Perfect loss of generalization due to noise in K=2 parity machines

1994 J. Phys. A: Math. Gen. 27 1917

(http://iopscience.iop.org/0305-4470/27/6/017)

J. Phys. A: Math. Gen. 27 (1994) 1917–1927. Printed in the UK

Perfect loss of generalization due to noise in $K = 2$ parity machines

Yoshiyuki Kabashima†
Department of Physics, Kyoto University, Kyoto 606, Japan

Received 2 August 1993

Abstract. Learning in a specific type of multilayer network referred to as a $K = 2$ parity machine is studied in the limit that both the system size N and the number of examples m become infinite while the ratio $\alpha = m/N$ remains finite. The machine consists of $K = 2$ hidden units with non-overlapping receptive fields each of size $N/2$. The output is the sign of the product of the two hidden units for each input. We investigate incremental learning by empirically using a least-action algorithm in the following two learning paradigms. In the first, it is assumed that each example is transmitted perfectly to a student. We show that an ability to generalize emerges as the rescaled length of the connection vector l reaches a critical value l_c. Further, we show that a student can identify the target exactly in the limit $\alpha \to \infty$, where the prediction error ε decreases to zero as $\varepsilon \sim 0.441\alpha^{-1/3}$. In the second paradigm, we examine what happens if each teacher signal is reversed to the opposite sign at a noise rate λ. For small λ, it is found that the prediction error converges to a finite value of $O(\sqrt{\lambda})$ in $O(\lambda^{-3/2})$ iterations. However, for a noise rate beyond a critical value $\lambda_c \sim 0.175$, the student cannot acquire any generalization ability even as $\alpha \to \infty$.

1. Introduction

For many years, great effort has been made to understand learning machines. Single-layer perceptrons have been an especially central target of research because of their simplicity, and several remarkable results have been obtained (e.g. see Seung *et al* 1992). However, as is well known, the class of problems solvable by simple perceptrons is limited. It is therefore of significant interest now to investigate multilayer networks.

In this paper, we examine learning in simple multilayer networks known as $K = 2$ parity machines. These machines consist of N input units, $K = 2$ hidden units and one output unit (figure 1). The input units are divided into two disjoint sets of the same size $N/2$. The kth hidden unit is connected to the kth set of input units via a connection vector $J_k = (j_{1k}, j_{2k}, \cdots, j_{(N/2)k})$, where $k = 1, 2$. For each input $x = (x_1, x_2) = (x_{11}, x_{21}, \cdots, x_{(N/2)1}, x_{12}, x_{22}, \cdots, x_{(N/2)2})$, the sign of the product of the two hidden units $\operatorname{sign}((J_1 \cdot x_1)(J_2 \cdot x_2))$ is returned.

This kind of multilayer network for general K was first introduced by Mitchison and Durbin (1989). They addressed the problem of how many random dichotomies can be implemented in such a network. For the fully connected version, they found an upper bound on the capacity which scales as $O(NK \log K)$ as $K \to \infty$. By using the replica trick, Hansel *et al* (1992) investigated $K = 2$ networks in which $m = N\alpha$ examples are memorized. They found a phase for $\alpha < \alpha_* = \pi^2/8$ where the student memorizes the entire

† Present address: Department of Physics, Nara Women's University, Nara 630, Japan.

0305-4470/94/061917+11$19.50

learning set without being able to generalize the rule from the examples. They explained the existence of this phase as follows. For even K, two machines represented by J and $-J$ provide exactly the same input–output relation. This property makes the free energy of the system an even function of the overlap q of the student with the teacher. This free energy consists of two terms. One is an entropy term and the other is a training energy due to mistakes made on the given set of examples. A balance of these two terms determines the thermodynamic state. For $\alpha < \alpha_*$, the entropy term is dominant because the training energy is small when there are a small number of examples. Thus, a paramagnetic solution $q = 0$ which maximizes the entropy is realized. This represents a rote memorization phase. However, this solution becomes unstable at $\alpha = \alpha_*$ because the training energy for the paramagnetic state increases as learning continues. Eventually, the symmetry is spontaneously broken and a non-zero replica symmetric solution is obtained. This corresponds to the onset of generalization. They also showed that this replica symmetric solution increases monotonically to 1 as $1-q \sim (1/\alpha)^2$ for $\alpha \to \infty$, which implies that the generalization error ε scales asymptotically as $\varepsilon \sim \mathrm{O}(1/\alpha)$, a result which is also obtained by many authors for general systems (e.g. see Amari *et al* 1992).

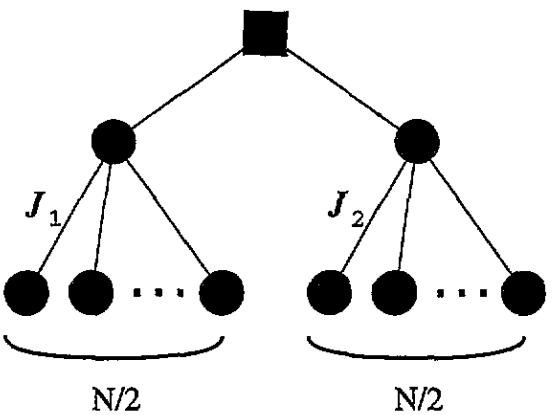

Figure 1. Schematic representation of a $K = 2$ parity machine.

