Bayesian Retrieval in Associative Memories with Storage Errors

Friedrich T. Sommer and Peter Dayan

***Abstract*—It is well known that for finite-sized networks, one-step retrieval in the autoassociative Willshaw net is a suboptimal way to extract the information stored in the synapses. Iterative retrieval strategies are much better, but have hitherto only had heuristic justification. We show how they emerge naturally from considerations of probabilistic inference under conditions of noisy and partial input and a corrupted weight matrix. We start from the conditional probability distribution over possible patterns for retrieval. This contains all possible information that is available to an observer of the network and the initial input. Since this distribution is over exponentially many patterns, we use it to develop two approximate, but tractable, iterative retrieval methods. One performs maximum likelihood inference to find the single most likely pattern, using the (negative log of the) conditional probability as a Lyapunov function for retrieval. In physics terms, if storage errors are present, then the modified iterative update equations contain an additional antiferromagnetic interaction term and site dependent threshold values. The second method makes a mean field assumption to optimize a tractable estimate of the full conditional probability distribution. This leads to iterative mean field equations which can be interpreted in terms of a network of neurons with sigmoidal responses but with the same interactions and thresholds as in the maximum likelihood update equations. In the absence of storage errors, both models become very similiar to the Willshaw model, where standard retrieval is iterated using a particular form of linear threshold strategy.**

***Index Terms*— Bayesian reasoning, correlation associative memory, graded response neurons, iterative retrieval, maximum likelihood retrieval, mean field methods, threshold strategies, storage errors, Willshaw model.**

I. Introduction

NEURAL associative memories with the capacity for pattern completion were first proposed as cybernetic models to relate psychological phenomena with processes in networks of nerve cells [1]–[6]. Such associative memories have a natural mapping onto parallel hardware, and can be used for information retrieval from large heterogenous databases [7], [8], and also to help understand information processing in strongly connected circuits in the cortex [9], [10]. Even though, since Hopfield's famous paper [11], they have been extensively analyzed using the methods of statistical physics, there remain many open questions.

The Willshaw net [2] (see Section II-D) is one of the most efficient associative memory models in terms of information stored per bit of memory. However, it has not been so widely used since its performance degrades significantly if there are errors in the initial patterns presented or if there are errors in the synaptic weight matrix [12]. Both sorts of error are highly likely in large-scale hardware implementations in silicon or optical devices and also in networks of biological neurons.

With the notable exception of [13], one of the main troubles with most existing theory on associative memories is that inference on the basis of the inputs is not treated in a systematically probabilistic way. In this paper, we attempt such a treatment, which offers the prospect of helping with the problems mentioned above. This paper presents the theory underlying the approach in the context of the finite sized autoassociative Willshaw net, to which it is particularly well suited; however, the same theory can be used for inference in other models, including heteroassociative memories. Comprehensive empirical studies will be needed to test forms of this approach.

Our treatment reveals the close relationship between iterative retrieval methods in associative memory and Bayesian reasoning and its mean-field approximation. *A priori* knowledge about the training patterns, errors in the initial pattern, and storage errors in the weight matrix lead to additional constraint terms in a Lyapunov function governing the retrieval in a binary neural network. Some of these terms are new, others justify heuristic additions that have already been made, and are known to improve retrieval performance significantly. The link between mean-field approximations to the reasoning process and graded response associative memories has not previously been made, and is a further key step in our method.

In the next section, we define the task for an autoassociative memory, and briefly describe the Willshaw associative memory model. By comparing the asymptotic capacities of different models, we will argue that iterative retrieval strategies hold substantial promise for Willshaw nets. In Section III, we consider the posterior probability distribution over the possible output patterns, given a particular input. In Section IV, we derive a Lyapunov function for iterative retrieval in the Willshaw net and show the influence on its different threshold strategies proposed in the literature. In Sections IV and V we discuss two iterative retrieval methods that arise from the conditional distribution derived in Section III: a maximum likelihood (ML) method that attempts to find the most likely single

Manuscript received May 23, 1997; revised February 19, 1998. The work of F. T. Sommer was supported by Grant SO352/3-1 from the Deutsche Forschungsgemeinschaft. The work of P. Dayan was supported by NSF Grant IBN-9 634 339 and the Surdna Foundation.

F. T. Sommer is with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA, on leave from Department of Neuroinformatics, University of Ulm, 89069 Ulm, Germany.

P. Dayan is with the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Publisher Item Identifier S 1045-9227(98)04730-4.

output, and a mean field method that tries to approximate the whole distribution. The methods use very similar equations, except that the mean field method employs sigmoidal units. In Section VI we summarize the consequences of the approach we propose.

II. Associative Memory Models

A. The Autoassociative Memory Task

The task for a binary *autoassociative memory* is to find among a set of the M stored *training patterns* $\{\boldsymbol{x}^\nu\colon \boldsymbol{x} \in \{0,1\}^n, \nu = 1,\cdots,M\}$, the one which is closest to a binary initial pattern $\tilde{x} \in \{0,1\}^n$. Autoassociative memory is a special case of heteroassociative memory where the stored associations are between training pattern pairs which can contain different patterns: $\boldsymbol{x}^\nu \rightarrow \boldsymbol{y}^\nu$. The metric in the space of binary patterns is usually the *Hamming distance* which is defined as the number of components for which two patterns disagree. If pattern $(\hat{\boldsymbol{x}}^\eta)$ is the ultimate estimate of a training pattern $(\boldsymbol{x}^\eta)$ we distinguish the two possible error types: a "miss" error converts a 1-entry in x_a^η to "0" and a "false alarm" error does the opposite. The number of active (1-components) in a pattern is called the *activity*: $|\boldsymbol{x}| = \Sigma_{i=1}^n\ x_i$. We will consider training sets for which all patterns have similiar activities, that is, $|\boldsymbol{x}^\nu| \approx b\ \forall\nu$. The training patterns are called *sparse* if their activities are much smaller than the dimension, i.e., $b \ll n/2$.

