponding targets y are generated by the response of a deterministic teacher network corrupted by additive Gaussian output noise of zero mean and variance σ_ν^2. The teacher network is also a SCM, characterized by input-to-hidden weights $\mathbf{w}_i^*$. The order parameters sufficient to close the dynamics, and to describe the network generalization error are overlaps between various input-to-hidden vectors

$$\mathbf{w}_i \cdot \mathbf{w}_k \equiv Q_{ik}, \quad \mathbf{w}_i \cdot \mathbf{w}_n^* \equiv R_{in}, \quad \text{and} \quad \mathbf{w}_n^* \cdot \mathbf{w}_m^* \equiv T_{nm} \quad . \tag{18}$$

Employing the learning rule given in (1) and using gradient descent on a squared error measure as weight update we get focusing on a $1/t$-annealing

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \frac{\eta_0}{t} \nabla_{\mathbf{w}} \frac{1}{2} \left[\rho(\mathbf{w}^*, \mathbf{x}) + \zeta - \rho(\mathbf{w}, \mathbf{x}) \right]^2 \tag{19}$$

as update rule where ζ is the disruption of the teacher output by the additive Gaussian noise of variance σ_ν^2. By calculating the corresponding dot products and averaging over the random training pattern distribution and the output noise, one easily gets update equations for the order parameter which become in the thermodynamic limit ($N \to \infty$) a deterministic set of coupled differential equations. These equations are given in closed form in [10], where only the constant learning rate η is examined, while in the current paper we concentrate on the annealed learning rate η_0/t. The general solution of the coupled differential equations can only be obtained through numerical integration. However, the asymptotic behaviour in the case of annealed learning *is* amenable to analysis, and this is one of the primary results of the paper.

Finally, the network performance is measured in terms of the generalization error

$$\epsilon_g \equiv \langle 1/2\,[\ \rho(\mathbf{w},\boldsymbol{x}) - y\ ]^2 \rangle_x\ , \tag{20}$$

which can be expressed in closed form in terms of the order parameters [10] (this is also the reason for the absence of the $\mathbf{w}$ average in comparison to (17)). The goal of this study is to solve the dynamics at late times and find the learning rate schedule which gives the optimal decay of the generalization error.

We assume an isotropic teacher $T_{nm} = \delta_{nm}$ and use task symmetry to reduce the system to a vector of four order parameters $u^T = (r, q, s, c)$ related to the overlaps by $R_{in} = \delta_{in}(1+r) + (1-\delta_{in})s$ and $Q_{ik} = \delta_{ik}(1+q) + (1-\delta_{ik})c$.

With learning rate annealing and $\lim_{t\to\infty} u \to 0$ we describe the dynamics in this vicinity by a linearisation of the equations of motion in [10] giving

$$\frac{d}{dt}\mathbf{u} = \eta M \mathbf{u} + \eta^2 \sigma_\nu^2 \mathbf{b}\,, \tag{21}$$

where σ_ν^2 is the noise variance, $\mathbf{b}^T = \frac{2}{\pi}\left(0, 1/\sqrt{3}, 0, 1/2\right)$, $\eta = \eta_0/t^p$, and

$$M = \frac{2}{3\sqrt{3}\pi}\begin{pmatrix} -4 & \frac{3}{2} & -\frac{9}{4}(K-1)\sqrt{3} & \frac{3}{4}(K-1)\sqrt{3} \\ 4 & -3 & \frac{3}{2}(K-1)\sqrt{3} & -\frac{3}{2}(K-1)\sqrt{3} \\ -\frac{3}{2}\sqrt{3} & \frac{3}{8}\sqrt{3} & -\frac{3\sqrt{3}}{2}(K-2) - 3 & 0 \\ 3\sqrt{3} & -\frac{9}{4}\sqrt{3} & 3\sqrt{3}(K-2) + 6 & -3\sqrt{3}(K-2) - 6 \end{pmatrix}\ . \tag{22}$$

The asymptotic equations of motion (21) were derived by dropping terms of order $\mathcal{O}(\eta||\mathbf{u}||^2)$ and higher, *and* terms of order $\mathcal{O}(\eta^2\,\mathbf{u})$. While the latter are linear in the order parameters, they are negligible in comparison to the $\eta\,\mathbf{u}$ and $\eta^2\sigma_\nu^2\mathbf{b}$ terms in (21) as $t \to \infty$.

