Higher-order Boltzmann machines and entropy bounds

1999 J. Phys. A: Math. Gen. 32 5529

(http://iopscience.iop.org/0305-4470/32/30/301)

J. Phys. A: Math. Gen. **32** (1999) 5529–5538. Printed in the UK PII: S0305-4470(99)98475-5

Higher-order Boltzmann machines and entropy bounds

Bruno Apolloni†, Egidio Battistini‡ and Diego de Falco†

† Dipartimento di Scienze dell'Informazione, Università di Milano, Italy

‡ Dipartimento di Matematica, Politecnico di Milano, Italy

Received 16 October 1998, in final form 29 April 1999

Abstract. We examine some aspects of the interface area between mathematical statistics and statistical physics relevant to the study of Boltzmann machines. The Boltzmann machine learning algorithm is based on a variational principle (Gibbs' lemma for relative entropy). This fact suggests the possibility of a scheme of successive approximations: here we consider successive approximations parametrized by the order of many-body interactions among individual units. We prove bounds on the gain in relative entropy in the crucial step of adding, and estimating by Hebb's rule, a new parameter. We address the problem of providing, on the basis of local observations, upper and lower bounds on the entropy. While upper bounds are easily obtained by subadditivity, lower bounds involve localization of Hirschman bounds on a dual quantum system.

1. Introduction

We consider the Boltzmann machine model in a rather more general setting than in Ackley *et al* (1985), allowing, as in Azencott (1992), for direct interaction between more than two neurons. We establish, first of all, our notation. For fixed integer ν, we set $\Lambda_\nu = \{1, 2, \ldots, \nu\}$. Points in Λ_ν will be called sites, or nodes, or neurons. Having set $S = \{-1, 1\}$, we indicate an element of S^ν by $\underline{\sigma} = (\sigma_1, \sigma_2, \ldots \sigma_\nu)$, $\sigma_i \in S$, $i = 1, 2, \ldots, \nu$. A probability measure on S^ν is determined by a density function ρ: $\underline{\sigma} \in S^\nu \to \rho(\underline{\sigma})$ satisfying $\rho(\underline{\sigma}) \geqslant 0$, $\forall \underline{\sigma} \in S^\nu$ and $\sum_{\underline{\sigma} \in S^\nu} \rho(\underline{\sigma}) = 1$.

We associate, with each subset M of Λ_ν, a subset variable σ_M defined by

$$\sigma_M = \prod_{i \in M} \sigma_i \qquad \text{if } M \text{ is nonempty, 1 otherwise.}$$

Any probability density ρ on S^ν can be written in the form

$$\rho(\underline{\sigma}) = \frac{1}{2^\nu} \sum_{M \subseteq \Lambda_\nu} s_M \sigma_M$$

where $s_M = E_\rho(\sigma_M)$, and $E_\rho(\cdot)$ indicates expectation with respect to the probability density ρ. The set $\{s_M, M \neq \emptyset\}$ of moments determines a coordinate system (the s-chart) on the $(2^\nu - 1)$-dimensional manifold $PM(S^\nu)$ of all probability measures on S^ν. If we restrict our attention to the set $PM_+(S^\nu)$ of strictly positive probability measures on S^ν, we can also write

$$\rho(\underline{\sigma}) = \frac{\exp(\sum_{M \subseteq \Lambda_\nu, M \neq \emptyset} \theta_M \sigma_M)}{Z(\theta)} \qquad \text{where} \quad Z(\theta) = \sum_{\underline{\sigma} \in \Lambda_\nu} \exp\left(\sum_{M \subseteq \Lambda_\nu, M \neq \emptyset} \theta_M \sigma_M \right).$$

In the above sense, the collection of coupling constants $\{\theta_M, M \neq \emptyset\}$ equips $PM_+(S^\nu)$ with another coordinate system (the θ-chart). The simple and deep relations between the s- and

0305-4470/99/305529+10$30.00

θ-charts have been reviewed and applied to problems in machine learning by Amari *et al* (1992).

The conventional Boltzmann machine has only two-body couplings ($\theta_M = 0$ for $|M| > 2$) and must allow for hidden nodes in order to compensate for this limitation. We wish to explore here the advantage of remaining in the exponential family of probability distributions by allowing for many-body couplings, without hidden nodes.

