EUROPHYSICS LETTERS 1 March 2000

Europhys. Lett., **49** (5), pp. 685–691 (2000)

Statistical mechanics of mutual information maximization

R. Urbanczik

Neural Computing Research Group, Aston University
Aston Triangle, Birmingham B4 7ET, UK

(received 14 May 1999; accepted in final form 14 December 1999)

PACS. 87.10.+e – General theory and mathematical aspects.
PACS. 05.50.+q – Lattice theory and statistics (Ising, Potts, etc.).
PACS. 64.60.Cn – Order-disorder transformations; statistical mechanics of model systems.

Abstract. – An unsupervised learning procedure based on maximizing the mutual information between the outputs of two networks receiving different but statistically dependent inputs is analyzed (Becker S. and Hinton G., *Nature*, **355** (1992) 161). By exploiting a formal analogy to supervised learning in parity machines, the theory of zero-temperature Gibbs learning for the unsupervised procedure is presented for the case that the networks are perceptrons and for the case of fully connected committees.

It has long been realized that information theory can provide a useful conceptual framework for unsupervised learning in neural networks. The basic idea is to treat the network as a channel of limited capacity and to adapt the parameters of the network to optimize the information transfer. Different optimality criteria exist, and a discussion of such criteria and some equivalences between them from the perspective of statistical physics is given in [1, 2].

In most cases, however, the information theoretic approach has led to useful learning algorithms only for single-layer networks. An exception is the proposal by Becker and Hinton (see [3], a review is given in [4]) which builds on ideas from computational linguistics. They assume two different but statistically dependent modes of input, ξ_1 and ξ_2, and each of these modes is processed by a different network. Due to the statistical dependency, some features of one input mode will be predictable given the other input mode, and the goal of training is to discover such mutually predictable features by maximizing the mutual information of the outputs of the two networks.

It is worthwhile mentioning that in the context of sensory processing the scenario considered by Becker and Hinton is by no means artificial. For instance, simultaneous auditory (ξ_1) and visual (ξ_2) sensations are statistically dependent since they may be caused by the same object, and such dependences obviously provide useful information about the nature of the object. Statistical dependences also arise within one sensory system at different times: In speech the current phoneme (ξ_1) is to a certain extent predictable from the preceding ones (ξ_2) and the same phenomenon re-occurs at the level of words.

In an application to vision, a simulation in [3] shows that two multilayer networks can learn higher-order features by maximizing the mutual information of their outputs. They learn to estimate the distance of an object from the stereo disparity which may arise when the object

is seen by two eyes. Formally, each network has to detect whether one half of its input is a shifted copy of its other half. Shift is not a linearly separable concept, and thus could not have been learned by unsupervised learning procedures such as principal-component analysis or competitive learning [5] which are restricted to single-layer networks.

While the proposal by Becker and Hinton provides a conceptually elegant "model of cortical self-organization" [4], practical experience shows that this learning scenario can be computationally quite difficult [6]. As a first step in theoretically analyzing mutual information maximization, this letter examines its sample complexity, that is the relationship between the performance of the networks on a training set which is a sample of the input distribution and their off-sample performance. We initially assume that the two networks are perceptrons and establish a formal analogy between mutual information maximization and a supervised learning problem for parity machines. Focusing on the case that some feature of one input is perfectly predictable given the other input, learning curves for zero-temperature Gibbs learning are calculated. In a second step, the analysis is extended to the situation when the two networks are fully connected committee machines.

We consider the case that the outputs $\sigma_i(\xi_i)$ of the two networks are discrete and then their mutual information $I(\sigma_1, \sigma_2)$ is given by

$$I(\sigma_1, \sigma_2) = \sum_{y_1, y_2} p_{\sigma_1 \sigma_2}(y_1, y_2) \ln \frac{p_{\sigma_1 \sigma_2}(y_1, y_2)}{p_{\sigma_1}(y_1) p_{\sigma_2}(y_2)} \,. \tag{1}$$

The sum runs over all possible output values, $p_{\sigma_1,\sigma_2}(y_1, y_2)$ is the joint probability that $(\sigma_1, \sigma_2) = (y1, y2)$ and p_{σ_i} denote the corresponding marginal probabilities, *e.g.*, $p_{\sigma_1}(y_1) = \sum_{y_2} p_{\sigma_1,\sigma_2}(y_1, y_2)$. Thus $I(\sigma_1, \sigma_2)$ is a distance measure (KL-divergence) between the joint distribution of the outputs and the product distribution of the marginals; it vanishes when σ_1 and σ_2 are statistically independent. In terms of the joint and marginal entropies the mutual information may be written as $I(\sigma_1, \sigma_2) = -H(\sigma_1, \sigma_2) + H(\sigma_1) + H(\sigma_2)$, where, *e.g.*, $H(\sigma_1) = -\sum_{y1} p_{\sigma_1}(y_1) \ln p_{\sigma_1}(y_1)$.