The truncations used to arrive at the asymptotic dynamics shed light on the approach to equilibrium that is not implicit in the master equation approach. In the latter, the dominant terms for the asymptotic behaviour of (8) were identified by the coefficient's time scale; there is no indication, in terms of system observables, for the onset of the asymptotic regime. In contrast, in the present approach, the conditions for validity of the asymptotic approximations are cast in terms of system observables directly by comparing $\eta^2\mathbf{u}$ versus $\eta\mathbf{u}$ and $\eta^2\sigma_\nu^2$.

The solution to Eq.(21) is

$$\mathbf{u}(t) = \gamma(t,t_0)\ \mathbf{u}_0 + \sigma_\nu^2\ \beta(t,t_0)\ \mathbf{b} \tag{23}$$

where $\mathbf{u}_0 \equiv \mathbf{u}(t_0)$,

$$\gamma(t,t_0) = \exp\left\{M\int_{t_0}^{t} d\tau\ \eta(\tau)\right\}\ \text{ and }\ \beta(t,t_0) = \int_{t_0}^{t} d\tau\ \gamma(t,\tau)\ \eta^2(\tau)\,. \tag{24}$$

Both matrices γ and β can be calculated in closed form, whereby each matrix element is a linear combination of the modes ϕ_i for β and θ_i for γ $(i = 1\ldots 4)$ with

$$\phi_i = \left(\frac{t}{t_0}\right)^{\alpha_i\eta_0} \quad \text{and} \quad \theta_i = -\frac{\eta_0^2}{1+\alpha_i\eta_0}\left[\frac{1}{t} - t^{\alpha_i\eta_0}t_0^{-(\alpha_i\eta_0+1)}\right]\ , \tag{25}$$

where α_i are the eigenvalues of the matrix M (Fig. 1(a)). Comparing this to (15) one sees, that the ϕ_i's and θ_i's have the same general structure as the terms in (15). Using this solutions, one obtains an explicit expression for the linearized generalization error (first order in $\mathbf{u}$)

$$\epsilon_l = \frac{K}{\pi}\left(\frac{1}{\sqrt{3}}(q-2r) + \frac{K-1}{2}(c-2s)\right)\,. \tag{26}$$

It turns out, that only two modes survive and we get

$$\epsilon_l = \sigma_\nu^2\left[c_1\theta_1(t) + c_2\theta_2(t)\right] + a_1\phi_1 + a_2\phi_2 \tag{27}$$

with the eigenvalues

$$\alpha_1 = -\frac{1}{\pi}\left(\frac{4}{\sqrt{3}} - 2\right)\ \text{ and }\ \alpha_2 = -\frac{1}{\pi}\left(\frac{4}{\sqrt{3}} + 2(K-1)\right)\,. \tag{28}$$

The constants c_1 and c_2 depend only on K while a_1 and a_2 depend also on the initial conditions. Obviously, the fastest decay one can obtain is $1/t$ when choosing $\eta_0 > \eta_0^{\text{crit}}$, which is (for $K \geq 2$)

$$\eta_0^{\rm crit} = \max\left(-\frac{1}{\alpha_1}, -\frac{1}{\alpha_2}\right) = \frac{\pi}{4/\sqrt{3}-2}\,. \tag{29}$$

In this case the modes ϕ_i in (27) decay faster then $1/t$ and can be ignored while the modes θ_i are dominated by the $1/t$-component; we thus have a $1/t$ decay independent of the initial conditions which is of the form

$$\epsilon_l = -\sigma_\nu^2\eta_0^2\left(\frac{c_1}{1+\alpha_1\eta_0} + \frac{c_2}{1+\alpha_2\eta_0}\right)\frac{1}{t} \;\equiv\; \sigma_\nu^2 f(\eta_0, K)\frac{1}{t}\;\;. \tag{30}$$

For optimal decay of the asymptotic error we have also to minimize the prefactor $f(\eta_0, K)$ in Eq.(30). The values of $\eta_0^{\rm opt}(K)$ for various values K are shown in Fig. 1(b), where the special case of $K = 1$ (see below) is also included: There is a significant difference between the values for $K = 1$ and $K = 2$ and a rather weak dependence on K for $K \geq 2$ which may be explained by the need to unlearn correlations between vectors associated with different hidden nodes which is absent in the case of single node systems. The sensitivity of the generalization error decay factor on the choice of η_0 is shown in Fig. 1(c).