In our framework, the objective of learning is not to memorize a given set of examples correctly but to extract a target rule from examples. A number of learning algorithms which aim to embed a given set of examples in a network have been proposed, although none of them is ensured to converge in a finite time for multilayer networks. Among such empirical methods, Mitchison and Durbin (1989) reported that a strategy called the 'least action algorithm' (LAA) exhibits very good performance for numerical experiments, which has to be explained theoretically. We will show that when each example is independently drawn from a uniform distribution, this algorithm enables a student machine to identify the target relation exactly in our problem unless teacher signals are disrupted by noise.

In the original version of LAA, a set of examples are cyclically shown to a student until the examples are memorized. Instead of proceeding in such a way, we will apply LAA incrementally. In the following, we consider the limit that both the system size N and the number of examples become infinite, keeping $\alpha = m/N$ finite. This then allows the learning process to be described by the dynamics of the following two order parameters: the overlap $q = (\boldsymbol{J} \cdot \boldsymbol{J}^0)/(|\boldsymbol{J}||\boldsymbol{J}^0|)$ and the length of the connection vector $l = |\boldsymbol{J}|/\sqrt{N/2}$, where $\boldsymbol{J}$ and $\boldsymbol{J}^0$ represent the connection vectors to a hidden unit of a student and a teacher,

respectively.

The results obtained in this paper are summarized as follows. When none of the examples are disrupted by noise, it is shown that a generalization ability emerges as the rescaled length of the vector reaches a critical value $l_c \sim 0.696$. However, for $l < l_c$, the generalization ability is instead lost as learning goes on. We also show that the student can extract the target rule exactly in the limit $\alpha \to \infty$, where the prediction error ε converges to zero as $\varepsilon \sim 0.441\alpha^{-1/3}$. Next, we investigate what happens when the teacher signal is reversed to the opposite sign at a rate λ. For small λ, it is shown that a prediction error converges not to zero but to a finite value of $O(\sqrt{\lambda})$ after $O(\lambda^{-3/2})$ examples are given. However, if the noise rate is greater than a critical value $\lambda_c \sim 0.175$, the student cannot acquire any generalization ability even as $\alpha \to \infty$.

2. Learning without noise

In this paper, student machines are assumed to have the same architecture as that of the teacher. Thus, it is possible that a student extracts the target rule exactly. In this section, we further assume that every teacher signal is correctly transmitted to the student. As mentioned above, only empirical algorithms are known for the memorization of a given set of examples in multilayer networks. Mitchison and Durbin (1989) found that a slight modification of Nilsson's strategy (1965) exhibits very good performance for general K parity machines and called it the 'least action algorithm' (LAA). For a given set of P examples $(\boldsymbol{x}^1, \sigma^1), (\boldsymbol{x}^2, \sigma^2), \ldots, (\boldsymbol{x}^P, \sigma^P)$, where $\boldsymbol{x}^\mu = (\boldsymbol{x}_1{}^\mu, \boldsymbol{x}_2{}^\mu)$ forms the input vector and $\sigma^\mu \in \{-1, +1\}$ is the teacher signal ($\mu = 1, 2, \ldots, P$), the original version of LAA works in the present machine as follows. (i) If the student gives a correct answer, the connection $\boldsymbol{J} = (\boldsymbol{J}_1, \boldsymbol{J}_2)$ remains unchanged and the next example is presented. (ii) If a wrong answer is returned, local fields $(\boldsymbol{J}_1 \cdot \boldsymbol{x}_1)$ and $(\boldsymbol{J}_2 \cdot \boldsymbol{x}_2)$ are computed and the connection vector $\boldsymbol{J}_k$, which corresponds to the hidden unit k ($k = 1$ or 2) of which the absolute value of local field is less than that of the other, is updated with the standard perceptron algorithm. (iii) Returns to step (i). This procedure is iterated cyclically until the student memorizes the entire set of examples or the number of iterations reaches a given upper bound.

Here, we apply the algorithm in an incremental way. Namely, an example is shown, the connection $\boldsymbol{J}$ is updated according to the LAA, and the example is never again referred to. We assume that the component of every input $\boldsymbol{x}_k$ ($k = 1, 2$) is drawn independently from a uniform distribution describing the unit sphere $S^{N/2}$, and denote the connection after m examples are presented as $\boldsymbol{J}^m = (\boldsymbol{J}_1{}^m, \boldsymbol{J}_2{}^m)$. Hence, the update rule for the $(m+1)$th example $(\boldsymbol{x}, \sigma)$ can be represented as

$$\begin{aligned} \boldsymbol{J}_k{}^{m+1} = \boldsymbol{J}_k{}^m &- \Theta(-\sigma \cdot (\boldsymbol{J}_1{}^m \cdot \boldsymbol{x}_1)(\boldsymbol{J}_2{}^m \cdot \boldsymbol{x}_2)) \\ &\times \Theta(|(\boldsymbol{J}_j{}^m \cdot \boldsymbol{x}_j)| - |(\boldsymbol{J}_k{}^m \cdot \boldsymbol{x}_k)|) \cdot \operatorname{sign}(\boldsymbol{J}_k{}^m \cdot \boldsymbol{x}_k) \cdot \boldsymbol{x}_k \end{aligned} \tag{1}$$

where $k, j = 1, 2$ ($k \neq j$) and $\Theta(y)$ is Heaviside's step function.