B. The Willshaw Model

A standard, correlational, Hebbian learning rule [14] takes the outer product of the training patterns. A well-known example is the Hopfield net [11]. Nonlinear functions of the outer product have been introduced to account for synaptic saturation and quantization effects. The Willshaw net [2] uses an extreme form of synaptic saturation, clipping each synapse at the value one. This makes the elements of the synaptic weight matrix

$$C_{ij}^H = \min\left(1, \sum_{\nu=1}^M x_i^\nu x_j^\nu\right) = 1 - \prod_{\nu=1}^M (1 - x_i^\nu x_j^\nu) = \sup_{\nu=1}^M\ x_i^\nu x_j^\nu. \tag{1}$$

For an empirical comparison of linear and clipped Hebbian learning for sparse training patterns, see [15]. Clearly, information about a high number of training patterns can only be extracted from the weight matrix if the training patterns are sparse enough in order to prevent $P[C_{ij}^H = 1] \rightarrow 1$.

Given an initial pattern $\boldsymbol{x}(0) = \tilde{\boldsymbol{x}}^\eta$, which is a corrupted version of the training pattern $\boldsymbol{x}^\eta$, the *retrieval process* in the memory is described by the update equations

$$x_j(t+1) = H\left(\sum_{i=1}^n c_{ij}^H x_i(t) - \Theta(t)\right) \quad \forall j = 1,\cdots,n \tag{2}$$

where $\Theta(t)$ is a global threshold, and $H(x)$ is the Heaviside function. In the original Willshaw model the update process (2) is not iterated. The model works efficiently if the training patterns are sufficiently sparse, i.e., the activities in the stored patterns are of the order of the log of the number of components, and the threshold is adjusted properly, i.e., $\Theta(0) = \Sigma_i\ \tilde{x}_i^\eta x_i^\eta$.

C. Capacity and Efficiency of a Memory Model

In order to understand the potential for our new statistical framework for retrieval, we must first outline what is known about the power of existing associative memory models. Capacity results in this area are notoriously confusing, because many different measures of capacity have been employed in many different ways.

Autoassociative memories only provide new information if the initial pattern is substantially changed during retrieval as the nearest training pattern is determined. Competent pattern completion requires the training patterns to have nonvanishing basins of attraction which allow retrieval from initial patterns containing a substantial level of noise. A given error criterion for retrieval and a maximum level of input noise will fix an upper bound M^+ on the possible number of patterns in the training set. The bound will decrease if either increased retrieval precision or increased input fault tolerance is required. One popular measure of performance is the ratio M^+/n between the number of stored pattern components M^+n and the number of required synapses n^2. However, this patterns-per-neuron ratio (sometimes called the critical capacity) does not allow a fair evaluation for sparse training patterns, each of which contains far fewer bits of information than its number of components. Even if the patterns are not sparse, the measure does not take into account the information loss due to retrieval errors and the information about the retrieved pattern which is already contained in the initial patterns. Rather the true amount of information about training patterns *gained during retrieval* should be considered. The *information capacity* of an associative memory is defined as this amount of information divided by the number of synapses. The information capacity measure, which is also popular, still ignores one important property of a model—namely, the number of bits required to represent each synaptic weight. Therefore, the *information efficiency* has been proposed which is defined as the information capacity divided by the minimum number of bits required per synapse, see [16]. Obviously, this dimensionless quantity can maximally assume the value of one and for binary synapses its numerical value coincides with the information capacity.

D. Asymptotic Capacity Results

As the number of units grows, the asymptotic capacity of the heteroassociative Willshaw model for vanishing retrieval errors is $\ln 2$ bits/synapse [2]. Changing the task from hetero- to autoassociative memory reduces the capacity for the two reasons described above: the information about the final pattern contained in the initial pattern presented, and the requirement for nonempty basins of attraction for the patterns.

If one thinks of the memory task as a form of information channel for the memory patterns, then the information capacity is bounded by the maximum capacity describing the learning process (10 which has been called learning bound. For heteroassociative memories, the learning bound has been

shown to coincide with the information capacity [16]. Since the synaptic matrix for an autoassociative memory is symmetric, it has half as many free parameters, and thus the learning bound is expected to be reduced by at least a factor of two. A particular method of extracting information from the autoassociative Willshaw matrix does actually achieve the maximum capacity of $(\ln 2)/2 = 0.35$ bits/synapse [17]. However, this method extracts the whole set of memory patterns by an exhaustive exploration of the entire space of sparse patterns. This is fine for recognition, but cannot be used to complete single patterns, as is required for an autoassociative memory. For the autoassociative task the information capacity of Willshaw model has been determined to be $(\ln 2)/4 = 0.17$ bits/synapse [16]. This is still high compared to the models discussed later, but is only half of what can be reached in a recognition task. Hence, the retrieval procedure prescribed by the Willshaw model is clearly a limiting factor, a fact which is also underpinned by theoretical and empirical capacity results of iterative retrieval in the Willshaw model [15], which exceed the value 0.17.

At first sight, the Willshaw learning procedure would also appear to be suboptimal: linear Hebbian learning without the clipping promises higher information capacity since the synapses could then carry more than one bit each. Surprisingly, analysis for the $\{0, 1\}$ Hopfield model with sparse training patterns yields an information capacity of $1/(8 \ln 2) = 0.18$ bits/synapse [18] (the result of the first reference has been transformed to give true autoassociative information capacity) [16], which is strikingly close to that achieved with clipped synapses. The resulting information efficiency is clearly below that of the Willshaw model even for four-state (two bit) synapses.

Another large class of associative memories that employ dense training patterns has been investigated. However, for these, the asymptotic capacity always goes to zero asymptotically for a error criterion for retrieval that demands vanishing errors. For instance, for the Hopfield model with linear learning, the number of patterns per neuron is $1/(2 \ln n)$, and for clipped learning, it is $1/(\pi \ln n)$ [19]. If a general nonlinear function is used in place of the Heaviside function for weight saturation, then, in general, the information capacity stays below that of linear learning [20]. With a small finite error criterion, the Hopfield model achieves an asymptotic pattern-per-neuron ratio of 0.14 [11], [21], and, in experiments with nonmonotonic retrieval dynamics (i.e., replacing the Heaviside function in (2) by a nonmonotonic function), it has reached 0.3 [22], [23]. Nevertheless, even the higher result corresponds to an information efficiency lower than that of the autoassociative Willshaw model.

Recently, [24] suggested a further definition of capacity that we mention for the sake of completeness. Instead of just the M training patterns, this measure counts *all* patterns that are stable and have some sufficiently large basin of attraction. This number grows exponentially with the number of units and includes all mixture or spurious states. These states are normally considered to reduce the sizes of the basins of attraction for the training patterns, and therefore to be undesirable. The ratio between this enhanced number of patterns and the number of neurons is called the basis rate, and, inspired by channel coding theory, the rate of exponential growth is called the capacity. Requiring a positive capacity is consistent with placing an upper bound on the basis rate. This idea had already been used as an alternative way of determining the pattern-per-neuron ratio in the Hopfield model: The result is close to the 0.14 cited above [25]—for a slightly different model, [24] showed that the basis rate is 0.17.

