Pergamon

Appl. Math. Lett. Vol. 9, No. 3, pp. 109–113, 1996
Copyright©1996 Elsevier Science Ltd
Printed in Great Britain. All rights reserved
0893-9659/96 $15.00 + 0.00

S0893-9659(96)00041-9

Synchronous Boltzmann Machines Can Be Universal Approximators

L. Younes
ENS Cachan, CMLA, 61 av. du Président Wilson, 94, Cachan, France

(Received May 1995; accepted June 1995)

Abstract—We prove in this paper that the class of reversible synchronous Boltzmann machines is universal for the representation of arbitrary functions defined on finite sets. This completes a similar result from Sussmann in the sequential case.

Keywords—Synchronous random fields, Cellular automata, Gibbs distributions, Boltzmann machines, Neural networks.

1. INTRODUCTION

Boltzmann machines are stochastic neural networks which are most of the time dedicated to classification tasks. Given an input, they run some specified probabilistic dynamics, which provides a random response. One may also say that they associate to each input a probability distribution on the output.

More precisely, the machine is composed of neurons which are divided into three (finite) sets, denoted I, H, R. Elements of I are input neurons, elements of R are response neurons, and H contains hidden neurons. Let $S = I \cup H \cup R$ be the whole set of neurons. For $s \in S$, a binary variable x_s indicates whether s is *active* (one also says *firing*), in which case $x_s = 1$, or *still*, in which case $x_s = 0$. During the evolution, the state x_s varies randomly, depending on the activities of the other neurons. The selection of the new state y_s is based on a transition probability of the kind

$$p_s(x_t, t \in S;\ y_s),$$

providing the probability that the new value at s is y_s given that the former configuration of the network is x. To perform classification, all neurons are updated, except the input neurons which are clamped to some configuration x_I. The joint probability distribution of the remaining neurons (in $H \cup R$) converges to some equilibrium distribution, which depends on x_I. By only looking at the response part of the network, we obtain the input-output behaviour. In general, a deterministic answer is required, and the network is fitted so that this output distribution is very close to a point-mass measure.

We denote by Ω_I the set of input configurations, and Ω_R the set of output configurations: $\Omega_I = F^I$, $\Omega_R = F^R$ with $F = \{0, 1\}$. We already have used notation x_I for a configuration in Ω_I, and we shall denote in a similar way configurations defined on other subsets of S; the response associated to x_I, which is a probability distribution on Ω_R, is denoted by $\psi(\ .\ ;\ x_I)$. It is the marginal distribution of a probability $\pi(\ .\ ;\ x_I)$ defined on the larger set $\Omega_H \times \Omega_R$, which is the equilibrium distribution of the whole network evolving with clamped input x_I.

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

The transition probabilities (p_s above) are parameterized by a high-dimensional parameter θ, and the aim of learning is to estimate θ so that the network accomplishes the classification task. This task may be represented by a function f which associates to each input the "correct answer" in output. We say that a network accomplishes a task f if the response associated to any input x_I, which is a probability measure, is close to the Dirac measure at $f(x_I)$. We write π_θ, ψ_θ, when it is necessary to emphasize the parameterization.

For the original model of sequential machines [1], only one neuron is updated at each time chunk. Synchronous Boltzmann machines (with which we are mainly concerned) update all their neurons simultaneously. By choosing a parametric form of the transitions, and specifying a dynamic (i.e., synchronous or sequential), one completely specifies a class of neural networks.

In this paper, we shall define such a system (namely reversible synchronous Boltzmann machines) and prove the following property.

PROPERTY [UNI]. *For given I and R, there exists a set H such that, for any function $f : \Omega_I \to \Omega_R$, there exists θ such that, for every $x_I \in \Omega_I$, $\psi_\theta(\,.\,; x_I)$ is close to the Dirac distribution $\delta_{f(x_I)}$, up to a given, but arbitrarily small error.*

Such a result has been proved in [2] in the case of sequential machines.

Among deterministic networks, it has been shown that perceptrons with one hidden layer provide such universal approximators. "Tasks" for these networks are continuous functions from R^d to R^k, and typical results state that, given f (let's say continuous with compact support) and a precision order ϵ, there exists a perceptron with one hidden layer which has an input/output behaviour given by f up to an error of ϵ [3,4]. This subject has been widely studied, providing extensions assuming only measurability of f [5], or estimating the number of hidden neurons needed to approximate a given class of functions f [6].

