Neural associative memory and the Willshaw-Palm probability distribution.

Andreas Knoblauch

2008

Preprint:

This is an accepted article published in SIAM Journal on Applied Mathematics. The final authenticated version is available online at: https://doi.org/[DOI not available]

NEURAL ASSOCIATIVE MEMORY AND THE WILLSHAW-PALM PROBABILITY DISTRIBUTION *

ANDREAS KNOBLAUCH†

Abstract. Previous asymptotic analyses of binary neural associative networks of the Willshaw or Steinbuch type relied on a binomial approximation of the neurons' dendritic potentials. This approximation has been proven to be good only if the stored patterns are extremely sparse, for example, when the mean number of active units k per pattern vector scales with the logarithm of the vector size n. Recent promising results concerning storage capacity and retrieval efficiency for larger pattern activities $k > \log n$ have been doubted because here the binomial approximation can lead to a massive overestimation of performance. In this work I compute and characterize the exact Willshaw-Palm distribution of the dendritic potentials for hetero-association, auto-association, and fixed and random pattern activity. Comparing the raw and central moments of the Willshaw-Palm distribution to the moments of the corresponding binomial probability reveals that, asymptotically, the binomial approximation becomes exact for almost any sublinear pattern activity, including $k = O(n/\log^2 n)$. This verifies, for large networks, the existence of a wide high-performance parameter range as predicted by the approximative theory.

Key words. neural network, Willshaw model, information retrieval, storage capacity, fault tolerance

AMS subject classifications. 92B20, 68T10, 60C05

1. Introduction. *Associative memories* are systems that contain information about a finite set of associations between pattern vector pairs $\{(\mathbf{u}^\mu \mapsto \mathbf{v}^\mu) : \mu = 1, ..., M\}$, where $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ are called *address* and *content* patterns, respectively [28]. Given a possibly noisy address pattern $\tilde{\mathbf{u}}$ the problem is to find a content pattern $\mathbf{v}^\mu$ for which the corresponding address pattern $\mathbf{u}^\mu$ is most similar to $\tilde{\mathbf{u}}$. This is a variant of the *Best Match Problem* in [31] and efficient solutions have widespread applications including object recognition and information retrieval [28, 36, 40, 3, 13, 20, 32, 42].

In *neural network implementations* the information about the associations is stored in the synaptic connectivity of one or more neuron populations [46, 16, 17, 37]. Besides the potential for technical applications, neural associative memories also play an important role in many *brain theories* (e.g., [14, 30, 5, 35, 16, 17, 11, 12, 27, 10, 15]), where the patterns correspond to attractors in the brain's neuronal state space.

One of the most efficient networks is the so-called Willshaw or Steinbuch model with binary neurons and synapses [44, 46, 34, 33, 8, 43]. In particular, it has been shown that the Willshaw model has a very high asymptotic storage capacity of $C = 0.7$ bits per synapse which exceeds the capacity of most alternative models. For example, the original Hopfield model achieves only $C = 0.14$ bps [16, 1, 2]. In general the classical work points out that high capacities can be obtained only if the stored patterns are extremely sparse, for example, when the mean number of active units k per pattern vector scales logarithmic with the vector size n.

For a number of reasons, a regime of larger pattern activity with $k/\log n \to \infty$ has recently gained increased attention: First, logarithmic $k = \log n$ is simply too sparse for many applications of distributed representations [41, 45, 42]. Second, activity patterns with extremely sparse activity $k \sim \log n$ appear inconsistent with neurophysiology because they are difficult to stabilize in a noisy regime where neurons

*Published in: SIAM Journal on Applied Mathematics 69(1), pp 169–196, 2008.

†Honda Research Institute Europe, Carl-Legien-Strasse 30, D-63073 Offenbach/Main, Germany (andreas.knoblauch@honda-ri.de).

have high rates of spontaneous activity [29]. Third, it has been argued that $k/\log n \to \infty$ can actually lead to a massive increase in storage capacity and retrieval efficiency if the network structures are adequately compressed ([22]; see also [18, 19, 20]). Fourth, $k/\log n \to \infty$ allows an efficient inhibitory implementation of the Willshaw model which implies new interpretations for inhibitory circuits in the brain [22, 24].

However, the viability of this regime with moderately sparse patterns, $k/\log n \to \infty$, has been doubted. On the one hand, here the established theories on Willshaw or Steinbuch type networks with fixed connectivity structure predict only a very low performance, for example zero storage capacity per synapse, such that both technical applicability and biological relevance seem unlikely . On the other hand, the extended theory considering structural changes and inhibitory implementations predicts high performance, but relied, similar to the established theories, on a *binomial approximation* of the neurons' dendritic potentials (e.g., [46, 34, 37, 33, 4, 43, 20]). This approximation may be inaccurate for large pattern activities $k \gg \log n$ and thus the corresponding high-performance regime illusory. Indeed, the convergence of the binomial approximation to the true potential distribution and thus the asymptotic correctness of the theory has been demonstrated only for some special cases involving very sparse activity patterns, where a binary pattern vector of n neurons contains only $k = \log n$ or $k \leq n^{1/3}$ active units [34, 38]. Another analysis showed that the binomial approximation becomes very inaccurate for linear $k \sim n$ [19, 21]. However, it remained unclear for precisely which $k(n)$ the binomial approximation converges to the true potential distribution.

In this work I have solved this problem. Section 2 gives an overview of the Willshaw model and the analysis employing the binomial approximation of the dendritic potentials. Section 3 then defines and computes the exact Willshaw-Palm distribution of the dendritic potentials which can be used to determine exact retrieval error probabilities and storage capacity. Section 4 characterizes the Willshaw-Palm probability by computing the raw and central moments. Finally, section 5 compares the Willshaw-Palm probability to the binomial probability and determines asymptotic conditions when the two probability distributions become identical.

2. Binary associative networks.

2.1. Learning and retrieving patterns. An attractive model of neural associative memory both for biological modeling and applications is the so-called Willshaw or Steinbuch model with binary neurons and synapses [46, 44, 34, 33, 37, 7, 4, 43, 20] illustrated in Fig. 2.1. Each address pattern $\mathbf{u}^\mu$ is a binary vector of length m containing k one-entries and $m-k$ zero-entries. Similarly, each content pattern $\mathbf{v}^\mu$ is a binary vector of length n containing l one-entries and $n-l$ zero-entries. Typically, the patterns are sparse, i.e., $k \ll m$ and $l \ll n$. For our analysis of storage capacity we will further assume that each pattern is randomly drawn from the sets of the $\binom{m}{k}$ potential address patterns and the $\binom{n}{l}$ potential content patterns.

The M pattern pairs are stored *hetero-associatively* in a binary *memory matrix* $\mathbf{A} \in \{0,1\}^{m\times n}$ with

$$A_{ij} = \min\left(1, \tilde{A}_{ij} + \sum_{\mu=1}^{M} u_i^\mu \cdot v_j^\mu\right) \in \{0,1\}\,, \tag{2.1}$$

where $\tilde{\mathbf{A}}$ is a binary noise matrix with each component being active independently with probability $\tilde{p}_1$.

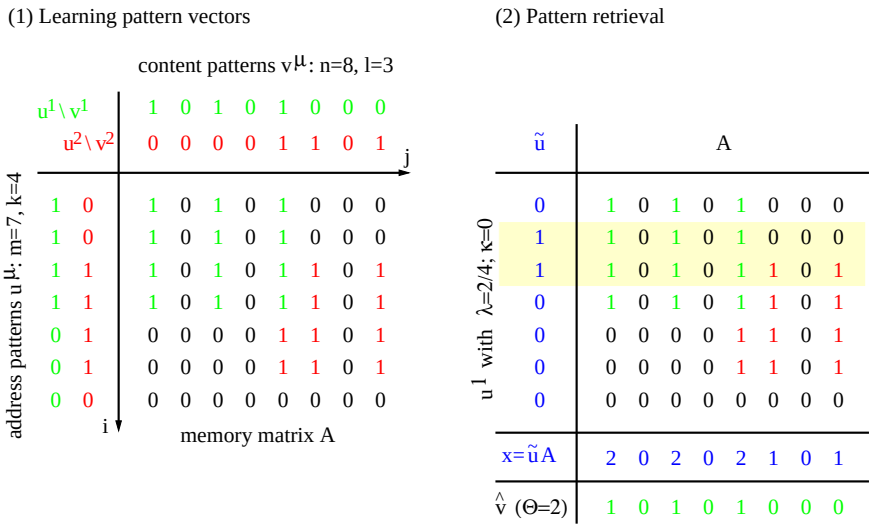

FIG. 2.1. *Example of the binary Willshaw associative memory for hetero-association.* **Left:** *During learning M associations between address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$ are stored in the binary memory matrix $\mathbf{A}$ representing binary synaptic weights of the connection from neuron population u to v. Initially all synapses are inactive ($\tilde{p}_1 = 0$). During learning of pattern associations, the synapses are activated according to Hebbian coincidence learning (eq. 2.1).* **Right:** *For retrieval an address pattern $\tilde{\mathbf{u}}$ is propagated through the network. Vector-matrix-multiplication yields the dendritic potentials $\mathbf{x} = \tilde{\mathbf{u}}\mathbf{A}$. To obtain the retrieval result $\hat{\mathbf{v}}$ (here equal to $\mathbf{v}^1$) a threshold Θ is applied. For pattern part retrieval with $\tilde{\mathbf{u}} \subseteq \mathbf{u}^\mu$ we can simply choose the Willshaw threshold $\Theta = |\tilde{\mathbf{u}}|$. Then the retrieval output is a superset of the original pattern, $\hat{\mathbf{v}} \supseteq \mathbf{v}^\mu$, that means $\hat{\mathbf{v}}$ contains no miss-errors.*

The *neural interpretation* is that of two neuron populations, an address population u consisting of m neurons and a content population v consisting of n neurons. The patterns $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ describe the activity states of the two populations at time μ, and A_{ij} is the strength of the Hebbian learned synaptic connection from neuron u_i to neuron v_j. Positive $\tilde{p}_1$ can be used to model noisy synaptic potentiation (e.g., the synapses that are already active before learning starts), noisy synaptic transmission, or incomplete connectivity [22, 23, 24].