The goal of the learning process is to find σ_i within given classes of networks Σ_i such that the joint distribution of $\sigma_1(\xi_1)$ and $\sigma_2(\xi_2)$ maximizes (1). As usual, learning is based on a training set of inputs with elements (ξ_1^μ, ξ_2^μ), $\mu = 1, \ldots, P$, drawn independently of the input distribution, and we shall focus on the learning strategy that chooses $\sigma_i \in \Sigma_i$ to maximize (1) on the empirical distribution given by the training set, that is for $p_{\sigma_1\sigma_2}(y_1, y_2) = P^{-1} \sum_{\mu=1}^{P} \prod_{i=1}^{2} \delta_{y_i, \sigma_i(\xi_i^\mu)}$.

An important property of (1) is that as long as one just considers a single instance of an input pair and a corresponding pair of outputs, it is not possible to decide whether this specific input/output relationship contributes to maximizing the mutual information. So, in contrast to other learning scenarios, $I(\sigma_1, \sigma_2)$ is not the average of some objective function L over all samples, $I(\sigma_1, \sigma_2) \neq \langle L(\sigma_1(\xi_1), \sigma_2(\xi_2)) \rangle_{(\xi_1, \xi_2)}$. To avoid this problem, we shall consider only the case of binary outputs, $\sigma_i \in \{-1, 1\}$ and make the following assumptions:

a) The marginals of the input distribution are point symmetric around 0, $p_{\xi_i}(x_i) = p_{\xi_i}(-x_i)$.

b) The function implemented by the networks are odd, $\sigma_i(-x_i) = -\sigma_i(x_i)$.

As a consequence of the assumptions, the marginal entropies of the outputs will be independent of the choice of σ_i and maximal, $H(\sigma_i) = \ln 2$. In general, controlling the marginal entropies for perceptron-like architectures is just a question of adjusting the bias, and one will

not expect this to dominate the complexity of the learning process. Rewriting (1) in terms of the conditional probability $p_{\sigma_1\sigma_2}(y_1|y_2)$ and using a) and b) yields

$$I(\sigma_1,\sigma_2) = \sum_{y_1} p_{\sigma_1\sigma_2}(y_1|1)\ln p_{\sigma_1\sigma_2}(y_1|1) + \ln 2\,. \tag{2}$$

So to maximize $I(\sigma_1,\sigma_2)$ one would ideally choose σ_i so that $p_{\sigma_1\sigma_2}(y_1|1)$ is 0 or 1, that is the relationship between $\sigma_1(\xi_1)$ and $\sigma_2(\xi_2)$ should be functional. But in general, due to the nature of the input distribution and/or the choice of network architecture, such a functional relationship may not be achievable. However, even in these cases $I(\sigma_1,\sigma_2)$ is maximized by maximizing (or, equivalently, minimizing) the expectation of the product $\sigma_1\sigma_2$. Hence the maximization of $I(\sigma_1,\sigma_2)$ becomes equivalent to a supervised learning problem for the parity architecture obtained by combining the two networks. The only difference to the standard case is that we are only given examples for which $\sigma_1(\xi_1^\mu)\sigma_2(\xi_2^\mu) = 1$.