In our situation, we consider discrete tasks: f is a relation between the set of inputs and the set of outputs which are finite. If one wishes to consider continuous tasks f, an approximation has to be done by a piecewise constant function (this *coding* is in fact crucial for the efficiency of practical applications; see [7] for an example).

Note that the universality property is a theoretical result. It is an almost necessary prerequisite for the usefulness of a neural network model. However, one must be aware that some tasks f may be so complex that they would require a very large set H, for which practical implementation is impossible. Another issue is that the function f is only known through a restricted *training set* of examples, i.e., a set $(x_I^k, f(x_I^k))$, $k = 1, \dots, K$ which can be considered as a small sample of the whole family $(x_I, f(x_I)), x_I \in \Omega_I$. So, even when Boltzmann machines large enough to represent f can be run, learning algorithms on the mere basis of the training set will amount to solutions which, although performing well for the training set, will be unable to induce the correct answer for unlearned x_I. This leads to questions which are linked to nonparametric statistics, which have been well explored in the case of feed-forward networks (see [8]), but which remain open for Boltzmann machines.

2. SYNCHRONOUS BOLTZMANN MACHINES

2.1. Model

The model we describe is the synchronous counterpart of the standard sequential Boltzmann machines. More details may be found in [7,9–11].

Consider a set of *interaction weights* (w_{st}, $s, t \in S$), and a set of *thresholds* ($h_s, s \in H \cup R$), with $w_{st} = w_{ts}$. Let $\theta = (w, h)$ be our parameter, and set

$$p_s(x, y_s) = \frac{\exp\left(y_s\left(\sum_{t\neq s} w_{st}x_t + h_s\right)\right)}{1 + \exp\left(\sum_{t\neq s} w_{st}x_t + h_s\right)}.$$

Assume that some input configuration z_I is clamped. At a given time, all neurons in $H \cup R$ are simultaneously updated according to the p_s; this leads to define (for $x, y \in \Omega_S$, with $x_I = y_I = z_I$)

$$P(x,y) = \prod_{s \in H \cup R} p_s(x, y_s). \tag{1}$$

When it is necessary to strengthen the dependence on z_I, we may write P^{z_I}. Since $P(x,y) > 0$, there exists a unique invariant distribution on $\Omega_{H \cup R}$, denoted $\pi(\,.\,;\, z_I)$. In fact, it may be checked that P is π-reversible; that is

$$\pi(x; z_I) P^{z_I}(x,y) = \pi(y; z_I) P^{z_I}(y,x). \tag{2}$$

Thanks to this property, it is possible to give a precise description of π, in terms of the two-step distribution at equilibrium. Consider the Markov chain (X^n) on Ω with initial distribution π, and transition $P = P^{z_I}$, and let μ be the joint distribution of (X^n, X^{n+1}); μ, as a distribution on $\Omega \times \Omega$, is given by $\mu(x, y;\, z_I) = \pi(x;\, z_I) P(x,y)$. It may be proved that

$$\mu(x,y) = \frac{\exp\left[\sum_{st} w_{st} x_s y_t + \sum_s h_s x_s + \sum_s h_s y_s\right]}{Z}. \tag{3}$$

The distribution $\pi(\,..;\, z_I)$ is the marginal of μ. We call π a reversible synchronous field.

3. UNIVERSALITY

We now show that property [UNI] holds for reversible synchronous Boltzmann machines. As in [2], we prove the following fact: if V is a finite set, there exists a finite set H, which does not intersect V, such that if ν is a positive probability distribution on Ω_V, there exists a reversible synchronous field π on $\Omega_{V \cup H} = \Omega_V \times \Omega_H$ with marginal ν on Ω_V. To obtain [UNI], just set $V = I \cup R$ and ν any positive distribution approaching an "ideal" distribution on Ω_V given by

$$\nu(x_I, x_R) = \nu_0(x_I) \delta_{F(x_I)}(x_R),$$

where ν_0 is any distribution on Ω_I.

Our technique is close to Sussmann's approach [2], and some of the following is inspired from his work. We begin with a lemma.

LEMMA 1. *Let ρ be a real number. Consider a fixed integer N, and binary variables $x_1, \ldots, x_N$, $x_i = 0$ or 1. One can find real numbers w and h such that:*

- *If $\rho \geq 0$,*

$$\log\{1 + \exp[w(x_1 + \cdots + x_N) + h]\} = \rho x_1 \ldots x_N + Q(x_1, \ldots, x_N).$$

- *If $\rho \leq 0$*

$$\log\{1 + \exp[w(x_1 + \cdots + x_{N-1} - x_N) + h]\} = \rho x_1 \ldots x_N + Q(x_1, \ldots, x_N),$$

Q being a polynomial of degree less than $N - 1$ in $x_1, \ldots, x_N$.