Besides the feed-forward interpretation, the Willshaw model can also be used to model *auto-association* or pattern completion where address population content population are identical, $u = v$, and consequently also $\mathbf{u}^\mu = \mathbf{v}^\mu$. Here the memory matrix $\mathbf{A}$ describes the recurrent synaptic connectivity within the neuron population.

For independently generated random patterns, there is a simple relation between the number M of stored associations and the so-called *memory load* p_1 defined as the fraction of one-entries in the memory matrix. The probability that a synapse is *not* activated by the association of one pattern pair is $1 - kl/mn$, therefore after learning M pattern associations,

$$p_1 = 1 - (1 - \tilde{p}_1)\left(1 - \frac{kl}{mn}\right)^M \geq \tilde{p}_1, \tag{2.2}$$

$$M = \frac{\ln \frac{1-p_1}{1-\tilde{p}_1}}{\ln(1 - kl/mn)} \approx -\frac{mn}{kl} \ln \frac{1 - p_1}{1 - \tilde{p}_1}, \tag{2.3}$$

where the approximation is valid for $kl \ll mn$. As we will see, the memory load p_1

will play an important role both for the exact analysis of the Willshaw model and for the binomial approximative analysis.

After learning, the stored information can be retrieved applying an address pattern $\tilde{\mathbf{u}}$. Vector-matrix-multiplication yields the dendritic potentials $\mathbf{x} = \tilde{\mathbf{u}}\mathbf{A}$ of the content neurons, and imposing a threshold Θ gives the (one-step) retrieval result $\hat{\mathbf{v}}$,

$$\hat{v}_j = \begin{cases} 1\,, & x_j = (\sum_{i=1}^m \tilde{u}_i A_{ij}) \geq \Theta \\ 0\,, & \text{otherwise} \end{cases} . \tag{2.4}$$

Choosing $\Theta = z := \sum_{i=1}^m \tilde{u}_i$ will be referred to as the *Willshaw threshold* and plays an important role both for more realistic *spiking* neuron networks [26, 19] and also for pattern part retrieval with $\tilde{\mathbf{u}} \subseteq \mathbf{u}^\mu$ as analyzed in section 2.3.

2.2. Retrieval errors and storage capacity. We have retrieval errors if the retrieval result $\hat{\mathbf{v}}^\mu$ is not identical to the originally learned pattern $\mathbf{v}^\mu$. For a closer analysis we can divide the neurons of the content population into two groups: The *lo-units* which correspond to the $n - l$ zero-entries of $\mathbf{v}^\mu$, and the *hi-units* which correspond to the l one-entries of $\mathbf{v}^\mu$. For an error-free retrieval result $\hat{\mathbf{v}}^\mu$ the potentials $\mathbf{x}$ of lo- and hi-units must be separable, i.e., the largest potential of a lo-unit must be smaller than the smallest potential of a hi-unit. If the two potential distributions have overlap two kinds of retrieval errors can occur. An *add-error* occurs if the potential of a lo-unit is above threshold Θ, and a *miss-error* occurs if the potential of a hi-unit is below threshold. If the probability distribution of a lo-unit i is known we can compute the probability p_{01} of an *add-error*. Similarly, for a hi-unit j we can compute the probability p_{10} of a *miss-error*. With $z = |\tilde{\mathbf{u}}|$ being the activity of the address pattern, we have

$$p_{01} = \mathrm{pr}(\hat{v}_i = 1 | v_i^\mu = 0) = \sum_{x=\Theta}^{z} \mathrm{pr}[x_i = x] \tag{2.5}$$

$$p_{10} = \mathrm{pr}(\hat{v}_j = 0 | v_j^\mu = 1) = \sum_{x=0}^{\Theta-1} \mathrm{pr}[x_j = x] \ . \tag{2.6}$$

Thus, the expected *Hamming distance* $h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) := \sum_{j=1}^n |v_j^\mu - \hat{v}_j^\mu|$ between learned and retrieved pattern is

$$Eh(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) = (n - l)p_{01} + lp_{10} \tag{2.7}$$

To enforce retrieval quality we bound the expected Hamming distance to be no more than a fraction ϵ of the content pattern activity l. Thus, we require

$$(n - l)p_{01} + lp_{10} \leq \epsilon l \tag{2.8}$$

where retrieval quality parameter ϵ is typically a small positive constant (e.g., $\epsilon = 0.01$). Because the minimal Hamming distance (optimizing Θ) is obviously increasing with M, we can finally define the *pattern capacity* M_ϵ

$$M_\epsilon := \max\{M : (n - l)p_{01} + lp_{10} \leq \epsilon l\} \tag{2.9}$$

being the maximal number of storable pattern associations fulfilling the retrieval quality requirement eq. 2.8. Considering the Shannon information of individual content patterns, we get the normalized *network storage capacity* in bits per synapse,

$$C_\epsilon := \frac{M_\epsilon T(\mathbf{v}^\mu; \hat{\mathbf{v}}^\mu)}{mn} \tag{2.10}$$

where $T(\mathbf{v}^\mu; \hat{\mathbf{v}}^\mu)$ is the transinformation (or mutual information) between learned and retrieved content pattern [9]. From the network capacity we can derive further performance measures such as *information capacity* C^I and *synaptic capacity* C^S making use of the compressibility of the memory matrix for a memory load $p_1 \neq 0.5$ (see section 2.4; for more details see [18, 19, 20, 22]).

2.3. Sketch of the binomial approximative analysis for random patterns. The approximative analysis of the Willshaw model relies on the assumption that the one-entries in the memory matrix are generated independently of each other. Although obviously not true for distributed patterns, this assumptions leads to seminal insights into the Willshaw model and, at least for certain parameter ranges, quite good approximations of the actual storage capacity (see section 3; see [22]).

Let us again assume that the retrieval address pattern $\tilde{u}$ contains $c = \lambda k$ correct and $f = \kappa k$ false one-entries of address pattern u^μ previously used for learning ($0 < \lambda \leq 1$, $\kappa \geq 0$). Assuming $\mathrm{pr}[A_{ij} = 1] = p_1$ independently of i, j, the dendritic potentials x_{lo} of a lo-unit and x_{hi} of a hi-unit are binomially distributed (eq. A.2),

$$\mathrm{pr}[x_{\mathrm{lo}} = x] = p_B(x; c+f, p_1), \quad x = 0, 1, \ldots, c+f \tag{2.11}$$

$$\mathrm{pr}[x_{\mathrm{hi}} = x] = p_B(x-c; f, p_1), \quad x = c, c+1, \ldots, c+f \ . \tag{2.12}$$

For purposes of clarity, in the following we restrict the analysis to the case of *pattern part retrieval* where the address pattern contains no add noise, that is, $f = 0$. For the general analysis see [43]. Here one can apply the Willshaw threshold $\Theta = c$ which will limit the retrieval errors to add noise. Thus, the retrieval error probabilities are

$$p_{01} = p(\hat{v}_i = 1 | v_i^\mu = 0) \approx {p_1}^{\lambda k}. \tag{2.13}$$

and $p_{10} = 0$. To enforce retrieval quality as described above (see eq. 2.8) we have to bound the error probability p_{01} by $p_{01\epsilon}$,

$$p_{01} \leq p_{01\epsilon} := \frac{\epsilon l}{n-l} \ . \tag{2.14}$$

The number of patterns that can be stored is limited to the point where $p_{01} = p_{01\epsilon}$ or, equivalently, where the memory load reaches

$$p_{1\epsilon} \approx \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda \cdot k}} \qquad \left(\Leftrightarrow k \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{\lambda \mathrm{ld} p_{1\epsilon}}\right) \ . \tag{2.15}$$

From eq. 2.3 we obtain the maximal number of stored patterns or *pattern capacity*

$$M_\epsilon \approx -\lambda^2 \cdot (\mathrm{ld} p_{1\epsilon})^2 \cdot \ln \frac{1-p_{1\epsilon}}{1-\tilde{p}_1} \cdot \frac{k}{l} \cdot \frac{mn}{(\mathrm{ld}\frac{n-l}{\epsilon \cdot l})^2}. \tag{2.16}$$