The variational basis of the Boltzmann machine learning algorithm (Ackley *et al* 1985) is in the following elementary considerations. Let ρ_0 and ρ_1 be two elements of $PM_+(S^\nu)$; Jensen's inequality shows that

$$E_{\rho_0}(\ln \rho_0) \geqslant E_{\rho_0}(\ln \rho_1) \qquad \text{with equality holding iff} \quad \rho_0 = \rho_1. \tag{1.1}$$

Call $\theta(i)$, $s(i)$ the coordinates of $\rho_i (i = 0, 1)$ in the two charts considered above. It is trivial to compute that, for each nonempty subset $M \subseteq \Lambda_\nu$, we have

$$\frac{\partial E_{\rho_0}(\ln \rho_1)}{\partial \theta_M(1)} = (s_M(0) - s_M(1)). \tag{1.2}$$

Therefore, for $\varepsilon > 0$, setting

$$\Delta\theta_M(1) = \varepsilon(s_M(0) - s_M(1)) \tag{1.3}$$

we have

$$\sum_{M \subseteq \Lambda_\nu, M \neq \emptyset} \frac{\partial E_{\rho_0}(\ln \rho_1)}{\partial \theta_M(1)} \Delta\theta_M(1) \geqslant 0. \tag{1.4}$$

Inequality (1.4) says that a 'small' updating, $\theta_M(1) \to \theta_M(1) + \Delta\theta_M(1)$, of the coupling parameters of ρ_1 produces a new probability measure which is closer to ρ_0 in the sense that $E_{\rho_0}(\ln \rho_1)$ gets closer to the upper bound that Jensen's inequality sets for it.

The (higher-order) Boltzmann machine learning algorithm can now be described by the following steps:

- with the current values of the machine parameters $\{\theta_M(1)\}$, a simulator is allowed to run according to a Glauber dynamics (Glauber 1963) having ρ_1 as its stationary distribution; this simulation is supposed to last long enough to allow the estimation, as ergodic means, of some of the moments $s_M(1)$;
- these moments are compared with the corresponding moments $s_M(0)$ estimated from the environmental signal consisting of a random sample drawn from the population ρ_0;
- the parameters of the simulator are updated according to (1.3);
- the above procedure restarts with the new parameters.

Inequality (1.4) suggests that repeated execution of this cycle makes ρ_1 closer to ρ_0, hopefully leading to a final ρ_1 'indistinguishable' from ρ_0. In this sense the simulator will have built a model of the environment.

This paper makes some preliminary steps towards an operationally meaningful assessment of the performance of the above procedure. We observe, in this respect, that, in terms of the relative entropy

$$I(\rho_0, \rho_1) = E_{\rho_0}(\ln \rho_0) - E_{\rho_0}(\ln \rho_1) = \sum_{\underline{\sigma} \in S^\nu} \rho_0(\underline{\sigma}) \ln \frac{\rho_0(\underline{\sigma})}{\rho_1(\underline{\sigma})} \tag{1.5}$$

the inequality $E_{\rho_0}(\ln \rho_0) \geqslant E_{\rho_0}(\ln \rho_1)$ amounts to a statement of Gibbs' lemma:

$$I(\rho_0, \rho_1) \geqslant 0 \qquad \text{equality holding iff} \quad \rho_0 = \rho_1. \tag{1.6}$$

As it is, in general, out of the question to obtain identity between model and environment (namely $I(\rho_0, \rho_1) = 0$), the performance of the learning procedure will be measured by its

capability of achieving a relative entropy below a preassigned positive threshold $I_{\min}$. Such a threshold will be determined, in each actual application, according to the standard techniques of the theory of hypotheses testing: in particular, inequality 3.1 (page 75) of Kullback (1959) shows the crucial role of $I(\rho_0, \rho_1)$ in relating the sample size to the errors of the first and second kind in a test of the hypothesis ρ_0 versus the alternative hypothesis ρ_1.

In the above setting, the problem of finding a 'good' model ρ_1 of the environmental distribution ρ_0 becomes the problem of singling out a class of parametric models within which an element can be found having, with reference to ρ_0, relative entropy smaller than $I_{\min}$. Reasons of computational feasibility make it obvious that such a parametric class must have less than the maximum number $2^\nu - 1$ of parameters. Having fixed a subset A of the set $2^{\Lambda_\nu} - \{\emptyset\}$ of all nonempty subsets of Λ_ν, we consider those models for which the coupling constants θ_M vanish for $M \notin A$. An elementary explicit computation shows that, for a probability density of the form:

$$\rho_A(\underline{\sigma}) = \frac{\exp\left(\sum_{M \in A} \theta_M \sigma_M\right)}{Z(\theta)} \tag{1.7}$$

we have

$$I(\rho_0, \rho_A) = H(\rho_A) - H(\rho_0) + \sum_{M \in A} \theta_M (E_{\rho_A}(\sigma_M) - E_{\rho_0}(\sigma_M)) \tag{1.8}$$

where $\{\theta_M, M \in A\}$ are those θ-coordinates of ρ_A which have not been set to zero, and $H(\rho) = -\sum_{\underline{\sigma} \in S^\nu} \rho(\sigma) \ln \rho(\underline{\sigma})$ is the entropy of $\rho \in PM_+(S^\nu)$.