The influence of the noise strength on the generalization error can be seen directly from (30): the noise variance σ_ν^2 is just a prefactor scaling the $1/t$ decay. Neither the value for the critical nor for the optimal η_0 is influenced by it.

The calculation above holds for the case $K = 1$ (where c and s and the mode θ_1 are absent). In this case

$$\eta_0^{\rm opt}(K=1) = 2\eta_0^{\rm crit}(K=1) = -\frac{2}{\alpha_2} = \frac{\sqrt{3}\pi}{2}\,. \tag{31}$$

Finally, for the general annealing schedule of the form $\eta = \eta_0/t^p$ with $0 < p < 1$ the equations of motion (23) can be investigated, and one again finds $1/t^p$ decay.

One other point that is worthwhile mentioning is that the exact asymptotic results obtained here are consistent with those obtained using a variational method aimed at finding the globally optimal learning rate at all times [14,15] but which requires a numerical solution of a set of coupled differential equations.

IV. DISCUSSION AND SUMMARY

We employed the master equation and order parameter approaches to study the convergence of on-line learning under different annealing schedules. For the $1/t$ annealing schedule, the small noise expansion provides a critical value for η_0 (Eq.9) in terms of the curvature, above which $\sqrt{t}\,v$ is asymptotically normal, and the misadjustment $E[|v|^2]$ decays as $1/t$. Though not developed here, the approach also tells us that to achieve the most rapid decay of the misadjustment, one should set $\eta_0 = G^{*-1}$. Further development of the dynamics suggests algorithms that automatically attain this optimal decay rate using only order $O(N)$ computation and storage [5,22].

The approach naturally extends to $1/t^p$ annealing with $0 < p < 1$, where we find the misadjustment decays as $1/t^p$. This behavior is independent of η_0; that is, there is no critical value of η_0 to obtain the asymptotically normal distribution on $\sqrt{t^p}\,v$. On naive inspection, this is a very curious result as it suggests that one can pick p arbitrarily close to unity, and obtain decay of the misadjustment arbitrarily close to the optimal $1/t$ rate. The caveat is that in deriving these results, we have truncated Eq.(8), retaining only the terms pertinent to the asymptotic distribution. The *approach* to the asymptotic distribution is not discussed at all. Clearly there is interesting dynamics in this pre-equilibrium regime, none of which is developed in this framework as it stands. The analysis has been carried beyond the lowest-order description of the fluctuation density for *constant* learning rate [17,23], by a perturbation expansion of the fluctuation density Π. Presumably a similar approach could be developed for annealed learning in order to discuss the *approach* to the equilibrium density. However, a numerical solution to the full non-linear order parameter equations would provide this information with less computation apparatus.

The analysis of learning dynamics through the master equation is completely general, placing no a priori constraints on the architecture or data distribution, but it requires knowledge of the jump moments in the asymptotic regime for calculating the relevant properties. These jump moments *are* of course architecture and data dependent. Since the analysis proceeds from the Kramers-Moyal expansion (3), which is an infinite order partial differential equation, it is *necessarily* perturbative in its approach.

In contrast, the order parameters approach begins by choosing appropriate order parameters, which are architecture dependent, making specific assumptions regarding the data distribution, and then writing down equations of motion in closed form. The latter are coupled, non-linear ordinary differential equations that can be solved numerically, or explored asymptotically using suitable linearization, as carried out here. The fact that the equations are ordinary differential equations, with finite number of terms, rather than infinite order partial differential equations holds obvious advantages for numerical investigation.

Using the order parameters approach we considered the task of a soft committee machine (architectural constraint) learning a teacher of the same architecture characterized by a set of isotropic teacher vectors with added noise (assumptions on data distribution). We obtain the dynamics in the asymptotic regime for any number of hidden nodes, and provide explicit expressions for the decaying generalization error and for the critical (Eq.29) and optimal learning rate prefactors for any number of hidden nodes K. Similar results have been obtained for the critical learning rate prefactors using both methods, and both methods have been used to study general $1/t^p$ annealing [24].