With this result we can also estimate the network capacity (eq. 2.10)

$$C_\epsilon = \frac{M_\epsilon T(l/n, p_{01\epsilon}, 0)}{m} \approx \lambda \cdot \mathrm{ld} p_{1\epsilon} \cdot \ln \frac{1-p_{1\epsilon}}{1-\tilde{p}_1} \cdot \eta \tag{2.17}$$

where $T(p, p_{01}, p_{10})$ is the transinformation (or mutual information) of a binary channel (see eq. A.1, [9]) and

$$\eta := \frac{T\left(\frac{l}{n}, \frac{\epsilon l}{n-l}, 0\right)}{-\frac{l}{n}\mathrm{ld}\frac{\epsilon l}{n-l}} \approx \frac{1}{1 + \frac{\ln \epsilon}{\ln(l/n)}}. \tag{2.18}$$

The approximation is valid for small $\epsilon, l/n \ll 1$ when $T \approx -(l/n)\mathrm{ld}(l/n)$: In that case $\eta \to 1$ for large $n \to \infty$. For $p_{1\epsilon} = 0.5$ and $\tilde{p}_1 = 0$ we have therefore $C_\epsilon \to \ln 2 \approx 0.69$ bits per synapse, the asymptotic storage capacity of the Willshaw model [46, 34, 43, 22]. Note that C_ϵ increases by factor $1/(1-\tilde{p}_1)$ if $1-\tilde{p}_1$ is interpreted as network connectivity (i.e., the chance that a potential synapse is actually realized; see [22, 8, 4]).

2.4. The asymptotic regimes of sparse and dense potentiation. The main conclusions from the binomial approximative analysis are that a very high storage capacity of almost 0.7 bits per synapse can be achieved for sparse patterns with $k \sim \log n$ and memory load $p_1 = 0.5$. Then we can store on the order of $M \sim mn/(\log n)^2$ pattern associations with high retrieval quality. From eqs. 2.15,2.17 it is easy to see that asymptotically

$$C_\epsilon > 0 \Leftrightarrow k \sim \log n \Leftrightarrow 0 < p_{1\epsilon} < 1. \tag{2.19}$$

Thus, the analysis suggests that neural associative memory is efficient ($C_\epsilon > 0$) only for logarithmically sparse patterns. For sub-logarithmic sparse patterns with $k/\log n \to 0$ we have $p_{1\epsilon} \to 0$ and for supra-logarithmic sparseness with $k/\log n \to \infty$ we have $p_{1\epsilon} \to 1$, both cases implying vanishing network storage capacity $C_\epsilon \to 0$. These results bear importance for both technical applications and biology, in particular with respect to the sparseness of postulated Hebbian cell assemblies in the real brain [14, 5, 35]. In the following we will refer to the three cases $p_{1\epsilon} \to 0/c/1$ as sparse, balanced, and dense synaptic potentiation, respectively.

I have argued elsewhere that these conclusion may be biased by the definition of network storage capacity, and that alternative definitions of storage capacity considering the compressibility of the network lead to different conclusions [18, 19, 20, 22]. For example, in technical implementations of the Willshaw model the memory matrix can be compressed for $p_1 \to 0/1$ and the storage capacity improves by factor $I(p_1) := -p_1\mathrm{ld}p_1 - (1-p_1)\mathrm{ld}(1-p_1)$. Similar arguments hold for biological networks where "compression" could be realized by synaptic pruning and structural plasticity (see [22] for more details). This has led to the definition of information capacity $C^I_\epsilon := C_\epsilon/I(p_{1\epsilon})$ and synaptic capacity $C^S_\epsilon := C_\epsilon/\min(p_{1\epsilon}, 1-p_{1\epsilon})$. Interestingly, and in contrast to network capacity C_ϵ, optimizing C^I_ϵ and C^S_ϵ reveals highest capacities for $p_{1\epsilon} \to 0$ and $p_{1\epsilon} \to 1$. Here, presuming the validity of the binomial theory, technical implementations could fully exploit the physical memory by storing $C^I_\epsilon \to 1$ bit information per memory bit. Similarly, biological networks could improve storage capacity to arbitrary large values $C^S_\epsilon \sim \log n \to \infty$ bit per synapse. By these results, the regimes with ultra-sparse and moderately sparse patterns (or cell assemblies) have gained increased attention. However, the convergence of the binomial approximations towards the exact values is questionable since this has been strictly proven only for some special conditions including $k \sim \log n$ [34, 38]. In particular, for dense potentiation with $p_{0\epsilon} = 1 - p_{1\epsilon} \to 0$, supra-logarithmic sparseness, $k/\log n \to \infty$, and

$$p_{1\epsilon} = (\frac{\epsilon l}{n-l})^{1/\lambda k} = e^{\frac{\ln(\epsilon l/(n-l))}{\lambda k}} \approx 1 - \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k}, \tag{2.20}$$

numerical simulations of the Willshaw model reveal that the real capacities can be massively overestimated by the binomial approximative analysis [22]. Therefore, in the following we conduct an exact analysis of the Willshaw model based on the exact potential distributions, and investigate conditions when the binomial probability distribution becomes a good approximation of the Willshaw-Palm distribution.

3. The Willshaw-Palm distribution of the dendritic potentials. For an exact analysis of the Willshaw model we have to compute the distribution of the neurons' dendritic potentials, i.e., the probability $\mathrm{pr}[X = x]$ that the potential X of a given hi- or lo-unit equals a certain value x (see eqs. 2.5,2.6). This probability distribution is also called *Willshaw-Palm distribution* for random pattern associations and random retrieval address pattern. In the following more formal definition we take into account different ways to generate random patterns.

DEFINITION 3.1. *(Willshaw-Palm probability) Let A be the memory matrix of a Willshaw associative memory after learning M random pattern associations and with synaptic noise $\tilde{p}_1$ as described in section 2.1. The associations are between address patterns u^μ with size m and mean activity k, and content patterns v^μ with size n and mean activity l ($\mu = 1, 2, \ldots, M$). Further let $\tilde{u}$ be a binary random address pattern with activity $z = |\tilde{u}|$. Then we define the Willshaw-Palm probability as the probability $\mathrm{pr}[(\tilde{u}A)_j = x]$ that a given content neuron v_j has potential x when retrieving with $\tilde{u}$. We distinguish between four relevant versions of the Willshaw-Palm probability depending on the generation of the random patterns:*

1. *$p_{\mathrm{Ph}}(x; k, l, m, n, M, \tilde{p}_1, z)$ for fixed address activity and hetero-association.*
2. *$p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, \sigma)$ for fixed address activity and auto-association.*
3. *$p_{\mathrm{Wh}}(x; k, l, m, n, M, \tilde{p}_1, z)$ for random address activity and hetero-association.*
4. *$p_{\mathrm{Wa}}(x; k, n, M, \tilde{p}_1, z, \sigma)$ for random address activity and auto-association.*

Auto-association means that address patterns and content patterns are identical, $u^\mu = v^\mu$. Fixed address activity means that each address pattern has exactly k active units. Random address activity means that a component of an address pattern is active, $u_i^\mu = 1$, with probability k/m independently of other components. For the hetero-associative cases, the content patterns can have either fixed activity l or random activity with mean l. The auto-associative cases require an additional parameter $\sigma := \mathrm{pr}[j \in \tilde{u}]$ denoting the probability that neuron j is among the z active address units.

We sometimes denote p_{W} briefly as the *Willshaw probability* since p_{Wh} has first been determined by Buckingham and Willshaw [7, 6]. Similarly, we denote p_{P} briefly as the *Palm probability* since some special cases of p_{Ph} have first been determined by Palm [34]. Note that the difference between the two variants is that the Palm model has address patterns with fixed activity and the Willshaw model has address patterns with fixed mean.

THEOREM 3.2. *The four Willshaw-Palm probabilities p_{Ph}, p_{Pa}, p_{Wh}, p_{Wa} are given by eqs. 3.22,3.34,3.39,3.41, respectively.*

The proof of the theorem follows in the next four subsections each determining one version of the Willshaw-Palm probability and the corresponding retrieval error probabilities.

