Local and Global Convergence of On-Line Learning

N. Barkai[1], H. S. Seung[2], and H. Sompolinsky[1,2]
[1]Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel
[2]AT&T Bell Laboratories, 600 Mountain Ave., NJ 07974, USA

October 24, 1994

Abstract

We study the performance of an on-line algorithm for learning dichotomies, with a dynamical error-dependent learning rate. The asymptotic scaling form of the solution to the associated Markov equations is derived, assuming certain smoothness conditions. We show that the system converges to the optimal solution and the generalization error vanishes inversely with the number of examples. The system is capable of escaping from local minima, and adapts rapidly to shifts in the target function. The general theory is illustrated for the perceptron and committee machine.

Much of learning theory has analyzed the paradigm of batch learning, in which the learner has free access to a fixed set of examples stored in memory. This paradigm leads naturally to an equilibrium statistical mechanical approach based on an energy function that is the learner's error on the training set. The theoretical advantage of this equilibrium formulation is that the learner's performance as a function of the number of examples can be studied without addressing the complex dynamic aspects of learning. The disadvantage is that it gives no information about the learner's performance as a function of training time. Many learning phenomena, especially biological ones, fit more naturally in an *on-line* learning paradigm, rather than a batch one. At each time step, the learner receives a single new example drawn at random, and is unable to store previous examples in memory (see e.g. [1]–[4]). In the simplest cases, the change in the learner depends only on the new example and its current state. Hence training time is proportional to the number of examples.

Interesting theoretical results have been derived for on-line learning of functions that are smooth functions of the system's parameters, using stochastic approximation theory[5] and statistical mechanical methods[1]–[4]. In this Letter, we focus on learning of dichotomies, to which most of this existing theory is inapplicable. We also address the trade-off between global and local convergence properties. If the learning rate is held constant, the learner can never converge to the global optimum of zero error, even in the infinite time limit. Hence various schedules for decreasing the learning rate with time have been proposed. However, if the learning rate decreases too quickly, the learner can become trapped in a local minimum. If the learning rate decreases slowly, the learner may escape from local minima, but convergence to the global optimum is slow. We propose a dynamical error-dependent learning rate that both overcomes local minima and provides fast asymptotic convergence to the correct target.

We consider dichotomies that are defined by thresholded smooth functions,

$$\sigma(\mathbf{s};\mathbf{w}) = \mathrm{sgn}(f(\mathbf{w};\mathbf{s})) \tag{1}$$

where f is a differentiable function of a set of parameters $\mathbf{w}$, called the weight vector of the system, and $\mathbf{s}$ is the input vector. The smooth function f represents the local field induced on the system's output unit by the input $\mathbf{s}$. The target function is the dichotomy, $\sigma_0(\mathbf{s})$ defined by $\sigma_0(\mathbf{s}) = \mathrm{sgn}(f(\mathbf{w}_0;\mathbf{s}))$, where the target weight vector $\mathbf{w}$ is assumed to be in the space of $\mathbf{w}$. For this class of functions, the following on-line learning rule is proposed[4]

$$\mathbf{w}' - \mathbf{w} = \eta \mathbf{g}(\mathbf{w};\mathbf{s}) \ , \quad \mathbf{g}(\mathbf{w};\mathbf{s}) \equiv \epsilon(\mathbf{w};\mathbf{s})\sigma_0(\mathbf{s})\nabla f(\mathbf{w};\mathbf{s}) \tag{2}$$

where $\mathbf{w}'$ and $\mathbf{w}$ are the new and current weight vectors respectively, ∇ denotes a gradient with respect to $\mathbf{w}$, and $\epsilon(\mathbf{s};\mathbf{w})$ denotes the error of the learner $\mathbf{w}$ on the input $\mathbf{s}$: $\epsilon = \theta(-\sigma(\mathbf{s};\mathbf{w})\sigma_0(\mathbf{s}))$, where $\theta(x)$ is the Heaviside step function. The inputs $\mathbf{s}$ are drawn at random using a probability measure $D\mathbf{s}$. The scalar prefactor η is called the learning rate. Thus, at each step, the weights are changed only if the current state $\mathbf{w}$ yields a wrong output for the new example $\mathbf{s}$. In this case, $\mathbf{w}$ is modified such that the local field f is changed in the direction of correcting the error on $\mathbf{s}$.

