Learning in a Two-Layer Neural Network of Edge Detectors

1990 Europhys. Lett. 13 567

(http://iopscience.iop.org/0295-5075/13/6/016)

EUROPHYSICS LETTERS 15 November 1990

Europhys. Lett., 13 (6), pp. 567-572 (1990)

Learning in a Two-Layer Neural Network of Edge Detectors.

H. SOMPOLINSKY (*) and N. TISHBY

AT&T Bell Laboratories - Murray Hill, NJ 07974

(received 3 August 1990; accepted in final form 17 September 1990)

PACS. 87.10 – General, theoretical, and mathematical biophysics (inc. logic of biosystems, quantum biology, and relevant aspects of thermodynamics, information theory, cybernetics, and bionics.
PACS. 02.50 – Probability theory, stochastic processes, and statistics.
PACS. 05.20 – Statistical mechanics.

Abstract. – Learning from examples to count domains in one-dimensional patterns is studied. Increasing the number of examples used for training a network to perform the task is equivalent to the annealing of a one-dimensional Ising model. The generalization error falls off exponentially with the number of examples per weight. The related contiguity problem, where the network discriminates between patterns with small and large number of domains, exhibits a first-order phase transition to perfect generalization at all temperatures. Monte Carlo simulations of both models are in very good agreement with the theoretical predictions.

An important class of pattern recognition problems can be solved by an integration of overlapping feature detectors each restricted to small segments of the pattern. Geometrical properties of binary images, such as total area, Euler number, and border length are examples of problems in this class [1]. Moreover, several biological systems, in particular in vision, appear to have similar architecture [2]. It is thus interesting to study dynamic processes under which systems with such architecture evolve to perform their task. Here we study this issue in the context of the dynamics of supervised learning.

Until now theoretical understanding of learning in neural networks has been restricted mostly to single-layer models [3-9] or to rather general upper bounds on the necessary number of examples [10]. In this paper we present simple two-layer models which can be solved in the limit of large inputs. For simplicity we consider here adapting only one layer of weights, namely those connecting the input to the hidden layer. The theory can be extended to training of both layers. Training of the network is assumed to contain a stochastic noise analogous to *temperature*. It has been recently shown that the generalization error decreases asymptotically like the inverse of the number of training examples whenever the measure of error is a smooth function of the network's weights [8]. First-order transition to perfect generalization is found, however, in some single-layer networks with binary

(*) Permanent address: Racah Institute of Physics, Hebrew University, Jerusalem, 91904 Israel.

weights [8, 9]. Here we find discrete networks that exhibit yet another behavior, the error decreases exponentially with the number of examples per weight. The above results apply to learning at finite temperature. We show that the properties of the learning process are sensitive to the details of the task such as whether the output of the network is Boolean or real valued.

We consider a two-layer network trained to count the number of domains in one-dimensional binary patterns, consisting of strings of N bits $S_i = \pm 1$, $i = 0, \ldots, N$, with periodic boundary conditions, $S_0 = S_N$. The task is to count the number of domains of, say, $+1$'s, in a pattern. For an input pattern $\boldsymbol{S}$ this number, denoted as N^+, is

$$N^+(\boldsymbol{S}) = (1/4) \sum_{i=1}^{N} (1 - S_i S_{i-1}) . \tag{1}$$

A two-layer network that solves this problem is shown in fig. 1 (inset). The network consists of a layer of hidden neurons that function as edge detectors. The number of + domains equals half the number of domain walls or edges. Each pattern contains equal number of two types of edges: a + edge at the site i corresponds to $S_{i-1} = -1$, $S_i = +1$, whereas a − edge to $S_{i-1} = +1$, $S_i = -1$. Thus the number of domains is determined from either the number of + or − edges. In the network shown in fig. 1 each hidden neuron $\sigma_i^{\mathrm{h}} = \pm 1$ receives two inputs via weights with alternating signs, $W_{2i} = 1$, $W_{2i-1} = -1$, and has a threshold 1. It thus acts as a detector of a + edge at site i, *i.e.* $\sigma_i^{\mathrm{h}} = +1$ when $S_i = 1$, $S_{i-1} = -1$ and $\sigma_i^{\mathrm{h}} = -1$, otherwise. The weights from the hidden layer to the output neuron σ are all set to 1/2. The output neuron is assumed to be linear: $\sigma = (1/2) \sum_{i=1}^{N} \sigma_i^{\mathrm{h}} + (1/2) N$, thus one obtains $\sigma(\boldsymbol{S}) = N^+(\boldsymbol{S})$. Changing the signs of all the weights $\{W_i\}$ yields an equivalent solution with σ_i^{h} acting as detectors of − edges.

