Learning curves for mutual information maximization

R. Urbanczik
Institut für Theoretische Physik, Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany
(Received 29 November 2002; published 11 July 2003)

An unsupervised learning procedure based on maximizing the mutual information between the outputs of two networks receiving different but statistically dependent inputs is analyzed [S. Becker and G. Hinton, Nature (London) **355**, 161 (1992)]. For a generic data model, I show that in the large sample limit the structure in the data is recognized by mutual information maximization. For a more restricted model, where the networks are similar to perceptrons, I calculate the learning curves for zero-temperature Gibbs learning. These show that convergence can be rather slow, and a way of regularizing the procedure is considered.

DOI: 10.1103/PhysRevE.68.016106 PACS number(s): 84.35.+i, 89.20.Ff, 64.60.Cn

I. INTRODUCTION

In unsupervised learning one often tries to find a mapping σ of a high dimensional signal X into a simple output space $\mathbb{Y}$ which preserves the interesting and important features of the signal. The statement of the problem is rather vague, and a wealth of algorithms exist for the task which often define the meaning of ''interesting and important'' in terms of the algorithm itself [1]. In search for a principled approach, it seems natural to turn to information theory and to require that the mutual information $I(X;\sigma(X))$ between the signal X and its encoding $\sigma(X)$ should be large. Unfortunately, this is often a trivial problem. If one component of X, say the first one, has a continuous distribution, the mutual information between X and this component is infinite, and so $I(X;\sigma(X))$ can be maximized by simply choosing σ to project X onto its first component.

To arrive at a meaningful task one has thus considered maximizing $I(X;\sigma(X+\eta))$, where η is isotropic Gaussian noise [2]. Then if σ is constrained to be linear and X is Gaussian, the problem becomes equivalent to principal component analysis, but one can also consider nonlinear choices for σ. The drawback of this approach is that, if one reparametrizes X, setting $\hat{X}=\psi(X)$, then maximizing $I(\hat{X};\sigma(\hat{X}+\eta))$ will in general yield quite different results even if ψ is a simple linear and volume preserving mapping. So in this approach the meaning of ''interesting and important'' is implicitly defined by the choice of a coordinate system for X.

It is much more natural to apply information theory when considering the related scenario that one has access to two signals X_1 and X_2 which are different but statistically dependent. For instance, X_1 might be a visual and X_2 the corresponding auditory signal. Then $I(X_1;X_2)$ is a reparametrization invariant measure of the statistical dependence of the two signals, and one can ask for a simple encoding of X_1 that preserves the mutual information of the two signals. So in this scenario one will look for a mapping σ_1 of X_1 into a simple output space $\mathbb{Y}_1$ for which $I(\sigma_1(X_1);X_2)$ is large. This is the basic idea of the information bottleneck method [3,4].

In the same setting, a more symmetric approach has been proposed by Becker and Hinton [5,6]. The idea is to look for simple encodings σ_1,σ_2 of both signals which yield a large value of $I(\sigma_1(X_1);\sigma_2(X_2))$. An attractive feature of this approach is that to compute the mutual information of the encodings one has to estimate probabilities only in the simple output spaces $\mathbb{Y}_1$ and $\mathbb{Y}_2$ and not in the high dimensional space of the signals themselves.

While the main thrust of this paper is to analyze Becker and Hinton's proposal using statistical physics, I shall first give some general characteristics of what can be learned by maximizing I for a large class of scenarios where the approach seems suitable. I then specialize to the case that the σ_i are perceptronlike architectures with discrete output values and set up a framework for analyzing learning from examples in the thermodynamic limit. Next, some learning curves obtained for specific cases are discussed, and I conclude by addressing the limitations of the presented approach and some insights gained from it.

II. GENERAL CHARACTERISTICS

In general terms the mutual information of X_1 and X_2 is the Kullback-Leibler divergence between the joint distribution of X_1 and X_2 and the product distribution of their marginal distributions. If the variables have probability densities this definition reads

$$I(X_1;X_2)=\int dx_1 dx_2 p(x_1,x_2)\log_2\frac{p(x_1,x_2)}{p(x_1)p(x_2)}. \quad (1)$$