The order parameter approach provides a potentially helpful insight on the passage into the asymptotic regime. Unlike the truncation of the small noise expansion, the truncation of the order parameter equations to obtain the asymptotic dynamics is couched in terms of system observables (c.f. the discussion following Eq.(22)). That is, one knows exactly which observables must be dominant for the system to be in the asymptotic regime. Equivalently, starting from the full equations, the order parameters approach can tell us when the system is close to the equilibrium distribution.

We see the two approaches as complimentary: the perturbative expansion of the master equation provides analytic results on the asymptotic behavior without reference to specific architecture or data distributions. In this respect, it is entirely general. However the technique is most facile when limited to the lowest order in perturbation, as used here, and this focuses attention on the asymptotic regime. In practice, much algorithmic effort is expended *outside* the asymptotic regime. The order parameter approach provides finite order equations of motion for specific systems, with restricted data distributions. However these equations of motion are convenient for numerical solution, and express the learning dynamics throughout the training.

Acknowledgements: DS and BS would like to thank the Leverhulme Trust for their support (F/250/K). TL thanks the International Human Frontier Science Program and the National Science foundation for support under grants SF 473-96 and ECS-9704094 respectively.

---

[1] V. Fabian, *Ann. Math. Statist.*, **39**, 1327 (1968).
[2] L. Goldstein, Technical Report DRB-306, Dept. of Mathematics, University of Southern California, LA, (1987).
[3] T. M. Heskes and B. Kappen, in *Mathematical Foundations of Neural Networks*, edited by J. Taylor, (Elsevier, Amsterdam, 1993), p 199.
[4] T. K. Leen and J. E. Moody, in *Advances in Neural Information Processing Systems*, edited by C.L. Giles, S.J. Hanson, and J.D. Cowan (Morgan Kaufmann, San Mateo, CA, U.S.A., 1993) Vol. 5, p. 451.
[5] T. K. Leen and G. B. Orr, in *Advances in Neural Information Processing Systems*, edited by J.D. Cowan, G. Tesauro, and J. Alspector (Morgan Kaufmann, San Francisco, CA, U.S.A., 1994) Vol. 6, p. 477.
[6] T. K. Leen and J. E. Moody, *Phys. Rev. E* **56**, 1262 (1997).
[7] W. Wiegerinck, A. Komoda, and T. Heskes *J. Phys. A*, **27**, 4425 (1994).
[8] T. Heskes and J. Coolen, *J. Phys. A*, **30**, 4983 (1997).
[9] M. Biehl and H. Schwarze, *J. Phys. A* **28**, 643 (1995).
[10] D. Saad and S.A. Solla *Phys. Rev. Lett.* **74**, 4337 (1995) and *Phys. Rev. E* **52** 4225 (1995).
[11] M. Biehl, P. Riegler, and C. Wöhler, J. Phys. A **29**, 4769 (1996).
[12] A.H.L. West and D. Saad, *Phys. Rev. E*, **56**, 3426 (1997).
[13] D. Saad and M. Rattray, *Phys. Rev. E*, **57**, 2170 (1998).
[14] D. Saad and M. Rattray, *Phys. Rev. Lett.*, **79**, 2578 (1997).
[15] M. Rattray and D. Saad, *Phys. Rev. E*, **58**, in press (1998).
[16] M. Rattray and D. Saad, *Jour. Phys. A*, **30**, L771 (1997).
[17] G. B. Orr, , Ph.D. thesis, Oregon Graduate Institute, 1996.
[18] N. Barkai, , Ph.D. thesis, Hebrew University of Jerusalem 1995.
[19] C.W.H. Mace and A.C.C. Coolen, *Statistics and Computing* **8**, 55 (1998).
[20] D. Barber, D. Saad, and P. Sollich, Europhys. Lett. **34**, 151 (1996).
[21] P. Riegler and M. Biehl *J. Phys. A* **28**, L507 (1995).
[22] G.B. Orr and Todd K. Leen, in *Advances in Neural Information Processing Systems*, edited by M. Mozer, M. Jordan and T. Petsche (The MIT Press, Cambridge, MA, U.S.A, 1997) Vol. 9, p. 606.
[23] T. K. Leen, in *On-Line Learning in Neural Networks*, edited by D. Saad (The Newton Institute Series, Cambridge University Press, Cambridge, UK, 1998), p. 43.
[24] For annealing as $1/t^p$ with $p \neq 1$, the order parameter equations are a bit more restrictive to deal with than the equations of motion for the fluctuation density. In particular, we were able to solve the asymptotic, linearized, order parameter