PROOF. First assume that $\rho \geq 0$ and set

$$\phi(x_1, \ldots, x_N) = \log\{1 + \exp[w(x_1 + \cdots + x_N) + h]\}. \tag{4}$$

Since every function of $x_1, \ldots, x_N$ can be expressed as a polynomial of partial degree one with respect to each variable, it suffices to find w and h such that the coefficient of the term of highest degree of ϕ is equal to ρ. This term is given by

$$g(w,h) = \sum_{k=0}^{N} (-1)^{N-k} \binom{N}{k} \log\left(1 + e^{kw+h}\right). \tag{5}$$

Now, fix $w > 0$ and h such that $Nw + h > 0$ and $(N-1)w + h < 0$. Then the function $g(tw, th)$ is 0 for $t = 0$, and tends to infinity when t tends to infinity, so that every positive ρ can be attained.

The case of $\rho < 0$ can be deduced from the other case with the change of variables $y_i = x_i$ for $i = 1 \ldots N-1$ and $y_N = 1 - x_N$. This ends the proof of Lemma 1.

LEMMA 2. *Let V be a finite set with N elements, and Ω_V the set of binary configurations with indices in V. For any positive probability measure ν on Ω_V, there exist a larger set $S = V \cup H$ and a probability measure π on Ω_S, described below, such that ν is the marginal of π over V.*

Assume that V is ordered in an arbitrary way. For $A \subset V$, denote by $s(A)$ the largest element in A. One can choose

$$H = \{A \subset V,\ \mathrm{card}(A) > 1\},$$

and

$$\pi(x_V, u_H) = \frac{1}{Z}\exp\left[\sum_{A\in H}\left(w_A u_A\left(\sum_{s\in A,\ s<s(A)} x_s + \epsilon_A x_{s(A)}\right) + h_A u_A\right) + \sum_{s\in V} h_s x_s\right], \tag{6}$$

where $\epsilon_A \in \{-1, +1\}$ is fixed.

PROOF. The proof is an iterative application of Lemma 1. Every positive distribution on Ω_V has the form $\nu(x) = \exp[E(x)]/Z$, and the energy $E(x)$ can be expressed as a polynomial in $x_s, s \in V$.

For every subset A of V with at least 2 elements, let $S^+_A(x_A)$ be $\sum_{s\in A} u_s$, and $S^-_A(x_A) = S^+_A(x_A) - 2x_{s(A)}$. For $\epsilon = +$ or $-$, define

$$\phi^\epsilon_A(u_A) = \log\{1 + \exp[w_A S^\epsilon_A(u_A)]\},$$

as in (4) with some numbers w_A, h_A. Each of these A is considered as a new neuron with state u_A. Now define the energy

$$\bar{E}(x, u) = \sum_A [w_A u_A(S^\epsilon_A(x_A)) + h_A u_A] + \text{linear terms},$$

and the associated Gibbs field $\pi(x, u) = \exp(\bar{E}(x, u))/Z$. In the preceding expression of $\bar{E}$, the ϵ in S^ϵ_A depends on A. A little calculation shows that in order to have

$$\nu(x) = \sum_u \pi(x, u),$$

it suffices that

$$\sum_A \phi^\epsilon_A(x_A) + \text{linear terms} = E(x).$$

To solve this, first apply Lemma 1 and find numbers w_V and h_V, to cancel the term of highest degree in E (with an adequate choice of $\epsilon = +$ or $-$). Then subtract from E the function Q which remains from Lemma 1. The remaining energy only contains terms of degree less than N. Lemma 1 can be applied again to find all ϕ^ϵ_A, for A with cardinality equal to $N - 1$, and the procedure can be iterated until only terms of degree one remain.

THEOREM 1. *Every probability distribution on Ω_V is the marginal distribution on Ω_V of a reversible synchronous random field over some larger set of sites. An upper bound for the number of additional sites is $2^N - N - 1$.*

PROOF. Take H and π as in Lemma 2, and define a distribution μ on $(\Omega_V \times \Omega_H)^2$ by (for $z = (x_V, u_H)$, $\bar{z} = (y_V, v_H)$ in $\Omega_V \times \Omega_H$)

$$\mu(z, \bar{z}) = \pi(x_V, v_H)\pi(y_V, u_H).$$

It is easy to convince oneself that μ takes the form given in (3), and therefore, that its marginal, μ_0, on $\Omega_V \times \Omega_H$ is a reversible synchronous field. Moreover, we have

$$\begin{aligned}\sum_{u_H} \mu_0(x_V, u_H) &= \sum_{u_H}\sum_{y_V, v_H} \pi(x_V, v_H)\pi(y_V, u_H)\\ &= \sum_{v_H} \pi(x_V, v_H) = \psi(x_V).\end{aligned}$$

REMARKS.