By averaging over $\mathbf{s}$ of the RHS of Eq. (2) it can be seen that the above dynamics is a stochastic approximation of the gradient descent rule given by $\mathbf{w}' - \mathbf{w} = -\eta\nabla E(\mathbf{w})$; E is the differentiable energy function

$$E(\mathbf{w}) = \langle \epsilon(\mathbf{w};\mathbf{s})|f(\mathbf{w};\mathbf{s})| \rangle \tag{3}$$

where $\langle ... \rangle$ denotes average with respect to the input distribution $D\mathbf{s}$. The performance of the learner is usually measured by the generalization error, defined as the probability of disagreement between the system and the target function on a random input

$$\epsilon_g(\mathbf{w}) = \langle \epsilon(\mathbf{w};\mathbf{s}) \rangle \tag{4}$$

We first discuss the behavior for small fixed η. Although $\mathbf{w} = \mathbf{w}_0$ is a fixed point of the dynamics, Eq. (2), it is unstable for any finite fixed value of η. Starting from $\delta\mathbf{w} \equiv \mathbf{w} - \mathbf{w}_0 \neq 0$ the errors induce random fluctuations in $\mathbf{w}$ that lead at long times to an equilibrium distribution of $\delta\mathbf{w}$ with a width proportional to η. This implies that as time n grows to ∞ the typical generalization error is $\epsilon_g(\infty) \propto \eta$. To show this we convert Eq. (2) into a Markov equation for the probability distribution, $P(\mathbf{w}, n)$,

$$P(\mathbf{w}, n+1) = \int d\mathbf{w}' T(\mathbf{w}|\mathbf{w}') P(\mathbf{w}', n) \tag{5}$$

where $T(\mathbf{w}|\mathbf{w}') = \langle \delta(\mathbf{w} - \mathbf{w}' - \eta \mathbf{g}(\mathbf{w}'; \mathbf{s})) \rangle$ is the transition matrix from $\mathbf{w}'$ to $\mathbf{w}$. Evaluating Eq. (5) in the limit of small fixed η, we find that the equilibrium distribution, P_∞, has the scaling form,

$$P_\infty(\mathbf{w}; \eta) = \eta^{-1} F(\eta^{-1} \delta\mathbf{w}) \tag{6}$$

where $F(\mathbf{x})$ is a distribution which obeys the equation

$$\hat{t}F(\mathbf{x}) \;\equiv\; \int d\mathbf{y}\, t(\mathbf{x}|\mathbf{y}) F(\mathbf{y}) = 0 \;. \tag{7}$$

The matrix $\hat{t}$ is the scaling form of $T(\mathbf{w}|\mathbf{w}') - \delta(\mathbf{w} - \mathbf{w}')$, and is given by

$$\begin{aligned} t(\mathbf{x}|\mathbf{y}) = & -\delta(\mathbf{y} - \mathbf{x}) \, \langle | \; \mathbf{x} \cdot \nabla f_0(\mathbf{s}) | \; \rangle_B \\ & \sum_{\sigma = \pm 1} \langle \; \delta(\mathbf{y} - \mathbf{x} - \sigma \nabla f_0(\mathbf{s})) \theta(\mathbf{y} \cdot (\mathbf{y} - \mathbf{x}))(\mathbf{y} \cdot (\mathbf{y} - \mathbf{x})) \; \rangle_B \end{aligned} \tag{8}$$