We consider learning the task in networks that are restricted to the architecture of fig. 1. We further fix the weights to the output neuron and the thresholds of the hidden layer, thus the only free parameters are the $2N$ weights $\{W_i\}$ which can be taken to be binary $W_i = \pm 1$. It is convenient to redefine these weights using the transformation $W_i \to (-1)^i W_i$, so that the two equivalent solutions to this task are $W_i = +1$, $i = 1, \ldots, 2N$ and $W_i = -1$, $i = 1, \ldots, 2N$. For any input $\boldsymbol{S}$ the states of the hidden neurons are $\sigma_i^{\mathrm{h}} = \mathrm{sgn}(W_{2i} S_i - W_{2i-1} S_{i-1} - 1)$ which can be expressed as

$$\sigma_i^{\mathrm{h}} = \frac{1}{2}(- W_{2i} W_{2i-1} S_i S_{i-1} + W_{2i} S_i - W_{2i-1} S_{i-1} - 1) . \tag{2}$$

The network is trained by selecting a random sample of P patterns $\boldsymbol{S}^l$, $l = 1, \ldots, P$, with the corresponding numbers of domains, using a training energy

$$E(\boldsymbol{W}) = \sum_{l=1}^{P} \varepsilon(\boldsymbol{W}; \boldsymbol{S}^l) , \tag{3}$$

where $\varepsilon(\boldsymbol{W}; \boldsymbol{S})$ is the error of the network $\boldsymbol{W}$ on the input $\boldsymbol{S}$. The error is chosen here as $(\sigma(\boldsymbol{W}; \boldsymbol{S}) - N^+(\boldsymbol{S}))^2$, yielding using (2)

$$\varepsilon(\boldsymbol{W}; \boldsymbol{S}) = (4N)^{-1} \left(\sum_{i=1}^{N} (1 - W_{2i} W_{2i-1}) S_i S_{i-1} + \sum_{i=1}^{N} (W_{2i} - W_{2i+1}) S_i \right)^2 , \tag{4}$$

where we have used the notation $W_{2N+1} \equiv W_1$. We consider here the case of uniform measure

on the input patterns, *i.e.* the patterns are drawn with the equal probability $P(\boldsymbol{S}) = 2^{-N}$. In this case for large N the distribution of N^+ is a Gaussian centered at $N/4$ with a width of order $\sqrt{N}$. For a typical $\boldsymbol{S}$ the network $\boldsymbol{W}$ generates approximately an equal number of mistakes by adding and subtracting edges, so that the typical value of $\sigma - N^+$ is of order $\sqrt{N}$, and ε of eq. (3) is of order 1. Finally, as in a general learning problem, the number of examples must scale as the number of free weights, *i.e.* $P = \alpha\, 2N$, so that the error $E(\boldsymbol{W})$ is of order N.

Given a fixed set of examples, we assume a stochastic training algorithm, similar to a finite-temperature Monte Carlo process. The long-time outcome of this learning process can be described by the Gibbs distribution $P(\boldsymbol{W}) = Z^{-1} \exp[-\beta E(\boldsymbol{W})]$ where $Z = \mathrm{Tr}_{\boldsymbol{W}} \exp[-\beta E(\boldsymbol{W})]$, and the temperature $T = \beta^{-1}$ is the amplitude of the stochastic noise in the training process [8].

The primary goal of the training is to reduce the *generalization error* $\varepsilon(\boldsymbol{W})$ which measures the error averaged over *all* the inputs, *i.e.* $\varepsilon(\boldsymbol{W}) \equiv 2^{-N} \sum_{\boldsymbol{S}} \varepsilon(\boldsymbol{W}; \boldsymbol{S})$. A straightforward averaging of eq. (4) yields $\varepsilon(\boldsymbol{W}) = 1 - m(\boldsymbol{W})$, where the order parameter m is

$$m(\boldsymbol{W}) = (2N)^{-1} \sum_{i=1}^{2N} W_i W_{i+1}. \tag{5}$$

Comparing eq. (5) with eq. (4), one observes that the average part of the training energy E, which is $2\alpha N(1 - m(\boldsymbol{W}))$, is identical with a 1D nearest-neighbor chain of W_i with a coupling constant $\beta\alpha$. However, the random part of E contains long-range coupling between all pairs of W_i's. In addition, whereas the average part is invariant under the transformation $\boldsymbol{W} \to -\boldsymbol{W}$, this is not so for the random part, although there are two equivalent optimal solutions, $W_i = 1$ and $W_i = -1$. Thus the behavior of the system depends upon the competition between the two qualitatively different components of E.

