Pattern Capacity of a Perceptron for Sparse Discrimination

Vladimir Itskov and L. F. Abbott

Department of Neuroscience, Department of Physiology and Cellular Biophysics, Columbia University Medical Center, New York, New York 10032-2695, USA

(Received 20 January 2008; published 30 June 2008)

We evaluate the capacity and performance of a perceptron discriminator operating in a highly sparse regime where classic perceptron results do not apply. The perceptron is constructed to respond to a specified set of q stimuli, with only statistical information provided about other stimuli to which it is not supposed to respond. We compute the probability of both false-positive and false-negative errors and determine the capacity of the system for not responding to nonselected stimuli and for responding to selected stimuli in the presence of noise. If q is a sublinear function of N, the number of inputs to the perceptron, these capacities are exponential in N/q.

DOI: 10.1103/PhysRevLett.101.018101 PACS numbers: 87.19.ll, 87.18.Sn

Sparse coding is a useful and widespread strategy for representing complex data [1]. Biological systems often generate a high-dimensional representation of sensory data at the initial receptor level but modify this to a sparse representation at later processing stages. For example, in the olfactory system of insects, Kenyon cells show odorant selectivity to a much higher degree than do the projection neurons providing their input or the olfactory receptor neurons that generate the initial olfactory response [2,3]. Here we analyze, in a general context, the capacity and discrimination properties of a simple and surprisingly robust mechanism for generating sparse, highly selective responses from high-dimentional, nonsparse inputs.

A simple model of neural selectivity is the single-layer perceptron [4]. It is based on an idealized neuron that receives N inputs, characterized by an N-dimensional activity vector u, via N synapses represented by a weight vector w. The total input to the neuron is approximated as $w \cdot u$, and the neuron fires when $w \cdot u \geq \theta$ for a positive threshold θ.

We explore the capacity of this model neuron for generating sparse responses. Ideally, we would like the neuron to respond to a small number q of specified inputs corresponding to ''selected'' stimuli and to no others. All of the input vectors, whether selected or not, are chosen from the same probability distribution, which represents the distribution of natural stimuli. We compute the probability of false-positive responses to nonselected input vectors chosen randomly from this distribution, and we compute the false-negative probability that the q selected stimuli fail to generate a response upon repeated presentation due to noise associated with their input vectors. From these results, we determine the capacities of the model for generating correct responses in the situation of *sparse coding* defined as $q \sim N^{\beta}$, with $\beta < 1$.

It is important to appreciate that the situation that we are considering is different from the classic perceptron problem. In the classic case, we are provided with a set of M input vectors divided into two groups (each such division is called a *dichotomy*). One group contains q inputs that are supposed to generate responses, and the other $M - q$ inputs that should not. If $M < 2N$, the probability that a weight vector can be found for almost all dichotomies approaches 1 in the limit $N \to \infty$ [5–7]. The problem that we consider differs in two important ways. First, we do not assume that all of the inputs to be discriminated are known prior to specifying the weight vector. Instead, we construct the optimal weight vector solely on the basis of knowledge of the q selected stimuli, with some knowledge of general statistical properties of other stimuli but without a complete list of what they are. We feel that this is more relevant to biological applications than the problem solved in the traditional approach. For example, an animal cannot be expected to have knowledge of all of the odorants that it will encounter during its lifetime prior to setting up its olfactory selectivity. The second difference is that we consider extremely sparse representations for which $q \sim N^{\beta}$, with $\beta < 1$. In the traditional approach, all of the dichotomies of M points are counted equally, and, when $M \sim N \to \infty$, almost all of them are concentrated around an almost even split into two groups of $q \simeq M/2$ elements each. The classical results thus apply only when q is of order N.

In the general framework for discrimination that we consider, each stimulus is associated with an N-dimensional input vector with components drawn independently from a Gaussian distribution with zero mean and unit variance. This applies to both selected and nonselected stimuli, that is, to stimuli that should evoke a response and to those that should not. We first construct a perceptron that guarantees that $w \cdot u \geq \theta$ for all q of the selected stimuli, designated by $u = \xi^1, \xi^2, \ldots, \xi^q$ for $q < N$, and that minimizes the probability that this condition is satisfied for any other stimulus. For a perceptron designed to respond to a set of selected stimuli ξ, we define the false-positive probability $P^{(r)}_{\mathrm{fp}|\xi}$ as the probability that at least one of r nonselected, randomly chosen stimuli generates a response. Although the perceptron is guaranteed to respond to the selected inputs ξ, it may fail to respond to repeated presentations of the same stimulus if the corresponding input

is corrupted by noise. We introduce noise by assuming that, when one of the q selected stimuli with original input ξ^i is presented again, it generates an input vector $\xi^i + \eta$, where the components of the noise vector η are drawn independently from a Gaussian distribution with zero mean and variance σ^2. The false-negative response probability $P^{(s)}_{\mathrm{fn}|\xi}$ is the probability that any one of the noisy selected input vectors fails to evoke a response on any one of s trials.

