Stability of the replica-symmetric solution for the information conveyed by a neural network

Simon Schultz[1] and Alessandro Treves[2]
[1]*Department of Experimental Psychology, South Parks Road, University of Oxford, Oxford OX1 3UD, United Kingdom*
[2]*Programme in Neuroscience, International School for Advanced Studies, Via Beirut 2-4, 34013 Trieste, Italy*
(Received 26 September 1997)

The information that a pattern of firing in the output layer of a feedforward network of threshold-linear neurons conveys about the network's inputs is considered. A replica-symmetric solution is found to be stable for all but small amounts of noise. The region of instability depends on the contribution of the threshold and the sparseness: for distributed pattern distributions, the unstable region extends to higher noise variances than for very sparse distributions, for which it is almost nonexistent. [S1063-651X(98)02003-0]

PACS number(s): 87.10.+e, 84.35.+i, 89.70.+c

I. INTRODUCTION

Advances in techniques for the formal analysis of neural networks [1–5] offer insight into the behavior of models of biological interest. Of particular interest are methods which allow the calculation of the information that can be conveyed by a given neural structure, as these offer both useful intuitions and the prospect of conducting pertinent experiments [6]. The replica trick [7] has been used to achieve this in the case of binary units [5] and threshold-linear units [8,9], by appealing to an assumption of replica symmetry. In the case of binary units with continuous inputs, the validity of the replica-symmetric ansatz is justified by the duality with the Gardner calculation of the storage capacity for continuous couplings [2,5,10]. We now analyze the stability of the replica-symmetric solution for mutual information in a network of threshold-linear units.

The model describes a feedforward network of threshold-linear units with partially diluted connectivity. This is a simpler version of the calculation described in [8,9]. In the calculation considered here, there is only one mode of operation (which we might call ''transmission''), as opposed to the division into storage and recall modes in that calculation. There are N cells in the input layer, and M (proportional to N) in the output layer. The limit of interest is $N\rightarrow\infty$.

$\{\eta_i\}$ are the firing rates of each cell i in the output layer. The probability density of finding a given firing pattern is taken to be

$$P(\{\eta_i\})\prod_i d\eta_i=\prod_i P_\eta(\eta_i)d\eta_i. \tag{1}$$

Each input cell is thus assumed to code independent information.

$\{\xi_j\}$ are the firing rates produced in each cell in the output layer. They are determined by the matrix multiplication of the pattern $\{\eta_i\}$ with the synaptic weights J_{ij}, followed by Gaussian distortion, thresholding, and rectification.

$$\xi_j=\left[\xi_0+\sum_i c_{ij}J_{ij}\eta_i+\epsilon_j\right]^+=[\tilde{\xi}_j]^+, \tag{2}$$

$$\langle(\epsilon_j)^2\rangle=\sigma_\epsilon^2. \tag{3}$$

Each output cell receives C_j (which we will take to be of the order of 10^4) connections from input layer cells:

$$c_{ij}\in\{0,1\},\quad \langle c_{ij}\rangle N=C_j\quad (C\equiv\langle C_j\rangle). \tag{4}$$

The mean value across all patterns of each synaptic weight is taken to be equal across synapses, and is therefore taken into the threshold term. The synaptic weights J_{ij} are thus of zero mean, and variance σ_J^2 (all that affects the calculation is the first two moments of their distribution),

$$\langle(J_{ij})^2\rangle=\sigma_J^2. \tag{5}$$

The average of the mutual information

$$I(\{\eta_i\},\{\xi_j\})=\int\prod_i d\eta_i\int\prod_j d\xi_j P(\{\eta_i\},\{\xi_j\})\ln\frac{P(\{\eta_i\},\{\xi_j\})}{P(\{\eta_i\})P(\{\xi_j\})} \tag{6}$$

over the quenched variables c_{ij}, J_{ij} is written using the replica trick as

$$\langle I(\eta,\xi)\rangle_{c,J}=\lim_{n\rightarrow 0}\frac{1}{n}\left\langle\int d\eta\, d\xi\, P(\eta,\xi)\left\{\left[\frac{P(\eta,\xi)}{P(\eta)}\right]^n-[P(\xi)]^n\right\}\right\rangle_{c,J}. \tag{7}$$

The calculation is valid only for nonzero noise variance, and it will be seen that the only region in which the solution is not well behaved is that of very low noise variance.

II. CALCULATION OF MUTUAL INFORMATION

First, introducing replica indices $\alpha=1,\ldots,n+1$, and breaking the integral over ξ into subthreshold and suprathreshold components, we observe that

$$\langle I(\eta,\xi)\rangle_{c,J}=\lim_{n\to 0}\frac{1}{n}\left\{\int d\eta\left[\frac{1}{P(\eta)}\right]^n\prod_\alpha\left(\int_{-\infty}^{0}d\tilde{\xi}^\alpha\langle P(\eta^\alpha,\tilde{\xi}^\alpha)|_{\eta^\alpha=\eta}\rangle_{c,J}\right)\right.$$
$$+\int d\eta\left[\frac{1}{P(\eta)}\right]^n\int_0^\infty d\tilde{\xi}\prod_\alpha\langle P(\eta^\alpha,\tilde{\xi}^\alpha)|_{\eta^\alpha=\eta,\tilde{\xi}^\alpha=\tilde{\xi}}\rangle_{c,J}-\prod_\alpha\left(\int_{-\infty}^{0}d\tilde{\xi}^\alpha\int d\eta^\alpha\langle P(\eta^\alpha,\tilde{\xi}^\alpha)\rangle_{c,J}\right)$$
$$\left.-\int_0^\infty d\tilde{\xi}\prod_\alpha\left(\int d\eta^\alpha\langle P(\eta^\alpha,\tilde{\xi}^\alpha)|_{\tilde{\xi}^\alpha=\tilde{\xi}}\rangle_{c,J}\right)\right\}. \tag{8}$$