Here $\nabla f_0(\mathbf{s})$ is the value of the gradient of $f(\mathbf{w}; \mathbf{s})$ at $\mathbf{w} = \mathbf{w}_0$, and $\langle \ldots \rangle_B$ means averaging over the decision boundary of $\mathbf{s}$, namely, $\int D\mathbf{s} \delta(f_0(\mathbf{s}))$. This result is valid provided the input distribution is smooth and nonvanishing near the decision boundary, and ∇f_0 is not identically zero on the decision boundary. Finally, from the definition of ϵ_g, Eq.(4), it follows that for small $\delta\mathbf{w}$, ϵ_g is homogeneous of order 1, and hence scales linearly with η. Thus, we obtain for the infinite time limit of the average ϵ_g at small η,

$$\epsilon_\infty = \epsilon_0 \eta; \qquad \epsilon_0 \equiv \int d\mathbf{x} F(\mathbf{x}) \, \langle | \; \mathbf{x} \cdot \nabla f_0(\mathbf{s}) | \; \rangle_B \tag{9}$$

It should be noted that, unlike other on-line learning problems [1, 2, 3], the equilibrium distribution in our case is not Gaussian.

One way to guarantee asymptotic convergence to $\mathbf{w}_0$ is to adopt a fixed power law schedule in which η is a decreasing function of time n: $\eta(n) = \eta_0 n^{-z}$ where $0 < z < 1$. In this case, $\mathbf{w}_0$ is a stable fixed point of the dynamics and $P(\mathbf{w}, n)$ at long times is of the form

$$P(\mathbf{w}, n) = \eta^{-1}(n) (F(\delta\mathbf{w} \eta^{-1}(n)) + n^{z-1} G(\delta\mathbf{w} \eta^{-1}(n))) \tag{10}$$

where F is the stationary distribution, given by Eq. (7). The coefficient of the correction term, G, solves the inhomogeneous equation

$$\eta_0^{-1} z (\mathbf{x} \cdot \nabla F + F(\mathbf{x})) = \hat{t} G(\mathbf{x}) \tag{11}$$

where the operator $\hat{t}$ is defined in Eq. (8). Thus, to leading order in inverse time, the system follows adiabatically the finite-η stationary distribution, and $\epsilon_g(n)$ vanishes asymptotically as $\epsilon_g(n) = \epsilon_0 \eta(n) \propto n^{-z}$, where ϵ_0 is given by Eq. (9). The optimal schedule is obtained for $\eta = \eta_0 n^{-1}$. In this case, the scaled distribution $F(\mathbf{x})$ solves the homogeneous equation

$$\eta_0^{-1}(\mathbf{x} \cdot \nabla F + F(\mathbf{x})) = \hat{t} F(\mathbf{x}) \tag{12}$$

This equation is expected to have a well-behaved solution at least for sufficiently large η_0, implying that $\epsilon_g \propto n^{-1}$ provided that η_0 is smaller than some critical value.

The above analysis assumes that the system is in the neighborhood of the global minimum $\mathbf{w}_0$, and thus ignores completely the effect of local minima. In fact, the above power-law schedule for the learning rate poses a major problem concerning the ability to escape from a local minimum of $E(\mathbf{w})$, since it leads with probability 1 to convergence to the nearest local minimum [5]. On the other hand, if η is constant, the fluctuations in the dynamics may be sufficient to escape from local minima so that at long time the system will spend most of its time in the basin of attraction of the global minimum [7, 8]. However, as noted above, with fixed η, the system will not reach the optimal solution, but will maintain a residual error of order η. As the escape time from local minima is expected to scale exponentially with η, a small decrease in the generalization error requires a large increase in time. A well known alternative is to choose $\eta(n) \sim 1/\log n$, which guarantees escape from local minima[6]. Unfortunately, the rate of decrease of ϵ_g then scales as $\sim 1/\log n$, which is extremely slow. An additional undesired feature of choosing a learning rate that is an explicit function of time is the introduction of an absolute origin of time. Consequently, as time passes, the system becomes less able to learn new functions or to adapt to changes in the input distribution.