$I(X_1;X_2)$ is non-negative and vanishes only if X_1 and X_2 are independent. So a positive value indicates statistical dependence, and the ideal scenario for Becker and Hinton's proposal is that this dependence is such that for suitable functions τ_1 and τ_2 we have $\tau_1(X_1)=\tau_2(X_2)$ for any possible joint occurrence of a pair (X_1,X_2). For instance, $\tau_1(X_1)$ might be the common cause of the two signals. I shall further assume that the knowledge of $\tau_1(X_1)$ [or $\tau_2(X_2)$] encapsulates the entire statistical dependency of the two signals, so that the joint density of (X_1,X_2) can be written as

$$p(x_1,x_2)=\frac{\delta_{\tau_1(x_1),\tau_2(x_2)}}{z_{\tau_1(x_1)}}p(x_1)p(x_2). \quad (2)$$

For brevity I have assumed that the τ_i take on discrete values, so δ refers to Kronecker's delta and the normalization is

$$z_k = \mathrm{Prob}[\tau_1(X_1)=k] = \mathrm{Prob}[\tau_2(X_2)=k]. \quad (3)$$

If the joint distribution of the signals is given by Eq. (2), it makes sense to ask whether the τ_i can be inferred by observing only (X_1, X_2). This naturally leads one to consider the mutual information because a simple calculation shows that $I(X_1;X_2) = I(\tau_1(X_1);\tau_2(X_2))$. In the Appendix I show, using standard information theoretic relations, that any two mappings σ_i which also preserve the mutual information $I(X_1;X_2) = I(\sigma_1(X_1);\sigma_2(X_2))$ are related to the τ_i in a simple way; namely,

$$\tau_i(x_i) = \phi_i(\sigma_i(x_i)) \quad (4)$$

holds identically for suitable mappings ϕ_i, and in this sense the τ_i provide a simplest description of the data. If the σ_i have the same number of output values as the τ_i, the ϕ_i can only be permutations. Of course, as an unsupervised learning procedure, maximizing $I(\sigma_1(X_1);\sigma_2(X_2))$ does not fix specific values for the output labels. Despite this, I shall sometimes call the τ_i teachers and take such trivial permutational symmetries into account only tacitly.

Realistically, one will not be able to choose the σ_i based on the knowledge of the entire distribution of (X_1, X_2), but only have access to a training set $\mathbb{D}$ of finitely many example pairs (X_1^μ, X_2^μ) sampled independently from (X_1, X_2). For a given $\sigma = (\sigma_1, \sigma_2)$, a pair of students, one will then compute the empirical frequencies

$$p_{u_1,u_2}(\mathbb{D},\sigma) = \frac{1}{m}\sum_{\mu=1}^{m}\prod_{i=1}^{2}\delta_{u_i,\sigma_i(X_i^\mu)}, \quad (5)$$

where m is the number of examples in $\mathbb{D}$. Then the discrete version of Eq. (1) allows us to determine the empirical mutual information $I(\mathbb{D},\sigma)$ of the student pair on the training set by

$$I(\mathbb{D},\sigma) = \sum_{u_1,u_2=1}^{K} p_{u_1,u_2}(\mathbb{D},\sigma)\log_2\frac{p_{u_1,u_2}(\mathbb{D},\sigma)}{p_{u_1,\cdot}(\mathbb{D},\sigma)p_{\cdot,u_2}(\mathbb{D},\sigma)}; \quad (6)$$

here K is the number of output classes and the explicit formula for the first marginal in Eq. (6) is $p_{u_1,\cdot}(\mathbb{D},\sigma) = \sum_{u_2=1}^{K} p_{u_1,u_2}(\mathbb{D},\sigma)$.

When learning, one has to restrict σ_1 and σ_2 to lie in a predefined set of functions and the obvious strategy is to choose a pair maximizing $I(\mathbb{D},\sigma)$. Of course, Eq. (4) will then only hold in the limit $m\to\infty$ of an infinite training set, and a key issue is to quantify the speed of this convergence. This seems especially important since the number of values taken on by the τ_i is in general not known. So it is quite possible that K is chosen too large. Then, even in the infinite training set limit, there can be many different function pairs where σ_i takes on all of the K values, $I(\sigma_1(X_1);\sigma_2(X_2))$ is maximized, but $\phi_i(\sigma_i(X_i)) = \tau_i(X_i)$ can satisfied by mappings ϕ_i which merge class labels. Thus one cannot expect that the number of classes in the data is automatically inferred by mutual information maximization and will have to experiment with different values of K, considerably increasing the risk of overfitting.