This allows us to treat both terms of Eq. (7) in the same manner. To obtain the probability density $\langle P(\eta^\alpha,\tilde{\xi}^\alpha)\rangle$, we use Dirac δ functions to implement the constraints defined by Eq. (2):

$$\langle P(\eta^\alpha,\tilde{\xi}^\alpha)\rangle_{c,J}=\left\langle\int\left[\prod_j D\left(\frac{\epsilon_j^\alpha}{\sigma_\epsilon}\right)\right]\left[\prod_{ij}D\left(\frac{J_{ij}}{\sigma_J}\right)\right]\prod_j\delta\left[\tilde{\xi}_j^\alpha-\xi_0-\sum_i c_{ij}J_{ij}\eta_i^\alpha-\epsilon_j^\alpha\right]P(\{\eta_i^\alpha\})^{n+1}\right\rangle_c, \tag{9}$$

where

$$Du=\frac{du}{\sqrt{2\pi}}e^{-u^2/2}. \tag{10}$$

Using the integral form of the Dirac δ function introduces a Lagrange multiplier x_j^α. The integrals over the noise and interaction distributions are performed, and the quenched average over the connections performed in the thermodynamic limit, so that

$$\langle P(\eta,\tilde{\xi})^{n+1}\rangle=\int\left(\prod_{j,\alpha}\frac{dx_j^\alpha}{2\pi}\right)\exp\left\{i\sum_{j,\alpha}x_j^\alpha(\tilde{\xi}_j^\alpha-\xi_0)-\frac{1}{2}\sum_{j,\alpha,\beta}x_j^\alpha x_j^\beta\left[\sigma_\epsilon^2\delta_{\alpha\beta}+\frac{\sigma_J^2C}{N}\sum_i\eta_i^\alpha\eta_i^\beta\right]P(\{\eta_i^\alpha\})^{n+1}\right\}, \tag{11}$$

where $\delta_{\alpha\beta}$ is the Kronecker delta. A Lagrange multiplier

$$z^{\alpha\beta}=\frac{1}{N}\sum_i\eta_i^\alpha\eta_i^\beta \tag{12}$$

is introduced using the integral form of the Dirac δ function via an auxiliary variable $\tilde{z}^{\alpha\beta}$. We then obtain

$$\langle P(\eta,\tilde{\xi})^{n+1}\rangle=\int\left(\prod_\alpha\frac{dz^\alpha d\tilde{z}^\alpha}{2\pi/N}\right)\left(\prod_{(\alpha\beta)}\frac{dz^{\alpha\beta}d\tilde{z}^{\alpha\beta}}{2\pi/N}\right)\exp\left\{iN\sum_\alpha z^\alpha\tilde{z}^\alpha+iN\sum_{(\alpha\beta)}z^{\alpha\beta}\tilde{z}^{\alpha\beta}-\sum_\alpha\tilde{z}^\alpha\sum_i(\eta_i^\alpha)^2-\sum_{(\alpha\beta)}\tilde{z}^{\alpha\beta}\sum_i\eta_i^\alpha\eta_i^\beta\right.$$
$$\left.-\frac{1}{2}\sum_{j\alpha\beta}(\tilde{\xi}_j^\alpha-\xi_0)E_{\alpha\beta}(\tilde{\xi}_j^\beta-\xi_0)-\frac{1}{2}\operatorname{Tr}\ln\mathbf{M}\right\}(2\pi)^{-(n+1)/2}P(\{\eta_i^\alpha\})^{n+1}, \tag{13}$$

where $\mathbf{M}=\sigma_\epsilon^2\mathbf{I}+\sigma_J^2C\mathbf{Z}$ and $\mathbf{E}=\mathbf{M}^{-1}$. $\mathbf{Z}$ is the matrix with elements $z^{\alpha\beta}$, and $(\alpha\beta)$ is the pair $\alpha\beta$, $\alpha\neq\beta$.

Thus

$$\langle I\rangle=\lim_{n\to0}\left\{\int\left(\prod_\alpha\frac{dz^\alpha d\tilde{z}^\alpha}{2\pi/N}\right)\exp\left[iN\sum_\alpha z^\alpha\tilde{z}^\alpha-N\mathcal{H}_A(\tilde{z}^\alpha)\right.\right.$$
$$\left.-M\mathcal{G}(z^\alpha,z^\alpha)\right]-\int\left(\prod_\alpha\frac{dz^\alpha d\tilde{z}^\alpha}{2\pi/N}\right)$$
$$\times\left(\prod_{(\alpha\beta)}\frac{dz^{\alpha\beta}d\tilde{z}^{\alpha\beta}}{2\pi/N}\right)\exp\left[iN\sum_\alpha z^\alpha\tilde{z}^\alpha+iN\sum_{(\alpha\beta)}z^{\alpha\beta}\tilde{z}^{\alpha\beta}\right.$$
$$\left.\left.-N\mathcal{H}_B(\tilde{z}^\alpha,\tilde{z}^{\alpha\beta})-M\mathcal{G}(z^\alpha,z^{\alpha\beta}),\right. \tag{14}$$

where

$$e^{-\mathcal{H}_A(\tilde{z}^\alpha)}=\int_\eta d\eta\,P(\eta)\exp\left(-\sum_\alpha\tilde{z}^\alpha\eta^2\right), \tag{15}$$

$$e^{-\mathcal{H}_B(\tilde{z}^\alpha,\tilde{z}^{\alpha\beta})}=\int_\eta\left(\prod_\alpha d\eta^\alpha P(\eta^\alpha)\right)\exp\left(-\sum_\alpha\tilde{z}^\alpha(\eta^\alpha)^2\right.$$
$$\left.-\sum_{(\alpha\beta)}\tilde{z}^{\alpha\beta}\eta^\alpha\eta^\beta\right), \tag{16}$$

$$e^{-\mathcal{G}}(z^{\alpha},z^{\alpha\beta})=e^{-(1/2)\mathrm{Tr}\ln \mathbf{M}}\left\{\int_0^{\infty}\frac{d\tilde{\xi}}{\sqrt{2\pi}}\exp-\frac{1}{2}(\tilde{\xi}-\xi_0)^2\times\sum_{\alpha\beta}E_{\alpha\beta}+\int_{-\infty}^{0}\left(\prod_{\alpha}\frac{d\tilde{\xi}^{\alpha}}{\sqrt{2\pi}}\right)\times\exp\left[-\frac{1}{2}\sum_{\alpha\beta}(\tilde{\xi}^{\alpha}-\xi_0)E_{\alpha\beta}(\tilde{\xi}^{\beta}-\xi_0)\right]\right\}. \tag{17}$$

III. REPLICA-SYMMETRIC SOLUTION

The assumption of replica symmetry can be written

$$z_A^{\alpha}=z_A^{\alpha\beta}=z_{0A}(n),\quad i\tilde{z}_A^{\alpha}=\tilde{z}_{0A}(n),$$
$$z_B^{\alpha}=z_{0B}(n),\quad i\tilde{z}_B^{\alpha}=\tilde{z}_{0B}(n), \tag{18}$$
$$z_B^{\alpha\beta}=z_1(n),\quad i\tilde{z}_B^{\alpha\beta}=-\tilde{z}_1(n).$$