The false-positive and false-negative response probabilities discussed in the previous paragraph refer to a particular set of selected input vectors $\xi^1, \ldots, \xi^q$. What we are really interested in, however, are the false-positive and false-negative probabilities when the ξ^i are randomly selected. Because the selected and nonselected stimuli are drawn independently, these probabilities are simply the expectation values $\langle P^{(r)}_{\mathrm{fp}|\xi}\rangle$ and $\langle P^{(s)}_{\mathrm{fn}|\xi}\rangle$ over the distribution of possible selected input vectors $\xi^1, \ldots, \xi^q$. Corresponding to these two response probabilities, we define two capacities for the perceptron: (i) the largest value of r such that $\langle P^{(r)}_{\mathrm{fp}|\xi}\rangle \to 0$ and (ii) the largest value of s such that $\langle P^{(s)}_{\mathrm{fn}|\xi}\rangle \to 0$, where the rightward arrows refer to the limit $N \to \infty$. We show that in the sparse case, which we define as $q \sim N^\beta$, with $\beta < 1$, these capacities grow exponentially in $N^{1-\beta}$.

The basic response condition for the model neuron that we consider can be given a simple geometric interpretation by rewriting the condition $w \cdot u \geq \theta$ as $(u - w_0) \cdot w_0 \geq 0$, with $w_0 = \theta|w|^{-2}w$. This inequality implies that the point specified by any response-generating input vector u is separated from the origin by a hyperplane that passes through the point w_0 and is tangent to the sphere of radius $d \stackrel{\mathrm{def}}{=} |w_0|$ around the origin. We wish to construct a hyperplane (or, equivalently, a vector w_0) that separates, as much as possible, the q input vectors corresponding to selected stimuli from all others.

To begin, we construct a hyperplane that passes through all of the selected stimuli $\xi^1, \xi^2, \ldots, \xi^q$. For $q < N$, this is always possible but it does not completely specify the hyperplane. However, we can determine the unique hyperplane that separates random nonselected stimuli from the selected stimuli with maximum probability. Because the components of random input vectors are normally distributed with zero mean and unit variance, their dot products with a fixed vector w_0 are normally distributed with zero mean and variance $d^2 = |w_0|^2$. The probability that any randomly selected input vector satisfies the threshold condition (which can be written as $w_0 \cdot u \geq d^2$) and generates a response is thus given by $\frac{1}{2}\mathrm{erfc}\sqrt{d^2/2}$. This is a decreasing function of d, so, to minimize the probability of a false-positive error, we should choose the plane furthest from the origin—i.e., with the longest possible vector w_0. Finding the maximum of $|w_0|^2$ with the constraints $(\xi^i - w_0) \cdot w_0 = 0$ for $i = 1, \ldots, q$ defines a unique hyperplane given by

$$w_0 = \left(\sum_{i,j=1}^{q} (C^{-1})_{ij}\right)^{-1} \sum_{i,j=1}^{q} (C^{-1})_{ij}\xi^j, \qquad (1)$$

where C is the matrix with elements $C_{ij} = \xi^i \cdot \xi^j$. Equation (1) defines a hyperplane that results in no false negatives (in the absence of noise) and minimizes the probability of a false-positive error in discriminating the selected stimuli.

Note that, if the ξ^i are linearly dependent, the matrix C is not invertible, and the hyperplane furthest from the origin containing the ξ^i is not defined by this formula. Moreover, in this situation the hyperplane may pass through the origin and thus produce a very poor discriminator. This is analogous to the failure of the single-layer perceptron for linearly inseparable problems, such as XOR. However, the probability of choosing linearly dependent ξ^i is zero given the Gaussian statistics that we consider.