III. STATISTICAL PHYSICS

Let us now assume that the τ_i are perceptrons which yield output values in $0,\ldots,K-1$, and each τ_i is characterized by an N-dimensional weight vector B_i of unit length and scalar biases κ_i^k, $k=1,\ldots,K-1$. On an N-dimensional input X_i the output of τ_i then is

$$\tau_i(x_i) = \sum_{k=1}^{K-1}\Theta(B_i^T x_i - \kappa_i^k), \quad (7)$$

where Θ is the 0,1 step function. While Eq. (7) is invariant with respect to permutations of the biases, for brevity, I shall always assume that the bias terms are in ascending order $(\kappa_i^k \leqslant \kappa_i^{k+1})$. The marginal densities $p(x_1)$ and $p(x_2)$ that are used to define the joint density of the data (2) are assumed to have independent Gaussian input components with zero mean and unit variance. Then, to satisfy condition (3), the biases of τ_1 and τ_2 must be equal, $\kappa_1^k = \kappa_2^k = \kappa^k$.

We assume that the general architecture of the teachers is known and focus on pairs of students σ_i performing a classification analogous to Eq. (7) but with weight vectors J_i and biases λ_i^k. Note that, while formally I assume that the number of biases is the same for teachers and students, this does not restrict generality. For instance, a scenario where the teachers have fewer output classes than the students is obtained by choosing some of the κ^k to be equal.

The performance of a student pair is then assessed using Eq. (6) to determine $I(\mathbb{D},\sigma)$. To investigate, in the thermodynamic limit, the typical properties of maximizing $I(\mathbb{D},\sigma)$, one has to fix a prior measure on the parameters of the students. For the weight vectors, we assume that the J_i are drawn from the uniform density dJ on the unit sphere. As there are only finitely many λ_i^k the results for $N\to\infty$ do not depend on the prior density $d\lambda$ on the biases as long as the density vanishes nowhere. One could now consider the partition function

$$Z = \int dJ \int d\lambda\, e^{\beta N I(\mathbb{D},\sigma)} \quad (8)$$

for the Gibbs weight $e^{\beta N I(\mathbb{D},\sigma)}$ on the space of students. But a key technical difference from many other learning paradigms is that this Gibbs weight does not factorize over the examples. There are, however, some special cases; namely, if there are just two output classes and no biases, where one can replace $I(\mathbb{D},\sigma)$ by an equivalent cost function which is just a sum over examples [7]. Then maximizing $I(\mathbb{D},\sigma)$ is closely related to a supervised learning problem for parity machines.

Here, I want to analyze more general scenarios and it is easier not to start with $e^{\beta N I(\mathbb{D},\sigma)}$ but to introduce target values t_{u_1,u_2} for the empirical frequencies $p_{u_1,u_2}(\mathbb{D},\sigma)$ which determine $I(\mathbb{D},\sigma)$. We now consider the partition function

$$Z=\int dJ \int d\lambda \prod_{u_1,u_2} \exp\left\{-\frac{\beta N}{2}[t_{u_1,u_2}-p_{u_1,u_2}(\mathbb{D},\sigma)]^2\right\}. \tag{9}$$

Analyzing the divergence of $\ln Z$ for $\beta\to\infty$ then tells us if the target values are feasible, i.e., whether student networks σ_i exist with $t_{u_1,u_2}=p_{u_1,u_2}(\mathbb{D},\sigma)$.

In the thermodynamic limit one expects to find two regimes. As long as the number of training examples m is small compared to N, it will be possible to find students that achieve the global maximum $\log_2 K$ of the mutual information. In terms of the target values this means that $t_{u_1,u_2}=K^{-1}\delta_{u_1,u_2}$ is feasible, and we need to study the partition function (9) for this choice of t_{u_1,u_2}. Once the ratio $\alpha=m/N$ becomes large enough, there will in general be no students σ such that $I(\mathbb{D},\sigma)=\log_2 K$, and we need to determine the achievable empirical frequencies by finding feasible target values of t_{u_1,u_2} using Eq. (9). We can then search for feasible target values that yield the maximal mutual information $I(\alpha)$.