3.1. Fixed pattern activity and hetero-association. Here we will determine the Willshaw-Palm probability $p_{\mathrm{Ph}}(x; k, l, m, n, M, z)$ of Def. 3.1. For brevity we identify patterns with sets of one-entries, e.g., $\mathbf{u} = 011001$ is identified with the index set $\mathbf{u} = \{2, 3, 6\}$. Generalizing Palm's definition of a predicate or condition C (see appendix 1 in [34]) for index sets Y, N ("yes!" and "no!") let

$$C(Y, N, j) := [\forall i \in Y : A_{ij} = 1] \cap [\forall i \in N : A_{ij} = 0] \tag{3.1}$$

i.e., condition $C(Y, N, j)$ means that content neuron j receives inputs from the subset Y of address pattern $\tilde{\mathbf{u}}$, but no input from subset N. We further assume that Y and N are disjunct random sets unrelated to the M stored pattern pairs. For Y equal to a further $M + 1$-th address pattern, i.e., $Y = \mathbf{u}^{M+1}$, the condition $C(Y, \emptyset, j)$ would

coincide with the definition of C in the appendix of [34]. Then C would be equivalent to the occurrence of an add-error at lo-unit j for retrieval with the noise-free address pattern $\tilde{\mathbf{u}} = \mathbf{u}^{M+1}$. We first compute the probability that $C(Y, \emptyset, j)$ holds after storing M pattern associations. Contrary to [34] we assume that $Y \subseteq \{1, \ldots, m\}$ is an arbitrary *subset* of address units unrelated to the M stored pattern associations.

$$\mathrm{pr}(C(Y,\emptyset,j)) = \mathrm{pr}([\forall i \in Y : A_{ij} = 1]) = 1 - \mathrm{pr}([\exists i \in Y : A_{ij} = 0]) \tag{3.2}$$

$$= 1 - \mathrm{pr}(\bigcup_{i \in Y} [A_{ij} = 0]) = 1 - \mathrm{pr}(\bigcup_{i=1}^{|Y|} [A_{ij} = 0]) \tag{3.3}$$

$$= 1 - \sum_{s=1}^{|Y|} (-1)^{s+1} \sum_{1 \le i_1 < \ldots < i_s \le |Y|} \mathrm{pr}(\bigcap_{h=1}^{s} [A_{i_h j} = 0]) \tag{3.4}$$

$$= 1 - \sum_{s=1}^{|Y|} (-1)^{s+1} \binom{|Y|}{s} \mathrm{pr}(\bigcap_{i=1}^{s} [A_{ij} = 0]) \tag{3.5}$$

For eq. 3.4 we used the formula of Sylvester-Poincaré eq. A.6. Note that for random patterns the probabilities that a given subcolumn of $\mathbf{A}$ has at least one zero-entry (eq. 3.3) or only zero-entries (see eq. 3.5) depend only on the subcolumn's size, but not on the specific indices. The latter probability writes

$$\mathrm{pr}(\bigcap_{i=1}^{s} [A_{ij} = 0]) = \mathrm{pr}(\bigcap_{i=1}^{s} [\tilde{A}_{ij} = 0] \cap \bigcap_{\mu=1}^{M} [1, \ldots, s \notin \mathbf{u}^{\mu} \vee j \notin \mathbf{v}^{\mu}]) \tag{3.6}$$

$$= (1 - \tilde{p}_1)^s (\mathrm{pr}[1, \ldots, s \notin \mathbf{u}^1 \vee j \notin \mathbf{v}^1])^M \tag{3.7}$$

where we used the facts that the entries of the noise matrix $\tilde{\mathbf{A}}$ and the address patterns are generated independently of each other and the probability that all entries of a subcolumn remain zero during learning of the μ-th pattern pair is independent of μ. The latter probability writes

$$\mathrm{pr}[1, \ldots, s \notin \mathbf{u}^1 \vee j \notin \mathbf{v}^1] \tag{3.8}$$

$$= \mathrm{pr}([1, \ldots, s \notin \mathbf{u}^1]) + \mathrm{pr}([j \notin \mathbf{v}^1]) - \mathrm{pr}([1, \ldots, s \notin \mathbf{u}^1 \wedge j \notin \mathbf{v}^1]) \tag{3.9}$$

$$= \frac{\binom{m-s}{k}}{\binom{m}{k}} + \frac{\binom{n-1}{l}}{\binom{n}{l}} - \frac{\binom{m-s}{k}\binom{n-1}{l}}{\binom{m}{k}\binom{n}{l}} \tag{3.10}$$

$$= B(m,k,s) + B(n,l,1) - B(m,k,s)B(n,l,1) = 1 - \frac{l(1 - B(m,k,s))}{n} \tag{3.11}$$

where $B(a,b,c) := \binom{a-b}{c} / \binom{a}{c} = \prod_{i=0}^{c-1} (a - b - i)/(a - i) = B(a,c,b)$, see [34] and eq. A.8 in appendix A. Thus

$$\mathrm{pr}(C(Y,\emptyset,j)) = \sum_{s=0}^{|Y|} (\tilde{p}_1 - 1)^s \binom{|Y|}{s} (1 - \frac{l}{n}(1 - B(m,k,s)))^M \tag{3.12}$$

With this result we can finally compute the general case with arbitrary, but disjunct $Y, N = \{N_1, N_2, \ldots\} \subseteq \{1, \ldots, m\}$, $Y \cap N = \emptyset$:

$$\mathrm{pr}(C(Y,N,j)) = \mathrm{pr}(C(Y,\emptyset,j)) - \mathrm{pr}(\bigcup_{i=1}^{|N|} C(Y \cup \{N_i\},\emptyset,j)) \tag{3.13}$$

$$= \mathrm{pr}(C(Y,\emptyset,j)) - \sum_{t=1}^{|N|}(-1)^{t+1} \sum_{1\le i_1<\ldots<i_t\le|N|} \mathrm{pr}(\bigcap_{h=1}^{t} C(Y \cup \{N_{i_h}\},\emptyset,j)) \tag{3.14}$$

$$= \mathrm{pr}(C(Y,\emptyset,j)) - \sum_{t=1}^{|N|}(-1)^{t+1}\binom{|N|}{t}\mathrm{pr}(C(Y \cup \{N_1,\ldots,N_t\},\emptyset,j) \tag{3.15}$$

$$= \sum_{t=0}^{|N|}(-1)^t\binom{|N|}{t}\sum_{s=0}^{|Y|+t}(-1)^s\binom{|Y|+t}{s}(1-\tilde{p}_1)^s(1-\frac{l}{n}(1-B(m,k,s)))^M \tag{3.16}$$

$$= \sum_{s=0}^{|Y|+|N|}(1-\tilde{p}_1)^s(1-\frac{l(1-B(m,k,s))}{n})^M \sum_{\substack{t=\max(0,\\ s-|Y|)}}^{|N|}(-1)^{s+t}\binom{|Y|+t}{s}\binom{|N|}{t} \tag{3.17}$$

$$= \sum_{s=|N|}^{|Y|+|N|}(-1)^{s-|N|}\binom{|Y|}{s-|N|}(1-\tilde{p}_1)^s(1-\frac{l}{n}(1-B(m,k,s)))^M \tag{3.18}$$

where for eq. 3.14 we used again eq. A.6 (Sylvester-Poincaré), and for the last equation we used eq. A.7. Thus, the (Willshaw-)Palm probability for hetero-association is

$$p_{\mathrm{Ph}}(x;k,l,m,n,M,z) = \mathrm{pr}\left(\bigcup_{Y\subseteq\tilde{u},|Y|=x,N=\tilde{u}-Y} C(Y,N,j)\right) \tag{3.19}$$

$$= \binom{z}{x}\mathrm{pr}(C(\{1,\ldots,x\},\{x+1,\ldots z\},j)) \tag{3.20}$$

$$= \binom{z}{x}\sum_{s=z-x}^{z}(-1)^{s-z+x}\binom{x}{s-z+x}(1-\tilde{p}_1)^s(1-\frac{l}{n}(1-B(m,k,s)))^M \tag{3.21}$$

$$= \binom{z}{x}\sum_{s=0}^{x}(-1)^s\binom{x}{s}(1-\tilde{p}_1)^{s+z-x}(1-\frac{l}{n}(1-B(m,k,s+z-x)))^M \tag{3.22}$$

for $0 \le x \le z$ and B as defined below eq. 3.11.

Now we are able to compute exact retrieval error probabilities when addressing with noisy patterns. For example, when addressing with a single address pattern containing c correct and f false one-entries and retrieving with threshold Θ, then the exact retrieval error probabilities p_{01} of a false one-entry and p_{10} of a missing one-entry are

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{Ph}}(x;k,l,m,n,M-1,\tilde{p}_1,c+f) \tag{3.23}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{Ph}}(x-c;k,l,m,n,M-1,\tilde{p}_1,f)\ . \tag{3.24}$$

Note that the situation is as if only $M-1$ patterns were stored since, as a precondition, the pattern to be retrieved does affect neither any of the synapses of a 0-neuron nor any of the synapses connecting add-noise to a 1-neuron.