The following considerations formalize the variational choice of the element of the given class of models that 'best approximates' a given ρ_0:

- For each choice of the set $\{s_M(0) \equiv E_{\rho_0}(\sigma_M), M \in A\}$ of 'environmental moments' associated to subsets in A, there exists a unique choice of coupling constants $\{\theta_M, M \in A\}$ in (1.7) such that equation

$$E_{\rho_A}(\sigma_N) = s_N(0) \qquad \forall N \in A \tag{1.9}$$

  is satisfied (lemma A 4.6 of Lanford 1973).
- For fixed ρ_0, we shall indicate by $\rho_{A,0}$ the corresponding solution of the form (1.7) of equation (1.9). We shall refer to $\rho_{A,0}$ as to the 'approximation of order A' to ρ_0, or the 'Boltzmann machine of order A' associated to ρ_0.
- It is easy to check, from (1.2), that $\rho_{A,0}$ minimizes $I(\rho_0, \rho)$ under the constraint that $\theta_N = 0$ for every nonempty N not belonging to A.
- Equation (1.8) shows that

$$I(\rho_0, \rho_{A,0}) = H(\rho_{A,0}) - H(\rho_0). \tag{1.10}$$

- It is easy to check, from (1.10), that $\rho_{A,0}$ maximizes $H(\rho)$ under the constraint that $E_\rho(\sigma_N) = E_{\rho_0}(\sigma_N), \forall N \in A$.

Thus, how close the Boltzmann machine of order A is, in relative entropy, to the environmental distribution depends only on the environmental entropy $H(\rho_0)$ and on the maximum entropy compatible with the given environmental expectations of subset variables for subsets in A. It will be notationally convenient, from now on, to drop the suffix '0' when referring to the assigned environmental distribution (which we shall, therefore, simply indicate by ρ) and to the associated Boltzmann machine of order A (which will be indicated by ρ_A).

As a final remark of this section, we observe that a training sample $\underline{\sigma}(1), \underline{\sigma}(2), \ldots, \underline{\sigma}(n)$ of fixed size does not provide knowledge of the actual environmental law $\rho(\underline{\sigma})$, but only of its

empirical estimate $\rho^*(\underline{\sigma}; \underline{\sigma}(1), \underline{\sigma}(2), \ldots, \underline{\sigma}(n))$, which is itself a random variable, function of the sample. A detailed study of the random variable $I(\rho^*, \rho_A)$ in its comparison with $I(\rho, \rho_A)$, leading under suitable hypotheses to the Akaike (1974) information criterion, can be found in Murata *et al* (1995). In the following, we shall, however, neglect all considerations of estimation errors in the numerical values of the parameters of ρ_A.

The paper is organized as follows. In section 2 we prove upper and lower bounds on the relative entropy decrement achieved through the introduction of one more parameter in the model. In section 3 we study 'realistic' entropy bounds, namely entropy inequalities in which only quantities which are known at the current level of approximation appear. In particular we specialize 'quantum' entropic inequalities to our models. Section 4 is devoted to discussion and open problems.

2. Reconfiguration of a Boltzmann machine

Having fixed the environmental distribution ρ, having fixed the order A, it may occur that the Boltzmann machine of order A (which will be indicated by ρ_A, its coordinates being indicated by $s(A)$ or $\theta(A)$) fails to give a value below the threshold $I_{\min}$ set according to the criteria discussed in the previous section. Such a situation opens the problem of choosing (at least) a new subset $M \notin A$ and, on the basis of the value of $E_\rho(\sigma_M)$ estimated from the environmental distribution, constructing the Boltzmann machine of order $B = A \cup \{M\}$ (which will be, of course, called ρ_B, its coordinates being indicated by $s(B)$ or $\theta(B)$).

This section is devoted to upper and lower bounds on the decrement of relative entropy achieved through this crucial step of 'adding one more parameter to the model' (and through its iterations). Equality (1.10) and the observation that ρ_A is the approximation of order A to ρ_B, make it clear that such a decrement is measured by $I(\rho_B, \rho_A)$, and that

$$I(\rho, \rho_B) = I(\rho, \rho_A) - I(\rho_B, \rho_A). \tag{2.1}$$

A lower bound on $I(\rho_B, \rho_A)$ is easily obtained through the following steps.