We have so far only considered capacity using Hebbian-type learning rules. The seminal work of [26] analyzed the potential capacity of autoassociative memories given the optimal settings of the weights. For nonsparse training patterns, the Gardner bound is 2 bits/synapse, which is much higher than the Hopfield capacity. However, more complicated learning rules, such as the delta rule, which require multiple presentations of the patterns are required. For sparse training patterns, the Gardner bound [26] coincides with the information capacity [18] for linear Hebbian learning. Therefore, for associative memories with sparse training patterns, retrieval is the bottleneck; for dense training patterns, the learning procedure must also be refined.

To summarize: 1) binary clipped learning as employed in the Willshaw model has the largest information efficiency (0.17); 2) for sparse training patterns, which are required for the Willshaw model to be optimally efficient, more sophisticated learning rules will not help; and 3) changing the retrieval strategy in the Willshaw model can potentially improve its efficiency by a factor of two.

E. Modifications to the Willshaw Net

Although the Willshaw net was one of the first associative memories to be suggested [1], it is only recently that modifications and improvements have been proposed. For one-step retrieval, statistical arguments have been adduced in favor of a more refined, site dependent threshold [27], [28]. Various methods for iterative retrieval have also been suggested on the basis of heuristic arguments [29], [30], [15]. Iterative retrieval in finite-sized systems reaches, and even slightly surpasses, the asymptotic capacity, typically achieving efficiencies up to \$0.2\$ with much lower retrieval errors even for moderate size systems (e.g., $n = 2000$, see [15]). It is clearly important to provide a strong theoretical framework for these iterative strategies to understand their basis and the scope for further improvement.

III. Retrieval by Probabilistic Inference

We consider Bayesian analysis of the process of autoassociative recall. The output of this is a posterior distribution over all possible patterns, expressing how likely it is that each pattern underlies the initial pattern that was presented. This posterior distribution depends on a variety of forms of prior information, as discussed below. Given a loss function, the particular pattern that minimizes the expected posterior loss can be chosen. Quite a range of behaviors is supported by different loss functions. This full posterior distribution is computationally intractable to manipulate. We therefore consider two approaches. One is maximum *a posteriori* (MAP) inference,

which requires finding the single pattern that maximizes the conditional probability. We will see in the next section that this leads to a retrieval method that is very close to existing suggestions, and also provides a theoretical framework to justify many of the otherwise heuristic modifications proposed for all sorts of associative memory models, not only the Willshaw net. The second approach to the full posterior distribution is to approximate it by a simpler mean field distribution.

Retrieval is cued by an initial pattern $\tilde{\boldsymbol{x}}$, which is a noisy version of one of the training patterns. We omit the index η where it is obvious. If the memory matrix is C, then retrieval in the associative memory should depend on the conditional or *a posteriori* probability $P[\boldsymbol{x}|C,\tilde{\boldsymbol{x}}] = P[\boldsymbol{x},C,\tilde{\boldsymbol{x}}]/P[C,\tilde{\boldsymbol{x}}]$, where

$$P[\boldsymbol{x},C,\tilde{\boldsymbol{x}}] = P[C|\boldsymbol{x}]P[\tilde{\boldsymbol{x}}|\boldsymbol{x}]P[\boldsymbol{x}] \tag{3}$$

under a reasonable probabilistic model for which C is independent of $\tilde{\boldsymbol{x}}$ given $\boldsymbol{x}$. In this section we derive expressions and approximations for the three factors on the right-hand side of (3), based on prior knowledge about the pattern activity, the noise in the initial pattern, and storage noise in the synaptic matrix. Based on this analysis, we go on to suggest two approximate iterative retrieval equations, one that finds the $\boldsymbol{x}^*$ that maximizes $P[\boldsymbol{x}|C,\tilde{\boldsymbol{x}}]$, using (3) as a Lyapunov function, and the other that finds a mean-field approximation to $P[\boldsymbol{x}|C,\tilde{\boldsymbol{x}}]$.

A. Biased Random Training Patterns

The last term in (3) is the prior probability that pattern $\boldsymbol{x}$ could have been one of the training patterns. We consider the case in which each component of a training pattern is generated independently with bias $p = P[x_i = 1] = b/n$, producing a set of training patterns with mean activity b. Therefore

$$P[\boldsymbol{x}] = p^{|\boldsymbol{x}|}(1-p)^{n-|\boldsymbol{x}|}. \tag{4}$$

Making a Gaussian approximation to the expression in (4), the log probability is given as

$$\begin{aligned}\log P[\boldsymbol{x}] &= -\frac{1}{2np(1-p)}\left(\sum_i\sum_j x_i x_j - 2np|\boldsymbol{x}|\right) + k \\ &= -\frac{1}{2np(1-p)}\left(\sum_i x_i - b\right)^2 + k\end{aligned} \tag{5}$$

where k is a generic constant with respect to the variable parameters that takes different values in each equation.

B. Noise in the Initial Pattern

The middle term in (3) quantifies the way that $\tilde{\boldsymbol{x}}$ could have been produced as a corrupted version of training pattern $\boldsymbol{x}$. Clearly, the farther $\tilde{\boldsymbol{x}}$ is from $\boldsymbol{x}$, the less likely it is that it was generated by $\boldsymbol{x}$ in the first place. We describe the *a priori* knowledge about these initial distortions using the conditional probabilities $r = P[\tilde{x}_i = 1|x_i = 0]$ and $s = P[\tilde{x}_i = 0|x_i = 1]$. In the case for which the process of corruption preserves the mean activity of the patterns, one of these error probabilities can be eliminated using the expression $s = r/g(p)$ with $g(p) = p/(1-p)$.