3.2. Fixed pattern activity and auto-association. The analysis for hetero-association in section 3.1 can be extended to auto-association where address and content population are identical, i.e., $m = n$, $k = l$, and $u^\mu = v^\mu$ (see also appendix 1 in [34]). Here the diagonal matrix elements A_{jj} have a much higher probability

$$\bar{p}_1 = 1 - (1 - \tilde{p}_1)(1 - k/n)^M \tag{3.25}$$

of being activated than non-diagonal elements (cf. eq. 2.2). We use again $C(Y, N, j)$ as defined in eq. 3.1, but now we have to care whether j is contained in Y or N. We first compute the special case $N = \emptyset$ and $j \notin Y$. The analysis for $\mathrm{pr}(C(Y, \emptyset, j \notin Y))$ starts the same way as for the hetero-associative case (see eqs. 3.2-3.7). Instead of eqs. 3.8-3.11 we have to write $\mathrm{pr}([1, \ldots, s \notin u^1 \vee j \notin u^1]) = \mathrm{pr}([1, \ldots, s \notin u^1]) + \mathrm{pr}([j \notin u^1]) - \mathrm{pr}([1, \ldots, s, j \notin u^1]) = B(n, k, s) + B(n, k, 1) - B(n, k, s+1) = 1 - \frac{k}{n}(1 - \frac{n}{n-s} B(n, k, s))$ and therefore

$$\mathrm{pr}(C(Y, \emptyset, j \notin Y)) = \sum_{s=0}^{|Y|} (\tilde{p}_1 - 1)^s \binom{|Y|}{s} (1 - \frac{k}{n}(1 - \frac{n}{n-s} B(n, k, s)))^M \ . \tag{3.26}$$

With this result we can again compute the general case with arbitrary, but disjunct $Y, N = \{N_1, N_2, \ldots\} \subseteq \{1, \ldots, m\}$, $Y \cap N = \emptyset$, but $j \notin Y \cup N$ (cf. eqs.3.13-3.18):

$$\begin{aligned} &\mathrm{pr}(C(Y, N, j \notin Y \cup N)) \\ &= \sum_{s=|N|}^{|Y|+|N|} (-1)^{s-|N|} \binom{|Y|}{s - |N|} (1 - \tilde{p}_1)^s (1 - \frac{k}{n}(1 - \frac{nB(n, k, s)}{n - s}))^M \end{aligned} \tag{3.27}$$

If we presume $N = \emptyset$ and $j \in Y$ then eq. 3.3 becomes $\mathrm{pr}(C(Y, \emptyset, j \in Y)) = 1 - \mathrm{pr}(\bigcup_{i=1}^{|Y|-1} [A_{ij} = 0]) - \mathrm{pr}[A_{jj} = 0](1 - \mathrm{pr}(\bigcup_{i=1}^{|Y|-1} [A_{ij} = 0] | [A_{jj} = 0]))$. Here the first probability on the right side evolves as before except for replacing $|Y|$ by $|Y| - 1$. The conditional probability is $1 - \mathrm{pr}(\cap_{i=1}^{|Y|-1} [A_{ij} = 1] | [A_{jj} = 0]) = 1 - \tilde{p}_1^{|Y|-1}$ because $A_{jj} = 0$ implies that the other synapses of neuron j can be activated only by noise. Thus with $\mathrm{pr}[A_{jj} = 0] = 1 - \bar{p}_1$ we obtain

$$\mathrm{pr}(C(Y, \emptyset, j \in Y)) = \mathrm{pr}(C(Y - \{j\}, \emptyset, j)) - (1 - \bar{p}_1)\tilde{p}_1^{|Y|-1} \tag{3.28}$$

This can be generalized to $N \neq \emptyset$ analogously to eqs. 3.13-3.18. Eq. 3.15 becomes

$$\mathrm{pr}(C(Y, N, j \in Y)) = \sum_{t=0}^{|N|} (-1)^t \binom{|N|}{t} \mathrm{pr}(C(Y \cup \{N_1, \ldots, N_t\}, \emptyset, j \in Y) \ . \tag{3.29}$$

Inserting eq. 3.28 yields two components. The first component equals eq. 3.27 except for replacing $|Y|$ by $|Y| - 1$. The second component is $\sum_{t=0}^{|N|} (-1)^t \binom{|N|}{t} (1 - \bar{p}_1)\tilde{p}_1^{|Y|-1+t} = (1 - \bar{p}_1)\tilde{p}_1^{|Y|-1}(1 - \tilde{p}_1)^{|N|}$ and therefore

$$\begin{aligned} &\mathrm{pr}(C(Y, N, j \in Y)) = -(1 - \bar{p}_1)\tilde{p}_1^{|Y|-1}(1 - \tilde{p}_1)^{|N|} \\ &+ \sum_{s=|N|}^{|Y|+|N|-1} (-1)^{s-|N|} \binom{|Y| - 1}{s - |N|} (1 - \tilde{p}_1)^s (1 - \frac{k}{n}(1 - \frac{nB(n, k, s)}{n - s}))^M \end{aligned} \tag{3.30}$$

We will also need the case $j \in N$. This case implies $A_{jj} = 0$ and therefore any other synapse of neuron j can only be activated by noise. Thus simply

$$\mathrm{pr}(C(Y, N, j \in N)) = (1 - \bar{p}_1)\tilde{p}_1^{|Y|}(1 - \tilde{p}_1)^{|N|-1} \tag{3.31}$$

With this we can finally determine the Palm probability for auto-association. If neuron j does not belong to the z address units then we can proceed as in eqs. 3.19-3.22 and obtain

$$p_{\mathrm{Pa}}(x; k, n, M, z, 0) = \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} (1 - \tilde{p}_1)^{s+z-x} (1 - \frac{k}{n}(1 - \frac{nB(n, k, s + z - x)}{n - z + x - s}))^M \tag{3.32}$$

If neuron j is among the z address units we have to split the union of eq. 3.19 into two disjunct components, $\bigcup_{Y \subseteq \tilde{u}, |Y|=x, N=\tilde{u}-Y, j \in Y} C$ and $\bigcup_{Y \subseteq \tilde{u}, |Y|=x, N=\tilde{u}-Y, j \in N} C$. Then we can proceed again with transformations similar to eqs. 3.19-3.22. With eq. 3.30, the first union corresponds to $p_{\mathrm{Pa}}(x-1; k, n, M, z-1, 0) - \binom{z-1}{x-1}(1-\bar{p}_1)\tilde{p}_1^{x-1}(1-\tilde{p}_1)^{z-x}$ With eq. 3.31, the second union becomes $\binom{z-1}{x}(1 - \bar{p}_1)\tilde{p}_1^x(1 - \tilde{p}_1)^{z-x-1}$. Adding the two components yields

$$p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, 1) = p_{\mathrm{Pa}}(x - 1; k, n, M, \tilde{p}_1, z - 1, 0) + (1 - \bar{p}_1)\left(p_B(x; z - 1, \tilde{p}_1) - p_B(x - 1; z - 1, \tilde{p}_1)\right) \tag{3.33}$$

and thus the general Palm probability for auto-association is

$$p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, \sigma) = (1 - \sigma)p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, 0) + \sigma p_{\mathrm{Pa}}(x; k, n, M, \tilde{p}_1, z, 1)\ . \tag{3.34}$$

When addressing with a single address pattern containing c correct and f false one-entries then $\sigma = f/(n - k)$ for a lo-unit, while $\sigma = 0$ for the f noisy inputs to the hi-units. Thus, retrieving with threshold Θ, the exact retrieval error probabilities p_{01} of a false one-entry and p_{10} of a missing one-entry are

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{Pa}}(x; k, n, M - 1, \tilde{p}_1, c + f, f/(n - k)) \tag{3.35}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{Pa}}(x - c; k, n, M - 1, \tilde{p}_1, f, 0) \tag{3.36}$$

3.3. Random pattern activity and hetero-association. For technical applications, the patterns to be stored have often fixed pattern activities k and l (e.g., see [34, 41, 17, 42]). However, for the biological interpretation we identify the pattern activities with the size of cell assemblies [14, 5, 35], and it seems not very plausible to assume that all cell assemblies had exactly the same size. Here it might be more realistic to assume that an address pattern component is 1 with probability k/m independently of each other (and similarly l/n for the content patterns). Then the mean assembly sizes are still k and l but the size of a given cell assemblies is a binomially distributed random variable.

The analysis can be conducted in analogy to section 3.1. Due to independently generated pattern components eq. 3.8 simplifies to

$$\mathrm{pr}[1, \ldots, s \notin \mathbf{u}^1 \vee j \notin \mathbf{v}^1] = 1 - l/n + (l/n)(1 - k/m)^s \tag{3.37}$$

$$= 1 - \frac{l}{n}(1 - (1 - \frac{k}{m})^s) \tag{3.38}$$

Thus in the further analysis of section 3.1 we can simply replace $B(m,k,s)$ by $(1-k/m)^s$. From eq. 3.22 we finally obtain the *Willshaw probability* for hetero-association

$$\begin{aligned} &p_{\mathrm{Wh}}(x;k,l,m,n,M,\tilde{p}_1,z) \\ &= \binom{z}{x}\sum_{s=0}^{x}(-1)^s\binom{x}{s}(1-\tilde{p}_1)^{s+z-x}(1-\frac{l}{n}(1-(1-\frac{k}{m})^{s+z-x}))^M \qquad (3.39) \\ &= \sum_{i=0}^{M} p_B(i;M,l/n)p_B(x;z,1-(1-\tilde{p}_1)(1-k/m)^i) \qquad (3.40) \end{aligned}$$

for $0 \le x \le z$. The retrieval error probabilities p_{01} and p_{10} are as in eqs. 3.23,3.24 replacing p_{Ph} by p_{Wh}. The second formula eq. 3.40 results from an alternative approach to obtain the Willshaw probability for random pattern activities (see [7, 6]). Here the first binomial is the probability that the considered content neuron has unit-usage i, i.e. that it has been activated i times during the learning of the M associations. Given unit usage i the term $1-(1-\tilde{p}_1)(1-k/m)^i$ is the probability that a given synapse on the content neuron has been potentiated or activated by noise. Thus, the second binomial is the probability that a content neuron receives x out of the z random inputs given a unit usage of i.