To analyze mutual information maximization, one thus may consider the partition function

$$Z = \int \mathrm{d}\sigma_1 \mathrm{d}\sigma_2 \exp\left[-\beta \sum_\mu \theta(-\sigma_1(\xi_1^\mu)\sigma_2(\xi_2^\mu))\right] \tag{3}$$

in the limit $\beta \to \infty$. Here $\mathrm{d}\sigma_i$ relates to some prior measure on the space of all networks Σ_i and θ is the Heaviside step function. We shall focus on realizable cases and assume that the statistical dependence between ξ_1 and ξ_2 is such that $\tau_1(\xi_1) = \tau_2(\xi_2)$ holds for suitable features $\tau_i \in \Sigma_i$. This means that the joint density of the inputs is related to the marginals via

$$p_{\xi_1\xi_2}(x_1,x_2) = 2\theta\left(\tau_1(x_1)\tau_2(x_2)\right) p_{\xi_1}(x_1) p_{\xi_2}(x_2). \tag{4}$$

For the moments of Z one then finds for $\beta \to \infty$

$$\begin{aligned}\langle Z^n\rangle &= \int \prod_{a=1}^n (\mathrm{d}\sigma_1^a \mathrm{d}\sigma_2^a) \left\langle \prod_{a=1}^n \theta(\sigma_1^a(\xi_1)\sigma_2^a(\xi_2)) \right\rangle_{(\xi_1,\xi_2)}^P = \\ &= \int \prod_{a=1}^n (\mathrm{d}\sigma_1^a \mathrm{d}\sigma_2^a) \left\langle 2\theta\left(\tau_1(\xi_1)\tau_2(\xi_2)\right) \prod_{a=1}^n \theta(\sigma_1^a(\xi_1)\sigma_2^a(\xi_2)) \right\rangle_{\xi_1,\xi_2}^P ,\end{aligned} \tag{5}$$

where the average on the RHS is over the joint distribution of ξ_1 and ξ_2 on the first line, and over the product of the marginals on the second line. The last average in (5) could also be written as $\langle \prod_{a=1}^n \theta(\sigma_1^a(\xi_1)\sigma_2^a(\xi_2)\tau_1(\xi_1)\tau_2(\xi_2))\rangle_{\xi_1,\xi_2}$ and this shows that the moments are the same as for the standard supervised learning problem for the corresponding parity network. To simplify the calculations, we shall assume that the τ_i are picked from the same distribution as the σ_i and perform an average of $\langle Z^n\rangle$ over the choice of τ_i to obtain

$$\langle\langle Z^n\rangle\rangle = \int \prod_{a=0}^n (\mathrm{d}\sigma_1^a \mathrm{d}\sigma_2^a) \left\langle 2\prod_{a=0}^n \theta(\sigma_1^a(\xi_1)\sigma_2^a(\xi_2)) \right\rangle_{\xi_1,\xi_2}^P , \tag{6}$$

where the τ_i are now the 0-th replica.

While the calculation of the partition function is very similar to the supervised case, in mutual information maximization we are interested in how well the two networks extract the features arising from the statistical dependence. Consequently we define the generalization error of the i-th network to be

$$\epsilon_i = \min\{\langle\theta(\sigma_i(\xi_i)\tau_i(\xi_i))\rangle_{\xi_i}, \langle\theta(-\sigma_i(\xi_i)\tau_i(\xi_i))\rangle_{\xi_i}\}, \tag{7}$$

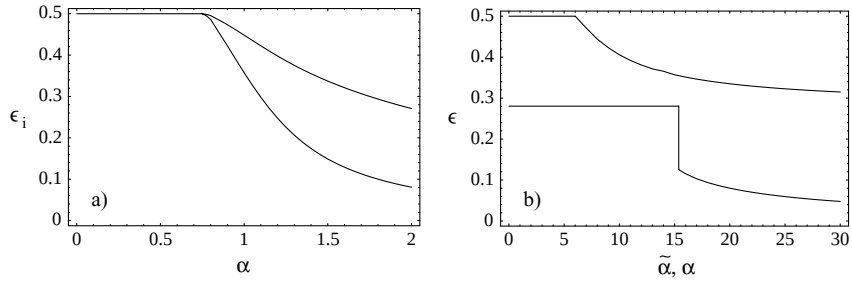

Fig. 1 – (a) Learning curves for the case when the two networks are perceptrons with $N_1/N_2 = 10$ and $P = \alpha N_1$. The upper curve shows ϵ_1, the lower one ϵ_2. (b) Learning curves for the case when the two networks are large committee machines. The upper curve is for $N_1 = N_2$ and $P = \tilde{\alpha} N_1$, the lower curve is for $K_1 N_1 = K_2 N_2$ and $P = \alpha K_1 N_1$. In both cases $\epsilon_1 = \epsilon_2 = \epsilon$.

reflecting the fact that the pair $(-\tau_1, -\tau_2)$ is equivalent to (τ_1, τ_2).