For both regimes the starting point is to factorize Eq. (9) over the patterns, linearizing the exponent by an integral transform with Gaussians L_{u_1,u_2} of zero mean and unit variance:

$$e^{-(\beta N/2)[t_{u_1,u_2}-p_{u_1,u_2}(\mathbb{D},\sigma)]^2} = \langle e^{iL_{u_1,u_2}\sqrt{\beta N}[t_{u_1,u_2}-p_{u_1,u_2}(\mathbb{D},\sigma)]}\rangle_{L_{u_1,u_2}}. \tag{10}$$

One now employs standard arguments to calculate the quenched average in the thermodynamic limit and finds, within a replica symmetric parametrization,

$$\lim_{N\to\infty} N^{-1}\langle \ln Z\rangle_{\mathbb{D}} = \max_{R,\lambda}\min_{q,L} \alpha G_0(L)+\alpha G_1(R,\lambda,q,L)+G_2(R,q),$$

$$G_0(L)=\sum_{u_1,u_2}\frac{L_{u_1,u_2}^2}{2\beta}+L_{u_1,u_2}t_{u_1,u_2},$$

$$G_2(R,q)=\frac{1}{2}\sum_i \frac{q_i-R_i^2}{1-q_i}+\ln(1-q_i). \tag{11}$$

Here $R_i=J_i^T B_i$ is the typical overlap with the teacher of a student picked from the Gibbs distribution (9) and q_i is the squared length of the thermal average of J_i. Further,

$$G_1(R,\lambda,q,l)=\left\langle f_{\{R_i q_i^{-1/2}\}}(y_1,y_2)\ln\sum_{u_1,u_2} e^{-L_{u_1,u_2}} \times\prod_i H_{u_i}(\lambda_i,q_i,y_i)\right\rangle_{y_1,y_2}, \tag{12}$$

where the y_i are independent Gaussians with zero mean and unit variance. Also,

$$f_{\{R_i q_i^{-1/2}\}}(y_1,y_2)=\sum_k \frac{1}{z_k}\prod_i H_k(\kappa,R_i q_i^{-1/2},y_i) \tag{13}$$

with

$$H_{u_i}(\lambda_i,q_i,y_i)=H\left(\frac{\lambda_i^{u_i}-q_i y_i}{\sqrt{1-q_i}}\right)-H\left(\frac{\lambda_i^{u_i+1}-q_i y_i}{\sqrt{1-q_i}}\right). \tag{14}$$

Here $H(z)$ is Gardner's H function and to define Eq. (14) for $u_i=0$ and $u_i=K-1$ we adopt the convention that $\lambda_i^0=-\infty$ and $\lambda_i^K=\infty$. The definition of $H_k(\kappa,R_i q_i^{-1/2},y_i)$ is entirely analogous, also using $\kappa^0=-\infty$ and $\kappa^K=\infty$.

Note that the physical interpretation of the auxiliary variables L_{u_1,u_2} is that a student pair σ picked from the Gibbs density will yield empirical frequencies $p_{u_1,u_2}(\mathbb{D},\sigma)=t_{u_1,u_2}+L_{u_1,u_2}/\beta$. Reasonably, one will consider only target values t_{u_1,u_2} for these frequencies which sum to 1, and then the stationary values of L_{u_1,u_2} must sum to 0. This can, of course, also be obtained by direct manipulation of Eq. (11).

We are mainly interested in evaluating Eq. (11) for $\beta\to\infty$. The stationarity conditions for the order parameters yield that the scaling of a conjugate L_{u_1,u_2} in this limit will depend on whether t_{u_1,u_2} is positive or zero. Denoting by S_t the support of t, i.e., the set of pairs $u=(u_1,u_2)$ for which $t_{u_1,u_2}>0$, the stationarity conditions yield that L_{u_1,u_2} diverges with β as $\ln\beta$ if $u\notin S_t$. But for positive t_{u_1,u_2}, if t is feasible, L_{u_1,u_2} diverges as $-\ln\beta$, while for two pairs $u,\hat{u}\in S_t$, the difference between the conjugates

$$L_{u_1,u_2}-L_{\hat{u}_1,\hat{u}_2}=l_{u_1,u_2}-l_{\hat{u}_1,\hat{u}_2} \tag{15}$$

stays finite for large β. Thus one obtains for the limit $\beta\to\infty$

$$\lim_{N\to\infty} N^{-1}\langle \ln Z\rangle_{\mathbb{D}} = \max_{R,\lambda}\min_{q,l} \alpha\sum_{u\in S_t} l_{u_1,u_2}t_{u_1,u_2}+\alpha\hat{G}_1(R,\lambda,q,l)+G_2(R,q),$$

$$\hat{G}_1(R,\lambda,q,l)=\left\langle f_{\{R_i q_i^{-1/2}\}}(y_1,y_2)\ln\sum_{u\in S_t} e^{-l_{u_1,u_2}} \times\prod_i H_{u_i}(\lambda_i,q_i,y_i)\right\rangle_{y_1,y_2}. \tag{16}$$