Quantities of interest include the generalization error and the training errors, $\varepsilon_t(T, P)$ and $\varepsilon_g(T, P)$, averaged over the Gibbs distribution $P(\boldsymbol{W})$ as well as over the quenched random patterns $\boldsymbol{S}^l$. Theoretical evaluation of quenched averages usually involves the difficult problem of calculating the average free energy $f = -T(2N)^{-1}[\ln Z]$, where $[\ldots]$ denotes averaging over the examples. Here we consider the *annealed approximation* to the problem which consists of replacing $[\ln Z]$ by $\ln[Z]$ [4, 8, 11].

Evaluating $\ln \mathrm{Tr}_{\boldsymbol{W}}[\exp[-\beta E(\boldsymbol{W})]]$ using a Gaussian transformation to linearize the quadratic form of eq. (4) and keeping only the leading terms in large N, we obtain

$$\beta f(m, K) = \frac{1}{2}\alpha \ln(1 + 2\beta(1 - m)) + mK - (2N)^{-1} \ln \mathrm{Tr}_{\boldsymbol{W}} \exp[-H(\boldsymbol{W})], \tag{6}$$

where the effective Hamiltonian of the weights, H, is the nearest-neighbor Ising Hamiltonian

$$H = -K \sum_{i=1}^{2N} W_i W_{i+1}. \tag{7}$$

Both m and K are determined by minimizing $f(m, K)$. Using the well-known properties of the 1D Ising model, we obtain

$$m_0 = (2N)^{-1} \sum_{i=1}^{2N} \langle W_i W_{i+1} \rangle - \mathrm{tgh}\, K_0, \tag{8}$$

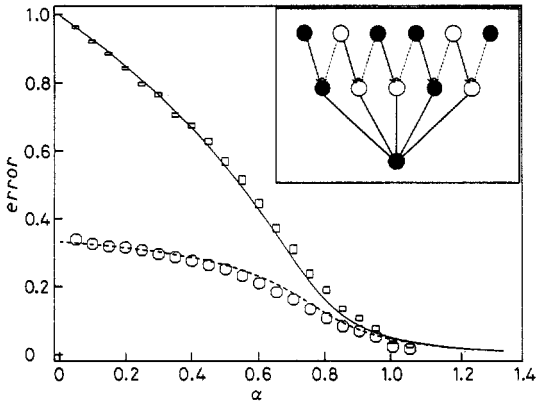

Fig. 1.

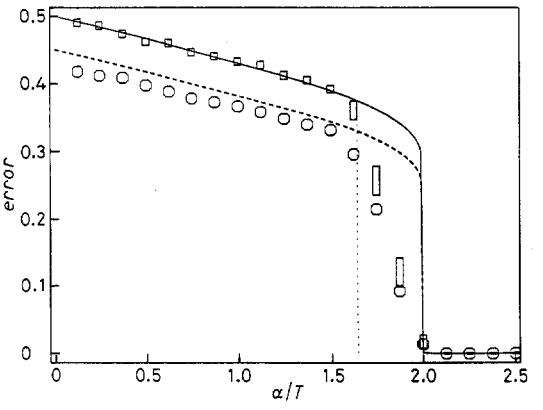

Fig. 2.

Fig. 1. – Generalization (□) and training errors (○) as functions of the number of examples per weight in the domain counting problem at $T = 1$. Data points are the results of Monte Carlo simulations with $N = 48$. The solid and dashed lines are the predictions of the annealed theory. The inset depicts one of the two optimal networks. The hidden units are connected with alternating positive (solid) and negative (dashed) weights. We show an example of an input pattern (S_0 is on the r.h.s.) with the corresponding states of hidden neurons. Filled circles in the first two layers represent +1 states and empty circles represent −1 states.

Fig. 2. – First-order transition to perfect generalization in the contiguity problem at $T = 5$. The lines are the results of the annealed theory for the training (○) and generalization (□) errors. The dotted line marks the position of the thermodynamic transition. The numerical data is for $N = 32$.

where $\langle \ldots \rangle$ denotes thermal averaging with the Ising Hamiltonian equation (7), and $K_0 = \alpha\beta/(1 + 2\beta(1 - m_0))$.