We first consider the case that the two networks are perceptrons, so $\sigma_i(x_i) = \mathrm{sign}(J_i^T x_i)$, where $J_i, x_i \in I\!R^{N_i}$. We assume that the prior on the class of networks is given by the uniform density on the unit sphere, $J_i^T J_i = 1$, and that the marginal densities of the inputs p_{ξ_i} are Gaussian with zero mean and variance N_i. Assuming replica-symmetry, it is now straightforward to evaluate (6) for large N_i in the limit $n \to 0$ and find

$$\langle\langle \ln Z \rangle\rangle \sim P \ln 2 + \max_{q_1, q_2} P\, G\left(\sqrt{\frac{q_1}{1-q_1}}, \sqrt{\frac{q_2}{1-q_2}} \right) + N_1 S(q_1) + N_2 S(q_2)\,. \tag{8}$$

Here $S(q) = \frac{1}{2} q + \frac{1}{2} \ln(1-q)$ and G is given by

$$\begin{aligned} G(Q_1, Q_2) &= 2 \left\langle \mathcal{H}(Q_1 z_1, Q_2 z_2) \ln \mathcal{H}(Q_1 z_1, Q_2 z_2) \right\rangle_{z_1, z_2}\,, \\ \mathcal{H}(z_1, z_2) &= H(z_1) H(z_2) + H(-z_1) H(-z_2)\,, \end{aligned} \tag{9}$$

where the z_i are real, independent and normally distributed, and $H(z) = \frac{1}{2} \mathrm{erfc}(z/\sqrt{2})$. The q_i represent the replica-symmetric weight vectors in different replicas, $q_i = J_i^{a\,T} J_i^b$, and since we averaged over the features τ_i, this is also the overlap between a student in version space and the weight vector defining τ_i. Hence the generalization errors are $\epsilon_i = \frac{1}{\pi} \arccos q_i$.

The balanced case $N_1 = N_2$ has already been considered in the literature on supervised learning in parity machines [7,8]. Due to the symmetry of the problem the two generalization errors are equal, $\epsilon_1 = \epsilon_2$. To achieve nontrivial behavior, the sample size P must scale with the system size, and we set $P = \alpha N_1$. The remarkable feature of the learning curve is that for $\alpha < \pi^2/4$ no generalization occurs and $\epsilon_i = \frac{1}{2}$. Above this critical value there is a continuous transition to nontrivial generalization and for large α the same asymptotic behavior as in realizable supervised learning for a single perceptron is found: $\epsilon_i = 0.62/\alpha$ in this limit.

In mutual information maximization, however, it also makes good sense to consider the case where one of the features is harder to learn than the other one, and a model for this is to assume that N_1 is larger than N_2. We set $N_1/N_2 = \kappa$ and consider the learning curve for the scaling $P = \alpha N_1$. The transition to nontrivial generalization now occurs at the critical value $\alpha = \frac{\pi^2}{4} \kappa^{-\frac{1}{2}}$. The learning curves for the case that $\kappa = 10$ (fig. 1a) show that the two generalization errors now behave differently. In particular beyond the transition $\epsilon_2 < \epsilon_1$ and ϵ_2 decays rather quickly with α. This behavior of ϵ_2 reflects the fact that the number of free parameters in the second networks is smaller than in the first network.

It is interesting to take this to extremes and consider the situation were $N_1 \gg N_2 \gg 1$. Since the entropy function S is quadratic for small arguments, a scaling $q_1 = O(\sqrt{N_2/N_1})$ and $q_2 = O(1)$ in eq. (8) yields entropy terms $N_1 S(q_1)$ and $N_2 S(q_2)$ of the same magnitude $O(N_2)$. Using this scaling for the order parameters yields that the energy term $P \ln 2 + PG(Q_1, Q_2)$ is also of order N_2 when $P = O(\sqrt{N_1 N_2})$. Hence the scale of the learning curve for the easier problem is not simply dominated by the harder one and the scaling $P = \alpha\sqrt{N_1 N_2}$, $q_1 = O(\sqrt{N_2/N_1})$ yields a well-defined extremal problem. On this scale a second-order transition to nontrivial generalization for ϵ_2 is found at $\alpha = \pi^2/4$. This is the same numerical value as in the balanced case since both results are related to the $q_i \to 0$ expansion of (8). Beyond the transition, however, ϵ_2 behaves quite differently and for large α a fast asymptotic decay of $\epsilon_2 = \frac{\pi^2}{4}\alpha^{-2}$ is found. Of course to leading order $\epsilon_1 = \frac{1}{2}$ for this scaling of P.