When the mutual information is maximized by marginally feasible target values realized by only a single pair of students, we need to consider the limit $q_i\to 1$ in Eq. (16). As usual, the sum over u in $\hat{G}_1$ is dominated by its largest term in this limit. Setting

$$H_{u_i}^*(\lambda_i,y_i)=2\lim_{q_i\to 1}(1-q_i)\ln H_{u_i}(\lambda_i,q_i,y_i),$$

$$u^g(y_1,y_2)=\underset{u\in S_t}{\operatorname{argmax}}\{g_{u_1,u_2}+H^*_{u_1}(\lambda_1,y_1)+H^*_{u_2}(\lambda_2,y_2)/\gamma\}, \quad (17)$$

where, for $q_i\to 1$, γ is the ratio $(1-q_1)/(1-q_2)$, and $g_{u_1,u_2}=l_{u_1,u_2}(1-q_1)$, one obtains

$$t_{u_1,u_2}=\langle f_{\{R_i\}}(y_1,y_2)\,\delta_{(u_1,u_2),u^g(y_1,y_2)}\rangle_{y_1,y_2},$$

$$1-R_i^2=-\alpha\langle f_{\{R_i\}}(y_1,y_2)H^*_{u_i^g(y_1,y_2)}(\lambda_i,y_i)\rangle_{y_1,y_2}. \quad (18)$$

The interpretation of the above equations is that the target values t_{u_1,u_2} are marginally feasible for some value of α if one can find R_i, λ_i, g_{u_1,u_2}, and γ such that Eq. (18) holds for $u_i=0,\ldots,K-1$ and $i=1,2$.

Using the above results, the learning curves for maximizing $I(\mathbb{D},\sigma)$ in the large N limit can be calculated. In the regime where $I(\alpha)=\log_2 K$, we use Eq. (11) with the target values $t_{u_1,u_2}=K^{-1}\delta_{u_1,u_2}$. But above a critical number of examples $I(\alpha)$ will be smaller than $\log_2 K$. Using Eq. (18) to find feasible targets t_{u_1,u_2} that maximize the mutual information amounts to solving a constrained optimization problem.

IV. LEARNING CURVES

Before considering example scenarios, some words on numerically solving Eq. (16) or (18) are in order. This turns out to be a nontrivial task since averages of functions have to be computed which are quite nonsmooth, once the q_i are close to 1 in Eq. (16), and become discontinuous for Eq. (18). To achieve reliable numerical results, I have found it necessary to explicitly divide the two-dimensional domain of integration into subregions where the integrand is both continuous and differentiable. The number of subregions one has to consider increases quite rapidly with K.

Further, I have generally assumed site symmetry, $R_i=R,\lambda_i=\lambda,q_i=q$, although I did numerically check the local stability of the solution thus obtained for some points on the learning curves.

The simplest case is that the students have $K=2$ output classes and it is useful to first consider a degenerate scenario where the teachers have just a single output. So $I(X_1,X_2)=0$ and the two signals are in fact independent. This is analogous to the random map problem in supervised learning, since nothing can be learned, and any pair of students will perform equally badly on the whole distribution of inputs. But for finite α, up to $\alpha=11.0$, one can find student pairs achieving the maximal value $I(\mathbb{D},\sigma)=1$, as shown in Fig. 1. Above this critical value the maximal empirical mutual information $I(\alpha)$ starts to decay to zero, the feasible target matrix t becomes nondiagonal, but the value of the bias $\lambda^{(1)}$ is still zero. While above $\alpha=11.0$ student pairs with a diagonal t do exist, and have a nonzero $\lambda^{(1)}$, these pairs do not maximize $I(\mathbb{D},\sigma)$.

The random map problem is relevant for learning since the students always have the option of ignoring the structure in the data. Formally, when $R=0$ a learning problem with $I(X_1,X_2)>0$ is equivalent to the $I(X_1,X_2)=0$ case. This is illustrated (also Fig. 1) by a scenario where the teachers have two output classes and $\kappa^{(1)}=1$. This yields the moderate value $I(X_1,X_2)=0.631$. But up to $\alpha=22.3$ the structure present in the data is not recognized at all, and we observe the same behavior as for random examples. At $\alpha=22.3$ a first order phase transition occurs where R and $\lambda^{(1)}$ jump from zero to values which are already close to 1.