The saddle-point method is utilized in the thermodynamic limit, yielding the saddle-point equations

$$z_{0A}=\langle\eta^2\rangle_{\eta}, \tag{19a}$$
$$\tilde{z}_{0A}=0, \tag{19b}$$
$$z_{0B}=\langle\eta^2\rangle_{\eta}, \tag{19c}$$
$$\tilde{z}_{0B}=0, \tag{19d}$$
$$z_1=-\int_{-\infty}^{\infty}Ds\left\langle\left(\eta^2+\frac{s\eta}{\sqrt{\tilde{z}_1}}\right)\exp\left(-\frac{\tilde{z}_1}{2}\eta^2-s\sqrt{\tilde{z}_1}\eta\right)\right\rangle_{\eta}\times\ln\left\langle\exp\left(-\frac{\tilde{z}_1}{2}\eta^2-s\sqrt{\tilde{z}_1}\eta\right)\right\rangle_{\eta}, \tag{19e}$$
$$\tilde{z}_1=-\sigma_J^2Cr\left\{\frac{\xi_0}{(p_B+q_B)^{3/2}}\sigma\left(\frac{\xi_0}{\sqrt{p_B+q_B}}\right)-\frac{1}{p_B}\phi\left(\frac{\xi_0}{\sqrt{p_B+q_B}}\right)+\int_{-\infty}^{\infty}Dt\left[1+\ln\phi\left(\frac{-\xi_0-t\sqrt{q_B}}{\sqrt{p_B}}\right)\right]\times\sigma\left(\frac{-\xi_0-t\sqrt{q_B}}{\sqrt{p_B}}\right)p_B^{-3/2}\left(\xi_0+\frac{t(p_B+q_B)}{\sqrt{q_B}}\right)\right\}, \tag{19f}$$

and the expression for the information per input cell

$$\langle i\rangle=rG(p_A,q_A)+\frac{1}{2}z_1\tilde{z}_1-rG(p_B,q_B)-\int_{-\infty}^{\infty}Ds\left\langle\exp\left(-\frac{1}{2}\tilde{z}_1\eta^2-s\sqrt{\tilde{z}_1}\eta\right)\right\rangle_{\eta}\times\ln\left\langle\exp\left(-\frac{1}{2}\tilde{z}_1\eta^2-s\sqrt{\tilde{z}_1}\eta\right)\right\rangle_{\eta}, \tag{20}$$

where

$$G(p,q)=\frac{p\xi_0}{2(p+q)^{3/2}}\sigma\left(\frac{\xi_0}{\sqrt{p+q}}\right)-\frac{1}{2}(1+\ln p)\phi\left(\frac{\xi_0}{\sqrt{p+q}}\right)+\int_{-\infty}^{\infty}Dt\,\phi\left(\frac{-\xi_0-t\sqrt{q}}{\sqrt{p}}\right)\ln\phi\left(\frac{-\xi_0-t\sqrt{q}}{\sqrt{p}}\right) \tag{21}$$

and

$$\langle x(\eta)\rangle_{\eta}=\int_{\eta}d\eta\,P(\eta)x(\eta),$$
$$\phi(x)=\int_{-\infty}^{x}Ds,$$
$$\sigma(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \tag{22}$$
$$p_A=\sigma_{\epsilon}^2,\quad p_B=\sigma_{\epsilon}^2+\sigma_J^2C(z_{0B}-z_1),$$
$$q_A=\sigma_J^2Cz_{0A},\quad q_B=\sigma_J^2Cz_1.$$

We refer to $r=M/N$ as the anatomical divergence.

This expression must in general be evaluated numerically. However, considering some limiting cases can give us some insight into the behavior of the solution. In particular, the limit of linear processing can be obtained by taking $\xi_0\to+\infty$. In this limit, Eq. (19f) reduces to

$$\tilde{z}_1\to\frac{\sigma_J^2Cr}{p_B}. \tag{23}$$

The information per neuron obtained in the linear limit is

$$\langle i\rangle\to\frac{1}{2}r\ln\frac{p_B}{p_A}+\frac{1}{2}z_1\tilde{z}_1-\int_{-\infty}^{\infty}Ds\left\langle\exp\left(-\frac{1}{2}\tilde{z}_1\eta^2-s\sqrt{\tilde{z}_1}\eta\right)\right\rangle_{\eta}\times\ln\left\langle\exp\left(-\frac{1}{2}\tilde{z}_1\eta^2-s\sqrt{\tilde{z}_1}\eta\right)\right\rangle_{\eta}. \tag{24}$$

The information obtained in this limit is bounded by that which would be obtained from a simple Gaussian channel calculation, where we consider the channel

$$\xi_j^*=\sum_i c_{ij}J_{ij}\eta_i+\epsilon_j, \tag{25}$$

and perform the annealed and quenched averages to obtain the signal variance $\sigma_J^2C(\langle\eta^2\rangle_{\eta}-\langle\eta\rangle_{\eta}^2)$, and information per input cell

$$I_{\text{Gauss}}=\frac{r}{2}\ln\left[1+\frac{\sigma_J^2 C(\langle\eta^2\rangle_\eta-\langle\eta\rangle_\eta^2)}{\sigma_\epsilon^2}\right]. \tag{26}$$

The Gaussian channel information provides an upper limit corresponding to the optimal η distribution (for transmitting maximal information given a constraint on the signal power), and no dependence upon the same inputs of the output cells.

Within the linear limit, we can consider the special case of high noise variance (low signal to noise ratio). As $\sigma_\epsilon^2\to\infty$,

$$\tilde{z}_1\sim\frac{\sigma_J^2 Cr}{\sigma_\epsilon^2}, \tag{27}$$

and

$$z_1\simeq\langle n\rangle^2+O(\tilde{z}_1). \tag{28}$$

The information therefore falls to zero as

$$\langle i\rangle\sim\frac{\sigma_J^2 Cr(\langle\eta^2\rangle_\eta-\langle\eta\rangle_\eta^2)}{2\sigma_\epsilon^2}, \tag{29}$$

i.e., inversely with noise variance, as one would expect. We thus can see that for linear neurons with low signal to noise ratio, the transmitted information approaches the Gaussian channel limit. [It can also be shown (we have done so for the case of a Gaussian η distribution), that as $r\to 0$, the Gaussian channel bound is also reached.]