To overcome the above problems we propose a simple model in which η is a dynamical variable that obeys the following update rule

$$\eta' - \eta = \lambda \eta \left(A \epsilon(\mathbf{w}; \mathbf{s}) - \eta \right) \tag{13}$$

where A, λ are constants. This equation means that η is effectively a running average of the past errors. Note that our model incorporates the two important limits: the case $A = 0$ corresponds to $\eta \approx 1/(\lambda n)$ whereas $\lambda = 0$ corresponds to constant η. The behavior of this rule is the following. If the system is in the basin of attraction of the global minimum, η may decrease to zero together with ϵ_g yielding fast convergence to $\mathbf{w}_0$. In fact, using the above scaling properties, Eq. (13) implies that near $\mathbf{w}_0$, $\eta \approx \eta_0 n^{-1}$, with

$$\eta_0 = \frac{1}{\lambda(1 - A\epsilon_0)} \tag{14}$$

which has to be evaluated self-consistently, through Eqs. (9) and (12). Thus, for sufficiently small A, the asymptotic behavior is the same as the optimal convergence rate. On the other hand, if the system is near a local minimum of E, then ϵ_g and therefore also η are nonzero, which will lead to the eventual escape from the local minimum[9]. Finally, if the target or the input distribution change in time, the error will increase, and with it the capability of the system to adapt fast.

We first demonstrate our results for on-line perceptron learning, $\sigma_0(\mathbf{s}) = \mathrm{sgn}(\mathbf{w}_0 \cdot \mathbf{s})$, where $\mathbf{w}_0, \mathbf{s} \in R^N$ and $\mathbf{w}_0 \cdot \mathbf{w}_0 = 1$. We assume that the input distribution is a Gaussian centered at zero with a unit variance in each direction. A perceptron $\mathbf{w}$ is trained by the algorithm Eq. (2), which reduces to the well-known perceptron rule $\mathbf{g} = \theta(-\mathbf{w} \cdot \mathbf{s}\sigma_0(\mathbf{s}))\sigma_0(\mathbf{s})\mathbf{s}$. To guarantee that the algorithm does not diverge we add a normalization at each time step so that $\mathbf{w} \cdot \mathbf{w} = 1$ at all times. The learning rate η is modified via Eq. (13). We consider a large N limit in which the constant NA in Eq. (13), and hence $N\eta$, remain finite. The change in the overlap $R = \mathbf{w} \cdot \mathbf{w}_0$ and in η at each iteration is thus of order $1/N$ and a change of order 1 occurs after presentation of order N examples. The generalization error, Eq. (4), is $\epsilon_g = \pi^{-1} \cos^{-1} R$.

Multiplying Eq. (2) by $\mathbf{w}_0$, after incorporation of the normalization operation, and averaging with respect to the input distribution, yields the following differential equations for $R(\alpha)$ and $\eta(\alpha)$ where $\alpha = n/N$,

$$\frac{dR}{d\alpha} = \eta \frac{1 - R^2}{\sqrt{2\pi}} - \eta^2 \frac{R \cos^{-1} R}{2\pi}. \tag{15}$$

$$\frac{d\eta}{d\alpha} = \lambda\eta(A\pi^{-1} \cos^{-1} R - \eta) \tag{16}$$

For simplicity we have replaced the order unity quantities NA and $N\eta$ here and in the following by A and η, respectively. Terms of order $\sqrt{\eta/N}$ have been neglected.

When η is held fixed ($\lambda = 0$), Eq. (15) has a stable fixed point at $R < 1$, and hence ϵ_g converges to an η-dependent nonzero value $\epsilon_\infty(\eta) = \eta/\sqrt{2\pi^3} + O(\eta^2)$.

For asymptotic convergence to the target perceptron, a time-dependent η is necessary. Solving Eqs. (15) and (16) for positive λ, we find that for $0 < A < \sqrt{2/\pi}$, the only stable fixed point is $\eta = 0, \epsilon_g = 0$. The asymptotic approach to this fixed point is given by

$$\eta \sim \frac{\eta_0}{\alpha}, \quad \epsilon_g \sim \frac{1}{2\pi^3} \frac{\eta_0^2}{(\eta_0 - \sqrt{2\pi})} \frac{1}{\alpha} \tag{17}$$

where η_0 solves the quadratic equation $0 = \eta_0^2 \lambda (1 - A/\sqrt{2\pi^3}) - \eta_0(\lambda + 1/\sqrt{2\pi}) + \sqrt{2\pi}$, which is the same as the general relation, Eq. (14). The optimal convergence is achieved for $4A/\pi = 2\sqrt{2\pi} - 1/\lambda$ in which case $\epsilon_g \sim 1.27/\alpha$. This is not far from the batch asymptotic [10] $\epsilon_g(\alpha) \sim 0.625/\alpha$.