If the input vectors are noisy, the perceptron given by Eq. (1) can generate false negatives. If the noise in the selected stimuli is Gaussian, the false-negative response probability on a single trial with a selected stimulus for the perceptron defined by Eq. (1) is equal to $\frac{1}{2}$. However, we will show that the false-negative probability can be greatly reduced by choosing a parallel hyperplane slightly closer to the origin. This is done by shifting $\theta \mapsto (1 - \kappa)\theta$, or, equivalently, $w_0 \mapsto w_\kappa = (1 - \kappa)w_0$, for $0 \leq \kappa < 1$. We define our perceptron for sparse discrimination as the perceptron obtained by this shift, with κ a free parameter.

For the shifted hyperplane, the dot product of any of the selected input vectors with w_κ is $w_\kappa \cdot \xi^i = (1 - \kappa)d^2$, where d is the distance of the unshifted ($\kappa = 0$) hyperplane from the origin. If noise is added to a selected vector, so that $u = \xi^i + \eta$, where η is Gaussian-distributed with zero mean and variance σ^2, the dot product $w_\kappa \cdot u = (1 - \kappa)d^2 + (1 - \kappa)w_0 \cdot \eta$ is Gaussian-distributed with mean $(1 - \kappa)d^2$ and variance $(1 - \kappa)^2 d^2\sigma^2$. To make the neuron fire, this dot product must be greater than $(1 - \kappa)^2 d^2$, which means that the false-negative response probability is $P^{(1)}_{\mathrm{fn}|\xi} = \frac{1}{2}\mathrm{erfc}\sqrt{\kappa^2 d^2/(2\sigma^2)} \stackrel{\mathrm{def}}{=} p_{\mathrm{fn}}(d^2)$. Similarly, the probability of a false-positive response to a randomly selected stimulus for the shifted plane is $P^{(1)}_{\mathrm{fp}|\xi} = \frac{1}{2}\mathrm{erfc}\sqrt{(1 - \kappa)^2 d^2/2} \stackrel{\mathrm{def}}{=} p_{\mathrm{fp}}(d^2)$.

For independently drawn stimuli and noise, the false-positive response probability for r trials and the false-negative probability for s trials can be computed in terms of the single-trial probabilities as $P^{(r)}_{\mathrm{fp}|\xi} = 1 - [1 - p_{\mathrm{fp}}(d^2)]^r$ and $P^{(s)}_{\mathrm{fn}|\xi} = 1 - [1 - p_{\mathrm{fn}}(d^2)]^s$, respectively. These equations apply to specific choices of selected input vectors, but the dependence is solely through the value of d, the maximum distance from the origin to a hyperplane passing through these points. To determine the probabilities $\langle P^{(r)}_{\mathrm{fp}|\xi}\rangle$ and $\langle P^{(s)}_{\mathrm{fn}|\xi}\rangle$, we must average over the distribution of d^2 obtained from random selected stimuli $\xi^1, \ldots, \xi^q$. From Eq. (1), we find that $d^2 = |w_0|^2$ is given by $d^2 = [\sum_{i,j=1}^{q}(C^{-1})_{ij}]^{-1}$. The distribution of the matrix C with elements $C_{ij} = \xi^i \cdot \xi^j$ for Gaussian-distributed vectors ξ^i is known as the Wishart distribution. Theorem 3.4.7

of Ref. [8] (p. 72) states that, for $q < N$ and a Wishart-distributed matrix C, the random variable $|a|^2/(a^T C^{-1} a)$, for any vector a, follows a χ^2 distribution with $N - q + 1$ degrees of freedom. If we set $a = (1, \ldots, 1)^T$, this random variable becomes qd^2, so qd^2 is distributed as χ^2_{N-q+1}, and it follows that the mean and the variance of d^2 are $\langle d^2 \rangle = (N - q + 1)/q$ and $\mathrm{Var}(d^2) = 2(N - q + 1)/q^2$.