The numerical solution of the mutual information expression, as a function of the noise variance, is shown in Fig. 1, both for the case of linear units and for units with a threshold of $\xi_0=-0.4$, representing threshold-linear behavior. This is shown for a binary pattern distribution of sparseness a, where the sparseness of a distribution is a mean-invariant measure of spread and is defined in general as

$$a=\frac{\langle\eta\rangle_\eta^2}{\langle\eta^2\rangle_\eta}. \tag{30}$$

This measure is ''more sparse'' for smaller a, and reduces to the fraction of units ''on'' in the case of a binary distribution. The Gaussian channel bound appears on the same graphs for comparison.

The mutual information should be bounded by the pattern entropy as the noise variance becomes very small. As the noise variance decreases, the replica-symmetric solution approaches this bound in both the linear and threshold-linear cases. It can be seen, however, that for very small noise variances, the replica-symmetric solution changes direction and crosses this physical boundary. Inspection of Eq. (21) reveals divergence of the mutual information solution in the limit $\sigma_\epsilon^2\to 0$; this is in keeping with our intuition from the beginning that the calculation should not be valid in the deterministic limit. However, for such low noise variance the information has essentially saturated in any case. For threshold-linear neurons, the solution is also unstable to replica-symmetry-breaking fluctuations for relatively low noise variance, as will be discussed in the next section.

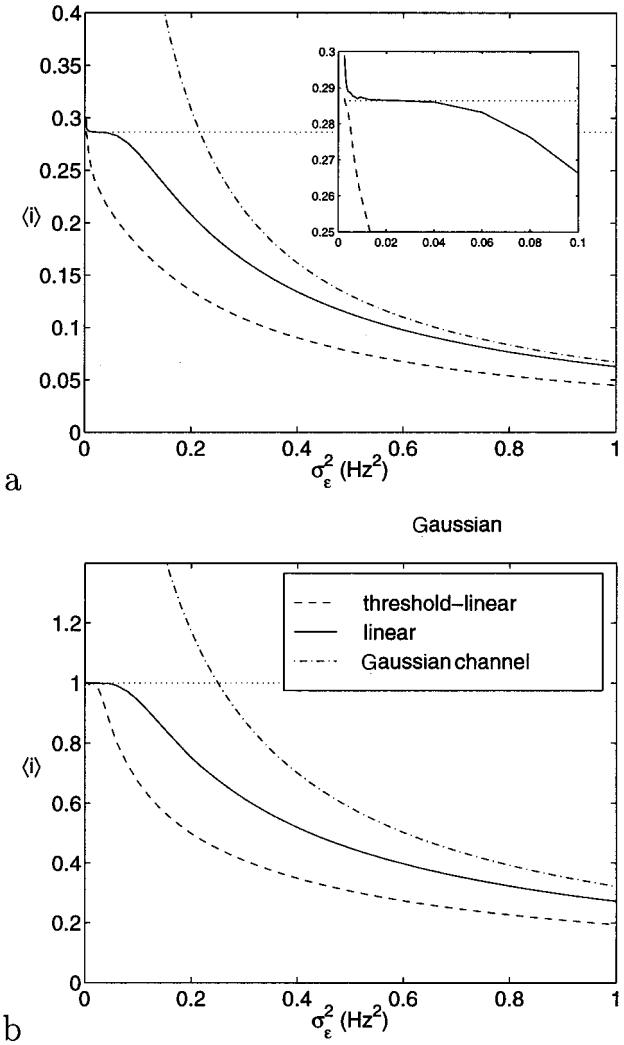

FIG. 1. Mutual information, measured in bits, as a function of noise variance. The dashed line is for a threshold $\xi_0=-0.4$, whereas the solid line is for the limit of linear neurons. The dot-dashed line indicates the simple Gaussian channel for comparison. The entropy of the input pattern distribution is indicated by the horizontal dotted line. (a) Input pattern distribution sparseness of 0.05. (b) Sparseness of 0.50.

IV. STABILITY OF THE REPLICA-SYMMETRIC SOLUTION

The stability of the replica-symmetric solution is analyzed after the style of de Almeida and Thouless [11]. For the solution for free energy this was addressed in the context of Hopfield-Little type autoassociative neural networks in [1], and for an autoassociator with threshold-linear units and for a threshold-linear variant of the Sherrington-Kirkpatrick model in [12]. For the solution for another quantity, the Gardner volume, this was addressed in [2] for Ising (± 1) neurons. In contrast, here we are determining the stability of the solution for mutual information in a network comprised of threshold-linear neurons, although the technique proceeds very similarly.

Fluctuations in the transverse (replica-symmetry breaking, RSB) and longitudinal (replica-symmetric, RS) directions are decoupled, and hence can be analyzed separately. Longitudi-

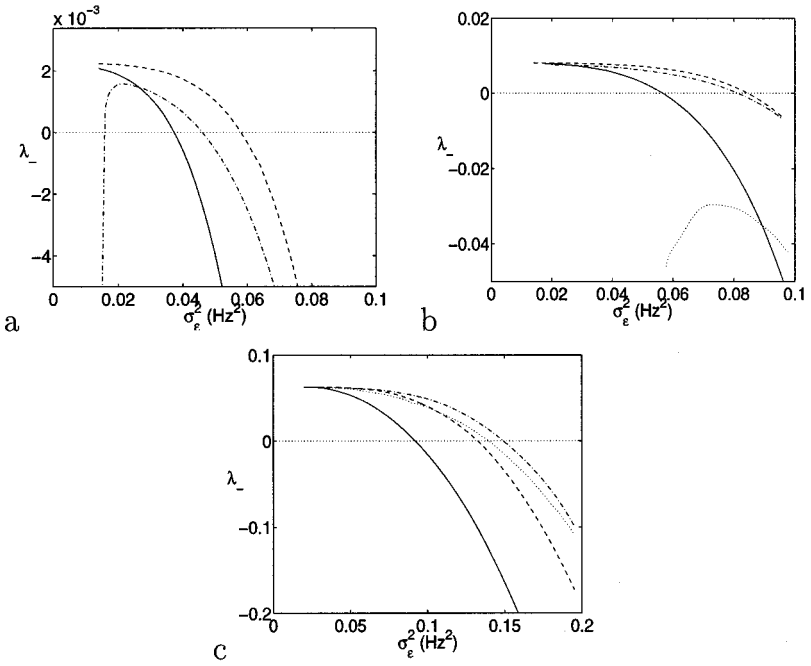

FIG. 2. The behavior of the replicon-mode eigenvalue λ_- as a function of noise variance. (a) Input sparseness $a=0.05$, (b) $a=0.10$, (c) $a=0.50$. In each of these graphs the solid line indicates the eigenvalue of threshold $\xi_0=-0.4$, the dashed curve $\xi_0=0.0$, the dot-dashed curve $\xi_0=0.4$, and the dotted curve $\xi_0=0.8$. The replica-symmetric solution is unstable in regions where these curves lie above the horizontal dotted line. In case (a), the $\xi_0=0.8$ line lies below the region examined in the graph.

nal fluctuations can be disregarded [11,13] if a unique saddle point is obtained, which appears to be the case. We will therefore concentrate upon transverse fluctuations.

