

A lacuna in the theory of asynchronous Boltzmann machine learning

Antonia J. Jones

Abstract. This note rectifies a logical gap in the derivation of the asynchronous Boltzmann machine learning algorithm.

Keywords: Asynchronous Boltzmann learning.

Communication regarding this paper should be addressed to:

Antonia J. Jones

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF WALES, CARDIFF
PO BOX 916
Cardiff CF2 4YN

Telephone: ..44-122-287-4812

Telefax: ..44-122-287-4598

Date/version: 27 September 2002

Published in: **III Simpósio Brasileiro de Redes Neurais**, São Carlos-USP, Brazil, pp19-27, 20 November 1996.

Copyright © 1996. Antonia J. Jones

A lacuna in the theory of asynchronous Boltzmann machine learning

CONTENTS

Introduction	1
Background on asynchronous Boltzmann machines	1
Asynchronous learning	2
The lacuna	3
Repairing the damage	6
Conclusion	8
References	9

A lacuna in the theory of asynchronous Boltzmann machine learning

by

Antonia J. Jones¹

Abstract. This note rectifies a logical gap in the derivation of the asynchronous Boltzmann machine learning algorithm.

Keywords: Asynchronous Boltzmann learning.

Introduction.

The aim of this note is to rectify a lacuna in the theory of asynchronous Boltzmann machine learning. The practical effect of this correction on what is known about asynchronous Boltzmann machine learning is nil: the fact is that the learning rule is correct and so all the papers that employ it are on safe theoretical ground. However, as we shall show, the mathematical *proof* of the rule is incomplete and one aim of this note is to provide the details of the missing piece. Another reason for clarifying the logic of the proof is that it sets the record straight and clears the way for the application of a similar analysis of learning to models whose mathematical details are rather different.

The learning algorithm referred to is basically due to Hinton and Sejnowski [Hinton 1983] but is closely related to the *Maximum Likelihood* model used to estimate unknown parameters of exponential distributions [Dempster 1977] or hidden Markov chains [Bahl 1983]. Maximum likelihood methods iteratively adjust the unknown parameters so as to increase the probability that the generic model will produce the observed data. There is a large literature within statistics on maximum likelihood estimation. The learning rule described in [Ackley 1985] actually has a close relationship to a method called *Expectation and Maximization*.

In this note it is assumed that the reader is familiar with the traditional presentation of asynchronous Boltzmann machine learning. The aim here is merely to identify the lacuna and offer a repair. The alternative would be to describe a complete proof of the learning algorithm and this would greatly increase the length of the note to little effect.

$$Q_T(\mathbf{x}) = q_T(\mathbf{h}|\mathbf{v})q(\mathbf{v}) \quad (1)$$

Background on asynchronous Boltzmann machines.

Notation. Let \mathbf{x} and \mathbf{v} be independent stochastic variables whose values are, respectively, the configurations and environmental (visible) configurations of a Boltzmann machine. We make the following definitions.

Let $P_T(\mathbf{x})$ be the free running equilibrium distribution and suppose $q(\mathbf{v})$ is the environmentally imposed probability distribution over the state vectors \mathbf{v} of visible units. We write where $q_T(\mathbf{h}|\mathbf{v})$ is the probability that vector \mathbf{h} will occur on the hidden units when \mathbf{v} is clamped on the visible units and the network of hidden units is allowed to run. Thus $Q_T(\mathbf{x})$ is the probability of observing global state \mathbf{x} over multiple runs in which successive vectors \mathbf{v} are clamped with probability $q(\mathbf{v})$. Also let $p_T(\mathbf{v})$ be the probability distribution over the states of the visible units when the network is running freely at temperature T .

¹ Department of Computer Science, University of Wales - Cardiff, PO Box 916, Cardiff CF2 4YN, UK.