- Relative entropy is monotonically increasing with respect to refinements of the partition in sample space. In particular, for every event $C \subseteq S^\nu$, we have (Kullback 1959):

$$P_B(C) \ln \frac{P_B(C)}{P_A(C)} + P_B(\bar{C}) \ln \frac{P_B(\bar{C})}{P_A(\bar{C})} \leqslant I(\rho_B, \rho_A) \tag{2.2}$$

  where we have set

$$P_A(C) = \sum_{\underline{\sigma} \in C} \rho_A(\underline{\sigma}) \qquad P_B(C) = \sum_{\underline{\sigma} \in C} p_B(\underline{\sigma}).$$

- Choose, in particular, for C the event '$\sigma_M = 1$', whose indicator function is $I_C(\underline{\sigma}) = \frac{1+\sigma_M}{2}$, so that $P_B(C) = \frac{1+s_M(B)}{2} = \frac{1+s_M}{2}$, where s_M is the environmental expectation of the subset variable σ_M, and $P_A(C) = \frac{1+s_M(A)}{2}$. (Notice that, in general, it will be $s_M(A) \neq s_M$, because $M \notin A$).
- Use Schuetzenberger's (1954) inequality, namely

$$p_0 \ln \frac{p_0}{p_1} + q_0 \ln \frac{q_0}{q_1} \geqslant 2(p_0 - p_1)^2 \qquad \text{for} \quad p_i \in (0, 1), \quad q_i = 1 - p_i, \quad (i = 0, 1) \tag{2.3}$$

  to conclude that

$$I(\rho_B, \rho_A) \geqslant \frac{(s_M(A) - s_M(B))^2}{2} \tag{2.4}$$

  (we remark that this inequality still holds if we exchange A with B).

In order to obtain an upper bound, we observe that the same reasoning as in (1.8) leads, for every choice of ρ' and ρ'' in $PM_+(S^\nu)$, with, respectively, coordinates s', θ' and $s''\theta''$, to the identity

$$I(\rho', \rho'') + I(\rho'', \rho') = \sum_{N \subseteq \Lambda_\nu} (\theta'_N - \theta''_N)(s'_N - s''_N) \tag{2.5}$$

which specializes, in our case, to

$$I(\rho_B, \rho_A) + I(\rho_A, \rho_B) = \theta_M(B)(s_M - s_M(A)). \tag{2.6}$$

We can, therefore, conclude that

$$\begin{aligned} I(\rho_B, \rho_A) &= \theta_M(B)(s_M - s_M(A)) - I(\rho_A, \rho_B) \\ &\leqslant \theta_M(B)(s_M - s_M(A)) - \tfrac{1}{2}(s_M - s_M(A))^2. \end{aligned} \tag{2.7}$$

Summarizing, for $B = A \cup \{M\}$, we have

$$\frac{1}{2}(s_M - s_M(A))^2 \leqslant I(\rho_B, \rho_A) \leqslant \frac{1}{2}(s_M - s_M(A))^2 \left(\frac{2\theta_M(B)}{s_M - s_M(A)} - 1 \right). \tag{2.8}$$

We make the following comments.

(1) Inequality (2.8) is 'epistemologically realistic', in that it estimates the decrement in relative entropy from below in terms of moments which can be estimated from an environmental sample and of the moments predicted by the model at the *current* order of approximation, and from above in terms of the same quantities and of the additional parameter of an attempted *subsequent* level of approximation.

(2) Inequality (2.8) implies, in particular, that, for $B = A \cup \{M\}$ with $M \notin A$,

$$\frac{\theta_M(B)}{s_M - s_M(A)} \geqslant 1. \tag{2.9}$$

This inequality amounts to a quantitative assessment of Hebb's (1949) rule in that it gives a bound on the increment that one has to give to the coupling constant associated to the subset M in order to bring the 'consensus' of the neurons in M (probability that $s_M = 1$) from the value $\frac{1+s_M(A)}{2}$ predicted by the model ρ_A, to the value $\frac{1+s_M(B)}{2} = \frac{1+s_M}{2}$ exhibited by the environment.

(3) Inequality (2.9) can be extended, by iteration, to the general case that new coupling constants are introduced for successively new subsets $M_1, M_2, \ldots, M_k$. It is convenient, in this case, to set $B = A \cup \{M_1, M_2, \ldots, M_k\}$, $A_i = A_{i-1} \cup \{M_i\}$ for $1 \leqslant i \leqslant k$, $A_0 = A$. As

$$I(\rho_B, \rho_A) = \sum_{i=1}^{k} I(\rho_A, \rho_{A_{i-1}}) \tag{2.10}$$

one can conclude that

$$\begin{aligned} &\frac{1}{2}\sum_{i=1}^{k}(s_{M_i}(B) - s_{M_i}(A_{i-1}))^2 \leqslant I(\rho_B, \rho_A) \\ &\qquad \leqslant \frac{1}{2}\sum_{i=1}^{k}(s_{M_i}(B) - s_{M_i}(A_{i-1}))^2 \left(\frac{2\theta_{M_i}(A_i)}{s_{M_i}(B) - s_{M_i}(A_{i-1})} - 1 \right). \end{aligned} \tag{2.11}$$