The probability of generating $\tilde{\boldsymbol{x}}$ as a corruption of x is

$$P[\tilde{\boldsymbol{x}}|\boldsymbol{x}] = \prod_i P[\tilde{\boldsymbol{x}}_i = 1|x_i]^{\tilde{x}_i} P[\tilde{x}_i = 0|x_i]^{1-\tilde{x}_i} \tag{6}$$

with $P[\tilde{x}_i = 1|x_i] = (1-s)^{x_i} r^{1-x_i}$ and $P[\tilde{x}_i = 0|x_i] = s^{x_i}(1-r)^{1-x_i}$. The second factor in (3) is therefore

$$\begin{aligned}&\log P[\tilde{\boldsymbol{x}}|\boldsymbol{x}] \\ &\quad = \sum_i -\log g(r)g(s)\tilde{x}_i x_i + \log\frac{s}{r}x_i + \log g(r)\tilde{x}_i + k.\end{aligned} \tag{7}$$

where k is a constant with respect to the variable parameters.

C. Matrix Elements with Storage Errors

The remaining factor in (3) is the probability $P[C|\boldsymbol{x}]$ that the synpatic weight matrix would be C if $\boldsymbol{x}$ had been one of the training patterns. This reflects the influence of three contributions: the effect of pattern $\boldsymbol{x}$ itself, the effect of all the other training patterns, and the effect of noise corrupting the perfect Hebbian matrix C^H. The storage process may be corrupted by two error types: "stuck-at-0" errors denote the case where elements of the Hebbian matrix with values one are converted into zero and "stuck-at-1" errors denote the inverse confusion. The imperfect storage process is again characterized by a pair of error probabilities: $\delta = P[C_{ij} = 0|C^H_{ij} = 1]$ quantifying the "stuck-at-0" errors and $\gamma = P[C_{ij} = 1|C^H_{ij} = 0]$, the "stuck-at-1" errors. The first factor in (3) can then be approximated by

$$P[C|\boldsymbol{x}] = \prod_{i,j} P[C_{ij} = 1|x_i,x_j]^{C_{ij}} P[C_{ij} = 0|x_i,x_j]^{1-C_{ij}} \tag{8}$$

with $P[C_{ij} = 1|x_i,x_j] = (1-\delta)^{x_i x_j}(1-q)^{1-x_i x_j}$ and $P[C_{ij} = 0|x_i,x_j] = \delta^{x_i x_j} q^{1-x_i x_j}$, where the probabilities that matrix elements C and C^H have not been changed from zero by other training patterns are

$$\begin{aligned}q &= P[C_{ij} = 0|x_i x_j = 0] = (1-\gamma)q' + \delta(1-q') \\ q' &= P[C^H_{ij} = 0|x_i x_j = 0] \simeq (1-p^2)^{M-1}.\end{aligned} \tag{9}$$

Equation (8) is only an approximation to the true probability, since we are ignoring dependencies between different elements of C^H that arise from the Hebbian storage process of the other memory patterns.

Equation (8) yields the log probability

$$\log P[C|\boldsymbol{x}] = \sum_i\sum_j \{i_q(C_{ij}) - i_\delta(C_{ij})\}x_i x_j - i_q(C_{ij}) \tag{10}$$

where

$$i_q(C_{ij}) = -C_{ij}\log(1-q) - (1-C_{ij})\log(q) \tag{11}$$

and similarly for $i_\delta(C_{ij})$. The terms $i_q(C_{ij})$ and $i_\delta(C_{ij})$ are the logarithmic probabilities of C_{ij} under the condition $x_i x_j = 0$ and $x_i x_j = 1$, respectively. If they are equal, for $\delta = q$ or

equivalently $\delta = 1 - \gamma$, we have the case of total information loss in C due to storage errors, where no information about the particular training pattern x is preserved in the corrupted matrix: $\log P[C|\boldsymbol{x}] = -\sum_{i,j} i_q(C_{ij})$ is independent of $\boldsymbol{x}$.

On the other hand, in the absence of storage errors, the $\log\delta$ term becomes dominating and we obtain

$$\log P[C|\boldsymbol{x}] = \log\delta \sum_i \sum_j \{1 - C_{ij}\} x_i x_j - i_{q'}(C_{ij})$$

as $\delta \to 0$.

IV. Approximate Map Inference

The MAP solution is the value of $\boldsymbol{x}$ that maximizes $P[\boldsymbol{x}|C, \tilde{\boldsymbol{x}}]$, or, equivalently, maximizes $P[\boldsymbol{x}, C, \tilde{\boldsymbol{x}}]$. Making the approximations derived in the previous section, $-\log P[\boldsymbol{x}, C, \tilde{\boldsymbol{x}}]$ can be used as a Lyapunov function for iterative retrieval. This leads to a modified update (23), which turns out to bear close relationships to modifications in the retrieval process that have previously been suggested. We discuss these modifications in turn and relate them to the MAP Lyapunov function.

A. Retrieval as Constraint Satisfaction

Iterative retrieval in the Willshaw model (2) can be described by the Lyapunov function

$$E^W(\boldsymbol{x}) = -\tfrac{1}{2} \sum_i \sum_j C^H_{ij} x_i x_j + \Theta(t) \sum_i x_i \qquad (12)$$

since it is easy to show that asynchronous application of the Willshaw update in (2) sets x_j to $1 - x_j$ if

$$E(x_1, \cdots, 1 - x_j, \cdots, x_n) - E(x_1, \cdots, x_j, \cdots, x_n) < 0. \qquad (13)$$

We call the method for setting the threshold $\Theta(t)$ during the course of iterations the *threshold strategy*.

The Lyapunov function

$$E(\boldsymbol{x}) = \frac{1}{2} \sum_i \sum_j \{1 - C^H_{ij}\} x_i x_j + \frac{\alpha}{2} \left(\sum_i x_i - \Theta(t) \right)^2 \qquad (14)$$

is equivalent (up to a constant factor and a constant offset) to (12) for $\alpha = -1$. The terms in (14) can be interpreted as constraint terms on $\boldsymbol{x}$: The first term punishes pairs of 1-components in $\boldsymbol{x}$ that coincide with a matrix element $C^H_{ij} = 0$. The α term in (14) is proportional to the quadratic deviation between the pattern activity and the threshold value. Since α is negative, using a constant threshold $\Theta(t) = \Theta\ \forall t$ does not stabilize the pattern activity near the threshold value, but rather drives the activity to its maximum or minimum value. Iterative retrieval can improve the performance, however, when combined with a linear threshold-setting strategy: $\Theta(t) = |\boldsymbol{x}(t)|$ [30], [15], [31]. Note that the linear threshold strategy just eliminates the influence of the α-term in (14) and hence is equivalent to a constant threshold strategy in a network with an additional antiferromagnetic interaction.