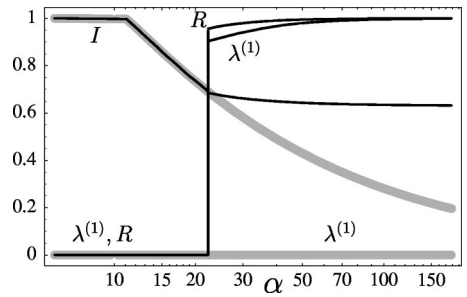

FIG. 1. Learning curves for students with $K=2$ output classes. The gray lines are for the random map problem, and the thin black lines for a pair of teachers with two output classes and $\kappa^{(1)}=1$.

When choosing $\kappa^{(1)}=0.5$, still for $K=2$, a different behavior is observed since $I(X_1,X_2)$ is now quite close to 1. The phase where $I(\alpha)=1$ is now a bit longer, extending up to $\alpha=11.1$. But already in this phase the order parameters show a nontrivial behavior. The value of R becomes positive above $\alpha=3.0$ but is not monotonic in α. So, while some structure is recognized in this phase due to entropic effects, the recognition is rather unreliable. This is also highlighted by the behavior of $\lambda^{(1)}$. While it is nonzero above $\alpha=3.0$, it initially even has very small negative values (not visible in Fig. 2). Above $\alpha=11.1$, when $I(\alpha)<1$, robust convergence of the order parameters to their asymptotic values sets in.

Turning to $K=3$ (outputs 0, 1, or 2), we again first consider the case of random examples. For all values of α the bias term satisfies the symmetry $\lambda^{(2)}=-\lambda^{(1)}$. The phase where $I(\alpha)$ has the maximal possible value, which now equals $\log_2 3$, is shorter than for $K=2$, extending to $\alpha=6.96$ as shown in Fig. 3. Above $\alpha=6.96$ the t matrix is still diagonal initially.

In this initial phase $\lambda^{(2)}$ decreases with increasing α; this narrows the gap between the output classes 0 and 2, making it easier to find a student pair with $t_{02}=0$. Remarkably, beyond $\alpha\approx 8$ one finds $\lambda^{(2)}=\lambda^{(1)}=0$ but $t_{11}>0$ as shown in

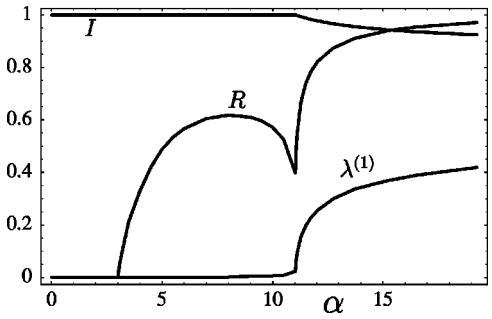

FIG. 2. Learning curves obtained when the students and the pair of teachers have two output classes but $\kappa^{(1)}=0.5$.

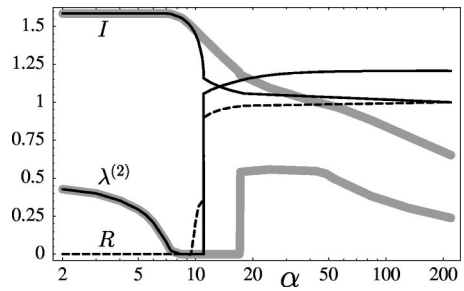

FIG. 3. Learning curves for students with $K=3$ output classes. The gray lines are for the random map problem, and the black lines for a pair of teachers with three output classes and $\kappa^{(2)}=-\kappa^{(1)}=1.21$.

Fig. 4. This verges on the paradoxical since by definition a student with $\lambda^{(2)}=\lambda^{(1)}$ will never produce the output label 1. However, we have taken the disorder average for $\lambda^{(1)}<\lambda^{(2)}$, so the observed result will naturally arise if the weight vectors of the optimal student pair satisfy $J_i^T X_i^\mu=0$ on a subset of $\mathbb{D}$. In addition, since we have take the thermodynamic limit first, $\lambda^{(2)}=\lambda^{(1)}$ may hold only in the large N limit and not for finite N.