The advantage of the learning rate dynamics (13) over fixed learning rate schedules is apparent when there is a sudden change in the target. Suppose that at time $\alpha_0 \gg 1$, the target perceptron $\mathbf{w}_0$ is replaced by a different, randomly chosen vector $\mathbf{w}_0'$. The rise in generalization error to $\epsilon_g \approx 0.5$ affects the learning rate dynamics (13), resulting in a quick reset of the learning rate η from $1/\alpha_0$ to order unity. Following this resetting time of order $\ln \alpha_0$, the dynamic will converge fast to the new target, as it did for the original one. On the other hand, with the fixed schedule $\eta = \eta_0/\alpha$, it takes time of the order $\alpha_0^{\eta_0/(\eta_0-\sqrt{2\pi})}$ to return to a $1/\alpha$ decrease in ϵ_g, as can be shown by analyzing Eq. (15). This is demonstrated in the computer simulations of Fig. 1.

To study convergence in the presence of local minima of the energy, we consider a target function generated by a two-layer network, called a committee machine. This is another instance of the general class of dichotomies Eq. (1) with

$$f(\mathbf{w};\mathbf{s}) = \sum_{l=1}^{K} \tanh(\beta \mathbf{w}^l \cdot \mathbf{s}). \tag{18}$$

The number of hidden units (committee members) is K and their gain constant is β. We assume that the target function is generated by the same architecture with K vectors obeying $\mathbf{w}_0^l \cdot \mathbf{w}_0^{l'} = \delta_{ll'}$. The K vectors $\{\mathbf{w}^l\}_{l=1}^K$ are trained according to the rule Eq. (2), with a normalization step that keeps $\mathbf{w}^l \cdot \mathbf{w}^l = 1$ at all times. The learning rate η obeys Eq. (13), and as before A is scaled with N so that $N\eta(n)$ remains finite in the large N limit. The energy, Eq. (3), in this case, depends on the system's weights through the overlap matrix $R_{kl} = \mathbf{w}_0^k \cdot \mathbf{w}^l$. There are $K!$ global minima of this energy, given by $\mathbf{w}^l = \mathbf{w}_0^{\pi(l)}$ where $\pi(l)$ is an arbitrary permutation of $l = 1, ..., K$. yielding perfect generalization. Due to the permutation symmetry of the system, the energy function posseses also a symmetric extremum with $R_{kl} = R$, for which the generalization error is finite. For sufficiently large β this extremum is a local minimum[11, 12].

The behavior of this system was studied by simulations with $N = 150$, $K = 3$ and $\beta = 1$, which posseses a symmetric local minimum. Figure 2 compares three fixed schedules for η with our Eq. (13). Only our algorithm both escapes from the local minimum and converges rapidly to the global minimum of zero error.

In summary, we have investigated on-line learning of general thresholded smooth functions with various schedules of learning rates. We have shown that simple update rules for the weights and the learning rate, Eqs. (2) and (13), yield not only local but also global convergence to the zero error state, with a convergence rate that is inversely proportional to the number of examples. This power-law is probably the optimal performance for learning of this class of functions, as suggested by known results on batch learning[10, 13]. More complex on-line rules may, however, improve the coefficient of the power-law, as suggested by the results of Ref. [14] regarding on-line perceptron learning. An important question is how the escape times of our algorithm grow with the size of the system. Another important issue is the extension of our results to the on-line learning of noisy data where the optimal state has a finite error [4].

Acknowledgements: We are grateful to Y. Freund, M. Kearns, and R. Schapire for helpful conversations. HS is partially supported by the Fund for Basic Research of the Israeli Academy of Arts and Sciences.

References

[1] T. Heskes and B. Kappen, Phys. Rev. **A44**, 2718 (1991).

[2] L. K. Hansen, R. Pathria, and P. Salamon, J. Phys. **A26**, 63 (1993).

[3] G. Radons, J. Phys. **A26** 3455 (1993).