(4) If some *a priori* upper bound $I(\rho, \rho_A) \leqslant h$ on the relative entropy $I(\rho, \rho_A) = H(\rho_A) - H(\rho)$ is available, inequality (2.11) can give, along a nested sequence of successive approximations $A = A_0 \subseteq A_1 \subseteq, \ldots, \subseteq A_m \subseteq \ldots$, a criterion for stopping the computationally expensive procedure of adding and estimating new parameters; knowing h, such a criterion will be of the form:

stop at the level m at which it is, for the first time, $h - \frac{1}{2}\sum_{i=1}^{m}(s_{M_i}(B) - s_{M_i}(A_{i-1}))^2 \leqslant I_{\min}$.

3. Entropy bounds

In order to be realistic, a stopping rule of the form outlined at the end of section 2 requires that the upper bound h be of the form $h = h(\{s_N; N \in A\})$, namely a function only of the experimental data $\{s_N; N \in A\}$ on the basis of which the approximation ρ_A of order A to the environment ρ has been constructed. As, in turn, it is $I(\rho, \rho_A) = H(\rho_A) - H(\rho)$, the above goal is achieved (with $h = h_2 - h_1$) once 'localized' inequalities of the form

$$h_1(\{E_\rho(\sigma_N); N \in A\}) \leqslant H(\rho) \leqslant h_2(\{E_\rho(\sigma_N); N \in A\}) \tag{3.1}$$

are established, for generic $\rho \in PM_+(S^\nu)$, and hence also for ρ_A.

The task of obtaining an *upper* bound of the above form is conceptually simple, being solved by the same idea underlying subadditivity of H: neglect dependence among subsets in A, and avoid overcounting of overlapping subsets.

For instance, in the case $A = \{N \subseteq \Lambda_\nu : 0 < |N| \leqslant k\}$, in which all interactions of up to k neurons are considered in the model, Han's inequality (Han 1978, Dembo and Cover 1991) gives

$$\frac{H(\rho)}{\nu} \leqslant \frac{1}{\binom{\nu}{k}} \sum_{N \subseteq \Lambda_\nu, |N|=k} \frac{H(\rho_N)}{k} \tag{3.2}$$

where $\rho_N(\underline{\sigma}) \equiv \sum_{\sigma_i = \pm 1, i \notin N} \rho(\underline{\sigma})$ is the marginal distribution of the variables localized in N. As $\rho_N(\underline{\sigma}) = \frac{1}{2^{|N|}} \sum_{M \subseteq N} s_M \sigma_M$, the right-hand side of (3.2) is, as needed, a function only of $\{s_N; N \in A\}$. An additional bonus of Han's inequality is that the r.h.s. is a monotonically decreasing function of k.

We concentrate, in the rest of this section, on the task of finding localized *lower* bounds on $H(\rho)$. The idea that $H(\rho)$ fails to equal the upper bound $\nu \ln 2$ by an amount which exactly measures how far ρ is from the centre of the simplex $PM_+(S^\nu)$ (independent, identically distributed σ, with $P(\sigma_i = 1) = \frac{1}{2}$) can be made precise through the form that Gross' (1975) inequality takes in our context:

$$H(\rho) \geqslant \nu \ln 2 - \sum_{i=1}^{\nu} \left(1 - \sum_{\underline{\sigma} \in S^\nu} \sqrt{\rho(\ldots \sigma_i = 1 \ldots)\rho(\ldots \sigma_i = -1 \ldots)}\right). \tag{3.3}$$

Shifting attention from individual sites to subsets requires us to follow the intuition that, if $\psi \equiv \rho^{1/2}$ is 'spread' on its domain S^ν, then its Fourier transform (namely the set of the coefficients of its expansion in terms of characters of the group $\{-1, 1\}^\nu$), defined as

$$\hat{\psi}(M) = \frac{1}{2^{\nu/2}} \sum_{\underline{\sigma} \in S^\nu} \rho(\underline{\sigma})^{1/2} \sigma_M \tag{3.4}$$

is concentrated on a small region of its domain $2^{\Lambda\nu}$.