Eq. 3.40 for $\tilde{p}_1 = 0$ has been found in 1991 by Buckingham and Willshaw [7, 6], while eq. 3.39 for $\tilde{p}_1 = 0$ has been derived from eq. 3.40 in 1999 by Sommer and Palm [43]. For numerical evaluations eq. 3.39 is particularly useful if z is small and M is large, while evaluating eq. 3.40 is more efficient for small M and large z. In cases where both M and z are large, evaluating the Willshaw probability can be computationally very expensive [22, 23].

Unfortunately, we do not know a formula for the exact Palm probability eq. 3.22 that is analogous to eq. 3.40. Thus, evaluating the exact error probabilities for the model variant with fixed assembly size is computationally cheap only for cases with small z. However, numerical investigations suggest that p_{W} quickly converges to p_{P} for large m, n and z and that the resulting retrieval error probabilities for fixed assembly sizes are smaller than for random assembly size [22].

3.4. Random pattern activity and auto-association. In analogy to the previous sections we can also investigate the auto-associative case with binomially distributed pattern activities where each pattern component is active with probability k/n independently of other components. Here $\mathrm{pr}(C(Y,N,j \notin Y \cup N))$ can be obtained in the same way as done in section 3.3 for hetero-association with $k = l$ and $m = n$. This corresponds to $\sigma = 0$ and leads to $p_{\mathrm{Wa}}(x;k,n,M,\tilde{p}_1,z,0) = p_{\mathrm{Wh}}(x;k,n,k,n,M,\tilde{p}_1,z)$. The remaining subtleties concerning autapses having a much higher activation probability $\bar{p}_1$ than other synapses (see eq. 3.25) can be handled in the same way as done in section 3.2 for fixed pattern activity. Thus, simply replacing $p_{\mathrm{Pa}}(x;k,n,M,\tilde{p}_1,z,0)$ by $p_{\mathrm{Wh}}(x;k,n,k,n,M,\tilde{p}_1,z)$ we obtain from eq. 3.34

$$\begin{aligned} p_{\mathrm{Wa}}(x;k,n,M,\tilde{p}_1,z,\sigma) &= (1-\sigma)p_{\mathrm{Wh}}(x;k,n,k,n,M,\tilde{p}_1,z) \\ &+\sigma p_{\mathrm{Wh}}(x-1;k,n,k,n,M,\tilde{p}_1,z-1) \\ &+\sigma(1-\bar{p}_1)\left(p_B(x;z-1,\tilde{p}_1) - p_B(x-1;z-1,\tilde{p}_1)\right) \qquad (3.41) \end{aligned}$$

When addressing with a single address pattern containing c correct and f false one-entries then the error probabilities for threshold Θ can be computed similarly as in

section 3.2,

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{Wa}}(x; k, n, M-1, \tilde{p}_1, c+f, \bar{\sigma}) \tag{3.42}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{Wa}}(x-c; k, n, M-1, \tilde{p}_1, f, 0) \tag{3.43}$$

for $0 \le x \le z$. For the lo-units σ has to be averaged over the constrained range of possible pattern activities k' with $c \le k' \le n-f$, thus, $\bar{\sigma} := (\sum_{k'=c}^{n-f} p_B(k'; n, k/n) f/(n-k'))/(\sum_{k'=c}^{n-f} p_B(k'; n, k/n))$. Note that computing the expected Hamming distance (see eq. 2.8) requires a similar adjustment. Note also that p_{10} is the same as for hetero-association with the corresponding parameters (see section 3.3).

3.5. Probabilities of add-errors for pattern part retrieval. For the particular case of pattern part retrieval, $c = \lambda k$ and $f = 0$ with $0 < \lambda \le 1$, we can use the Willshaw threshold $\Theta = \lambda k$. Then the probability of miss-errors in the retrieval outputs is generally $p_{10} = 0$. For fixed pattern activity the probability of an add-error is

$$p_{01,\mathrm{Ph}} = \sum_{s=0}^{\lambda k} (\tilde{p}_1 - 1)^s \binom{\lambda k}{s} [1 - \frac{l}{n}(1 - B(m, k, s))]^{M-1} \tag{3.44}$$

$$p_{01,\mathrm{Pa}} = \sum_{s=0}^{\lambda k} (\tilde{p}_1 - 1)^s \binom{\lambda k}{s} [1 - \frac{k}{n}(1 - \frac{n}{n-s} B(n, k, s))]^{M-1} \tag{3.45}$$

$$\tag{3.46}$$

for hetero-association and auto-association, respectively. For random pattern activity, the corresponding error probabilities are

$$p_{01,\mathrm{Wh}} = \sum_{s=0}^{\lambda k} (\tilde{p}_1 - 1)^s \binom{\lambda k}{s} [1 - \frac{l}{n}(1 - (1 - k/m)^s)]^{M-1} \tag{3.47}$$

$$= \sum_{i=0}^{M-1} p_B(i; M-1, l/n)(1 - (1 - \tilde{p}_1)(1 - k/m)^i)^{\lambda k} \tag{3.48}$$

$$\ge [1 - (1 - \tilde{p}_1)(1 - kl/mn)^{M-1}]^{\lambda k} = p_1^{\lambda k} \tag{3.49}$$

$$p_{01,\mathrm{Wa}} = p_{01,\mathrm{Wh}}|_{l=k, m=n}, \tag{3.50}$$

where p_B is again the binomial probability (see below eq. 2.12). Here the error probabilities are essentially the same for auto-association and hetero-association with $k = l$, $m = n$. Eq. 3.49 corresponds to the binomial approximation eq. 2.13 as used in section 2.3. The bound can be obtained from Jensen's inequality $Ef(y) \ge f(Ey)$ (e.g., [9]) for convex $f(y) := (1-y)^{\lambda k}$ with random variable $y := (1 - \tilde{p}_1)(1 - k/m)^i$. Here the expectation $Ey = 1 - p_1$ can be computed from eq. A.9 using $J = 1$.

Although I could not prove this strictly, numerical experiments suggest $p_{01,\mathrm{Pa}} \le p_{01,\mathrm{Ph}} \le p_{01,\mathrm{Wh}} = p_{01,\mathrm{Wa}}$ [22]. The binomial approximation eq. 3.49 can strongly underestimate p_{01}. Palm and Sommer [34, 38] give some asymptotic conditions when the true potential distribution converges to the corresponding binomial distribution, however, only for relatively small $k \sim \log n$ and $k \le n^{1/3}$, respectively. In section 5 we will see that the parameter range of convergence is actually much larger.

3.6. Numerical evaluations. Theorem 3.2 and the resulting retrieval error probabilities have been verified by extensive numerical simulations of the Willshaw model [22]. Some data is shown in Table 3.1.

	Θ	ϵ	p_{01}	mean	s.e.	p_{10}	mean	s.e.
S p_{Ph}	3	0.871142	0.200514	0.200473	0.000040	0.403276	0.403387	0.000049
p_{Pa}	3	0.824469	0.149855	0.149827	0.000036	0.474807	0.474726	0.000050
p_{Wh}	3	0.937330	0.223047	0.223043	0.000045	0.416887	0.416905	0.000054
p_{Wa}	4	0.974194	0.067171	0.067197	0.000027	0.817462	0.817423	0.000042
A p_{Ph}	3	1.023875	0.107831	0.107822	0.000031	0.538635	0.538590	0.000050
p_{Wh}	3	1.121372	0.127232	0.127211	0.000036	0.548828	0.548834	0.000057

TABLE 3.1

Results from numerical simulations of retrieval in the Willshaw model with $m = 10$, $k = 3$, $M = 5$, $\tilde{p}_1 = 0.1$ when addressing with patterns containing $c = 2$ correct and $f = 2$ false one-entries. Upper rows (S) show results for "symmetric" networks with $n = m$ and $l = k$ (cf. Fig. 3.1, left panel). Lower rows (A) show results for "asymmetric" networks with $n = 11$ and $l = 2$. The columns show optimal retrieval threshold Θ, output noise ϵ, and the error probabilities p_{01} and p_{10} for add-noise and miss-noise as well as the corresponding average values (mean) and standard errors (s.e.) from the simulation experiments (evaluating $N \approx 10^8$ retrievals in each case). The experimental values closely match the theoretical values and thus verify Theorem 3.2.