[4] Partial, preliminary results of this work have been reported in N. Barkai, H. S. Seung and H. Sompolinsky, *Advances in Neural Infromation Processing System 6*, to be published.

[5] H. J. Kushner and D. S. Clark, *Stochastic approximation methods for constrained and unconstrained systems*, (Springer, Berlin, 1978).

[6] H. J. Kushner, SIAM J. Appl. Math **47**, 169 (1987)

[7] G. B. Orr and T. K. Leen, *Advances in Neural Infromation Processing System 5* (San Mateo, CA: Morgan Kaufmann Publishers, 1993).

[8] T. Heskes, E. T. P. Slijpen, and B. Kappen, Phys. Rev. **A 46**, 5221 (1992).

[9] This argument assumes that the noise generated by the fluctuations in the force $\mathbf{g}(\mathbf{w};\mathbf{s})$, Eq. (2), are sufficiently large relative to its mean. To guarantee it in general one may add to $\mathbf{g}$ a term $\epsilon(\mathbf{w};\mathbf{s})\mathbf{z}$ where $\mathbf{z}$ is an uncorrelated Gaussian noise. As long as its variance is sufficiently small, it will preserve the asymptotic properties of the system.

[10] H. S. Seung, H. Sompolinsky, and N. Tishby, Phys. Rev. **A 45**, 6056 (1992).

[11] K. Kang, J.-H. Oh, C. Kwon, and Y. Park, Phys. Rev. **E 48**, 4805 (1993); H. Schwarze, J. Phys. **A 26**, 5781 (1993).

[12] On-line learning of a two-layer network with an *unstable* symmetric state has been recently studied, M. Biehl and H. Schwarze, preprint. See also Y. Kabashima, J. Phys. A27, 1917 (1994).

[13] S.-I. Amari, N. Fujita, and S. Shinomoto, Neural Comput. 4, 605 (1992); S.-I. Amari, Neural Networks **6**, 161 (1993).

[14] M. Biehl and P. Riegler, preprint, 1994.