Let the activation potential of unit i be

$$net_i(\mathbf{x}) = \sum_{j \neq i} w_{ij} x_j - \theta_i \quad (2)$$

where w_{ij} is the weight attached to the connection from unit j to unit i and θ_i is the bias or threshold. We assume that units are not connected to themselves, so that $w_{ii} = 0$, for $1 \leq i \leq n$.

The stochastic unit is defined so that the unit outputs $x_i = 0$ or 1 with probability

$$p(x_i) = \frac{1}{1 + e^{-(2x_i - 1)net_i(\mathbf{x})/T}} \quad (3)$$

Units are selected randomly for updating asynchronously.

The energy E of a given state \mathbf{x} is given by

$$E(\mathbf{x}) = E(x_1, x_2, \dots, x_n) = -\frac{1}{2} \sum_{i \neq j} w_{ij} x_i x_j + \sum_i \theta_i x_i \quad (4)$$

If a single unit changes state then, assuming $w_{ij} = w_{ji}$, for all i, j , the change of energy ΔE due to Δx_i is given by

$$\Delta E = \frac{\partial E}{\partial x_i} \Delta x_i = - \left(\sum_{\substack{j \\ j \neq i}} w_{ij} x_j - \theta_i \right) \Delta x_i = - net_i(\mathbf{x}) \Delta x_i \quad (5)$$

At low temperatures the model approaches the Hopfield model: the dynamics become progressively more deterministic and equation (5) can be used to show that, provided only one unit updates at a given time, the network will settle to a local energy minima. At higher temperatures the stochastic update rule provides a mechanism which enables the algorithm to avoid becoming trapped in a local minimum in its search for the global energy minimum.

For the asynchronous model the stochastic update rule for each unit implies that in equilibrium the distribution of the possible states of the system is the Boltzmann-Gibbs distribution, i.e. the probability that the system is in state $\mathbf{x} = (x_1, \dots, x_n)$ is given by

$$P_T(\mathbf{x}) = P_T(x_1, \dots, x_n) = \frac{1}{Z(T)} e^{-E(x_1, \dots, x_n)/T} \quad (6)$$

where $Z(T)$ is a normalisation constant (the *partition function*) chosen so that the probabilities sum to one.

Asynchronous learning.

The aim of asynchronous Boltzmann machine learning is to reduce the difference between the two observable distributions p_T and q by performing gradient descent in weight space on a suitable measure of their difference.

An information theoretic measure [Kullback 1959] of the ‘distance’ between these two probability distributions is given by

$$D(p_T|q) = \sum_{\mathbf{v}} q(\mathbf{v}) \log_e \left(\frac{q(\mathbf{v})}{p_T(\mathbf{v})} \right) \quad (7)$$

Lemma 1. For $T > 0$ we have $D(p_T|q) \geq 0$ and $D(p_T|q) = 0$ if and only if $p_T \equiv q$.

Proof. This is a straightforward exercise. \blacksquare

In order to perform gradient descent on D we need to know the gradient with respect to the w_{ij} . The required result is

$$\begin{aligned}\frac{\partial D}{\partial w_{ij}} &= -\frac{1}{T}(q_{ij} - p_{ij}) \\ \frac{\partial D}{\partial \theta_i} &= \frac{1}{T}(q_i - p_i)\end{aligned}\tag{8}$$

where q_{ij} is the probability, averaged over all environmental inputs and measured at equilibrium, that the i th and j th units are both on, and p_{ij} is the corresponding probability when the network is free running. More precisely

$$\begin{aligned}q_{ij} &= \sum_{\mathbf{x} \in X} x_i x_j Q_T(\mathbf{x}) \quad (\text{if } i \neq j), & q_i &= \sum_{\mathbf{x} \in X} x_i Q_T(\mathbf{x}) \\ p_{ij} &= \sum_{\mathbf{x} \in X} x_i x_j P_T(\mathbf{x}) \quad (\text{if } i \neq j), & p_i &= \sum_{\mathbf{x} \in X} x_i P_T(\mathbf{x})\end{aligned}\tag{9}$$

Hence to (locally) minimise D , it is therefore sufficient to observe p_{ij} and q_{ij} at thermal equilibrium and to change each weight according to the formula

$$\begin{aligned}\Delta w_{ij} &= \eta(q_{ij} - p_{ij}) \\ \Delta \theta_i &= -\eta(q_i - p_i)\end{aligned}\tag{10}$$

where $\eta > 0$ is the learning rate.