(1) From the proof of Lemma 1, we see that we can always take $h_A = -(N-1/2)w_A$ so that the number of added parameters in $\pi(x,u)$ is also $2^N - N - 1$. The total number of parameters is then $2^N - 1$, which is precisely the dimension of the space of all probabilities on Ω_V. Moreover, if we know that the initial random field ν is local, i.e., its energy $E(x)$ only includes terms of degree less than $r < N$, then one only needs to add $\sum_{k=2}^{r} \binom{N}{k}$ parameters. In particular, when $r = 2$ (which corresponds to sequential Boltzmann machines), this number is $N(N-1)/2$.

(2) It is clear that while proving Lemma 2, we also proved the universality of sequential Boltzmann machines: it suffices to stop while reducing the degree of the energy when only quadratic terms remain instead of linear terms. The method used in [2] is quite similar, but our proof is shorter thanks to the simplicity of the basic Lemma 1.

REFERENCES

1. D.H. Ackley, G. Hinton, and T. Sejnowski, A learning algorithm for Boltzmann machines, *Cognitive Sci.* **9**, 147–169, (1985).
2. H.J. Sussmann, On the convergence of learning algorithms for Boltzmann machines, Preprint, Rutgers University, (1988).
3. R. Hecht-Nielsen, Theory of the back-propagation neural network, In *Proceeding of the International Joint Conference on Neural Networks,* San Diego, pp. I:593–608, SOS Printing, (1989).
4. G. Cybenko, Approximation by superpositions of sigmoidal functions, *Mathematics of Control, Signal, and Systems* **2**, 303–314, (1989).
5. K. Hornik, M. Stinchcombe and H. White, Multilayer feedforward networks are universal approximators, *Neural Networks* **2**, 359–366, (1989).
6. A. Barron, Approximation and estimation bounds for artificial neural networks, *Machine Learning; Special Issue on the Fourth ACM Workshop on Computational Learning Theory*, (1991).
7. R. Azencott, A. Doutriaux and L. Younes, Synchronous Boltzmann machines and curve identification tasks, Preprint ENS-DIAM, (1992); *Network: Computations in Neural Networks* **4**, 461–480, (1992).
8. A. Barron, Complexity regularization with application to artificial neural networks, In *Nonparametric Functional Estimation and Related Topics*, (Edited by G. Roussas), pp. 561–576, Kluwer, (1991).
9. W.A. Little, The existence of persistent states in the brain, *Math. Biosci.* **19**, 101,120, (1974).
10. P. Peretto and J.J. Niez, Collective properties of neural networks, In *Disordered Systems and Biological Organisation*, (Edited by E. Bienenstock *et al.*), pp. 171–185, Springer-Verlag, Berlin, (1986).
11. R. Azencott, High-order interactions and synchronous learning, In *Proceedings of Stochastic Models, Statistical Methods and Algorithms in Image Analysis, Lecture Notes in Statistics*, (Edited by P. Barone and A. Frigessi), Springer, (1991).
12. R. Azencott, Synchronous Boltzmann machines and Gibbs fields, In *Neurocomputing*, NATO ASI Series, Vol. F 68, (Edited by F. Folgerman-Hierault), Springer-Verlag, (1990).
13. R. Hecht-Nielsen, Kolmogorov's mapping neural network existence theorem, In *IEEE First International Conference on Neural Networks,* San Diego, pp. III:11–14, SOS Printing, (1987).
14. O. Koslov and N. Vasilyev, Reversible Markov chains with local interactions, In *Multicomponent Random Systems*, (Edited by R.L. Dobrushin and Ya.G. Sinai), pp. 415–469, Dekker, New York, (1980).
15. V.Y. Kreinowich, Arbitrary non-linearity is sufficient to represent all functions by neural networks: A theorem, *Neural Networks* **4**, 381–383, (1991).
16. A. Trouvé, Partially parallel simulated annealing: Low and high temperature approach of the invariant measure, In *Proceedings of the US-French Workshop on Applied Stochatic Analysis, Lecture Notes*, (Edited by I. Karatzas and D. Ocone), Springer-Verlag, (1992).