These considerations are made precise by the following discrete form of Hirschman's uncertainty principle (Hirschman 1957, Dembo and Cover 1991):

$$H(\rho_\theta) + H(r_\theta) \geqslant \nu \ln 2 \tag{3.5}$$

where the suffix θ refers to the parameters in

$$\rho_\theta = \frac{(\exp \sum_{M \subseteq \Lambda_\nu; M \neq \emptyset} \theta_M \sigma_M)}{Z(\theta)}$$

and

$$H(r_\theta) = -\sum_{M \subseteq \Lambda_\nu} r_\theta(M) \ln(r_\theta(M)) \tag{3.6}$$

$$r_\theta(M) = (\hat{\psi}_\theta(M))^2 \tag{3.7}$$

$$\hat{\psi}_\theta(M) = \frac{1}{2^{\nu/2}} \sum_{\underline{\sigma} \in S^\nu} \rho_\theta(\underline{\sigma})^{1/2} \sigma_M. \tag{3.8}$$

Our first aim is to rewrite (3.5) in a form in which, instead of $H(r_\theta)$, a function of the moments $s_M(\theta) = E_{\rho_\theta}(\sigma_M)$ appears. We recall, first of all, that for $\beta > 0$

$$\frac{\mathrm{d}}{\mathrm{d}\beta} H(\rho_{\beta\theta}) = -\beta \sum_{\substack{M \subseteq \Lambda_\nu \\ N \subseteq \Lambda_\nu}} \theta_M \mathrm{cov}_{\rho\beta\theta}(\sigma_M, \sigma_N) \theta_N \leqslant 0 \tag{3.9}$$

so that $H(\rho_\theta) \geqslant H(\rho_{2\theta})$.

We can, therefore, write

$$H(\rho_\theta) + H(r_{2\theta}) \geqslant H(\rho_{2\theta}) + H(r_{2\theta}) \geqslant \nu \ln 2. \tag{3.10}$$

Now

$$\begin{aligned} \hat{\psi}_{2\theta}(M) &= \frac{1}{2^{\nu/2}} \sum_{\underline{\sigma} \in S^\nu} \rho_{2\theta}(\underline{\sigma})^{1/2} \sigma_M = \frac{1}{2^{\nu/2}} \sum_{\underline{\sigma} \in S^\nu} \frac{\exp \sum_{N \subseteq \Lambda_\nu, N \neq \emptyset} \theta_N \sigma_N}{Z(2\theta)^{1/2}} \sigma_M \\ &= \frac{1}{2^{\nu/2}} \frac{Z(\theta)}{Z(2\theta)^{1/2}} \sum_{\underline{\sigma} \in S^\nu} \rho_\theta(\underline{\sigma}) \sigma_M = C_\nu(\theta) s_M(\theta). \end{aligned} \tag{3.11}$$

The quantity $C_\nu(\theta)$ (constant w.r.t. M) is easily computed from the observation that, because of Parseval's identity, it must be $1 = \sum_{M \subseteq \Lambda_\nu} (\hat{\psi}_{2\theta}(M))^2 = C_\nu(\theta)^2 \sum_{M \subseteq \Lambda_\nu} s_M(\theta)^2$.

It is, therefore,

$$r_{2\theta}(M) = \frac{s_M(\theta)^2}{\sum_{N \subseteq \Lambda_\nu} s_N(\theta)^2} \tag{3.12}$$

and we can conclude, dropping from now on the suffixes θ, that

$$H(\rho) \geqslant \nu \ln 2 - H(r) \tag{3.13}$$

with

$$r(M) \equiv \frac{s_M^2}{\sum_{N \subseteq \Lambda_\nu} s_N^2} \tag{3.14}$$

$$H(r) = -\sum_{M \subseteq \Lambda_\nu} r(M) \ln(r(M)) \tag{3.15}$$

where $s_M = E_\rho(\sigma_M)$ for $M \neq \emptyset$, $s_\emptyset \equiv 1$.

Inequality (3.13) translates the problem of finding lower bounds of the form $H(\rho) \geqslant h_1(\{E_\rho(\sigma_N); N \in A\})$ into the problem of finding upper bounds of the form $H(r) \leqslant h_3(\{E_\rho(\sigma_N); N \in A\})$.

This problem can be easily solved by looking at the function r as a probability density on the subsets of Λ_ν: from the knowledge of *only* the moments $\{E_\rho(\sigma_N); N \in A\}$ one cannot, of course, calculate the probabilities in (3.14) (the denominator and some of the numerators in (3.14) being, in this case, unknown), but one *can* compute the conditional densities with respect to the event $A \cup \{\emptyset\}$, defined by

$$r(M|A \cup \{\emptyset\}) = \frac{s_M^2}{\sum_{N \in A \cup \emptyset} s_N^2} \qquad \text{if } M \in A \text{ or } M = \emptyset, \text{ 0 otherwise.}$$

$H(r)$ can, therefore, be bounded from above by the supremum of $H(t)$ as t varies in the set of all probability densities on 2^{Λ_ν} which, under conditioning with respect to the event $A \cup \{\emptyset\}$, give $t(M|A \cup \{\emptyset\}) = r(M|A \cup \{\emptyset\})$, $\forall M \in 2^{\Lambda_\nu}$.