Let us investigate the dynamics of the learning rule (1) in the limit that both the system size N and the number of examples m go to infinity, keeping $\alpha = m/N$ finite. We define the overlap of the student with the teacher as $q_k = (\boldsymbol{J}_k \cdot \boldsymbol{J}_k{}^0)/(|\boldsymbol{J}_k||\boldsymbol{J}_k{}^0|)$ and also define a rescaled length of the vector as $l_k = |\boldsymbol{J}_k|/\sqrt{N/2}$ ($k = 1, 2$). For simplicity, let us abbreviate the following random variables as

$$\begin{aligned} \sqrt{N/2}(\boldsymbol{J}_1 \cdot \boldsymbol{x})/|\boldsymbol{J}_1| &= u \qquad \sqrt{N/2}(\boldsymbol{J}_1{}^0 \cdot \boldsymbol{x})/|\boldsymbol{J}_1{}^0| = v \\ \sqrt{N/2}(\boldsymbol{J}_2 \cdot \boldsymbol{x})/|\boldsymbol{J}_2| &= s \qquad \sqrt{N/2}(\boldsymbol{J}_2{}^0 \cdot \boldsymbol{x})/|\boldsymbol{J}_2{}^0| = t. \end{aligned} \tag{2}$$

As a consequence of the central limit theorem, these pairs of random variables (u, v) and (s, t) obey two-variable normal distributions

$$\begin{aligned} P_q(u,v) &= \frac{1}{2\pi\sqrt{(1-q^2)}}\exp\left[-\frac{(u^2+v^2-2quv)}{2(1-q^2)}\right] \\ P_q(s,t) &= \frac{1}{2\pi\sqrt{(1-q^2)}}\exp\left[-\frac{(s^2+t^2-2qst)}{2(1-q^2)}\right] \end{aligned} \tag{3}$$

in the limit $N \to \infty$. We assume that $q_1 = q_2 = q$ and $l_1 = l_2 = l$ from the symmetry of the indices $k = 1$ and 2. Under these assumptions, a set of stochastic difference equations

$$\begin{cases} {l_{m+1}}^2 = {l_m}^2 + (2/N)[E(q_m) - 2F(q_m)\, l_m] + \mathrm{O}(1/N)\text{fluctuation} \\ q_{m+1} = [l_m q_m - (2/N)G(q_m)]/l_{m+1} + \mathrm{O}(1/N)\text{fluctuation} \end{cases} \tag{4}$$

are derived from equations (1) and (3), where $E(q) = \langle 1\rangle_q$, $F(q) = \langle \mathrm{sign}(u)u\rangle_q$ and $G(q) = \langle \mathrm{sign}(u)v\rangle_q$, where

$$\langle \ldots\rangle_q = \int_{\substack{(u\cdot s)(v\cdot t)<0,\\ |u|<|s|}} du\, dv\, ds\, dt\, P_q(u,v)P_q(s,t)(\ldots). \tag{5}$$

The first and second of equations (4) are obtained directly by squaring the learning rule (1) and by taking its projection along the direction of the teacher J^0, respectively. The effects of random variables x are separated into the expectations $E(q)$, $F(q)$ and $G(q)$, and the fluctuation factors. In equation (5), the domain of integration corresponds to the condition of modification that an answer is wrong ($(u\cdot s)(v\cdot t) < 0$) and the absolute value of the local field in question is less than that of the other ($|u| < |s|$). The expectations $E(q)$, $F(q)$ and $G(q)$ are plotted in figure 2. We find that $E(q)$ and $F(q)$ are even functions, and that $G(q)$ is an odd function with respect to q. The paramagnetic state $q = 0$ corresponds to the case in which the student gives an answer at random, and then the connection is updated with probability 1/2. Since both J_1 and J_2 are updated with equal probability, $E(0)$ becomes $1/2 \times 1/2 = 1/4$. For q sufficiently close to 1, namely $q = 1 - \epsilon$, the probability of a mistake is proportional to $\cos^{-1}(q) \sim \mathrm{O}(\epsilon^{1/2})$ and such mistakes occur only for $u \sim v \sim \cos^{-1}(q) \sim \mathrm{O}(\epsilon^{1/2})$. We therefore find that these expectations converge to zero as $E(1-\epsilon) \sim \mathrm{O}(\epsilon^{1/2})$, $F(1-\epsilon) \sim \mathrm{O}(\epsilon^{1/2}) \times \mathrm{O}(\epsilon^{1/2}) \sim \mathrm{O}(\epsilon)$ and $G(1-\epsilon) \sim \mathrm{O}(\epsilon^{1/2}) \times \mathrm{O}(\epsilon^{1/2}) \sim \mathrm{O}(\epsilon)$ as $\epsilon \to 0$. In the limit $N \to \infty$ and $m \to \infty$, keeping $\alpha = m/N$ finite, the fluctuation terms in equations (4) vanish as $O(1/\sqrt{N})$, and we finally obtain a pair of differential equations:

$$\begin{cases} \mathrm{d}l/\mathrm{d}\alpha = [E(q) - 2F(q)\, l]/l \\ \mathrm{d}q/\mathrm{d}\alpha = -[E(q) - 2(F(q) - G(q)/q)\, l]q/l^2. \end{cases} \tag{6}$$