We are in a situation where we know $P_T(\mathbf{x})$ and we want to infer some facts about $p_T(\mathbf{v})$. We can make an explicit link between these two distributions by observing that the configuration space X can be decomposed into the direct sum $X = H \oplus V$, where H and V are the configuration space of the hidden and visible units, respectively. Another way to put this is that for any $\mathbf{x} \in X$ there exist unique vectors $\mathbf{h} \in H$ and $\mathbf{v} \in V$ such that $\mathbf{x} = \mathbf{h} + \mathbf{v}$, moreover \mathbf{h} and \mathbf{v} are orthogonal. This observation allows us to express the following connection between the two distributions, viz.

$$p_T(\mathbf{v}) = \sum_{\mathbf{h} \in H} P_T(\mathbf{h} + \mathbf{v}) = \sum_{\mathbf{h} \in H} P_T(\mathbf{x})\tag{11}$$

Each term in the sum represents the probability of an $\mathbf{x} = \mathbf{h} + \mathbf{v}$, i.e. an \mathbf{x} whose projection on V is \mathbf{v} . The sum is over the exhaustive and mutually exclusive ways in which this can occur.

The lacuna.

It is also critical to know the equilibrium behaviour of the system when a vector \mathbf{v} is clamped onto the visible units, i.e. we need some information regarding $Q_T(\mathbf{x})$. In order to complete the proof of (10) one uses

$$p_T(\mathbf{v}) Q_T(\mathbf{x}) = q(\mathbf{v}) P_T(\mathbf{x})\tag{12}$$

where $\mathbf{x} = \mathbf{h} + \mathbf{v}$, to prove (8) - and it is precisely here that the problem arises.

The classical proof proceeds as follows. First it is observed that

$$\begin{aligned}P^+(V_\alpha \wedge H_\beta) &= P^+(H_\beta | V_\alpha) P^+(V_\alpha) \\ P^-(V_\alpha \wedge H_\beta) &= P^-(H_\beta | V_\alpha) P^-(V_\alpha)\end{aligned}\tag{13}$$

where the relationship between the two notations is summarised in Table 1.

Table 1 Relationship between the two notations.

Original notation	Present notation $\mathbf{x} = \mathbf{h} + \mathbf{v}$
$P^+(V_\alpha \wedge H_\beta)$	$Q_T(\mathbf{x})$
$P(V_\alpha \wedge H_\beta)$	$P_T(\mathbf{x})$
$P^+(V_\alpha)$	$q(\mathbf{v})$
$P(V_\alpha)$	$p_T(\mathbf{v})$

Here superscript ‘-’ refers to the free running case, superscript ‘+’ to the clamped case and $\mathbf{x} = \mathbf{h} + \mathbf{v}$.

It is then asserted that

$$P^-(H_\beta|V_\alpha) = P^+(H_\beta|V_\alpha) \quad (14)$$

which on using (13) is seen to be equivalent to

$$P_T(\mathbf{x})/p_T(\mathbf{v}) = Q_T(\mathbf{x})/q(\mathbf{v}) \quad (15)$$

(i.e. to (12)), provided we insist that both $p_T(\mathbf{v})$ and $q(\mathbf{v})$ are strictly positive for all \mathbf{v} . This condition causes no difficulty since we are free to place the condition $q(\mathbf{v}) > 0$ on the environmental distribution and, since $T > 0$, the stochastic update rule will ensure that every state has some positive probability of occurring.