This supremum is, in fact, attained (as it is fairly intuitive to guess, and easy to prove by conventional Lagrange multiplier methods) by a density t^* corresponding to a total probability mass α distributed among the elements of $A \cup \{\emptyset\}$ proportionally to $r(\cdot|A \cup \{\emptyset\})$, and to a total probability mass, $\beta = 1 - \alpha$, uniformly distributed on the remaining subsets of Λ_ν.

The result is easily expressed in terms of the entropy

$$h_A \equiv H(r(\cdot|A \cup \{\emptyset\})) = -\sum_{M \subseteq A \cup \{\emptyset\}} r(M|A \cup \{\emptyset\}) \ln(r(M|A \cup \{\emptyset\}))$$

of the conditional density, a function only of $\{s_M; M \in A\}$:

$$\alpha = \frac{e^{h_A}}{|\bar{A}| - 1 + e^{h_A}}. \tag{3.16}$$

This leads to the inequality

$$H(r) \leqslant H(t^*) = \alpha h_A + \beta \ln |\bar{A}| - \alpha \ln \alpha - \beta \ln \beta \tag{3.17}$$

where $\bar{A}$ is the complement of A.

As a concluding remark, we wish to summarize the strategy outlined above, in order to provide an intuitive understanding of the approach that we propose.

$H(\rho)$ is the *unknown* entropy of the source that has generated the training sample. At the current level A of approximation, we suppose that the moments $\{s_M; M \in A\}$ have been estimated with negligible sampling error, and that equation (1.9) has been solved with respect to the coupling constants $\{\theta_M; M \in A\}$. The issue is: is $I(\rho, \rho_A)$ smaller than the threshold $I_{\min}$?

$I_{\min}$ is determined here by the requirement that a sample of the size that will be drawn from the simulator in the actual application have a preassigned, large, probability of being classified as coming from the environment. Han's inequality sets an upper bound on $H(\rho_A)$ of the form $H(\rho_A) \leqslant h_2(\{s_M; M \in A\}) \equiv h_2(A)$ whose right-hand side can be computed at the current level of approximation. Hirschman's inequality sets a lower bound on $H(\rho)$ of the form: $H(\rho) \geqslant \nu \ln 2 - \alpha h_A + \beta \ln |\bar{A}| - \alpha \ln \alpha - \beta \ln \beta \equiv h_1(A)$ whose right-hand side can also be computed at the current level of approximation.

The qualitative meaning of this inequality is the following: having explored a large enough A and having found there many moments s_M of absolute value 'small' with respect to $s_\emptyset \equiv 1$, so that the conditional probability density $r(M|A \cup \{\emptyset\})$ is far from uniform and therefore h_A is 'small', one can draw the conclusion that $H(\rho)$ is 'large'. As

$$I(\rho, \rho_A) = H(\rho_A) - H(\rho) \leqslant h_2(A) - h_1(A)$$

if $h_2(A) - h_1(A) \leqslant I_{\min}$ one can draw the conclusion that the required accuracy has been attained with the current set of parameters.

If, on the contrary, $h_2(A) - h_1(A) \geqslant I_{\min}$ it may be necessary to enlarge A by a new subset M_1. Having estimated s_{M_1}, inequality (2.4) says that $\rho_{A \cup \{M_1\}}$ will be, in relative entropy, by at least an amount $\frac{(s_{M_1} - s_{M_1}(A))^2}{2}$, closer than ρ_A to ρ.

Even before solving the method of moments equations

$$\rho_{E_{\rho_{A \cup \{M_1\}}}}(\sigma_N) = s_N \qquad \forall N \in A \cup \{M_1\}$$

inequality (2.9) says that θ_{M_1} will go from the value 0 it currently has in ρ_A to a new value $\theta_{M_1}(A \cup M_1)$ satisfying

$$\frac{\theta_{M_1}(A \cup M_1)}{s_{M_1} - s_{M_1}(A)} \geqslant 1.$$

The initial point given by the parameters of ρ_A, supplemented by $s_{M_1} - s_{M_1}(A)$ as initial guess for θ_{M_1} suggests itself as natural in an iterative search of the solution of the method of moments equations.

4. Discussion and open problems

We observe that the parameter updating (by gradient descent) of a Boltzmann machine, reviewed in the introduction, proceeds in a way strongly reminiscent of Hebb's (1949) neurophysiological postulate: 'When an axon of cell A is near enough to excite a cell B and repeatedly and persistently takes part in firing it, some growth process or metabolic change takes place in one or both cells, such that A's efficiency as one of the cells firing B is enhanced'. In this suggestive language, our considerations of section 2, in particular inequality (2.8), can be read as a variational motivation and an *a priori* estimate of the amount of this enhancement.