We wish to consider small deviations in the saddle-point parameters about the replica-symmetric saddle point,

$$z^{\alpha\beta}=z_1+\delta z^{\alpha\beta},$$

$$\tilde{z}^{\alpha\beta}=\tilde{z}_1+\delta\tilde{z}^{\alpha\beta}. \tag{31}$$

Quadratic fluctuations in the function

$$\mathcal{B}(z^\alpha,\tilde{z}^\alpha,z^{\alpha\beta},\tilde{z}^{\alpha\beta})=iN\sum_\alpha z^\alpha\tilde{z}^\alpha+iN\sum_{(\alpha\beta)} z^{\alpha\beta}\tilde{z}^{\alpha\beta} - N\mathcal{H}_{\mathcal{B}}(\tilde{z}^\alpha,\tilde{z}^{\alpha\beta})-M\mathcal{G}(z^\alpha,z^{\alpha\beta}) \tag{32}$$

give us the stability matrix

$$\mathbf{\Gamma}=\begin{bmatrix} \dfrac{\partial^2\mathcal{B}}{\partial z^{\alpha\beta}\partial z^{\gamma\delta}} & \dfrac{\partial^2\mathcal{B}}{\partial z^{\alpha\beta}\partial(i\tilde{z}^{\gamma\delta})} \\ \dfrac{\partial^2\mathcal{B}}{\partial(i\tilde{z}^{\alpha\beta})\partial z^{\gamma\delta}} & \dfrac{\partial^2\mathcal{B}}{\partial(i\tilde{z}^{\alpha\beta})\partial(i\tilde{z}^{\gamma\delta})} \end{bmatrix} =\begin{bmatrix} A^{(\alpha\beta)(\gamma\delta)} & \delta_{(\alpha\beta)(\gamma\delta)} \\ \delta_{(\alpha\beta)(\gamma\delta)} & B^{(\alpha\beta)(\gamma\delta)} \end{bmatrix}, \tag{33}$$

where $\delta_{(\alpha\beta),(\gamma\delta)}=\delta_{\alpha\gamma}\delta_{\beta\delta}+\delta_{\alpha\delta}\delta_{\beta\gamma}$. In contrast to previous calculations based on quantities such as free energy, the expression for mutual information involves $n+1$ replicas. There are $n(n+1)/2$ independent variables $z^{\alpha\beta}$, and the same number of independent $\tilde{z}^{\alpha\beta}$. $\mathbf{\Gamma}$ is thus an $n(n+1)\times n(n+1)$ matrix.

The transverse eigenvalues of this matrix are given by the eigenvalues of the matrix

$$\begin{pmatrix} \lambda_A & 1 \\ 1 & \lambda_B \end{pmatrix}, \tag{34}$$

where λ_A and λ_B are the transverse eigenvalues of the submatrices $A^{(\alpha\beta)(\gamma\delta)}$ and $B^{(\alpha\beta)(\gamma\delta)}$, respectively. Calculation of these involves consideration of the symmetry properties of the submatrices, and is detailed in the Appendix. The eigenvalue equations reduce to

$$\lambda_A+c=\lambda,$$

$$1+c\lambda_B=c\lambda. \tag{35}$$

We thus have the two replicon-mode eigenvalues

$$\lambda_\pm=\tfrac{1}{2}(\lambda_A+\lambda_B)\pm\sqrt{\tfrac{1}{4}(\lambda_A-\lambda_B)^2+1}. \tag{36}$$

For stability, the product of the eigenvalues must be nonnegative. A further subtlety is introduced here. λ_+ can be seen to be >0 irrespective of σ_ϵ^2 or a. λ_-, on the other hand, changes sign, moving from negative to positive for smaller σ_ϵ^2. However, intuitively we expect, from the analogy of the noise with the ''temperature'' parameter in other models of neural networks [1] and physical systems [14] that if replica-symmetry breaking is to set in, it will do so at low noise variances. This is confirmed by the eminently sensible behavior of the mutual information curves of Fig. 1 at me-

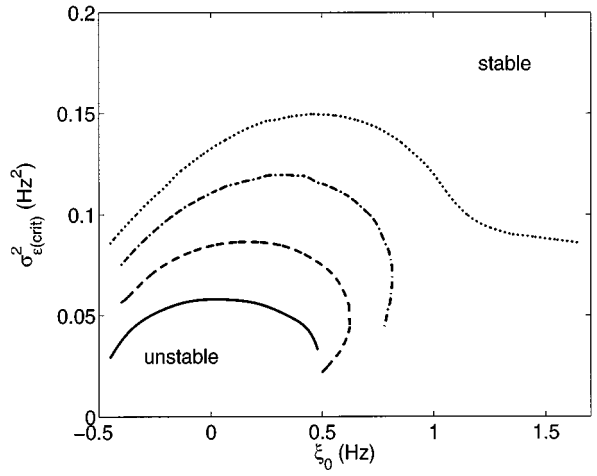

FIG. 3. A phase diagram showing the critical noise variance as a function of the threshold parameter ξ_0—the larger ξ_0 is, the more linear the regime. Solid curve, sparseness $a=0.05$; dashed curve, $a=0.10$; dot-dashed curve, $a=0.20$; dotted curve, $a=0.50$.

dium to high noise, but nonphysical behavior at very low noise values. It can be concluded that, as occurs in [1,12], a sign reversal has been introduced due to the integration contour, which must be corrected.