B. Pattern Activity Constraint

If all the training patterns have activity b, (14) suggests the very simple modification of setting $\alpha > 0$ and using $\Theta(t) = b$. In this case, the second component of the Lyapunov function (14) is a constraint term encouraging $|\boldsymbol{x}| = b$. For retrieval of sparse patterns, a constraint of this type was introduced into the Hopfield model [32]. For training patterns having varying activities this constraint term should be replaced by $\alpha\Sigma_k\ P[k](|\boldsymbol{x}| - k)^2$, where $P[k] = P[|\boldsymbol{x}| = k]$. Nevertheless, if the patterns are generated with a binomial distribution, i.e., $P[k] = \binom{n}{k} p^k (1-p)^{n-k}$ with $p = b/n$, as prescribed in Section III-A, then the resulting energy can be transformed into the form of (14), again requiring the same, constant, threshold $\Theta(t) = b$. However, this analysis does not suggest how large α should be—i.e., how important is this constraint compared with the term coming from the matrix.

In the MAP model, information about the activity of the training patterns is one of the three components of $P[\boldsymbol{x}, C, \boldsymbol{x}]$. Comparing (5) with the right-most part of (14), we can see that the appropriate setting of α is

$$\alpha = \frac{1}{np(1-p)} \qquad (15)$$

which is just the inverse variance of the activity distribution in the training patterns.

C. Storage Errors in the Synaptic Matrix

In the most interesting case for the memory, C is allowed to be a distorted version of the perfect learning matrix C^H. This requires replacing the first term on the right-hand side of (14) with a term such as

$$\frac{\zeta_0}{2} \sum_i \sum_j \{1 - C_{ij}\} x_i x_j + \frac{\zeta_1}{2} \sum_i \sum_j C_{ij} x_i x_j - \frac{1}{2} \sum_i \sum_j \{\zeta_0 - (\zeta_0 - \zeta_1) C_{ij}\} x_i x_j. \qquad (16)$$

The term multiplied by ζ_0 is just the same as that in (14). The term multiplied by ζ_1 is helpful if there are many "stuck-at-1" errors. It prevents undue advantage being given to patterns $\boldsymbol{x}$ for which $x_i = x_j = 1$, at least to the extent that $C_{ij} = 1$ can occur erroneously.

Comparing (10) with (16), the MAP model suggests particular values for the constraint coefficients

$$\zeta_0 = -2 \log \frac{\delta}{q} = 2 \log \left(1 - q' \left(1 - \frac{1-\gamma}{\delta} \right) \right) \qquad (17)$$

$$\zeta_1 = -2 \log \frac{1-\delta}{1-q} = 2 \log \left(1 - q' \left(1 - \frac{\gamma}{1-\delta} \right) \right). \qquad (18)$$

Equation (10) also includes the additive constant $-\sum_i \sum_j i_q(C_{ij})$ which does not depend on $\boldsymbol{x}$. The coefficients are nonnegative if $\delta \leq 1 - \gamma$. It can be observed that the constraint term is independent of C_{ij}, i.e., $\zeta_0 = \zeta_1$, only when the coefficients disappear. This happens either for $q' = 0$, if C^H contains no information about the training pattern due to crosstalk, or $\delta = 1 - \gamma$, the case of total information loss in C due to storage errors.

D. The Influence of the Initial Pattern

Typically in associative memories, the only influence of the initial pattern to the retrieval outcome is to determine the initial value of $\boldsymbol{x}$ for the dynamics. However, the initial pattern is the only piece of data pertaining to which of the memories should be retrieved, and it is also usually highly correlated with the correct memory. Therefore, retrieval may be improved by restricting the search to the vicinity of the initial pattern by its direct influence to the system dynamics. This is provided by an additional constraint term in the Lyapunov function in (14)

$$\begin{aligned}\beta_1 \sum_i \tilde{x}_i(1-x_i) + \beta_2 \sum_i (1-\tilde{x}_i)x_i \\ = \beta_1 \sum_i \tilde{x}_i + \beta_2 \sum_i x_i - (\beta_1+\beta_2)\sum_i \tilde{x}_i x_i \end{aligned} \tag{19}$$

where the first and the second term on the LHS punish deviations from 1- and 0-components of the initial pattern, respectively. Again, arbitrary offsets depending on the initial pattern $\beta(\tilde{\boldsymbol{x}})$ do not change the system dynamics. Comparing (7) with (19), the MAP Lyapunov function suggests the constraint coefficients

$$\beta_1 = -\log \frac{r}{1-s} \tag{20}$$

$$\beta_2 = -\log \frac{s}{1-r} \tag{21}$$

with the offset $\beta(\tilde{\boldsymbol{x}}) = \log(1-s/1-r)\,\Sigma_i\,\tilde{x}_i$. The resulting coefficients are positive if the initial pattern is closer to the training pattern than the inverted initial pattern, i.e., $r \leq 1-s$.

Modifications similar to this have previously been proposed for other memory models on a heuristic basis. For instance, for the Hopfield net, retrieval with a persistent field aligned with the initial pattern is known to improve retrieval performance [33]–[35]. Another modification, which is already prescribed by more expensive storage strategies such as pseudoinverse learning [36] to restrict the search space to the vicinity of the initial pattern is to add a positive offset in all diagonal weights C_{ii}. In other words, this adds a uniform self-interaction, or equivalently a transfer function with hysteresis [37]–[39].

E. Modified Retrieval Equations

Putting together the results of this section in (5), (6), and (10), the Lyapunov function derived by the MAP treatment can be written as

$$\begin{aligned} E^{\mathrm{MAP}}(\boldsymbol{x}) &= -\log P[\boldsymbol{x}, C, \tilde{\boldsymbol{x}}] + k \\ &= \tfrac{1}{2}\sum_i \sum_j \{V - UC_{ij}\}x_i x_j \\ &\quad - \sum_i (R + S\tilde{x}_i)x_i + k \end{aligned} \tag{22}$$

where k is again a generic constant.