The following argument is then advanced to support (14):

□ *The equation must hold because the probability of a hidden state given some visible state must be the same in equilibrium whether the visible units were clamped in that state or arrived there by free running.* [Ackley 1985]

Notice that no mention is made of the particular form of the equilibrium distribution. Hence we might expect (12) to hold regardless of the precise equilibrium distribution. However, for systems with an equilibrium distribution different from (6) equation (12) does not apply and the equivalent relationship may be much more complicated.

The case in which all units update synchronously was first studied in a model proposed by W. A. Little in 1974 (see [Little 1974] and [Little 1978]). Somewhat later Peretto [Peretto 1984] showed that the equilibrium distribution $P_T(\mathbf{x})$ is not described by the standard Boltzmann-Gibbs distribution using energy, but by a similar distribution in which energy is replaced in (6) by an appropriate Hamiltonian. Specifically

$$P_T(\mathbf{x}) = \frac{1}{Z(T)} \prod_m 2 \cosh\left(\frac{net_m(\mathbf{x})}{2T}\right) e^{A(\mathbf{x}, m)/2T} \quad (16)$$

where $Z(T)$ is a normalisation constant and

$$A(\mathbf{x}, i) = net_i(\mathbf{x}) - 2\theta_i x_i \quad (17)$$

Note the stochastic update rule does not change: precisely the same activation rule (3) is used as before. However, now all units are attempting to change state simultaneously and so (5) no longer holds. In fact at low temperature it is no longer energy that is being minimised but a more complicated function.

In the synchronous case one can derive (but this perhaps is not an appropriate place) information, similar to (12), regarding $Q_T(\mathbf{x})$ in the form

$$Q_T(\mathbf{x}) = q_T(\mathbf{h}|\mathbf{v}) q(\mathbf{v}) \quad (18)$$

where

$$q_T(\mathbf{h}|\mathbf{v}) = \frac{1}{Z_H(\mathbf{v}, T)} \beta(\mathbf{v}, \mathbf{h}) P_T(\mathbf{x}) \quad (19)$$

where $\mathbf{x} = \mathbf{h} + \mathbf{v}$, $Z_H(\mathbf{v}, T)$ is a normalisation factor and

$$\beta(\mathbf{v}, \mathbf{h}) = \frac{\exp\left(\frac{1}{T} \sum_{m \in U_H} h_m \sum_{j \in U_V} w_{mj} v_j\right)}{\prod_{k \in U_V} 2 \cosh\left(\frac{net_k(\mathbf{x})}{2T}\right)} e^{A(\mathbf{x}, k)/2T} \quad (20)$$

One can also readily verify these assertions by simulation experiments.

The same simulations (or further tedious algebra if one seeks a formal proof) can be used to show that (18) is incompatible with the $Q_T(\mathbf{x})$ given by (12). For example, let us call the formula for $Q_T(\mathbf{x})$ derived from (12) Theory A, and the formula for $Q_T(\mathbf{x})$ derived from (18) using (16), (17), (19) and (20) Theory B. Table 2 gives simulation results for a three unit synchronous network in which the first unit is the visible unit and the second and third units are hidden. The weights and thresholds are

$$\begin{aligned} \theta_1 &= -1 & w_{12} &= 1 & w_{13} &= -1 \\ w_{21} &= 1 & \theta_2 &= 1 & w_{23} &= 2 \\ w_{31} &= -1 & w_{32} &= 2 & \theta_3 &= -3 \end{aligned} \quad (21)$$

and the first unit is clamped to 1. The network was run for 20,000 updates at temperature $T = 2$.

Table 2 Simulation results for a three node synchronous network.

Unit values		Theory A	Theory B	Observed
\mathbf{v}	\mathbf{h}	$Q_T(\mathbf{h} + \mathbf{v})$	$Q_T(\mathbf{h} + \mathbf{v})$	$Q_T(\mathbf{h} + \mathbf{v})$
1	0 0	0.04851	0.05073	0.04874
1	1 0	0.09320	0.11446	0.11419
1	0 1	0.30520	0.25638	0.25738
1	1 1	0.55311	0.57843	0.57961

Although not a proof, this is fairly compelling evidence in favour of Theory B and the conclusion then follows that (12) does not apply in this case. Ergo, the argument \square advanced in support of (12) must be logically inadequate.