We will make a rough sketch of the dynamics of equations (6) before proceeding further. First, their solutions have reversal symmetry $q \leftrightarrow -q$ because these equations are invariant under the reversal transformation with respect to q. We therefore can assume $q > 0$ without loss of generality. Next, assume that a vector of finite length is chosen as an initial state ($l \sim 0$). We find that the overlap q decreases during the course of learning until l reaches $E(q)/2(F(q) - G(q)/q)$. This implies that the generalization ability is now lost by learning. This can be explained as follows. Whether q increases or decreases is determined by a balance of two factors in the second equation. One is a factor represented by $E(q) - 2F(q)\, l$. When a vector is modified in some random direction, information of a specific direction is lost. These terms represent just such a decrease in the factor q by random modification of a vector when l is small. The other factor causes q to grow via

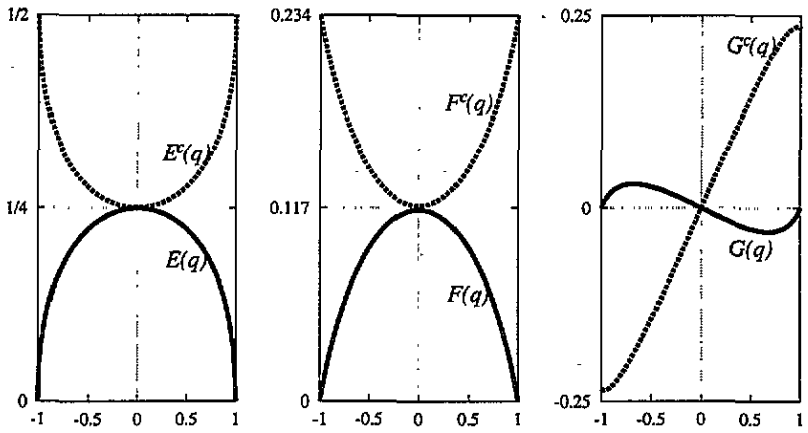

Figure 2. Expectations $E(q)$, $F(q)$ and $G(q)$ and their complements $E^c(q)$, $F^c(q)$ and $G^c(q)$ (see section 3).

learning, represented by the term $-2G(q)/ql$. For simple perceptrons, this factor always overcomes the decreasing factor even for small l because $G(q)$ becomes $O(1)$ and no deterioration of generalization occurs (Baum 1990). On the other hand, the symmetry under $q \leftrightarrow -q$ reduces $G(q)$ to $O(q)$ which is the same order of the decreasing factor in the present machines, and the relative size of the two factors reverses when l is small. This causes a rapid loss of generalization. The paramagnetic state $q = 0$ becomes unstable as the time l reaches $E(q)/2(F(q) - G(q)/q)$. This growth of q would be suppressed again if the condition $E(q)/2F(q) < l < E(q)/2(F(q) - G(q)/q)$ were satisfied. However, this inequality never holds because $G(q)/q$ is a non-positive function of q, and q finally grows to 1. If the length l is initially set greater than $E(q)/(2F(q))$, it continues to decrease until the condition $l < E(q)/(2F(q))$ holds. Once this inequality holds, l increases to infinity. In this case, q continues to grow for all time. Thus, we find that the perfect generalization state $q = 1$ and $l = \infty$ is obtained in the limit $\alpha \to \infty$, for any initial condition.

Let us investigate the details of these dynamics. If a vector of finite length is randomly chosen as an initial state, we can assume that l and q are initially placed in the vicinity of $(l, q) \sim (0, 0)$. Around $q \sim 0$, equations (6) are approximated as

$$\begin{cases} \mathrm{d}l/\mathrm{d}\alpha = \left[\dfrac{1}{4} - \dfrac{\sqrt{2}-1}{\sqrt{\pi}}l\right]/l \\ \mathrm{d}q/\mathrm{d}\alpha = -\left[\dfrac{1}{4} - \dfrac{2}{\pi^{3/2}}l\right]q/l^2. \end{cases} \tag{7}$$

From the condition $\mathrm{d}q/\mathrm{d}\alpha > 0$, we find that the student can start to generalize at the time that the rescaled length reaches a critical value

$$l_c = \frac{\pi^{3/2}}{8} \sim 0.696. \tag{8}$$

We plot the results of our numerical simulations of the problem, along with the theoretical predictions in figure 3. The data are consistent with the theoretical curves. In figure 3(a), both the numerical data and the theoretical curve are unremarkable at $l_c \sim 0.696$, but at $l_c' = \sqrt{\pi}/[4(\sqrt{2}-1)] \sim 1.069$, which corresponds to the condition $\mathrm{d}l/\mathrm{d}\alpha < 0$, a significant