These equations have been numerically solved for λ_-. Figure 2 shows the behavior of λ_- for a range of sparseness and thresholds. Where the eigenvalue passes above the zero axis (dotted line), a phase of RS instability is indicated. Figure 2(a) is for the situation of quite sparse coding of the patterns. As the noise is reduced from the high noise region, in which the RS solution is stable, the eigenvalue changes sign, and an unstable region is entered. In the case of threshold $\xi_0=0.4$, which represents only a very small degree of thresholdlike behavior, the eigenvalue can be seen to curve back and change sign again at lower noise values still. Due to nonconvergence of numerical integration, it is not possible to examine extremely small noise values; therefore it is not clear from this diagram whether the eigenvalue also falls below zero again for the other curves plotted in this figure, or if it instead has a finite value at zero noise. However, any region of RS stability at noise variances this low would obviously be irrelevant for the same numerical reasons.

It is apparent from Figs. 2(b) and 2(c) that as the input distribution is made less sparse (a is increased), the critical amount of noise below which instability arises increases. This will be discussed again shortly. Another effect that can be seen in Figs. 2(a) and 2(b) is that, as the neurons are made more linear (ξ_0 is increased), the critical noise first rises, then falls. This becomes more clear after plotting a phase diagram of noise against ξ_0 (Fig. 3). For low a (sparse distributions),

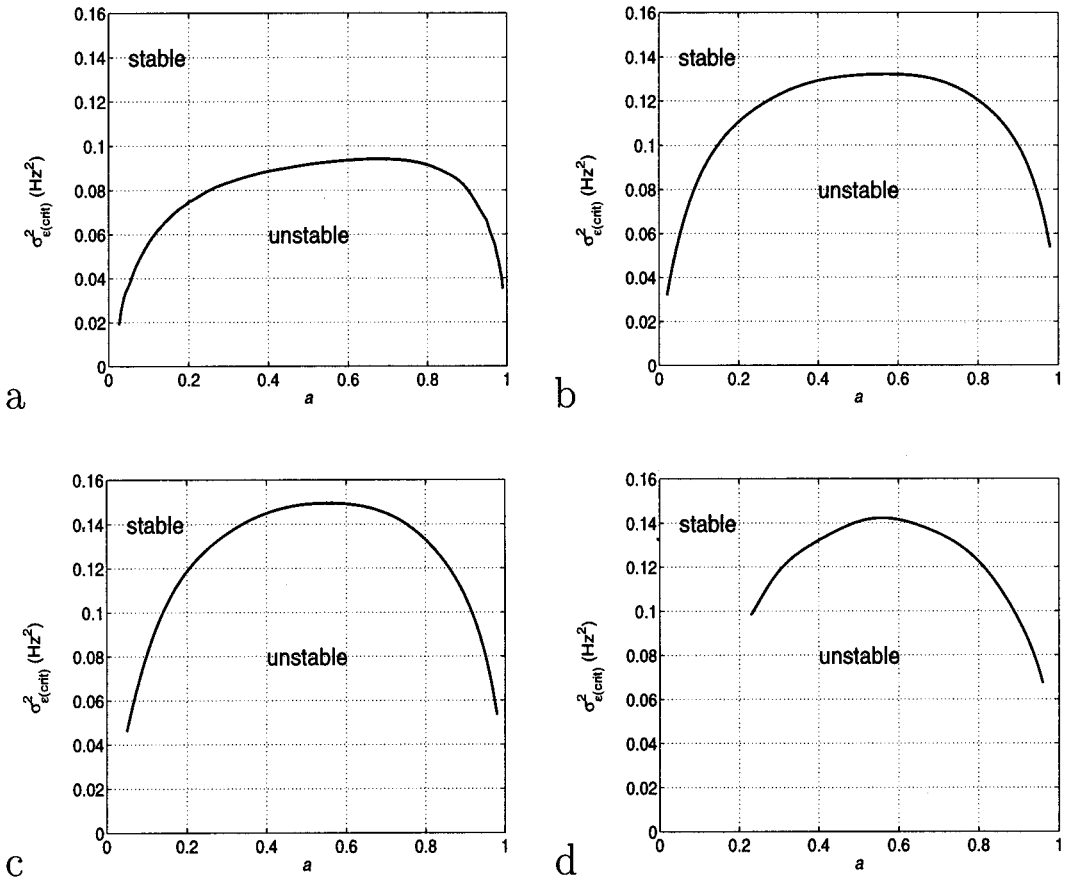

FIG. 4. The phase diagram for information transmission, for $r=2$ and $\sigma_J^2=1/C$. (a) Threshold $\xi_0=-0.4$. (b) Threshold $\xi_0=+0.0$. (c) Threshold $\xi_0=+0.4$. (d) Threshold $\xi_0=+0.8$.

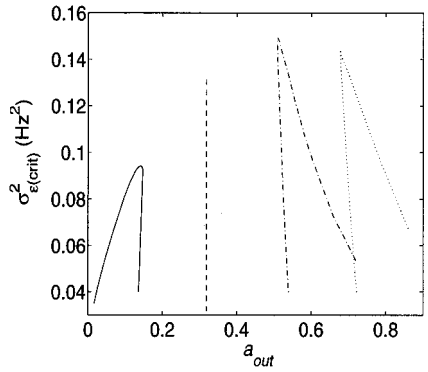

FIG. 5. The marginal noise variance as a functon of the sparseness of the *output* distribution. The solid line represents the curve for $\xi_0=-0.40$ [the same situation as Fig. 4(a), the dashed curve $\xi_0=0.0$, the dot-dashed curve $\xi_0=+0.40$, and the dotted curve $\xi_0=+0.80$. Note that for $\xi_0=0.0$ the output sparseness is fixed at $1/\pi$, as explained in the text, so this particular line is not informative about the relative region of instability.

the critical noise rises, falls, and then curves back around on itself—after the neurons become sufficiently linear, there is no more region of instability. As the pattern code becomes less sparse, at first the region of instability merely expands. When a reaches a certain value, however, the edge of the unstable region no longer curls in on itself, but extends outwards. At a sparseness of 0.5, for instance, the critical noise thus first rises with increasing linearity, taking longer to reach its peak than for more sparse distributions, then falls, and finally levels off and decreases slowly. The sparseness at which this change in behavior is exhibited is independent of the parameters of the system, and can be seen from Fig. 3 to lie somewhere between 0.2 and 0.5.

In the special case of the linear limit, in which $\xi_0\rightarrow\infty$, λ_A disappears (see the Appendix), and stability is assured. For finite ξ_0 and above the coefficient of sparseness referred to in the preceding paragraph, though, there is a distinct and reasonably large region of instability.