Repairing the damage.

It emerges that in the asynchronous case (12) can be regarded as resulting from an elegant identity derived from the Boltzmann-Gibbs distribution (6). In this section we deduce the relevant identity and derive (12).

Let $U = U_H \cup U_V$, where U_H and U_V are the set of hidden and visible units respectively. If a vector \mathbf{v} is clamped on

the visible units and remains there, what is the effect on a hidden unit?

If we write the state of the network as $\mathbf{x} = \mathbf{h} + \mathbf{v}$, the activation function for the hidden unit $i \in U_H$ is

$$net_i(\mathbf{x}) = \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j + \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j - \theta_i \quad (22)$$

(the condition $j \neq i$ in the second sum is superfluous but helps to have it there later). Now since \mathbf{v} is constant we can collect together the last two terms and view them as an *effective threshold* Θ_i , where

$$\Theta_i = \theta_i - \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j \quad (23)$$

The network of hidden units then behaves like a Boltzmann machine with its own interconnecting weights and thresholds Θ_i . This means that in principle we know the probability of any particular state \mathbf{h} of the hidden units; it will be determined by the Boltzmann-Gibbs distribution. To use this fact we shall need to know the relationship between the internal energy of the subnet of hidden units, operating with effective thresholds Θ_i , and the energy of the whole network. The next theorem makes this relationship explicit by means of an algebraic identity.

Theorem 1. Let

$$E_H(\mathbf{h}|\mathbf{v}) \equiv -\frac{1}{2} \sum_{i \in U_H} h_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j + \sum_{i \in U_H} \Theta_i h_i \quad (24)$$

where the Θ_i ($i \in U_H$) are defined by (23), and

$$E_V(\mathbf{v}) \equiv -\frac{1}{2} \sum_{i \in U_V} v_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j + \sum_{i \in U_V} \theta_i v_i \quad (25)$$

then

$$E(\mathbf{x}) = E_H(\mathbf{h}|\mathbf{v}) + E_V(\mathbf{v}) \quad (26)$$

where $\mathbf{x} = \mathbf{h} + \mathbf{v}$.

Remark. We can identify $E_H(\mathbf{h}|\mathbf{v})$ as the energy of the subnet U_H in state \mathbf{h} , when vector \mathbf{v} is clamped on U_V . However, the term $E_V(\mathbf{v})$ is the energy of subnet U_V in state \mathbf{v} when it is completely disconnected from the units of U_H .

Proof. By definition

$$E(\mathbf{x}) = E(x_1, \dots, x_n) = -\frac{1}{2} \sum_{i \in U} x_i \sum_{\substack{j \in U \\ j \neq i}} w_{ij} x_j + \sum_{i \in U} \theta_i x_i \quad (27)$$

Since $U = U_H \cup U_V$ we can split the inner sum of the first term into a sum over U_H and a sum over U_V to obtain

$$\begin{aligned}
 E(\mathbf{x}) &= -\frac{1}{2} \sum_{i \in U} x_i \left(\sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j + \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j \right) + \sum_{i \in U} \theta_i x_i \\
 &= -\frac{1}{2} \sum_{i \in U} x_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j - \frac{1}{2} \sum_{i \in U} x_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j + \sum_{i \in U} \theta_i x_i
 \end{aligned} \tag{28}$$

Now we again replace the sums over U by separate sums over U_H and U_V to obtain

$$\begin{aligned}
 E(\mathbf{x}) &= -\frac{1}{2} \sum_{i \in U_H} h_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j - \frac{1}{2} \sum_{i \in U_V} v_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j \\
 &\quad - \frac{1}{2} \sum_{i \in U_H} h_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j - \frac{1}{2} \sum_{i \in U_V} v_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j \\
 &\quad + \sum_{i \in U_H} \theta_i h_i + \sum_{i \in U_V} \theta_i v_i
 \end{aligned} \tag{29}$$

Notice that the second and third terms are equal - they are the same sum written two different ways.