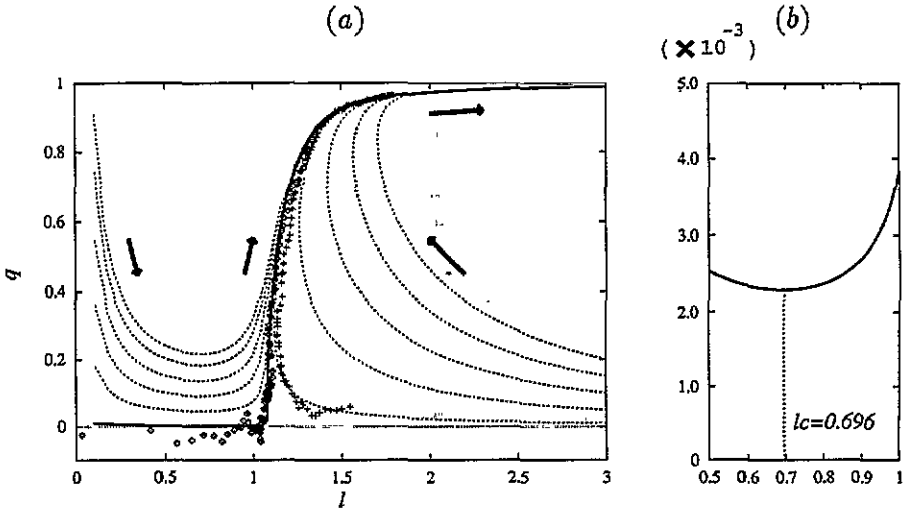

Figure 3. Evolution of the order parameters l and q by the least action algorithm. Numerical experiments are carried out for a system where $N = 2000$. (a) Markers represent the results of the experiments. They are placed for every interval $\Delta\alpha = 0.5$. Random vectors with length $l = 0.03$ and 1.57 are chosen as initial states. Lines are theoretical predictions using equations (6) for various initial conditions. (b) The region around the critical value l_c is enlarged. We find that q is switched to increase around $l_c \sim 0.696$ for the curve with the initial condition $(l, q) = (0.1, 0.01)$.

change occurs. However, q does indeed begin to increase at l_c, as is plotted in figure 3(b), but it is difficult to detect this critical point in a numerical experiment because statistical fluctuations of $O(1/\sqrt{N})$ for a finite system should be added to equations (6).

Next, we consider the asymptotic behavior in the limit $\alpha \to \infty$ in order to investigate how fast the student extracts the target rule. For $q = 1 - \epsilon$, equations (6) are approximated as

$$\begin{cases} \mathrm{d}l/\mathrm{d}\alpha = \left[\dfrac{\sqrt{2}}{\pi}\epsilon^{1/2} - \dfrac{2}{\sqrt{2\pi}}\epsilon l\right]/l \\ \mathrm{d}\epsilon/\mathrm{d}\alpha = \left[\dfrac{\sqrt{2}}{\pi}\epsilon^{1/2} - \dfrac{4}{\sqrt{2\pi}}\epsilon l\right]/l^2. \end{cases} \tag{9}$$

These equations have an asymptotic solution

$$\epsilon \sim 0.240\alpha^{-2/3}. \tag{10}$$

The prediction error for future examples ε is calculated from the overlap q as

$$\varepsilon = 2\left(\frac{\cos^{-1} q}{\pi}\right) - 2\left(\frac{\cos^{-1} q}{\pi}\right)^2. \tag{11}$$

The second term reflects the fact that a correct answer is finally produced if both local fields are discrepant with those of the teacher for $K = 2$ parity machines. From equations (10) and (11), we find that the prediction error converges to zero, e.g.

$$\varepsilon \sim 0.441\alpha^{-1/3} \tag{12}$$

as $\alpha \to \infty$.

If the student goes on memorizing the entire set of examples exactly during the course of learning, it is known that the prediction error ε converges to zero as $\varepsilon \sim O(\alpha^{-1})$ (for example, see Amari *et al* (1992)). The convergence of equation (12) seems very slow compared with this. However, this convergence is remarkably fast in the context of how much CPU time is required to obtain the prediction error ε. Hansel *et al* (1992) also reported the result of numerical experiments in their letter. In their experiments, a set of $m = N\alpha$ examples is drawn from a uniform distribution and embedded in the present $K = 2$ machine using the original version of LAA. The examples are shown cyclically to a student until the entire set is memorized. Although $\varepsilon \sim O(N/m)$ is obtained after memorization, the required number of iterations is not bounded for their strategy. As mentioned below, LAA approaches the perceptron algorithm (Rosenblatt 1962) in the limit $\varepsilon \to 0$. The number of updates which the perceptron algorithm must make to memorize a given set of examples is $O(1/d^2)$, where d is the minimum distance from an example to the classifying hyperplane (see, e.g., Minsky and Papert 1989). This indicates that the required number of iterations is bounded for a realization of an example set. However, when examples are chosen from a distribution, the expectation of this number becomes infinite because d can be arbitrarily small with a finite probability. This is one reason why computational time of the original LAA is not bounded in our learning paradigm. In addition, the correlation of the position of examples sometimes prevents the algorithm from determining a solution in a finite time for multilayer networks. In their experiment, Hansel *et al* found that such divergence is frequently observed around $\alpha \sim 3$. In order to avoid wasting computational resources for divergent cases, they stopped the algorithm when the number of iterations reached some fixed upper bound, even if the entire set was not yet memorized. This makes it impossible to obtain the exact convergence $\varepsilon \to 0$. On the other hand, the convergence of equation (12) states that we can obtain ε accuracy after $O(N/\varepsilon^3)$ examples are shown. The required time for this calculation is found to scale as $O(N^2/\varepsilon^3)$ in a serial computer. Solving a linear programming problem is another approach to find the target connection. Karmarkar's algorithm is known as a standard method for solving such problems. In order to obtain the same accuracy, this algorithm requires time of order $\sim O(N^{5.5}/\varepsilon^2)$ in a serial computer (Karmarkar 1984). Compared with these facts, convergence (12) is actually fast for large N.