At $\alpha=9.2$ a continuous phase transition occurs with the t matrix becoming nondiagonal (Fig. 4). It then has the form

$$t=\begin{pmatrix} a & 0 & b \\ 0 & c & 0 \\ b & 0 & a \end{pmatrix}. \tag{19}$$

This is followed by a first order phase transition at $\alpha=17.2$ with $\lambda^{(2)}$ jumping from 0 to 0.55. While the t matrix keeps its shape (19), the values of c and a change drastically. The class of solution the network is now exploring stays stable with increasing α and has a simple interpretation since the values of a and b converge. This means that from the point of mutual information there is no difference between outputs 0 and 2. In effect, the three-output-class architecture is emulating perceptrons, which have just two output values but use the nonmonotonic output function $\Theta(\lambda^{(2)}-|J_i^T\xi|)$. While perhaps not quite as powerful as the reversed-wedge perceptron [8], this architecture will have a very high storage capacity, and this leads to a remarkably slow convergence of $I(\alpha)$ to its asymptotic value of 0.

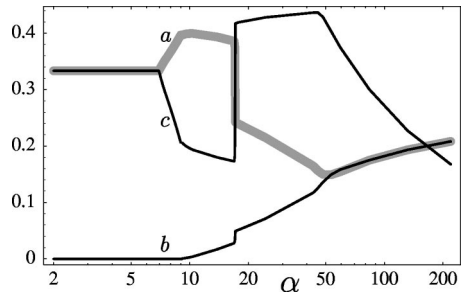

FIG. 4. Feasible t values for three output labels and random examples $a=t_{00}, b=t_{02}, c=t_{11}$ as in Eq. (19).

The slow convergence for random examples suggests that it may be useful to regularize mutual information maximization and one way of doing this is considered in Fig. 3. The teachers have three output classes and biases $\kappa^{(2)}=-\kappa^{(1)}=1.21$ yielding $I(X_1,X_2)=1$. The students also have three output classes but the training is regularized by choosing students that maximize the mutual information under the constraint that the t matrix be diagonal, so the outputs of the two students must be identical on the training set. The constraint becomes noticeable at $\alpha=9.4$, where the achievable $I(\alpha)$ is now lower than for the unconstrained case with $R=0$, i.e., the random problem discussed above. Due to the constraint there is a continuous phase transition to positive R at this point. Next, at $\alpha=10.9$, a first order phase transition to the asymptotic regime occurs, and the structure in the data is recognized well. At this point the biases become nonzero and satisfy the symmetry $\lambda^{(2)}=-\lambda^{(1)}$. Note that up to $\alpha=43$ the achievable $I(\alpha)$ is smaller than for the unconstrained random map problem. So regularizing the learning by constraining the student outputs to be equal is essential for the good generalization observed for α values in the range $[10.9,\ldots,43]$.

V. CONCLUSION

We have seen that mutual information maximization provides a principled approach to unsupervised learning. Interestingly, from a biological perspective, it emphasizes the role of multimodal sensor fusion in perception. In contrast to many other unsupervised learning schemes such as principal component analysis, mutual information maximization can capture very complex statistical dependencies in the data, if the architecture chosen for the two networks is powerful enough.

For the generic data model given by Eq. (2), I have shown that the structure in the data is recognized by mutual information maximization if the training set is sufficiently large, i.e., the procedure is consistent in a statistical sense. However, the detailed statistical physics calculations yield that many examples are needed to reach this asymptotic regime and that the learning process is complicated by many phase transitions. One reason for this is that a seemingly simple architecture such as a perceptron with three output classes can, from an information theoretic point of view, be equivalent to a perceptron which has just two output classes but uses a nonmonotonic activation function.

Of course, when considering the number of examples needed for reliable generalization, one has to keep in mind that examples are often much cheaper in unsupervised than in supervised learning. On the other hand, the detailed calculations have been for cases where the students are just perceptrons and there are only a few output classes. When increasing the number of output classes or when more powerful networks are used, one expects an even slower convergence. So, in applications, it may be necessary to compromise the generality of Becker and Hinton's approach by using suitable regularizations. We have considered one way of doing this, namely, constraining the two networks to give the same output on the examples in the training set.

A major limitation of the above statistical physics analysis is that I have considered only the replica symmetric theory. It is, however, evident that in many of the above scenarios replica symmetry will be broken. A case in point is the random map problem for two output classes where maximizing the mutual information yields a critical value $\alpha = 11.0$ up to which $I(\alpha) = 1$. This value is equal to the storage capacity of the tree parity machine with two hidden units [9], as one would expect, by the equivalence of the two problems in the unbiased case [7]. But one step of replica symmetry breaking, considered in [9] for the tree parity machine, shows that the critical capacity is in fact some 25% smaller.