Substituting (23) into the definition of E_H in (24) we have

$$\begin{aligned}
 E_H(\mathbf{h}|\mathbf{v}) &= -\frac{1}{2} \sum_{i \in U_H} h_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j + \sum_{i \in U_H} \left(\theta_i - \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j \right) h_i \\
 &= -\frac{1}{2} \sum_{i \in U_H} h_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j - \sum_{i \in U_H} h_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j + \sum_{i \in U_H} \theta_i h_i
 \end{aligned} \tag{30}$$

Hence from (29), (30) and (25) we have

$$\begin{aligned}
 E(\mathbf{x}) - E_H(\mathbf{h}|\mathbf{v}) - E_V(\mathbf{v}) \\
 = \sum_{i \in U_H} h_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j - \frac{1}{2} \sum_{i \in U_V} v_i \sum_{\substack{j \in U_H \\ j \neq i}} w_{ij} h_j - \frac{1}{2} \sum_{i \in U_H} h_i \sum_{\substack{j \in U_V \\ j \neq i}} w_{ij} v_j = 0
 \end{aligned} \tag{31}$$

identically, since the first term cancels the last two terms. \blacksquare

Note that when \mathbf{v} is clamped on U_V , $E_V(\mathbf{v})$ is constant. This makes the calculation of the probability of any particular vector \mathbf{h} on the hidden units particularly straightforward. Let us denote the probability that \mathbf{h} will occur when \mathbf{v} is clamped on U_V (and the subnet U_H is allowed to run) by $q_T(\mathbf{h}|\mathbf{v})$. The only effect of U_V on U_H is to cause U_H to run with effective thresholds Θ_i . Under these circumstance $q_T(\mathbf{h}|\mathbf{v})$ is governed by the Boltzmann-Gibbs distribution and we have the following corollary of Theorem 1.

Corollary 1. If the free running distribution of the system is described by the Boltzmann-Gibbs distribution of (6), then

$$q_T(\mathbf{h}|\mathbf{v}) = \alpha(\mathbf{v}, T) P_T(\mathbf{x}) \tag{32}$$

where $\alpha(\mathbf{v}, T)$ is a positive *constant* depending only on \mathbf{v} and T , and $\mathbf{x} = \mathbf{h} + \mathbf{v}$.

This corollary asserts that $q_T(\mathbf{h}|\mathbf{v})$ is proportional to the probability of the global state $\mathbf{x} = \mathbf{h} + \mathbf{v}$.

Proof. We write

$$\begin{aligned} q_T(\mathbf{h}|\mathbf{v}) &= \frac{1}{Z_H} e^{-E_H(\mathbf{h}|\mathbf{v})/T} = \frac{1}{Z_H} e^{-E(\mathbf{x})/T} e^{E_V(\mathbf{v})/T} \\ &= \frac{Z}{Z_H} e^{E_V(\mathbf{v})/T} \frac{1}{Z} e^{-E(\mathbf{x})/T} = \alpha(\mathbf{v}, T) P_T(\mathbf{x}) \end{aligned} \quad (33)$$

where Z and Z_H are appropriate normalisation constants and α depends only on \mathbf{v} and T . \blacksquare

This leads immediately to the following

Lemma 2. (Crucial lemma). If the free running distribution of the system is described by the Boltzmann-Gibbs distribution of (6), then from the above definitions we have

$$p_T(\mathbf{v}) Q_T(\mathbf{x}) = q(\mathbf{v}) P_T(\mathbf{x}) \quad (34)$$

where $\mathbf{x} = \mathbf{h} + \mathbf{v}$.