We next turn to the case that the two networks are fully connected committee machines with K_i hidden units. So $\sigma_i(x_i) = \operatorname{sign}(\sum_{j=1}^{K_i} \operatorname{sign}(J_{ij}^T x_i))$ for $J_{ij}, x_i \in I\!R^{N_i}$. The target features τ_i are assumed to have the same structure as the students, and we assume, as in the case of the perceptron, Gaussian marginal distributions of the inputs with zero mean and variance N_i. In contrast to the case of the perceptron the learning curves will now depend on the choice of the weight vectors $B_{ij} \in I\!R^{N_i}$ for the target features τ_i, and in particular on their overlaps $B_{ij}^T B_{ik}$. We shall consider the orthogonal case here and thus assume that the B_{ij} are picked from the uniform density on the set of vectors with $B_{ij}^T B_{ik} = \delta_{jk}$. We then need to make the same assumption about the prior distribution of student weight vectors J_{ij} to preserve the symmetry which allows a simple calculation of the moments of the partition function (6). This may seem slightly artificial, and one might instead wish to consider for the students the less restrictive prior of the uniform distribution on the set $J_{ij}^T J_{ij} = 1$. However, for supervised learning in the committee machine with an orthogonal teacher, it was found in [9] that at zero temperature the weight vectors of a typical student in version space are in fact orthogonal as well. Similarly, in the present case it is possible to calculate the quenched average of $\ln Z$ for the less restrictive prior using eq. (5). This yields that for orthogonal B_{ij} a solution of the saddle point equations arising from (5) is given by orthogonal students ($J_{ij}^{a\,T} J_{ik}^a = \delta_{jk}$). So for orthogonal τ_i one will not expect the choice between the two priors on the students to influence the learning curves.

To simplify the exposition, we thus assume the more restrictive prior, $J_{ij}^T J_{ik} = \delta_{jk}$ and we further assume the replica- and site-symmetric parametrization of the version space

$$J_{ij}^{a\,T} J_{ik}^b = p_i/K_i + \delta_{jk} q_i\,, \quad \text{for } a \neq b. \tag{10}$$

We consider the limit of many hidden units ($1 \ll K_i \ll N_i$) since (10) then implies that the joint distribution of the fields $K_i^{-\frac{1}{2}} \sum_{j=1}^{K_i} \operatorname{sign}(J_{ij}^{a\,T} x_i)$ will become Gaussian [10]. In this limit the average of $\ln Z$ is given by

$$\begin{aligned} \langle\langle \ln Z \rangle\rangle \sim P \ln 2 \; + \; & \max_{q_1, q_2} \; P\, G\left(\sqrt{\frac{q_1^{\mathrm{e}}}{1 - q_1^{\mathrm{e}}}}, \sqrt{\frac{q_2^{\mathrm{e}}}{1 - q_2^{\mathrm{e}}}} \right) + \\ & + \; \sum_{i=1}^{2} N_i \left((K_i - 1) S(q_i) + S(p_i + q_i) \right) \,. \end{aligned} \tag{11}$$

Here $q_i^{\mathrm{e}} \equiv \frac{2}{\pi}(p_i + \arcsin q_i)$ and the expressions for G and S are the same as in the case of the perceptron (8). Further the generalization errors are given by $\epsilon_i = \frac{1}{\pi} \arccos q_i^{\mathrm{e}}$.

For the large committee the learning curves show a nontrivial behavior on different scales and we shall exclude some of the more extreme cases by assuming that the number of free

parameters in one network is much larger than the number of input dimensions of the other network, $K_i N_i \gg N_j$.