The resulting phase diagrams are shown in Fig. 4. Figure 4(a) shows the situation for $\xi_0=-0.4$, which corresponds to threshold-linear behavior. As ξ_0 is increased [Figs. 4(b)–4(d); the neurons are made progressively ''more linear''], the critical noise variance at which instability of the RS solution sets in first increases, and then decreases, as would be expected from Fig. 3. In Fig. 4(d), the line of critical noise variance abruptly stops at $a\sim 0.23$: at this point, the replicon-mode eigenvalue passes below the zero axis, and stability is assured. In all cases, it is apparent that in particular for very sparse distributions, the replica-symmetric equations are valid down to quite low noise. For less sparse coding, where the pattern entropy is significantly higher, the replica-symmetry-broken solution would seem to be relevant for higher noise variances.

It should be noted that the sparseness of the distribution of outputs is not the same as that of the inputs. This can be determined by

$$a_{\text{out}}=\frac{\langle\xi\rangle^2_{\xi^+}}{\langle\xi^2\rangle_{\xi^+}}, \tag{37}$$

where

$$\langle x(\xi)\rangle_{\xi^+}=\int_0^\infty \frac{d\xi}{\sqrt{2\pi\sigma^2_\xi}}\,x(\xi)\exp-\frac{(\xi-\xi_0)^2}{2\sigma^2_\xi},$$

$$\sigma^2_\xi=\sigma^2_\epsilon+\sigma^2_J C(\langle\eta^2\rangle_\eta-\langle\eta\rangle^2_\eta). \tag{38}$$

The lines of marginal stability for $\xi_0=-0.4$, 0.0, 0.4, and 0.80 are replotted in Fig. 5 against the output sparseness. Although the phase diagrams look fairly similar when plotted as a function of input sparseness, they occupy different regions of the output-sparseness domain because of the thresholding. It is also worth noting that because of the mapping performed by Eq. (37), the boundaries of the regions in Fig. 4 do not necessarily form the boundaries of the regions in the output-sparseness plane, which in some instances constitute points from inside the above curves.

For neurons operating in the threshold-linear regime (left curve, $\xi_0<0.0$), where output sparseness is effectively constrained by the thresholding, the stability characteristics are qualitatively as has been described earlier. For $\xi_0=0.0$, it is apparent from Eqs. (37) and (38) that the output sparseness is constant (regardless of the input sparseness) at a value of $1/\pi$. As ξ_0 is increased above zero, the output becomes less sparse, and the line of marginal stability is flipped horizontally (because in this range the entropy is higher for smaller a_{out}; right curves). Assuming that the sparseness of coding in connected sets of neurons in the brain tends to be similar, the former curve (for threshold-linear behavior) might be considered the more biologically applicable, with the threshold in this model incorporating functionally the constraint on the degree of neural activity.

V. CONCLUSIONS

This paper has detailed the replica-symmetric solution for the information transmitted by a feedforward network of threshold-linear neurons, and examined its stability to fluctuations in the direction of replica-symmetry breaking. It appears that for sparse pattern distributions, replica-symmetry breaking only sets in at noise variances sufficiently small that we might reasonably consider them to be ''beyond the realm of biological interest,'' at least for noisy cortical cells. We believe that, quite importantly, there is every reason to expect that these results carry over to the slightly more complicated ''Schaffer collateral'' calculation described in [8,9]. There is thus reason to feel confidence in the replica-symmetric assumption when analyzing neural networks in areas such as the hippocampus which are known to code sparsely.

When more distributed (less sparse) encoding is used, the mutual information solution is prone to instability to replica-symmetry-breaking fluctuations at higher amounts of noise than in the sparse case. It is not clear from the current analysis what the quantitative effect of broken replica symmetry might be, or what the form of the exact solution would be in that case (e.g., the Parisi ansatz [15]). Care should therefore

be taken when analyzing the information conveyed by networks using more distributed encoding.

ACKNOWLEDGMENTS

We would like to thank S. Panzeri, F. Battaglia, and C. Fulvi-Mari for useful discussions. In particular, we would like to thank E. T. Rolls for his role in the collaborative research environment that allowed this work to be undertaken. S.S. would also like to thank the Oxford McDonnell-Pew Centre for Cognitive Neuroscience for financial support.

APPENDIX

In this appendix the transverse eigenvalues of the submatrices $A^{(\alpha\beta)(\gamma\delta)}$ and $B^{(\alpha\beta)(\gamma\delta)}$ are calculated. Both $A^{(\alpha\beta),(\gamma\delta)}$ and $B^{(\alpha\beta),(\gamma\delta)}$ have three different types of matrix elements depending on whether none, one, or two replica indices of the pair $(\alpha\beta)$ equal those of the pair $(\gamma\delta)$. The three possible values $A^{(\alpha\beta),(\gamma\delta)}$ can take are

$$P=\frac{\partial^2\mathcal{B}}{\partial z^{\alpha\beta}\partial z^{\alpha\beta}}=\frac{\sigma_J^4 Cr}{4W}\frac{(q^2+2pq)^2}{p^4(p+q)^4}\left\{\int_0^\infty\frac{d\xi}{\sqrt{2\pi}}\times(\xi-\xi_0)^4\exp\left[-\frac{(\xi-\xi_0)^2}{2(p+q)}\right]+\int_{-\infty}^\infty\frac{dt}{\sqrt{2\pi}}\times[(\xi-\xi_0)^2]^2_{\xi^-}(\tfrac{1}{2},t)\right\},$$

$$Q=\frac{\partial^2\mathcal{B}}{\partial z^{\alpha\beta}\partial z^{\alpha\gamma}}=\frac{\sigma_J^4 Cr}{4W}\frac{(q^2+2pq)^2}{p^4(p+q)^4}\left\{\int_0^\infty\frac{d\xi}{\sqrt{2\pi}}\times(\xi-\xi_0)^4\exp\left[-\frac{(\xi-\xi_0)^2}{2(p+q)}\right]+\int_{-\infty}^\infty\frac{dt}{\sqrt{2\pi}}\times[(\xi-\xi_0)^2]_{\xi^-}(\tfrac{1}{3},t)[(\xi-\xi_0)]^2_{\xi^-}(\tfrac{1}{3},t)\right\}$$