The average generalization error is simply $\varepsilon_g(\alpha, \beta) = 1 - m_0$. The average training error per example, derived by differentiating βf with respect to β and dividing by α, is $\varepsilon_t(\alpha, \beta) = \varepsilon_g/(1 + 2\beta\varepsilon_g)$.

Qualitatively, the above theory implies that the system behaves as an Ising model with a renormalized coupling K_0. For any fixed $\beta < \infty$ increasing the number of examples per weight, α, is equivalent to annealing the Ising model. For large α the generalization error decreases as

$$\varepsilon_g \sim 2 \exp[-2\beta\alpha], \quad \alpha \to \infty, \tag{9}$$

characteristic of the low-temperature behavior of the 1D Ising model. The type of errors that occur when α is large can be seen by evaluating the correlation function between a pair of weights W_i's. The above theory yields

$$\langle W_i W_j \rangle = \exp[-|i - j|/\xi], \tag{10}$$

where the correlation length ξ is $\xi = \ln m_0 \sim 2 \exp[-2\beta\alpha]$, $\alpha \gg 1$. Thus the competition between the two equivalent optimal solutions results in imperfect networks whose spatial structure alternates between the two. Long sections of σ_i^h's (of length $\sim \xi$) that detect + edges are interposed by sections that detect − edges.

In a previous work we showed that, in general, the annealed theory becomes exact only in the limit of $\beta \to 0$, $\alpha\beta$ finite [8]. It is interesting that in the present case this theory yields a

good quantitative approximation even in a temperature range where strong deviations from the high-T limit are observed. This is shown by Monte Carlo simulations of the system with the energy E at finite T. We find a very good agreement between the annealed theory and the simulations for all temperatures above ~ 0.3. The results for $T=1$ are given in fig. 1. Note that the theory reproduces nicely the substantial difference between the training and the generalization errors, whereas in the high-T limit both errors have the same value.

At low temperatures ($0 < T \lesssim 0.4$) the annealed theory predicts a first-order transition at $\alpha_c(T)$, where ε_g drops discontinuously to a low but finite value. This transition results from a singularity in the renormalized coupling K_0. According to this theory at $T=0$ ε_g drops to zero discontinuously at $\alpha_c(0)=1.16$. This implies that even within the annealed approximation there are strong cooperative effects at low temperatures generated by the *long-range* random interactions between the W_i's. The predicted discontinuous drop of ε_g is accurately reproduced in the simulations in the range $0.3 \lesssim T \lesssim 0.4$. At lower temperatures, however, we observe substantial deviations from the annealed results. The analysis of the low-T behavior is complicated due to the presence of spin-glass phenomena [8, 12] which are currently studied.

The properties of the training process depend on the definition of the task, such as the nature of the system's output. To demonstrate this point, we discuss the case where the network's task is not to count the number of domains but rather to identify those patterns whose number of domains is larger than some threshold N_0. This problem is known as the *contiguity problem* and has been recently studied as a benchmark model of learning [13].

The contiguity problem can be solved with the same architecture as in fig. 1. The only difference is that the output neuron is a threshold device,

$$\sigma = \operatorname{sgn}\left(\frac{1}{2}\sum_i \sigma_i^h + \frac{1}{2}N - N_0\right). \tag{11}$$

We consider again the case of uniform measure on the patterns. To obtain an interesting limit of the problem at large N, the threshold number N_0 must equal the average value, *i.e.* $N_0 = [N^+(\boldsymbol{S})] = N/4$. Otherwise the desired label will almost always have the same value. The appropriate measure of error for this problem is

$$\varepsilon(\boldsymbol{W};\boldsymbol{S}) \equiv \theta(-\sigma(\boldsymbol{W};\boldsymbol{S})\,\sigma_0(\boldsymbol{S})), \tag{12}$$

where $\theta(x>0)=1$; $\theta(x\leqslant 0)=0$ and the desired output σ_0 is $\sigma_0(\boldsymbol{S}) \equiv \operatorname{sgn}(N^+(\boldsymbol{S}) - N/4)$.