Finally, from (22) and (13), the modified update equation can be derived

$$x_i(t+1) = H\left(-\sum_j \{V - UC_{ij}\}x_j(t) + R + S\tilde{x}_i\right) \tag{23}$$

with the constants in (22) and (23)

$$U = \zeta_0 - \zeta_1 = 2\log g(q)/g(\delta) \tag{24}$$

$$V = \alpha + \zeta_0 = 1/b(1-p) - 2\log \delta/q \tag{25}$$

$$R = \alpha b - \beta_2 = 1/(1-p) + \log s/(1-r) \tag{26}$$

$$S = \beta_1 + \beta_2 = -\log g(r)g(s). \tag{27}$$

Without storage errors, i.e., for $U \rightarrow \infty$ the synaptic constraint term $\Sigma_j\,\{1 - C_{ij}\}x_j(t)$ dominates in the argument of the Heaviside function. In this case, the ML approach is close to the linear threshold strategy adopted in the original Willshaw model, except that all the other terms in the Lyapunov function decide among those patterns that satisfy the hard constraint. In practical applications this could be realized by a high finite ζ_0.

In the case of total information loss in C, i.e., for $U = 0$ the argument is only determined by the additional constraint terms. In all other cases, because $H(\boldsymbol{x}) = H(z\boldsymbol{x})\ \forall z > 0$, we can normalize the argument in (23) by U. The behavior of the modified update equations (23) is therefore influenced by three independent coefficients: $V/U > 1$ implies an additional antiferromagnetic interaction. $R/U \neq 0$ represents a constant threshold offset consisting of a positive component which is growing with the activity of the training patterns and a second component compensating deleted 1-entries in the initial pattern. It vanishes if the probabilities that a "0" in the initial pattern is correct or has been produced by the noise are equal. $S/U > 0$ introduces a site dependent threshold representing a sustained bias toward the initial pattern during the iterative retrieval.

V. Approximate Mean Field Inference

The idea of a mean field treatment of the retrieval problem is to find a good approximation to the conditional distribution $P[\boldsymbol{x}|C, \tilde{\boldsymbol{x}}]$ which makes for tractable computation. The obvious approximate distribution that has frequently been used for associative memories treats all the components of x_i as being independent, Bernoulli variables

$$Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] = \prod_i \mu_i^{x_i}(1-\mu_i)^{1-x_i} \tag{28}$$

with means $P[x_i = 1] = \mu_i \in [0, 1]$ which are adjustable free parameters. The process of retrieval is the process of finding μ_i which minimize the Kullback–Leibler divergence between Q and P

$$\begin{aligned} KL[Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] || P[\boldsymbol{x}|C, \tilde{\boldsymbol{x}}]] \\ = \sum_{\boldsymbol{x}} Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] \log \frac{Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}]}{P[\boldsymbol{x}|C, \tilde{\boldsymbol{x}}]}. \end{aligned} \tag{29}$$

Since $\log P[\tilde{\boldsymbol{x}}, C]$ does not depend on $\boldsymbol{x}$

$$\sum_{\boldsymbol{x}} Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] \log P[\tilde{\boldsymbol{x}}, C] = \log P[\tilde{\boldsymbol{x}}, C]$$

and so we can equally well choose $\{\mu_i\}$ to minimize

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}}) &= KL[Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}) || P[\boldsymbol{x}|C, \tilde{\boldsymbol{x}}]] \\ &\quad - \sum_{\boldsymbol{x}} Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] \log P[\tilde{\boldsymbol{x}}, C] \\ &= -\sum_{\boldsymbol{x}} Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] \log P[\boldsymbol{x}, C, \tilde{\boldsymbol{x}}] \\ &\quad + \sum_{\boldsymbol{x}} Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}] \log Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}]. \end{aligned} \tag{30}$$

In physics terms, $\mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}})$ is the free energy. The first term in (30) is the energy, and the second term is the negative entropy of distribution $Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}]$. See [40] for a development of this expression.

Employing the factorial distribution in (28) and the log joint distribution in (22), we obtain

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}}) &= \tfrac{1}{2} \sum_i \sum_j \{V - UC_{ij}\} \mu_i \mu_j \\ &\quad - \sum_i \{[R + S\tilde{x}_i]\mu_i + i(\mu_i)\} + k \end{aligned} \tag{31}$$

where $i(x) = -x \log x - (1 - x) \log(1 - x)$ and k is again a generic constant.

The identity $i(\mu_i) = -\int_0^{\mu_i} \sigma^{-1}(\mu)\, d\mu$ with $\sigma(x) = (1 + e^{-x})^{-1}$ implies an interpretation of the free energy as a function of the mean field variables

$$E^{MF}(\boldsymbol{\mu}) = \mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}}) = E^{\mathrm{MAP}}(\boldsymbol{\mu}) + \sum_i \int_0^{\mu_i} \sigma^{-1}(\mu)\, d\mu \tag{32}$$

where $E^{\mathrm{MAP}}(\boldsymbol{x})$ is the Lyapunov function of the network with binary neurons from (22) that we developed in Section III: The integral term in $E^{MF}(\boldsymbol{\mu})$ just represents the additional term introduced into the Lyapunov function by a transition from a network of binary neurons to a network of neurons with the graded response function $\sigma(x)$ [41].

The best choice of μ_i in the distribution in (28) is found by minimizing the new Lyapunov function

$$\frac{\partial(E^{MF}(\boldsymbol{\mu}))}{\partial \mu_i} = 0, \qquad \forall i. \tag{33}$$

This leads to the mean field equations

$$\mu_i = \sigma\left(-\sum_j \{V - UC_{ij}\}\mu_j + R + S\tilde{x}_i\right) \quad \forall i. \tag{34}$$

with the coefficients given by (24)–(27). These equations can either be seen as consistency conditions that the true minima of $-\mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}}$ must satisfy, or, when applied asynchronously, a method of coordinate-wise descent in $\mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}}$. $\mathcal{L}(\boldsymbol{\mu}; C, \tilde{\boldsymbol{x}})$ can have many local minima, so such simple descent methods are only guaranteed to find local MAP solutions.

The outcome of iterating (34)—initialized with $\boldsymbol{\mu} = \tilde{\boldsymbol{x}}$—to convergence at $\boldsymbol{\mu}^*$ is an approximation $Q[\boldsymbol{x}; C, \tilde{\boldsymbol{x}}]$ to the true conditional distribution $P[\boldsymbol{x}|C, \tilde{\boldsymbol{x}}]$. One could find the MAP pattern from this distribution (in which case the mean field procedure is mostly a heuristic optimization strategy for ML inference) or it could be used in conjunction with other loss functions. Hinton (personal communication) has suggested a better, but computationally more expensive, method for finding an approximate binary MAP solution from the mean field distribution by successively clamping units to zero or one and resolving the mean field equations.