$$(\beta\neq\gamma), \tag{A1}$$

$$R=\frac{\partial^2\mathcal{B}}{\partial z^{\alpha\beta}\partial z^{\gamma\delta}}=\frac{\sigma_J^4 Cr}{4W}\frac{(q^2+2pq)^2}{p^4(p+q)^4}\left\{\int_0^\infty\frac{d\xi}{\sqrt{2\pi}}(\xi-\xi_0)^4\exp\left[-\frac{(\xi-\xi_0)^2}{2(p+q)}\right]+\int_{-\infty}^\infty\frac{dt}{\sqrt{2\pi}}[(\xi-\xi_0)]^4_{\xi^-}(\tfrac{1}{4},t)\right\}\quad(\alpha\neq\gamma,\beta\neq\delta),$$

where $[x(\xi)]_{\xi^-}(k,t)$ is defined as

$$[x(\xi)]_{\xi^-}(k,t)=\int_{-\infty}^0\frac{d\xi}{\sqrt{2\pi}}x(\xi)\exp\left[-\frac{k}{2p}(\xi-\xi_0)^2+kt\left(\frac{q}{p(p+q)}\right)^{1/2}(\xi-\xi_0)-\frac{kt^2}{2}\right], \tag{A2}$$

which can be considered to be a weighted average of $x(\xi)$ over the subthreshold values of ξ. k is used to normalize the weight factor over the t integral in each of Eqs. (A1). Also,

$$W=\phi\left(\frac{\xi_0}{\sqrt{p+q}}\right)+\int_{-\infty}^\infty\frac{dt}{\sqrt{2\pi}}\phi\left[-\frac{\xi_0}{\sqrt{p}}-t\left(\frac{q}{p+q}\right)^{1/2}\right]\times\exp\left[-\frac{t^2(2p+q)}{4(p+q)}\right]\sqrt{p}, \tag{A3}$$

and p,q are here p_B and q_B from Eq. (22).

We have to solve the eigenvalue equation

$$\mathbf{A}\psi=\lambda\psi. \tag{A4}$$

The eigenvectors ψ have the column-vector form

$$\psi=(\{\delta z^{\alpha\beta}\})\quad(\alpha<\beta=1,\ldots,n+1). \tag{A5}$$

We now proceed as described in [11]. There are three classes of eigenvectors (and corresponding eigenvalues)—those invariant under interchange of all indices, those invariant under interchange of all but one index, and those invariant under interchange of all but two indices. These last describe the transverse mode, in which we are interested.

Let us consider fluctuations of the form

$$\delta z^{\alpha\beta}=\Delta^{\alpha\beta}\quad(\alpha<\beta=1,...,n+1), \tag{A6}$$

with

$$\Delta^{\alpha\beta}=\Delta,\quad\alpha,\beta\neq\alpha_0,\beta_0$$

$$\Delta^{\alpha_0\beta}=\Delta^{\alpha\beta_0}=\frac{2-n}{2}\Delta,\quad\alpha\neq\alpha_0,\beta_0 \tag{A7}$$

$$\Delta^{\alpha_0\beta_0}=\frac{(2-n)(1-n)}{2}\Delta$$

ensuring orthogonality between the eigenvectors describing RS and RSB fluctuations. As with [11], we have for $A^{(\alpha\beta),(\gamma\delta)}$ an eigenvalue

$$\lambda_A=P-2Q+R, \tag{A8}$$

with in this case $[\frac{1}{2}(n+1)(n-2)]$-fold degeneracy, and P, Q, and R as described above.

For $B^{(\alpha\beta),(\gamma\delta)}$, we consider fluctuations

$$\delta\tilde{z}^{\alpha\beta}=c\Delta^{\alpha\beta}\quad(\alpha<\beta=1,\ldots,n+1) \tag{A9}$$

and obtain similarly the eigenvalue

$$\lambda_B=P'-2Q'+R', \tag{A10}$$

where

$$P' = \frac{\partial^2 \mathcal{B}}{\partial(i\tilde{z}^{\alpha\beta})\partial(i\tilde{z}^{\alpha\beta})} = \int_{-\infty}^{\infty} Dt[\eta^2]_\eta^2(\tfrac{1}{2},t),$$

$$Q' = \frac{\partial^2 \mathcal{B}}{\partial(i\tilde{z}^{\alpha\beta})\partial(i\tilde{z}^{\alpha\gamma})} = \int_{-\infty}^{\infty} Dt[\eta^2]_\eta(\tfrac{1}{3},t)[\eta]_\eta^2(\tfrac{1}{3},t), \tag{A11}$$

$$R' = \frac{\partial^2 \mathcal{B}}{\partial(i\tilde{z}^{\alpha\beta})\partial(i\tilde{z}^{\gamma\delta})} = \int_{-\infty}^{\infty} Dt[\eta]_\eta^4(\tfrac{1}{4},t),$$

and $[x(\eta)]_\eta(k,t)$, the weighted pattern average, is defined as

$$[x(\eta)]_\eta(k,t) = \int_\eta d\eta\ P(\eta)x(\eta)\exp\left[-\frac{k}{2}\tilde{z}_1\eta^2 - kt\sqrt{\tilde{z}_1}\eta\right]. \tag{A12}$$

[1] D. Amit, H. Gutfreund, and H. Sompolinsky, Ann. Phys. (N.Y.) **173**, 30 (1987).
[2] E. Gardner, J. Phys. A **21**, 257 (1988).
[3] W. Bialek and A. Zee, Phys. Rev. Lett. **61**, 1512 (1988).
[4] A. Treves, J. Phys. A **23**, 2631 (1990).
[5] J.-P. Nadal and N. Parga, Network **4**, 295 (1993).
[6] A. Treves, C. A. Barnes, and E. T. Rolls, in *Perception, Memory and Emotion: Frontier in Neuroscience*, edited by T. Ono *et al.* (Elsevier, Amsterdam, 1996), Chap. 37, pp. 567–579.
[7] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
[8] A. Treves, J. Comput. Neurosci. **2**, 259 (1995).
[9] S. Schultz, S. Panzeri, E. T. Rolls, and A. Treves, in *Information Theory and the Brain*, edited by R. Baddeley, P. Földiák, and P. Hancock (Cambridge University Press, Cambridge, U.K., 1997).
[10] J.-P. Nadal and N. Parga, Neural Comput. **6**, 491 (1994).
[11] J. R. L. de Almeida and D. J. Thouless, J. Phys. A **11**, 983 (1978).
[12] A. Treves, J. Phys. A **24**, 2645 (1991).
[13] E. Gardner and B. Derrida, J. Phys. A **21**, 271 (1988).
[14] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[15] G. Parisi, J. Phys. A **13**, L115 (1980).