To compute the full false-positive and false-negative probabilities, we must average the expressions for $P^{(r)}_{\mathrm{fp}|\xi}$ and $P^{(s)}_{\mathrm{fn}|\xi}$ over values of d^2 that are distributed as χ^2 with the appropriate number of degrees of freedom. These integrals cannot be computed explicitly in terms of known functions; moreover, their asymptotic behavior (as $N \to \infty$) is difficult to evaluate unless an additional condition is satisfied. To make these integrals tractable, we restrict our attention to cases where self-averaging applies, meaning that we can approximate

$$\begin{aligned} \langle P^{(r)}_{\mathrm{fp}|\xi} \rangle &\approx 1 - [1 - p_{\mathrm{fp}}(\langle d^2 \rangle)]^r, \\ \langle P^{(s)}_{\mathrm{fn}|\xi} \rangle &\approx 1 - [1 - p_{\mathrm{fn}}(\langle d^2 \rangle)]^s. \end{aligned} \tag{2}$$

We require that this approximation becomes exact in the limit $N \to \infty$.

For the approximations of Eq. (2) to be valid in the single-trial case (i.e., $r = s = 1$), the functions $p_{\mathrm{fp}}(d^2)$ and $p_{\mathrm{fp}}(d^2)$ must be relatively constant across the width of the peak of the χ^2 distribution for d^2. By assuming that N is large, $q < N$, and that σ and κ are independent of N, it can be shown, in the limit of large N, that the product of the derivative of either of these two functions evaluated at the point $\langle d^2 \rangle$ is much smaller than the width $\sqrt{\mathrm{Var}(d^2)}$ of the χ^2 distribution. Below, we prove the validity of (2) rigorously in the more general context of multiple trials.

Under the approximations (2), the average single-trial false-positive and false-negative response probabilities are given, in the limit of large N, by $\langle P^{(1)}_{\mathrm{fp}|\xi} \rangle \approx \frac{1}{2} \mathrm{erfc} \sqrt{(1 - \kappa)^2 (N - q)/(2q)}$ and $\langle P^{(1)}_{\mathrm{fn}|\xi} \rangle \approx \frac{1}{2} \mathrm{erfc} \sqrt{\kappa^2 (N - q)/(2q\sigma^2)}$. If $q \sim N$, this indicates performance that is independent of N. Because the probability of an error can only increase with multiple trials and we have defined the capacity of the discrimination perceptron in terms of a zero-error limit as $N \to \infty$, the capacity is zero in this case, and we do not consider it further. Much more impressive performance occurs if $q \sim N^\beta$ for $\beta < 1$. Then, for large N, by using the asymptotic expansion $\mathrm{erfc}(\sqrt{x}) \to \exp(-x)/\sqrt{\pi x}$ as $x \to \infty$, we find

$$\begin{aligned} \langle P^{(1)}_{\mathrm{fp}|\xi} \rangle &\approx p_{\mathrm{fp}}(\langle d^2 \rangle) \approx \sqrt{\frac{q}{2\pi(1 - \kappa)^2 N}} \exp\left(- \frac{(1 - \kappa)^2 N}{2q} \right), \\ \langle P^{(1)}_{\mathrm{fn}|\xi} \rangle &\approx p_{\mathrm{fn}}(\langle d^2 \rangle) \approx \sqrt{\frac{q\sigma^2}{2\pi\kappa^2 N}} \exp\left(- \frac{\kappa^2 N}{2q\sigma^2} \right). \end{aligned} \tag{3}$$

These probabilities go to zero exponentially in $N^{1-\beta}$.

If the approximations of (2) are valid, N is large, and $q \sim N^\beta$, with $\beta < 1$, the multitrial false-positive and false-negative probabilities are obtained by substituting the results of (3) into Eqs. (2). It is then straightforward to derive the maximum values of r and s that ensure that these probabilities go to zero as $N \to \infty$ (using the observation that if functions $g(N)$ and $g(N)t(N)$ approach zero in the limit $N \to \infty$, then $[1 - g(N)]^{t(N)} \to 1$). The following conditions are sufficient:

$$\ln(r) < \frac{(1 - \kappa^2)N}{2q} - N^\delta \quad \text{and} \quad \ln(s) < \frac{\kappa^2 N}{2q\sigma^2} - N^\delta \tag{4}$$

for any small constant $\delta > 0$. These are not the sharpest possible upper bounds, but, as we now show, they are the best limits that we can derive before the approximation of Eqs. (2) breaks down.