Remark. We call this the ‘crucial lemma’ because it is the bedrock of the Boltzmann machine learning algorithm. The lemma gives the relationship between $Q_T(\mathbf{x})$ and $P_T(\mathbf{x})$ in terms of the observables $q(\mathbf{v})$ and $p_T(\mathbf{v})$, where $\mathbf{x} = \mathbf{h} + \mathbf{v}$. It shows that in the particular case of the Boltzmann-Gibbs distribution this relationship has a simple ratio form.

Proof. Suppose \mathbf{v} is clamped onto U_V and remains fixed. Summing (32) over all possible \mathbf{h} yields

$$\sum_{\mathbf{h} \in H} q_T(\mathbf{h}|\mathbf{v}) = \sum_{\mathbf{h} \in H} \alpha(\mathbf{v}, T) P_T(\mathbf{x}) = \alpha(\mathbf{v}, T) \sum_{\mathbf{h} \in H} P_T(\mathbf{x}) = 1 \quad (35)$$

since $\alpha(\mathbf{v}, T)$ is independent of \mathbf{h} and the sum of the probabilities must be 1. Now from (35) using (11) we have

$$\alpha(\mathbf{v}, T) p_T(\mathbf{v}) = 1 \quad (36)$$

So that Corollary 1 yields

$$p_T(\mathbf{v}) q_T(\mathbf{h}|\mathbf{v}) = p_T(\mathbf{v}) \alpha(\mathbf{v}, T) P_T(\mathbf{x}) = P_T(\mathbf{x}) \quad (37)$$

Since from (1), $Q_T(\mathbf{x}) = q_T(\mathbf{h}|\mathbf{v})q(\mathbf{v})$, where $\mathbf{x} = \mathbf{h} + \mathbf{v}$, (37) gives

$$p_T(\mathbf{v}) Q_T(\mathbf{x}) = p_T(\mathbf{v}) q_T(\mathbf{h}|\mathbf{v}) q(\mathbf{v}) = P_T(\mathbf{x}) q(\mathbf{v}) \quad (38)$$

which is the required result. \blacksquare

The proof of (8) and the derivation of (10) can now proceed in the usual way.

Conclusion.

These observations mainly serve to set the logical analysis straight. In all fairness it should be made clear that I did not discover this lacuna. It was discovered by Dr. Tom Westerdale of Birkbeck College, University of London, who some years ago convinced me (initially with some difficulty) that a theorem I thought I had proved was incorrect. I had been led astray by precisely the point I have sought to clarify. Regrettably, although the lacuna was readily fixed, the theorem I was seeking to prove had to be consigned to oblivion. However, recently Ursula Iturraran and I have succeeded in a modified version of my original endeavour and it seemed appropriate to report the lacuna and its rectification at this time.

References

- [Ackley 1985] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, *A learning algorithm for Boltzmann machines*, *Cognitive Science* **9**:147-169, 1985.
- [Bahl 1983] R. L. Bahl, F. Jelinek and R. L. Mercer. *A maximum likelihood approach to continuous speech recognition*. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **5**:179-190, 1983.
- [Dempster 1977] A. P. Dempster, A. M. Laird and D. B. Rubin. *Maximum likelihood from incomplete data via the EM algorithm*. *Journal of the Royal Statistical Society*, **39**:1-38, 1977.
- [Hinton 1983] G. E. Hinton and T. J. Sejnowski. *Optimal perceptual inference*. *Proc. IEEE Conference on Computer Vision and Pattern Recognition.*, Washington DC, 448-453, 1983.
- [Kullback 1959] *Information theory and statistics*, Wiley, N.Y.
- [Little 1974] W. A. Little. *The existence of persistent states in the brain*. *Mathematical Biosciences* **19**:101-120, 1974.
- [Little 1978] W. A. Little. and G. L. Shaw. *Analytic study of the memory storage capability of a neural network*. *Mathematical Biosciences* **39**:281-290, 1978.
- [Peretto 1984] P. Peretto. *Collective properties of neural networks: a statistical physics approach*. *Biological Cybernetics* **50**:51-62, 1978.