A similar result was also obtained by Baum (1990) for simple perceptrons. He showed that for this machine the prediction error ε is obtained in polynomial time by using the perceptron algorithm incrementally. The required CPU time for this learning is $O(N^2/\varepsilon^3)$ in a serial computer, which is the same result as ours. The reason for this is as follows. For $q = 1 - \epsilon$, the region of the input space in which the sign of local field is different from that of the teacher becomes small as $O(\epsilon^{1/2})$ in each receptive field in $K = 2$ parity machines. In this case, a wrong answer is given almost only if one receptive field receives an input from such a discrepant region and the other receives an input in region $O(1)$ in which the sign of local field is consistent with that of the teacher. This ensures that, by using LAA, we can modify the connection vector which corresponds to the sign of local field that differs from that of the teacher with a probability of almost 1, namely by modifying a vector corresponding to the smaller local field. This gives almost the same effect on each connection vector as that of perceptron learning. As a result, convergence similar to that of Baum's result is achieved asymptotically by our system.

3. Learning with noise

In this final section, we investigate the effect of noise on learning. We assume that the teacher signal is inverted at a rate λ for each example, where the noise is uncorrelated with

the position of the input. The problem of finding a decision boundary for an originally stochastic binary relation is another example of learning where a student must infer the optimal parameters from stochastic teacher signals (Kabashima and Shinomoto 1992, 1993). However, this kind of learning is of a rather different nature than that of learning taken up here. When a target relation is originally stochastic, the teacher signal $s = +1$ or -1 is drawn from a conditional probability which is a continuous function of the position of the input. In such cases, it would seem to a student that the teacher signals are disrupted by noise which is correlated with the position of the input. The level of this noise is maximized in the vicinity of the decision boundary, where $s = +1$ and -1 are equally generated. This makes it difficult to infer the decision boundary. A naive strategy to find the optimal machine parameters is to minimize the empirical error. Although this exhaustive search requires much computation and memory capacity, the mean square error of estimation exhibits slow power-law convergence, with exponent 2/3. This convergence can be accelerated by elaborate methods of inference. However, it is shown that we cannot obtain the fastest convergence with exponent 1 for this type of problem (Kawanabe and Amari 1993), although such convergence is achievable by the error minimum strategy if the noise is uncorrelated with the position of the input. Thus, the effect of the noise which is correlated with the position of the input will require a different approach.

With noise, a correct answer is regarded as wrong with probability λ, and vice versa. As a result, the connection is updated at a rate λ for a correct answer and at a rate $1-\lambda$ for a wrong answer. This replaces the expectations $E(q)$, $F(q)$ and $G(q)$ in equations (6) with

$$\begin{aligned}\widetilde{E}_\lambda(q) &= (1-\lambda)E(q) + \lambda E^{\mathrm{c}}(q)\\ \widetilde{F}_\lambda(q) &= (1-\lambda)F(q) + \lambda F^{\mathrm{c}}(q)\\ \widetilde{G}_\lambda(q) &= (1-\lambda)G(q) + \lambda G^{\mathrm{c}}(q)\end{aligned} \tag{13}$$

where $E^{\mathrm{c}}(q) = \langle 1\rangle_q{}^{\mathrm{c}}$, $F^{\mathrm{c}}(q) = \langle \mathrm{sign}(u)u\rangle_q{}^{\mathrm{c}}$ and $G^{\mathrm{c}}(q) = \langle \mathrm{sign}(u)v\rangle_q{}^{\mathrm{c}}$, where

$$\langle\ldots\rangle_q{}^{\mathrm{c}} = \int_{\substack{(u\cdot s)(v\cdot t)>0,\\ |u|<|s|}} \mathrm{d}u\,\mathrm{d}v\,\mathrm{d}s\,\mathrm{d}t\; P_q(u,v)P_q(s,t)(\ldots). \tag{14}$$

In equation (14), the domain of integration corresponds to the condition of misled modification due to noise that an answer is *right* ($(u\cdot s)(v\cdot t) > 0$) and the absolute value of the local field in question is less than that of the other ($|u| < |s|$). We plot these complementary expectations $E^{\mathrm{c}}(q)$, $F^{\mathrm{c}}(q)$ and $G^{\mathrm{c}}(q)$ together with $E(q)$, $F(q)$ and $G(q)$ in figure 2. They are not independent of each other because of the complementarity between the conditions $(u\cdot s)(v\cdot t) > 0$ and $(u\cdot s)(v\cdot t) < 0$.