Figure 3.1 gives examples for the Willshaw-Palm distribution illustrating the differences between the four probability versions and the binomial approximation. For small networks the difference between the four versions of the Willshaw-Palm distribution is significant. In comparison to the binomial approximation the Willshaw-Palm probability can have a much *larger variance* and *oscillatory modulations* [19, 21]. The difference in variance is computed in section 5.2 (see eq. 5.5), and conditions where the variances and higher-order moments become identical are computed in section 5.4. The oscillatory modulations can be understood from eq. 3.40 writing p_{Wh} as a superposition of $M + 1$ binomials. They occur if the binomials $p_B(x; z, 1 - (1 - \tilde{p}_1)(1 - k/m)^i)$ around mean unit usage $i \approx Ml/n$ have a small standard deviation $\sqrt{z(1 - \tilde{p}_1)(1 - k/m)^i(1 - (1 - \tilde{p}_1)(1 - k/m)^i)}$ compared to the mean distance $z(1 - \tilde{p}_1)((1 - k/m)^i - (1 - k/m)^{i+1})$ between two neighboring binomials, i.e., if

$$(1 - \tilde{p}_1)\frac{zk^2}{m^2}(1 - \frac{k}{m})^{Ml/n} \gg 1 \ . \tag{3.51}$$

4. Expectation, variance, and higher-order moments of the Willshaw-Palm distribution. In this section we investigate the moments of the Willshaw-Palm probability distribution. Here we will focus on the more simple case of random pattern activity, i.e., on the Willshaw probabilities p_{Wh} and p_{Wa} (see definition 3.1 and theorem 3.2). The analysis for fixed pattern activity is more difficult, but it is plausible to assume that the basic (asymptotic) properties for the Palm probabilities p_{Ph}, p_{Pa} are similar to p_{Wh}, p_{Wa}. At least the expectation values of Willshaw and Palm probabilities are the same: Because the dendritic potential is $x_j = \sum_{i \in \tilde{u}} A_{ij}$, the expectation for hetero-association is identical to the corresponding binomial expectation (see section 2.3),

$$E_{p_{\mathrm{Wh}}}(x_j) = E_{p_{\mathrm{Ph}}}(x_j) = E_{p_B}(x_j) = zp_1 \tag{4.1}$$

$$E_{p_{\mathrm{Wa}}}(x_j) = E_{p_{\mathrm{Pa}}}(x_j) = zp_1 + \sigma(\bar{p}_1 - p_1) \tag{4.2}$$

where p_1 is the memory load eq. 2.2. The expectation for auto-association follows similarly from $E_{p_{\mathrm{Wa}}}(x_j) = (z-1)p_1 + \sigma\bar{p}_1 + (1-\sigma)p_1$, where σ is the probability that

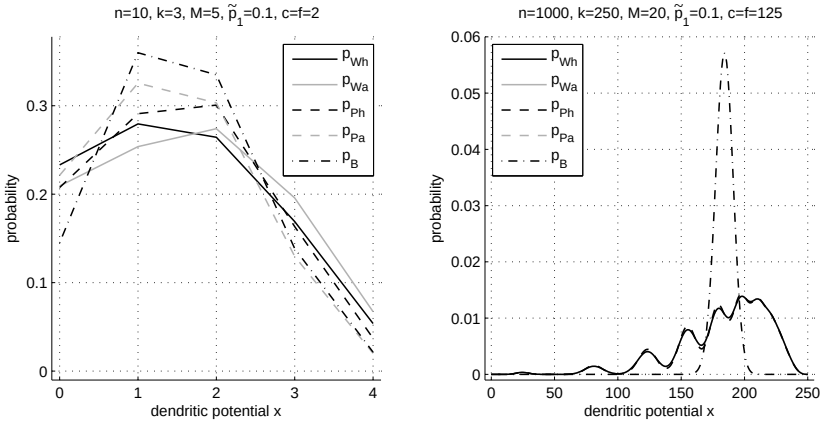

FIG. 3.1. *Examples of the Willshaw-Palm distributions (see Theorem 3.2) and the corresponding binomial approximation (eq. 2.11) for a small network (left panel) and a larger network (right panel). The plots show the distribution of the lo-units when addressing with c correct and f false units in symmetric networks ($m = n$ and $k = l$). The plots indicate that the binomial approximation can be very inaccurate.*

j is among the z active units of address pattern $\tilde{u}$, and $\bar{p}_1$ is the probability that A_{jj} is active (see eq. 3.25).

In the following text we will sometimes write $p_0 := 1 - p_1$, $\bar{p}_0 := 1 - \bar{p}_1$, and $\tilde{p}_0 := 1 - \tilde{p}_1$ for the sake of brevity.

4.1. Moment generating functions. The moment generating function of a random variable X with probability function p is defined by $G_p(t) := E_p(e^{tX})$ (e.g., see [39]). The following theorem shows that the moment generating functions of the Willshaw-Palm probabilities for *random* pattern activity k can be obtained from the generating function of the binomial probability (eq. A.3).

THEOREM 4.1. *The moment generating functions $G_{p_{\mathrm{Wh}}}(t; k, l, m, n, M, \tilde{p}_1, z)$ and $G_{p_{\mathrm{Wa}}}(t; k, n, M, \tilde{p}_1, z, \sigma)$ of the Willshaw probability functions p_{Wh} for hetero-association (eq. 3.39) and p_{Wa} for auto-association (eq. 3.41) are*

$$G_{p_{\mathrm{Wh}}}(t) = \sum_{i=0}^{M} p_B(i; M, l/n) G_{p_B}(t; z, 1 - \tilde{p}_0(1 - k/m)^i) \tag{4.3}$$

$$\begin{aligned} G_{p_{\mathrm{Wa}}}(t; \ldots, z, \sigma) = {} & (1-\sigma) G_{p_{\mathrm{Wh}}}(t; \ldots, z) + \sigma e^t G_{p_{\mathrm{Wh}}}(t; \ldots, z-1) \\ & + \sigma \bar{p}_0 (1 - e^t) G_{p_B}(t; z-1, \tilde{p}_1) \end{aligned} \tag{4.4}$$

Proof. By definition it is $G_{p_{\mathrm{Wh}}}(t) := E_{p_{\mathrm{Wh}}} e^{tX} = \sum_{x=0}^{z} e^{tx} \sum_{i=0}^{M} p_B(i; M, l/n) \cdot p_B(x; z, 1 - \tilde{p}_0(1 - k/m)^i) = \sum_{i=0}^{M} p_B(i; M, l/n) \sum_{x=0}^{z} e^{tx} p_B(x; z, 1 - \tilde{p}_0(1 - k/m)^i)$. Here the second sum is the moment generating function of a binomial eq. A.3 with $N = z$ and $P = 1 - (1 - k/m)^i$. This shows eq. 4.3. Similarly, the auto-associative moment generating function $G_{p_{\mathrm{Wa}}}(t)$ follows with eq. 3.41 because moment generating functions $G_{p(x)}(t)$ are linear in $p(x)$ and have the shifting property $G_{p(x-1)}(t) = \sum_x p(x-1) e^{tx} = \sum_x p(x) e^{t(x+1)} = e^t G_{p(x)}(t)$. ∎

4.2. Higher order moments. The d-th *raw moment* of a random variable X with probability function p is defined by the expectation $E_p X^d$ and can be computed

from the moment generating function $G_p(t) := E_p(e^{tX})$, where the d-th derivative $G_p^{(d)}(t)$ at $t = 0$ yields the d-th moment (e.g., [39]). Then the d-th *central moment* (or moment about the mean) is defined by the expectation $E_p(X-\mu)^d$ where $\mu := E_p X$ is the mean value. The following theorem computes the moments of the Willshaw probabilities from the moments of the binomial probability.

THEOREM 4.2. *Let $p_{\mathrm{Wh}}(x; k, l, m, n, M, \tilde{p}_1, z)$ be the Willshaw probability for hetero-association (eq. 3.39) and $p_B(x; z, 1-p_0)$ the corresponding binomial approximation with $p_0 := 1-p_1$ (see eqs. A.2,2.2). Then the raw and central moments of the Willshaw probability can be computed from the binomial moments (see eqs. A.4-A.5) by formally substituting powers p_0^j by numbers $p_0^{(j)}$ defined as*

$$p_0^{(j)} := \tilde{p}_0^j (1 - \frac{l}{n}(1-(1-\frac{k}{m})^j))^M \ . \tag{4.5}$$

where $\tilde{p}_0 := 1-\tilde{p}_1$. For example, the raw and central Willshaw moments for hetero-association, $\mathfrak{m}_{r,p_{\mathrm{Wh}}}(d; k, l, m, n, M, \tilde{p}_1, z)$ and $\mathfrak{m}_{c,p_{\mathrm{Wh}}}(d; k, l, m, n, M, \tilde{p}_1, z)$, can be obtained from

$$E_{p_{\mathrm{Wh}}}(X-\mu)^d = \sum_{j=0}^{d} p_0^{(j)} (-1)^j \binom{z}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (z-\mu-i)^d \tag{4.6}$$

which is true for an arbitrary offset μ. The raw and central moments follow with $\mu = 0$ and $\mu = zp_1$, respectively.

Similarly, the raw and central Willshaw moments for the auto-associative probability p_{Wa} (see eq. 3.41), $\mathfrak{m}_{r,p_{\mathrm{Wa}}}(d; k, n, M, \tilde{p}_1, z, \sigma)$ and $\mathfrak{m}_{c,p_{\mathrm{Wa}}}(d; k, n, M, \tilde{p}_1, z, \sigma)$, follow from

$$\begin{aligned} E_{p_{\mathrm{Wa}}}(X-\mu)^d = & \sum_{j=0}^{d} p_0^{(j)} (-1)^j \binom{z}{j} (1-\frac{\sigma j}{z}) \sum_{i=0}^{j} (-1)^i \binom{j}{i} (z-\mu-i)^d \ + \\ & \sigma \bar{p}_0 \sum_{j=0}^{d} \tilde{p}_0^j (-1)^j \binom{z-1}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} ((z-\mu-i-1)^d - (z-\mu-i)^d)) \end{aligned} \tag{4.7}$$

using $\mu = 0$ and $\mu = zp_1 - \sigma(\bar{p}_1 - p_1)$, respectively.