As to section 3, we observe that inequality 3.13 does have a physical meaning: taking the original σ as components of Heisenberg's quantum spins in a given space direction (say direction 3), $H(r_\theta)$ is the entropy of the distribution of the components of the same spins in a direction orthogonal to the previous one, say direction 1.

$\hat{\psi}(M) = \frac{1}{2^{\nu/2}} \sum_{\underline{\sigma} \in S^\nu} \rho(\underline{\sigma})^{1/2} \sigma_M$ is, indeed, in a suitable representation of the Pauli spin operators $\underline{S}(j)$ in which the $S_3(j)$'s are diagonal, the scalar product of the wavefunction $\psi(\underline{\sigma}) = \rho(\underline{\sigma})^{1/2}$ with the simultaneous eigenstate $\frac{1}{2^{\nu/2}}\sigma_M$ of the operators $S_1(j)$ belonging to the eigenvalue -1 if j is in M, to the eigenvalue $+1$ if j is not in M.

For systems in which the signals are written on quantum carriers (Feynman 1985, Deutsch 1985), this dual model might be physically accessible, see Apolloni *et al* (1989) for a preliminary exploration (restricted to combinatorial optimization) of the computational capabilities of a Heisenberg chain. We are working on the problem of extending such an analysis to the learning problem.

Our final remark concerns the hardware implementability of the higher-order models considered here: they do sacrifice some of the simplicity of the more conventional second-order models (with only two-body interactions plus 'hidden nodes') to the important requisite of existence and uniqueness of ρ_A for given ρ and A. We are exploring the possibility of implementing higher-order models on the p-RAM architecture (Clarkson *et al* 1992).

The p-RAM architecture realizes *on silicon* exactly the higher-order models we are interested in. The coordinates of the model which are actually accessible to updating (the memory contents) define, however, a chart different from the s- and θ-charts considered here. In a previous paper (Apolloni *et al* 1997) we have shown how to implement, in this new chart, the covariant learning rule of Amari *et al* (1992, 1998), and have examined some entropic rules for the optimization of the connection layout.

References

Ackley D H, Hinton G E and Sejnowski T J 1985 A learning algorithm for Boltzmann machines *Cognitive Sci.* **9** 147–69

Akaike H 1974 A new look at the statistical model identification *IEEE Trans. Automatic Control* **19** 716–23

Amari S 1998 Natural gradient works efficiently in learning *Neural Comput.* **10** 251–76

Amari S, Kurata K and Nagaoka H 1992 Information geometry of Boltzmann machines *IEEE Trans. Neural Networks* **3** 260–71

Apolloni B, Caravalho C and de Falco D 1989 Quantum stochastic optimisation *Stochastic Process. Appl.* **33** 234–44

Apolloni B, de Falco D and Taylor J G 1997 pRAM layout optimization *Neural Networks* **10** 1709–16

Azencott R 1992 Boltzmann machines: higher order interactions and synchronous learning *Stochastic Models in Image Analysis* ed P Barone and A Frigessi (Berlin: Springer)

Clarkson T G, Gorse D and Ng C K 1992 Learning probabilistic RAM nets using VLSI structures *IEEE Trans. Comput.* **41** 1552–61

Dembo A and Cover T M 1991 Information theoretic Inequalities *IEEE Trans. Inf. Theory* **37** 1501–18

Deutsch D 1985 Quantum theory, the Church-Turing principle and the universal quantum computer *Proc. R. Soc.* A **400** 97–117

Feynman R 1985 Quantum mechanical computers *Opt. News* **11** 11–20
Glauber R J 1963 Time dependent statistics of the Ising model *J. Math. Phys.* **4** 294–307
Gross L 1975 Logarithmic Sobolev inequalities *Am. J. Math.* **97** 1061–83
Han T S 1978 Non negative entropy measures of multivariate symmetric correlations *Inf. Contr.* **36** 133–56
Hebb D O 1949 *The Organisation of Behaviour* (New York: Wiley)
Hirschman I I 1957 A note on entropy *Am. J. Math.* **79** 152–6
Kullback S 1959 *Information Theory and Statistics* (New York: Wiley)
Lanford O E 1973 Entropy and equilibrium states in classical statistical mechanics *Statistical Mechanics and Mathematical Problems* ed A Lenard (Berlin: Springer)
Murata N, Yoshizawa S and Amari S 1995 Network information criterion—determining the number of hidden units for an artificial neural network model *IEEE Trans. Neural Networks* **5** 865–72
Schuetzenberger M P 1954 Contribution aux applications statistiques de la theorie de l'information *Publ. Inst. Stat. Univ. Paris* **1** 3–117