We now prove that the approximations (2) become exact in the limit of large N as long as the bounds of Eqs. (4) are satisfied with $\delta > \max(3/4 - \beta, 0)$. The magnitude of the error in the approximations of Eqs. (2) can be written as $|\Phi(\langle d^2 \rangle) - \langle \Phi(d^2) \rangle|$, where $\Phi(d^2) \stackrel{\mathrm{def}}{=} (1 - \frac{1}{2} \mathrm{erfc} \sqrt{b d^2/2})^t$, with $b = (1 - \kappa)^2$ and $t = r$ for the first equation in (2) and $b = \kappa^2/\sigma^2$ and $t = s$ for the second equation. To derive the maximum values of r and s for which these approximations are valid, we must determine the largest value of t for which this error approaches zero as $N \to \infty$. The function Φ satisfies $0 < \Phi(x) < 1$ for all x and also $\Phi'(x) \le \sqrt{b/(8\pi x)} \exp[-bx/2 + \ln(t)] \stackrel{\mathrm{def}}{=} \Psi(x)$. The function $\Psi(x)$, which we use as a bound on the derivative of $\Phi(x)$, is a monotonically decreasing function of x. Any differentiable function, such as Φ, that satisfies the conditions $0 \le \Phi(x) \le 1$ and $|\Phi'(x)| \le \Psi(x)$, for some continuous monotonically decreasing function $\Psi(x)$, obeys the bound $|\Phi(\bar{X}) - \langle \Phi(X) \rangle| \le \mu^{-2} \mathrm{Var}(X) + \mu \Psi(\bar{X} - \mu)$ for any $\mu > 0$ and any random variable X with mean $\bar{X} = \langle X \rangle$ and variance $\mathrm{Var}(X)$. We derive this bound via the sequence of inequalities

$$\begin{aligned} |\Phi(\bar{X}) - \langle \Phi(X) \rangle| &= |\langle [H(|X - \bar{X}| - \mu) \\ &\quad + H(\mu - |X - \bar{X}|)][\Phi(\bar{X}) - \Phi(X)] \rangle| \\ &\le \langle |\Phi(\bar{X}) - \Phi(X)| H(|X - \bar{X}| - \mu) \rangle \\ &\quad + \langle |\Phi(\bar{X}) - \Phi(X)| H(\mu - |X - \bar{X}|) \rangle \\ &\le \langle H(|X - \bar{X}| - \mu) \rangle + \langle |\Phi(\bar{X}) \\ &\quad - \Phi(X)| H(\mu - |X - \bar{X}|) \rangle \\ &\le P(|X - \bar{X}| \ge \mu) \\ &\quad + \langle |(X - \bar{X}) \Phi'[y(X)]| H(\mu - |X - \bar{X}|) \rangle \\ &\le \mu^{-2} \mathrm{Var}(X) + \mu \max_{|y - \bar{X}| \le \mu} |\Phi'(y)| \\ &\le \mu^{-2} \mathrm{Var}(X) + \mu \Psi(\bar{X} - \mu). \end{aligned}$$

The initial equality above follows from the fact that $H(|X - \bar{X}| - \mu) + H(\mu - |X - \bar{X}|) = 1$, where $H(x)$ is the Heaviside function. The first inequality is based on the observation that the absolute value of an average is

less than or equal to the average of the absolute value. The second inequality relies on replacing $|\Phi(\bar{X}) - \Phi(X)|$ by 1, because it is less than or equal to 1. The first term after the third inequality of this sequence follows from the definition of the expectation value, and the second term is a result of the intermediate value theorem $\Phi(X) - \Phi(\bar{X}) = (X - \bar{X})\Phi'[y(X)]$, where $y(X)$ satisfies $|y(X) - \bar{X}| \leq |X - \bar{X}|$. Finally, for the final two inequalities, we recall that $P(|X - \bar{X}| \geq \mu) \leq \mu^{-2}\mathrm{Var}(X)$ for any random variable X and that $\Psi(x)$ is a decreasing function.

In our case, $X = d^2$. Recalling that, for large N, $\langle d^2 \rangle \simeq N/q - 1$ and $\mathrm{Var}(d^2) \simeq 2q^{-2}(N - q)$, choosing $\mu = N^{3/4}/q$ and keeping the leading terms in N gives

$$|\Phi(\langle d^2 \rangle) - \langle \Phi(d^2) \rangle| \leq \frac{2}{\sqrt{N}} + \sqrt{\frac{bN^{1/2}}{8\pi q(1 - q/N - N^{-1/4})}} \times \exp\left(-\frac{b(N - q - N^{3/4})}{2q} + \ln(t)\right). \tag{5}$$

If $q \sim N$ and $t \sim 1$, the right side of this inequality goes to zero in the limit of large N, thus proving the validity of Eqs. (2) in the single-trial case. If $q \sim N^{\beta}$, with $\beta < 1$, the right side of this equation goes to 0 as $N \to \infty$ provided that, for some $\delta > \max(3/4 - \beta, 0)$, $\ln(t) < bN/(2q) - N^{\delta}$. Substituting the appropriate values of b generates the bounds (4). These are therefore our capacity bounds. Recalling that $N/q \sim N^{1-\beta}$, we find that both capacities are exponential in $N^{1-\beta}$.