We first investigate how the asymptotic behaviour represented by equation (10) is modified if an infinitesimal noise rate λ is introduced. From equations (6) and (13), order parameters for $q = 1-\epsilon$ turn out to be subject to the following equations

$$\begin{cases}\mathrm{d}l/\mathrm{d}\alpha = \left[\dfrac{\sqrt{2}}{\pi}\epsilon^{1/2} - 2\left(\dfrac{1}{\sqrt{2\pi}}\epsilon + \lambda\dfrac{\sqrt{2}-1}{\sqrt{\pi}}\right)l\right]/l\\ \mathrm{d}\epsilon/\mathrm{d}\alpha = \left[\dfrac{\sqrt{2}}{\pi}\epsilon^{1/2} - \dfrac{4}{\sqrt{2\pi}}\epsilon\, l\right]/l^2\end{cases} \tag{15}$$

where unimportant terms, which are irrelevant to the leading terms with respect to λ in the solution, are omitted. In the first equation, we find that the effect of noise decreases with l according to $-\mathrm{O}(\lambda)\,l$. This makes l and ϵ converge to finite values in the limit $\alpha \to \infty$. The fixed point of equations (15) can be scaled as

$$(l_0, \epsilon_0) \sim (\mathrm{O}(\lambda^{-1/2}), \mathrm{O}(\lambda)). \tag{16}$$

We will calculate the prediction error by using equation (11) assuming that the prediction ability of the student is examined in an environment without noise. Equation (16) indicates that

$$\varepsilon_0 \sim \mathrm{O}(\lambda^{1/2}) \tag{17}$$

holds even though $\alpha \to \infty$. In other words, the student cannot extract the target rule perfectly in this case.

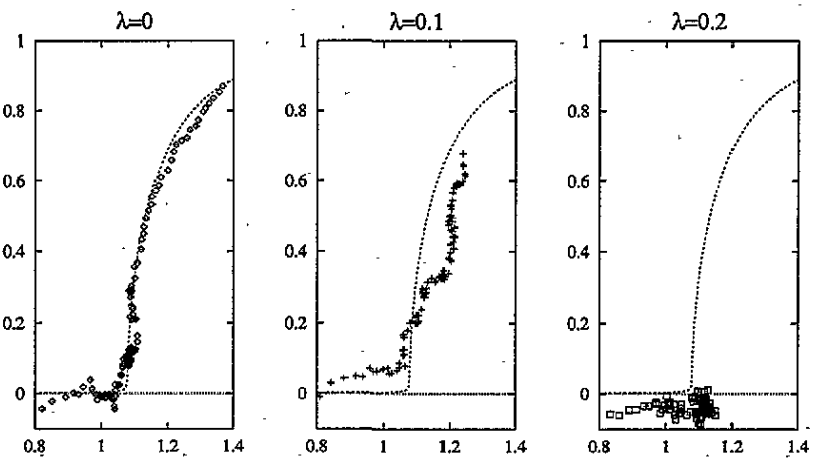

Figure 4. Deterioration of learning by noise. Every teacher signal is changed to the opposite sign by noise at a rate λ during the course of learning. Numerical experiments are carried out for a system where $N = 2000$ and with noise rates $\lambda = 0, 0.1$ and 0.2 (markers). The abscissa and the ordinate represent the rescaled length l and the overlap q, respectively. Markers are placed for every interval $\Delta\alpha = 0.5$. We find that a student cannot acquire any generalization ability for $\lambda = 0.2$, which is consistent with the theoretical prediction $\lambda_c \sim 0.175$. Lines represent the theoretical curves for $\lambda = 0$.

Next we investigate how fast this convergence is attained. We again consider the asymptotic behaviour in the limit $\alpha \to \infty$. By rescaling the variables l and ϵ as $l = \lambda^{-1/2}\tilde{l}$ and $\epsilon = \lambda\tilde{\epsilon}$, and linearizing equations (15) around the fixed point, we find that the deviation u from the the rescaled fixed point $(\tilde{l}_0, \tilde{\epsilon}_0)$ obeys a linearized equation

$$\mathrm{d}u/\mathrm{d}\alpha = -\begin{pmatrix} \mathrm{O}(\lambda^{3/2}) & 0 \\ \mathrm{O}(\lambda^{1/2}) & \mathrm{O}(\lambda^{1/2}) \end{pmatrix} u. \tag{18}$$

The matrix in equation (18) is found to have two positive eigenvalues $\beta_1 \sim \mathrm{O}(\lambda^{3/2})$ and $\beta_2 \sim \mathrm{O}(\lambda^{1/2})$, which means the asymptotic form of the solution is

$$\begin{aligned} l &\sim \mathrm{O}(\lambda^{-1/2})\left(1 + \mathrm{O}(\mathrm{e}^{-a\lambda^{3/2}\alpha})\right) \\ \epsilon &\sim \mathrm{O}(\lambda)\left(1 + \mathrm{O}(\mathrm{e}^{-a\lambda^{3/2}\alpha})\right). \end{aligned} \tag{19}$$

Equations (17) and (19) state that $\varepsilon \sim \mathrm{O}(\lambda^{1/2})$ is obtained in $\mathrm{O}(N/\varepsilon^3)$ iterations, which gives the same order of computation $\mathrm{O}(N^2/\varepsilon^3)$ in a serial computer as that of learning without noise. This safely reproduces the scaling relation (12) in the limit $\lambda \to 0$. In this sense, we can regard this learning strategy as robust for small levels of noise.