If the response function (i.e., the entropy) term is negligible in (32), i.e., if at least one of the constants U, V, R or S is large, then the attractors are the same patterns as for the binary system. This holds since then the energy is a linear function of a single μ_i, and must therefore achieve its minima at corners of the hypercube that limits the state space. On the other hand, if the entropy term dominates in (32), there is only a single stable state $\mu_i = 1/2\ \forall i$. As long as the entropy has some influence, the continuous response model has fewer stable states than the binary model but still each stable state corresponds to an attractor of the binary model [41].

The introduction of a nonzero temperature and thus graded response neurons is, of course, well known as a barrier method, and its link to mean field is also well understood. However, this is the first application to associative memory of which we are aware. In general, mean field methods can be expected to reduce the number of spurious states, without perturbing the training patterns too greatly. The extent to which this is true depends on a host of factors that are hard to determine, particularly the quality of the mean field approximation, in terms of the final value of the Kullback–Leibler divergence in (29).

VI. Discussion

A. Model Limitations and Simple Extensions

Our analysis so far is predicated on the particular formulation of the probabilistic model in Section III. This formulation contains assumptions which might be inappropriate in some cases of technical interest. For instance, we assumed for the training patterns and the initial pattern, uniform probabilities that do not depend on the component index. If training patterns are randomly generated with the probabilities $P[x_i = 1] = p_i$, (5) has to be replaced by

$$\log P[\boldsymbol{x}] = \sum_i x_i \log \frac{p_i}{1 - p_i} + k \tag{35}$$

introducing additional site-dependent threshold terms in the update prescriptions (23) and (34). Distortions in the initial pattern can also be modeled in a site-dependent manner by replacing parameters r and s by $\{r_i\colon\ i = 1, \cdots, n\}$ and $\{s_i\colon\ i = 1, \cdots, n\}$. From this a similiar modification with additional site dependent threshold terms results for both the models.

There might be other situations where *a priori* knowledge about the storage errors should be described in a different way than in Section III-C. In the deterministic case where storage errors are completely absent or their exact locations are known, only the first constraint term on the left-hand side of (16) should be used. As discussed in Section IV-E this leads to a diverging ζ_0. Of course, known storage errors can be

exactly compensated by an additional threshold term using the difference matrix $T_i = \zeta_0 \sum_k x_k(t)(C^H_{ki} - C_{ki})$.

If different probabilities for storage errors can be specified for different parts of the matrix i.e., δ_{ij} and γ_{ij} are given instead of δ and γ, the coefficients ζ_0 and ζ_1 in our model have to be replaced by the matrices ζ_0^{ij} and ζ_1^{ij}, where the matrix elements are determined from δ_{ij} and γ_{ij} as in (17) and (18).

There have been various recent improvements to the basic mean field method that we outlined, which are mostly based on more general convexity properties [42]. These can also be applied to the associative memory task and, by improving the quality of the approximation to the distribution, should improve recall. The price is mild in terms of computational complexity.

B. Summary and Conclusion

The original Willshaw model was designed as a one-shot, heteroassociative memory rather than as an iterative, autoassociative memory. As an autoassociator, it is hampered, since its Lyapunov function contains a counterproductive constraint term which comes from the threshold. If this threshold is kept constant, then the network is encouraged to recall just one of two patterns—with all components being either on or off. Linear threshold strategies [30], [15], [31], for which the threshold depends linearly on the previous activity, merely eliminate the influence of this harmful constraint.

Along with [13], we considered retrieval of information from an autoassociative memory as a prime case for probabilistic inference. Our retrieval strategies were determined explicitly from prior information about the the mean activity of the stored patterns, and the error rates in the initial pattern and in the synaptic matrix. The resulting iterative schemes are close to those that have been suggested in the literature, and so provide interpretation for (and suggestions for the values of) their otherwise purely heuristic parameters.

Compared with the Willshaw model the Lyapunov function derived from MAP inference from the posterior probability contains additional constraint terms. In particular, it includes a constant antiferromagnetic interaction and a constant, but site dependent threshold strategy. The choice $\alpha \neq -1$ in (14) introduces an antiferromagnetic interaction (and a constant threshold) in the Willshaw model which has heuristically been proposed earlier [43]. The terms $\beta_1, \beta_2 > 0$ in (19) introduce a threshold which consists of both a constant component and a site dependent component that is aligned with the initial pattern. In the Hopfield model, allowing the initial pattern to have a persistent influence has been shown to improve the retrieval in experiments [37], [33]–[35], [38], [39].

For pattern completion in the Hopfield model, connections with maximum entropy inference have been explored by MacKay [13]. His approach does not explain the antiferromagnetic terms, but also calls for the introduction of constant and site dependent thresholds given knowledge about the activities of the training patterns and the similarity between the testing pattern and the possible training patterns.

We also developed a mean field treatment of the probabilistic inference process, using a parameterized but tractable approximation to the true posterior distribution. Retrieval in this model requires changing the values of the parameters to reduce the consequent Kullback-Leibler divergence between approximate and true distributions. This leads to iterative mean field equations that can be interpreted as a neural network with similar interconnections and threshold values as prescribed by the modified MAP update equation, but with sigmoid neural response functions. The mean field equations should thereby reduce the number of spurious states [41].

ACKNOWLEDGMENT

The authors are grateful to five anonymous referees for valuable comments and suggestions as to how to improve the manuscript.