Averaging the error in eq. (12) over all $\boldsymbol{S}$, we find in the limit $N \to \infty$

$$\varepsilon(\boldsymbol{W}) \equiv \varepsilon(m^+, m^-) = \pi^{-1} \arccos\left(\frac{m^+}{\sqrt{(3-2m^-)}}\right), \tag{13}$$

where the two order parameters are

$$m^{\pm} \equiv N^{-1} \sum_{i=1}^{N} W_{2i} W_{2i\pm 1}. \tag{14}$$

Note that m^+ measures the average overlap between weights that share the same input, whereas m^- measures the overlap between weights that share the same output (see fig. 1).

The annealed theory for this problem yields

$$\beta f = \frac{1}{2}\alpha \ln(1 - \tau\varepsilon(m^+, m^-)) + m^+ K^+ + m^- K^- - \ln \operatorname{Tr}_W \exp[-H(\boldsymbol{W})], \tag{15}$$

where $\tau \equiv 1 - \exp[-\beta]$ and

$$H = -K^{+} \sum_{i=1}^{N} W_{2i} W_{2i+1} - K^{-} \sum_{i=1}^{N} W_{2i} W_{2i-1}. \tag{16}$$

The order parameters $m^{\pm}$ and the coupling constants $K^{\pm}$ are evaluated by minimization of f. One difference between eqs. (15), (16) and eqs. (6), (7) is that in this problem there are two different coupling constants between the chain of W_i's. However the most important difference is the singular dependence of ε, eq. (15), on $m^{\pm}$ at $m^{\pm} = 1$. This singularity leads to the appearance, at all T, of a discontinuous transition at a critical value of α from a high value of ε_g to perfect generalization, $\varepsilon_g = 0$. The mechanism for this transition is similar to that found in a single-layer perceptron with binary weights [8, 9]. This prediction is confirmed by Monte Carlo simulations of the contiguity problem at high T (fig. 2). Due to the small size of the system, the transition is not discontinuous. It occurs in a range of α between the thermodynamic transition, *i.e.* the point where the free energies of the poor and perfect generalization states are equal, and the point where the state of poor generalization becomes unstable. Note that in the contiguity problem already at $T = 5$ the quantitative deviations of the simulation results from the annealed theory are larger than in the counting problem, although in both cases it becomes exact as $T \to \infty$. Nevertheless the qualitative picture provided by the theory is correct except for low T, where in an intermediate range of α spin-glass phenomena are present.

In conclusion we remark that, although we assumed binary weights, our qualitative results hold also in the case of continuous weights as long as the hidden neurons have binary output. Furthermore, our approach is expected to apply to other systems of local feature detectors, including systems with two-dimensional input.

* * *

Useful discussions with D. HANSEL, D. A. HUSE and H. S. SEUNG are gratefully acknowledged.

REFERENCES

[1] See, *e.g.*, HORN B. K. P., *Robot Vision* (MIT Press, Cambridge, MA) 1986.
[2] See, *e.g.*, BARLOW H. B., *Proc. R. Soc. London, Ser. B*, **212** (1981) 1.
[3] MINSKY M. A. and PAPERT S. A., *Perceptrons* (MIT Press, Cambridge, MA) 1988.
[4] GARDNER E. and DERRIDA B., *J. Phys. A*, **22** (1989) 1983.
[5] KRAUTH W., MÉZARD M. and NADAL J.-P., *Complex Systems*, **2** (1988) 387.
[6] HANSEL D. and SOMPOLINSKY H., *Europhys. Lett.*, **11** (1990) 687.
[7] GYÖRGYI G. and TISHBY N., in *Neural Networks and Spin Glasses*, edited by W. K. THEUMANN and R. KÖBERLE (World Scientific, Singapore) 1990.
[8] SOMPOLINSKY H., TISHBY N. and SEUNG H. S., *Phys. Rev. Lett.*, **65** (1990) 1683.
[9] GYÖRGYI G., *Phys. Rev. A*, **41** (1990) 7097.
[10] BAUM E. B. and HAUSSLER D., *Neural Computation*, **1** (1989) 151.
[11] TISHBY N., LEVIN E. and SOLLA S. A., in *Proceedings of the International Joint Conference on Neural Networks, Washington D.C.*, **2** (1989) 403; LEVIN E., TISHBY N. and SOLLA S. A., *Special Issue on Neural Networks of the Proceedings of the IEEE* (1990).
[12] MÉZARD M., PARISI G. and VIRASORO M. A., *Spin Glass Theory and Beyond* (World Scientific, Singapore, Teaneck NJ, Hong Kong) 1987.
[13] SCHWARTZ D. B., SAMALAM V. K., DENKER J. S. and SOLLA S. A., *Neural Computation* (1990).