Note that the results obtained above hold only for low noise levels. The results of numerical experiments of learning with finite noise are plotted in figure 4. We find that the student does not acquire any generalization ability for $\lambda = 0.2$, although a fairly large

overlap is obtained for $\lambda = 0.1$. This loss of generalization due to noise is explained as follows. Around $q \sim 0$, the dynamics of the order parameters for the noise rate λ are subject to the equations

$$\begin{cases} \mathrm{d}l/\mathrm{d}\alpha = \left[\dfrac{1}{4} - \dfrac{\sqrt{2}-1}{\sqrt{\pi}} l\right]/l \\ \mathrm{d}q/\mathrm{d}\alpha = -\left[\dfrac{1}{4} - (1-2\lambda)\dfrac{2}{\pi^{3/2}} l\right] q/l^2. \end{cases} \tag{20}$$

This set of equations state that the paramagnetic state $q = 0$ becomes unstable when l becomes greater than a critical length $l_c = \pi^{3/2}/[8(1-2\lambda)] \sim 0.696(1-2\lambda)^{-1}$ which is obtained from the condition $\mathrm{d}q/\mathrm{d}\lambda > 0$, while l is going to converge to an attracting point $l_c' = \sqrt{\pi}/[4(\sqrt{2}-1)] \sim 1.069$, which is determined by the condition $\mathrm{d}l/\mathrm{d}\alpha < 0$. When l_c is less than l_c', l is finally attracted into the region $l > l_c$ for any initial condition. Then, q begins to grow. Equations (20) cannot describe the dynamics correctly after q grows to a finite value. However, the growth of q shifts the attracting point of l to a larger value, which maintains the condition for the growth of q, and a finite value of q is finally obtained. This is the case for a small noise rate. On the other hand, for a noise rate beyond

$$\lambda_c = \left[1 - \left(\frac{\sqrt{2}-1}{2}\right)\pi\right]/2 \sim 0.175 \tag{21}$$

which is determined by the condition $l_c' < l_c$, l converges to l_c' before q reaches a finite value. This makes the paramagnetic state $q = 0$ a stable fixed point. As a result, the student cannot acquire any generalization ability even in the limit $\alpha \to \infty$.

For a noise rate just below λ_c, the reversal symmetry $q \leftrightarrow -q$ causes the overlap obtained in the limit $\alpha \to \infty$ to scale as $q \sim \mathrm{O}((\lambda_c - \lambda)^{1/2})$. This is similar to what happens in second-order phase transitions in ferromagnetic spin systems. While this analogy is not exact, it is useful to compare these systems in the following way. In the previous section, we showed that the learning rule (1) reduces the prediction error ε to its minimum value (zero) as $\alpha \to \infty$ unless the teacher signal is changed by noise. On the other hand, when the sign of each teacher signal is reversed, the student adapts himself to the completely opposite rule which has the maximum prediction error, $\varepsilon = 1$. This means that the inversion of the teacher signal causes the prediction error ε to increase. By introducing a noise rate λ, the learning rule (1) accepts those updates that increase the 'energy' ε with a probability which is proportional to λ. This kind of dynamics is analogous to what happens in physical systems with a finite temperature, where the acceptance rate for an increase in the energy is the 'temperature' of the system. At a high temperature, the system is free to go anywhere in the phase space because energy barriers are easily surmounted. As a result, the thermal average of the overlap goes to zero. On the other hand, at a low temperature below the critical value, the system cannot move from one state to a state which has an overlap of the opposite sign because it takes an infinitely long time to climb over the energy barrier between them, which yields non-zero overlap. This is an explanation of phase transition in physical systems. Such an interpretation, however, also holds in our system by regarding the noise rate λ as a temperature. In this sense, the loss of generalization due to noise at λ_c is identical to a ferro-paramagnetic phase transition in spin systems.

Acknowledgments

The author thanks Shigeru Shinomoto, Toshihiro Tanizawa, Michael Crair and Glenn Paquette for helpful discussions, comments and advice. The author was a fellow of the

Japan Society for the Promotion of Science for Japanese Junior Scientists. This work was partly supported by Grant-in-Aid no 2333 for Scientific Research by the Ministry of Education, Science and Culture, Japan.

References

Amari S, Fujita N and Shinomoto S 1992 Four types of learning curves *Neural Comp.* 4 605–18

Baum E B 1990 The perceptron learning is fast for nonmalicious distributions *Neural Comp.* 2 248–60

Hansel D, Mato G and C Meunier 1992 Memorization without generalization in a multilayered neural network *Europhys. Lett.* **20** 471–6

Kabashima Y and Shinomoto S 1992 Learning curves for error minimum and maximum likelihood algorithms *Neural Comp.* 4 712–9

Kabashima Y and Shinomoto S 1993 Accererlation of learning in binary choise problems *Proc. 6th ACM Conf. on Computational Learning Theory* pp 446–52

Karmarkar N 1984 A new polynomial time algorithm for linear programming *Combinatorica* 4 373–95

Kawanabe M and Amari S 1993 Estimation of neuronal parameters by the semiparametric statistical method *Preprint*

Minsky M and Papert S 1989 *Perceptrons and Introduction to Computational Geometry* 2nd ed (Cambridge, MA: MIT)

Mitchison G J and Durbin R M 1989 Bounds on the learning capacity of some multi-layer networks *Biol. Cybern.* **60** 345–56

Nilsson N.J 1965 *Learning Machines* (New York: McGraw-Hill)

Rosenblatt F 1962 *Principles of Neurodynamics* (New York: Spartan Books)

Seung H S, Sompolinsky H and Tishby N 1992 Statistical mechanics of learning from examples *Phys. Rev.* A **45** 6056–91