REFERENCES

[1] K. Steinbuch, "Die Lernmatrix," *Kybernetik*, vol. 1, pp. 36–45, 1961.
[2] D. J. Willshaw, O. P. Buneman, and H. C. Longuet-Higgins, "Nonholographic associative memory," *Nature*, vol. 222, pp. 960–962, 1969.
[3] W. A. Wickelgren, "Context-sensitive coding, associative memory and serial order in (speech) behavior," *Psych. Rev.*, vol. 76, pp. 1–15, 1969.
[4] K. Nakano, "Associatron—A model of associative memory," *IEEE Trans. Syst., Man. Cybern.*, vol. SMC-2, pp. 380–388, 1972.
[5] S.-I. Amari, "Learning patterns and pattern sequences by self-organizations of threshold elements," *IEEE Trans. Comput.*, vol. C-21, pp. 1197–1206, 1972.
[6] T. Kohonen, "Correlation matrix memory," *IEEE Trans. Comput.*, vol. C-21, pp. 353–359, 1972.
[7] ——, *Self-Organization and Associative Memory*. Berlin, Germany: Springer-Verlag, 1983.
[8] H. J. Bentz, M. Hagström, and G. Palm, "Information storage and effective data retrieval in sparse matrices," *Neural Networks*, vol. 2, pp. 289–293, 1989.
[9] D. J. Amit, *Modeling Brain Function*. Cambridge, U.K.: Cambridge Univ. Press, 1989.
[10] ——, "The Hebbian paradigm reintegrated: Local reverberations as internal representations," *Behavioral and Brain Sci.*, vol. 18, pp. 617–657, 1995.
[11] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Nat. Academy Sci., USA*, vol. 79, 1982.
[12] U. Rückert, and H. Surmann, "Tolerance of a binary associative memory toward stuck-at-faults," in *Artificial Neural Networks*, T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, Eds. New York: Elsevier, 1991.
[13] D. J. C. MacKay, "Maximum entropy connections: Neural networks," in *Maximum Entropy and Bayesian Methods*, W. T. Grandy and L. Schick, Eds. Boston, MA: Kluwer, 1991.
[14] D. O. Hebb, *The Organization of Behavior*. New York: Wiley, 1949.
[15] F. Schwenker, F. T. Sommer, and G. Palm, "Iterative retrieval of sparsely coded associative memory patterns," *Neural Networks*, vol. 9, no. 3, pp. 445–455, 1996.
[16] G. Palm, "Memory capacities of local rules for synaptic modification," *Concepts in Neurosci.*, vol. 2, pp. 97–128, 1991.
[17] G. Palm and F. T. Sommer, "Information capacity in recurrent McCulloch–Pitts networks with sparsely coded memory states," *Network*, vol. 3, pp. 1–10, 1992.
[18] M. V. Tsodyks and M. V. Feigelman, "The enhanced storage capacity in neural networks with low activity level," *Europhys. Lett.*, vol. 6, pp. 101–105, 1988.
[19] R. J. McEliece, E. C. Posner, E. R. Rodemich, and S. S. Venkatesh, "The capacity of the hopfield associative memory," *IEEE Trans. Inform. Theory*, vol. 33, pp. 461–482, 1987.
[20] C. Mazza, "On the storage capacity of nonlinear neural networks," *Neural Networks*, vol. 10, no. 4, pp. 593–597, 1997.
[21] D. Amit, H. Gutfreund, and H. Sompolinsky, "Statistical mechanics of neural networks near saturation," *Ann. Phys.*, vol. 173, pp. 30–67, 1987.
[22] M. Morita, "Associative memory with nonmonotone dynamics," *Neural Networks*, vol. 6, pp. 115–126, 1993.
[23] H.-F. Yanai and S.-I. Amari, "Autoassociative memory with two-stage dynamics of nonmonotonic neurons," *IEEE Trans. Neural Networks*, vol. 7, pp. 803–815, 1996.

[24] P. Whittle, "Artificial memories: Capacities, basis rate and inference," *Neural Networks*, vol. 10, no. 9, pp. 1619–1626, 1997.
[25] E. Gardner, "Structure of metastable states in the Hopfield model," *J. Phys. A*, vol. 19, pp. L1047–L1052, 1986.
[26] ———, "The space of interactions in neural-network models," *J. Phys. A*, vol. 21, pp. 257–270, 1988.
[27] J. Buckingham and D. Willshaw, "On setting unit thresholds in an incompletely connected associative net," *Network*, vol. 4, pp. 441–459, 1993.
[28] B. Graham and D. Willshaw, "Improving recall from an associative memory," *Biol. Cybern.*, vol. 72, pp. 337–346, 1995.
[29] A. R. Gardner-Medwin, "The recall of events through the learning of associations between their parts," in *Proc. Roy. Soc. London B*, vol. 194, 1976, pp. 375–402.
[30] W. G. Gibson and J. Robinson, "Statistical analysis of the dynamics of a sparse associative memory," *Neural Networks*, vol. 5, pp. 645–662, 1992.
[31] H. Hirase and M. Recce, "A search for the optimal thresholding sequence in an associative memory," *Network*, vol. 7, no. 4, pp. 741–756, 1996.
[32] D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Information storage in neural networks with low levels of activity," *Phys. Rev. A*, vol. 35, no. 5, pp. 2293–2303, 1987.
[33] A. Engel, H. Englisch, and A. Schütte, "Improved retrieval in neural networks with external fields," *Europhys. Lett.*, vol. 8, pp. 393–399, 1989.
[34] D. J. Amit, G. Parisi, and S. Nicolis, "Neural potentials as stimuli for attractor neural networks with low levels of activity," *Network*, vol. 1, pp. 75–88, 1990.
[35] H. W. Yau and D. J. Wallace, "Enlarging the attractor basins of neural networks with noisy external fields," *J. Phys. A (Math. General)*, vol. 24, no. 23, pp. 5639–5650, 1991.
[36] I. Kanter and H. Sompolinsky, "Associative recall of memory without errors," *Phys. Rev. A*, vol. 35, no. 1, pp. 380–392, 1987.
[37] G. R. Gindi, A. F. Gmitro, and K. Parthasarathy, "Hopfield model associative memory with nonzero-diagonal terms in the memory matrix," *Appl. Opt.*, vol. 27, pp. 129–134, 1988.
[38] K. Araki and T. Saito, "An associative memory including time-invariant self-feedback," *Neural Networks*, vol. 7, no. 8, pp. 1267–1271, 1994.
[39] P. De Wilde, "The magnitude of the diagonal elements in neural networks," *Neural Networks*, vol. 10, no. 3, pp. 499–504, 1996.
[40] R. M. Neal and G. E. Hinton, "A new view of the EM algorithm that justifies incremental and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed.. Boston, MA: Kluwer, 1998.
[41] J. J. Hopfield, "Neurons with graded responses have collective computational properties like those of two-state neurons," in *Proc. Nat. Academy Sci., USA*, vol. 81, 1984, reprinted in [44].
[42] M. I. Jordan, Z. Ghahramani, T. S. Jaakola, and L. K. Saul, "An introduction to variational methods for graphical models," in *Learning in Graphical Models*, M. I. Jordan, Ed. Boston, MA: Kluwer, 1998.
[43] D. Golomb, N. Rubin, and H. Sompolinsky, "Willshaw model: Associative memory with sparse coding and low firing rates," *Phys. Rev. A*, vol. 41, pp. 1843–1854, 1990.
[44] J. A. Anderson and E. Rosenfeld, Eds., *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press, 1988.