To write down the one-step symmetry breaking equations for mutual information maximization is a straightforward task. But given the numerical difficulties already encountered in solving the replica symmetric equations, the numerics of one step of replica symmetry breaking are daunting. While one expects that some of the quantitative findings described above will change when replica symmetry breaking is taken into account, one can reasonably assume that more qualitative aspects such as the nature of the phase transitions are described correctly by the present theory.

ACKNOWLEDGMENTS

It is a pleasure to acknowledge many stimulating discussions with Georg Reents and Manfred Opper. This work was supported by the Deutsche Forschungsgemeinschaft.

APPENDIX

Our goal is to show that if the joint density of X_1 and X_2 satisfies Eq. (2), then $I(X_1;X_2) = I(\sigma_1(X_1);\sigma_2(X_2))$ implies Eq. (4). We shall need two facts from information theory (see, e.g., [10]). The first is the data processing inequality (DPI), which states that for any mapping σ

$$I(X_1;X_2) \geqslant I(X_1;\sigma(X_2)), \tag{A1}$$

processing cannot increase information. The second is the chain rule, which allows one to decompose the mutual information of a random variable X_1 with a pair of random variables (X_2,X_3) via

$$I(X_1;X_2,X_3) = I(X_1;X_3) + I(X_1;X_2|X_3), \tag{A2}$$

where the last term denotes the mutual information of the conditional distribution of (X_1,X_2) given a value of X_3, averaged over X_3.

Now, assuming Eq. (2) and

$$I(X_1;X_2) = I(\sigma_1(X_1);\sigma_2(X_2)), \tag{A3}$$

we have

$$\begin{aligned} I(X_1;X_2) &= I(X_1;\tau_2(X_2),\sigma_2(X_2)) \\ &= I(X_1;\sigma_2(X_2)) + I(X_1;\tau_2(X_2)|\sigma_2(X_2)) \\ &= I(X_1;X_2) + I(X_1;\tau_2(X_2)|\sigma_2(X_2)). \end{aligned} \tag{A4}$$

Here the first equality is a consequence of the DPI and Eq. (A3), the second is the chain rule, and the third is again the DPI and Eq. (A3).

So $I(X_1;\tau_2(X_2)|\sigma_2(X_2)) = 0$, and this means that X_1 and $\tau_2(X_2)$ are conditionally independent, given $\sigma_2(X_2)$. In other words,

$$p(X_1,\tau_2(X_2)|\sigma_2(X_2)) = p(X_1|\sigma_2(X_2))p(\tau_2(X_2)|\sigma_2(X_2)) \tag{A5}$$

or

$$p(X_1,\tau_2(X_2),\sigma_2(X_2)) = p(X_1,\sigma_2(X_2))p(\tau_2(X_2)|\sigma_2(X_2)). \tag{A6}$$

But from the definition of the joint density (2) we see that $p(X_1,\tau_2(X_2),\sigma_2(X_2))$ can be nonzero only if $\tau_1(X_1) = \tau_2(X_2)$ and in this case equals $p(X_1,\sigma_2(X_2))$. So $p(\tau_2(X_2)|\sigma_2(X_2))$ is either zero or 1 and this means that $\tau_2(X_2)$ is a function of $\sigma_2(X_2)$. By symmetry, this is also true of $\tau_1(X_1)$ and $\sigma_1(X_1)$.

[1] R. Duda, P. Hart, and D. Stork, *Pattern Classification* (Wiley, New York, 2001).
[2] G. Deco and D. Obradovic, *An Information-Theoretic Approach to Neural Computing* (Springer, New York, 1996).
[3] N. Tishby, F. Pereira, and W. Bialek, in 37th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, 1999 (unpublished).
[4] N. Slonim and N. Tishby, in *NIPS 12*, edited by S. Solla, T. Leen, and K. Müller (MIT Press, Cambridge, MA, 1999), pp. 617–623.
[5] S. Becker and G. Hinton, Nature (London) **355**, 161 (1992).
[6] S. Becker and G. Hinton, in *Backpropagation*: *Theory*, *Architectures and Applications*, edited by Y. Chauvin and D. Rumelhart (Lawrence Erlbaum, Hillsdale, NJ, 1994), pp. 313–349.
[7] R. Urbanczik, Europhys. Lett. **49**, 685 (2000).
[8] G.J. Bex, R. Serneels, and C. Van den Broeck, Phys. Rev. E **51**, 6309 (1995).
[9] E. Barkai, D. Hansel, and I. Kanter, Phys. Rev. Lett. **65**, 2312 (1990).
[10] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).