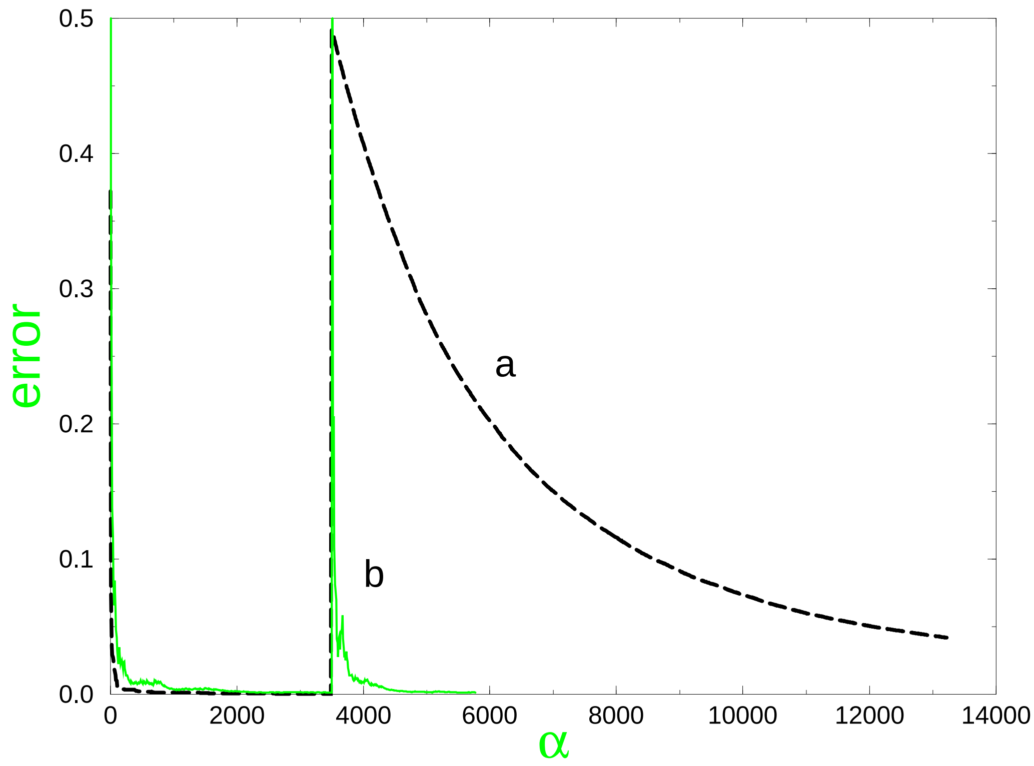

Figure 1: On-line learning of a perceptron rule, with $N = 20$. The direction of the target rule is changed to a random direction at $\alpha = 3500$, where $\alpha = n/N$. (a) A fixed schedule $\eta(\alpha) \propto \alpha^{-1}$. (b) The schedule of Eq. (13). Both schedules are with optimal parameters.

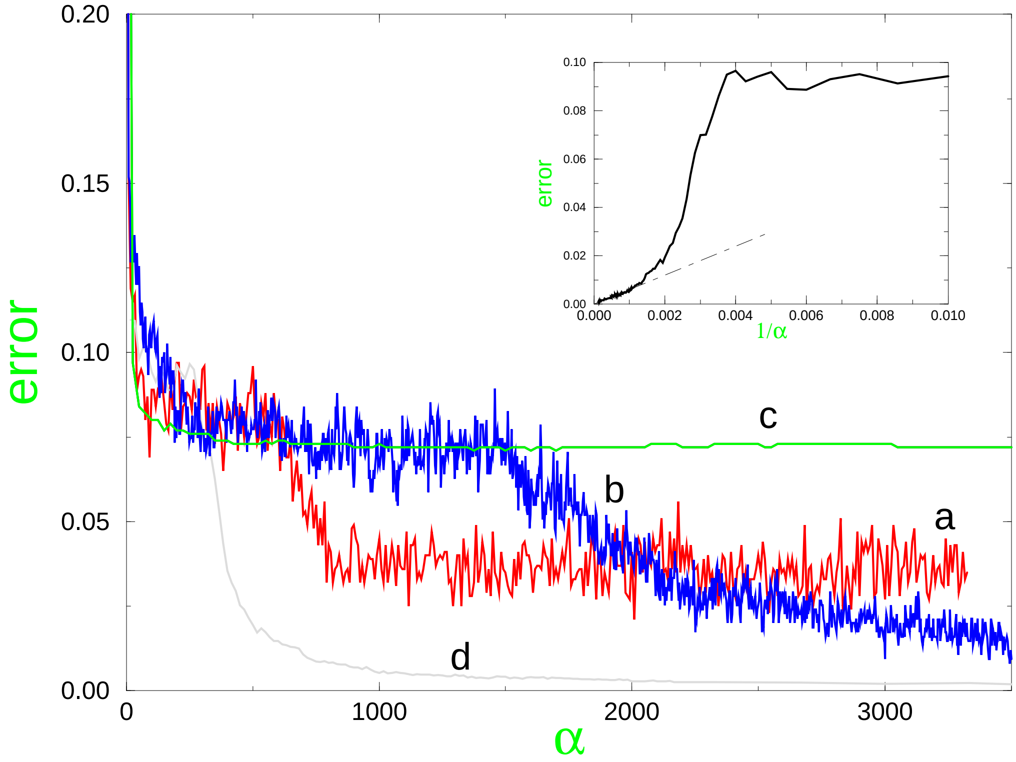

Figure 2: On-line learning of committee machine Eq. (18) with $N = 150$, $K = 3$, and $\beta = 1$. For small α the learner is drawn into the basin of attraction of the symmetric local minimum, with $\epsilon_g \approx 0.07$. (a) $\eta = 0.2$: The system escapes from this state but does not achieve zero error. (b) $\eta = 1/\log(1.01 + \alpha)$: The system achieves zero error but converges extremely slowly. (c) $\eta = 2.5/\alpha$: The system stays at the local minimum. (d) Eq. (13) with $A = 3$ and $\lambda = 1$: the system escapes from the local minimum and converges to the global minimum rapidly. The Inset exhibits the power-law convergence of this algorithm.