Because we are used to thinking, working, and living in spaces of small dimension, it seems remarkable that an exponentially large number of random points can be separated from a set of selected points, even when both sets are drawn from the same distribution (for a similarly surprising result in signal transmission, see [9]). Moreover, the discrimination perceptron that we introduced is very robust to noise even if the noise is drawn from the same distribution as the selected stimuli (i.e., $\sigma = 1$). The key to our results is that stimuli are represented by high-dimensional input vectors. This is true of most sensory systems, including olfaction, which can involve up to thousands of different receptor types. If we assume $N = 400$ (roughly the value for olfactory receptors types in humans), $\beta = \kappa = 0.5$, and $\sigma^2 = 1$, this would allow for single-trial false-positive and false-negative probabilities of 1.5%.

The results that we have reported were derived by assuming that input vectors were drawn from a spherically symmetric Gaussian distribution. They can easily be extended to correlated Gaussian distributions simply by constructing the weight vector to compensate for the structure of the correlation matrix, a process called decorrelation. The extension to non-Gaussian distributions would, of course, involve different capacity estimates, but similar general properties should hold as long as those distributions are unimodal.

The calculation that we have performed involves an average over sets of input vectors. This is similar to averages over interaction matrices in spin-glass calculations or over memory patterns in analyses of associative memories [10]. In our case, the validity of the approximations in Eqs. (2) amounts to imposing strong self-averaging, which makes the replica-symmetry-breaking methods used in these other computations unnecessary.

Finally, the perceptron with a nonzero threshold might seem prone to errors if the "intensity" of a stimulus is increased. In other words, one might worry that a nonselected stimulus u satisfying $w \cdot u < \theta$ for positive θ might get pushed above the threshold if u is multiplied by a constant λ greater than 1. Actually, this constant has to be surprisingly large before selectivity fails. The probability of misclassifying a nonselected stimulus drawn from the stimulus distribution that we have been using and then multiplied by a constant λ can be computed in a manner similar to Eq. (3) yielding

$$\langle P^{(1)}_{\mathrm{fp}|\xi} \rangle \approx \sqrt{\frac{q\lambda^2}{2\pi(1 - \kappa)^2 N}} \exp\left(-\frac{(1 - \kappa)^2 N}{2q\lambda^2}\right).$$

This indicates that, for a stimulus to be misclassified with a significant probability, it needs to be multiplied by a λ of order $N^{(1-\beta)/2}$. This suggests that, for large N, a form of intensity-invariant selectivity can be realized by the sparse perceptron that we have considered.

We thank Bill Bialek and Haim Sompolinsky for helpful comments. This research was supported by the Swartz Foundation and by the NIH.

[1] B. A. Olshausen, in *Probabilistic Models of the Brain: Perception and Neural Function*, edited by R. P. N. Rao, B. A. Olshausen, and M. S. Lewicki (MIT Press, Cambridge, MA, 2002), pp. 257–272.
[2] J. Perez-Orive, O. Mazor, G. C. Turner, S. Cassenaer, R. I. Wilson, and G. Laurent, Science **297**, 359 (2002).
[3] R. I. Wilson and Z. F. Mainen, Annu. Rev. Neurosci. **29**, 163 (2006).
[4] F. Rosenblatt, Psychol. Rev. **65**, 386 (1958).
[5] T. M. Cover, IEEE Trans. Electron. Comput. **ec-14**, 326 (1965).
[6] S. S. Venkatesh, in *Proceedings of Neural Networks for Computing*, AIP Conf. Proc. No. 151 (AIP, New York, 1986), pp. 440–445.
[7] E. Gardner, J. Phys. A **21**, 257 (1988).
[8] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis* (Academic Press, London, 1979).
[9] C. E. Shannon, Proc. IRE **37**, 10 (1949).
[10] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).