We first assume that the sample size is relatively small and set $P = \tilde{\alpha} N_1$. On this scale the system is in a permutation symmetric phase, $q_i = 0$, since $P \ll K_i N_i$. As for the perceptron the learning curve is initially flat but the transition to nontrivial generalization now occurs at the higher critical value of $\tilde{\alpha} = \frac{\pi^4}{16}\kappa^{-\frac{1}{2}}$, where again $\kappa = N_1/N2$. In the limit of large $\tilde{\alpha}$, assuming $\kappa = O(1)$, the generalization errors decay to a next plateau value, $\epsilon_i = \frac{1}{\pi}\arccos\frac{2}{\pi}$. An example of this behavior is shown in fig. 1b.

For larger sample sizes, when $P \gg N_i$, the order parameters satisfy the relation $p_i = 1 - q_i$ to order $O(N_i/P)$. Hence the generalization behavior depends on the architecture of the networks only via the number of free parameters $K_i N_i$.

Specializing to the case $K_1 N_1 = K_2 N_2$ and setting $P = \alpha K_1 N_1$ yields that initially the generalization errors are constant at the plateau value found in the $\tilde{\alpha} \to \infty$ limit (cf. fig. 1b). At $\alpha = 15.4$ a discontinuous transition to better generalization occurs, and for large α the asymptotic behavior is to leading order $\epsilon_i = 1.25/\alpha$. This is the same asymptotics as in the case of realizable supervised learning for the committee [9].

Remarkably, for the extremely unbalanced case $K_1 N_1 \gg K_2 N_2$, this imbalance is not reflected in the scale of the learning curve for ϵ_2. Setting $P = \alpha K_2 N_2$, one finds that $\epsilon_1 = \epsilon_2 = \frac{1}{\pi}\arccos\frac{2}{\pi}$ holds only initially. At $\alpha = 20.4$ a discontinuous transition to a lower value occurs in ϵ_2, while the value of ϵ_1 is constant for all α. Asymptotically ϵ_2 then decays to zero as $\epsilon_2 = 3.52/\alpha$. The reason for this behavior is that the first network provides useful information about the target features since $\epsilon_1 < \frac{1}{2}$. The present scenario is however not equivalent to the case where the second network is trained with a noisy teacher. This would lead to frustration in the second network and its generalization error would not decay to zero when training at zero temperature [10]. The difference to the case of a fixed noisy teacher is that in the present scenario the version space of the first network is shrinking with increasing α.

The above analysis shows that while the asymptotic decay of the generalization errors can be quite fast, in the initial stages of learning long plateaus occur. This may explain some of the computational difficulties experienced when applying mutual information maximization [6]. Such algorithmic issues have also been addressed in the context of supervised online learning in parity machines [11]. It was found that even for the optimal, within the class of algorithms considered, learning procedure no generalization occurs when the sample size is on the order of the system size unless knowledge about the target rule is already given to the networks by choosing nonrandom initial conditions. So it will be important to consider a wider class of algorithms for this problem.

A further issue is the extension of the present analysis from binary to m-ary classification. In the present case it was reasonable to consider a situation where the marginal entropies of the outputs are independent of the weights. This special feature of the binary case makes it possible to express the optimization problem for the mutual information in terms of the sample average of an objective function. For $m > 2$ this no longer holds, and an interesting question is how this technical difference is reflected in the learning behavior.

* * *

Part of this work was carried out during the Statphys-seminar at the Max Planck Institute for the Physics of Complex Systems in Dresden.

REFERENCES

[1] Nadal J. P. and Parga N., *Network: Comput. Neural Syst.*, **5** (1994) 565.
[2] Nadal J. P., Brunel N. and Parga N., *Network: Comput. Neural Syst.*, **9** (1998) 207.
[3] Becker S. and Hinton G., *Nature*, **355** (1992) 161.
[4] Becker S., *Network: Comput. Neural Syst.*, **7** (1996) 7.
[5] Hertz J., Krogh A. and Palmer R. G., *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City) 1991.
[6] Becker S. and Hinton G., in *Backpropagation: Theory, Architectures and Applications*, edited by Y. Chauvin and D. Rumelhart (Lawrence Erlbaum) 1994, pp. 313-349.
[7] Hansel D., Mato G. and Meunier C., *Europhys. Lett.*, **20** (1992) 471.
[8] Opper M., *Phys. Rev. Lett.*, **72** (1994) 2113.
[9] Schwarze H., *J. Phys. A*, **26** (1993) 5781.
[10] Urbanczik R., *J. Phys. A*, **28** (1995) 7097.
[11] Simonetti R. and Caticha N., *J. Phys. A*, **29** (1996) 4859.