Proof. The d-th *raw moment* $E_{p_{\mathrm{Wh}}} X^d$ equals the d-th derivative $G_{p_{\mathrm{Wh}}}^{(d)}(t)$ at $t = 0$. From eq. 4.3 we obtain

$$G_{p_{\mathrm{Wh}}}^{(d)}(0) = \sum_{i=0}^{M} p_B(i; M, l/n) G_{p_B}^{(d)}(0; z, 1-\tilde{p}_0(1-k/m)^i)$$

where $G_{p_B}^{(d)}(0; N, 1-Q) = \mathfrak{m}_{r,p_B}(d, N, 1-Q) = \sum_{j=0}^{d} c_j^{(d)}(N) Q^j$ is the d-th raw moment of the binomial probability (see eq. A.5). For brevity we have defined coefficients $c_j^{(d)}(N) := (-1)^j \binom{N}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (N-k)^d$. Applying eq. A.9 we obtain

$$\begin{aligned} E_{p_{\mathrm{Wa}}} X^d &= \sum_{i=0}^{M} p_B(i; M, l/n) \sum_{j=0}^{d} c_j^{(d)}(z) \tilde{p}_0^j (1-k/m)^{ij} \\ &= \sum_{j=0}^{d} c_j^{(d)}(z) \tilde{p}_0^j \sum_{i=0}^{M} p_B(i; M, l/n)(1-k/m)^{ij} = \sum_{j=0}^{d} c_j^{(d)}(z) p_0^{(j)} \ . \end{aligned}$$

This proves the formulae for the raw moments $\mathfrak{m}_{r,p_{\mathrm{Wh}}}$, for example eq. 4.6 for $\mu = 0$. The general moment eq. 4.6 with arbitrary offset μ follows then from inserting the raw moments into $E(X-\mu)^d = \sum_{i=0}^{d} \binom{d}{i}(-\mu)^{d-i} EX^i$, where we used the binomial sum (see below eq. A.9) and the linearity of the expectation operator. Inserting $\mu = zp_1$ (see eq. 4.1) finally yields the central moments $\mathfrak{m}_{c,p_{\mathrm{Wh}}}$ for the hetero-associative Willshaw probability (see also eq. A.5; cf. [25]).

Similarly, the general moment eq. 4.7 for auto-association follows with eq. 3.41 because moments $E_{p(x)}(X-\mu)^d$ are linear in $p(x)$ and have the shifting property $E_{p(x-1)}(X-\mu)^d = \sum_x p(x-1)(x-\mu)^d = \sum_x p(x)(x-\mu+1)^d = E_{p(x)}(X-(\mu-1))^d$. In particular, summing the two Willshaw terms in eq. 3.41 leads to the factor $(1-\sigma)\binom{z}{j} + \sigma\binom{z-1}{j} = \binom{z}{j}(1-\frac{\sigma j}{z})$ in eq. 4.7. The raw and central moments $\mathfrak{m}_{r,p_{\mathrm{Wa}}}$ and $\mathfrak{m}_{c,p_{\mathrm{Wa}}}$ then follow from inserting $\mu = 0$ and $\mu = E_{p_{\mathrm{Wa}}}X$ (see eq. 4.2). □

The following lemma gives a more detailed characterization of the numbers $p_0^{(j)}$ that have been used to compute the moments of the Willshaw probability.

LEMMA 4.3. *Let $p_0^{(j)}$ as defined in theorem 4.2. For $0 < P < 1$ we have*

$$R_j(P) := \frac{1}{P}\sum_{i=2}^{j}\binom{j}{i}(-P)^i = \frac{(1-P)^j - 1 + Pj}{P} \geq 0 \tag{4.8}$$

$$p_0 := 1 - p_1 = \tilde{p}_0(1-\frac{kl}{mn})^M \tag{4.9}$$

$$p_0^{(j)} := \tilde{p}_0^j(1-\frac{l}{n}(1-(1-\frac{k}{m})^j))^M = \tilde{p}_0^j(1-\frac{kl}{mn}(j-R_j(\frac{k}{m})))^M \approx p_0^j \tag{4.10}$$

For $j = 0,1$ we have $R_j(P) = 0$ and $p_0^{(j)} = p_0^j$. For $j \geq 2$ we have $R_j(P) > 0$. For sufficiently small $P \to 0$ the bound $R_j(P) < \binom{j}{2}P$ becomes true. Furthermore, for $j \geq 2$ we have the bounds

$$p_0^j < p_0^{(j)} < p_0^{j-R_j(k/m)} \tag{4.11}$$

$$0 < \frac{p_0^{(j)} - p_0^j}{p_0^j} < p_0^{-R_j(k/m)} - 1 < -(e-1)R_j(k/m)\ln p_0 \tag{4.12}$$

where the latter bound in eq. 4.12 is true only for $-R_j(k/m)\ln p_0 < 1$. In particular, the relative difference between $p_0^{(j)}$ and p_0^j vanishes for $R_j(k/m)\ln p_0 \to 0$. Finally, let $p := k/m \to 0$, $q := l/n$, $M = \ln p_0/\ln(1-pq)$ (see eq. 2.3). Then for $j^2 p(1-\ln p_0) \to 0$, fixed $\tilde{p}_1$, and using the asymptotic Θ notation as defined in appendix A we have

$$p_0^{(j)} - p_0^j = -\binom{j}{2}p(1-q)p_0^j\ln p_0 + \Theta(j^3p^2p_0^j(1-j\ln p_0)\ln p_0)\ . \tag{4.13}$$

Proof. Eq. 4.8 follows from the binomial sum (see below eq. A.9). Eq. 4.9 is simply rewriting eq. 2.2 with $\tilde{p}_0 := 1-\tilde{p}_1$ for the sake of completeness. Eq. 4.10 follows from simple transformations of the definitions eqs. 4.5,4.8. The claims for $j = 0,1$ follow trivially. $R_j(P) > 0$ for $j \geq 2$ follows from $(1-P)^j > (1-Pj)$ (see eq. A.11). $R_j(P) < \binom{j}{2}P$ for $P \to 0$ follows directly from the definition of R_j. The lower bound in eq. 4.11 follows from $(1-kl/mn)^j = 1-(kl/mn)(j-R_j(kl/mn))$ because $R_j(P)$ is monotonically increasing for $0 < P < 1$. The upper bound in eq. 4.11 follows from $(1-pq(j-R_j)) < (1-pq)^{j-R_j}$ (see eq. A.11 with $j-R_j > j-R_j(1) = 1$). Eq. 4.12

follows from eq. 4.11 and eq. A.12. We finally prove the asymptotic approximation eq. 4.13: For $p \to 0$, $M = \ln p_0 / \ln(1-pq)$ (see eq. 2.3) we have with eqs. A.13-A.14

$$M = \frac{-\ln p_0}{pq + \Theta(p^2q^2)} = \frac{-\ln p_0}{pq}\frac{1}{1+\Theta(pq)} = \frac{-\ln p_0}{pq}(1+\Theta(pq)) \tag{4.14}$$

$$Mpq = -\ln p_0 + \Theta(pq \ln p_0) \tag{4.15}$$

$$R_j(p) = \binom{j}{2}p - \binom{j}{3}p^2 + \ldots = \binom{j}{2}p + \Theta(j^3p^2) \to 0 \text{ for } j^2p \to 0 \tag{4.16}$$

The final purpose of this is to find a close approximation for

$$p_0^{(j)} - p_0^j = p_0^j\left(\frac{p_0^{(j)}}{p_0^j} - 1\right) \quad \text{with} \quad \frac{p_0^{(j)}}{p_0^j} = e^{M(\ln(1-pq(j-R_j))-j\ln(1-pq))} \ . \tag{4.17}$$

For $R_j \to 0$ the term in the outer brackets of the exponential writes $\ln(1-pq(j-R_j))-j\ln(1-pq)) = -pq(j-R_j) - \frac{p^2q^2(j-R_j)^2}{2} + \Theta(p^3q^3j^3) - j(-pq - \frac{p^2q^2}{2} + \Theta(p^3q^3)) = pqR_j - \frac{p^2q^2}{2}\left((j-R_j)^2 - j\right) + \Theta(p^3q^3j^3)$. Here we have $(j-R_j)^2 - j = j^2 - j + \Theta(jR_j) = j^2 - j + \Theta(j^3p)$ and therefore

$$\frac{p_0^{(j)}}{p_0^j} = e^{M(pqR_j - 0.5p^2q^2(j^2-j) + \Theta(p^3q^2j^3))} \tag{4.18}$$

$$= e^{M(p^2q\binom{j}{2}(1-q)+\Theta(p^3qj^3))} = e^{Mpq(p\binom{j}{2}(1-q)+\Theta(p^2j^3))} \tag{4.19}$$

$$= e^{(-\ln p_0 + \Theta(pq\ln p_0))(p\binom{j}{2}(1-q)+\Theta(p^2j^3))} = e^{-\binom{j}{2}p(1-q)\ln p_0 + \Theta(j^3p^2\ln p_0)} \tag{4.20}$$

$$= 1 - \binom{j}{2}p(1-q)\ln p_0 + \Theta(j^4p^2\ln^2 p_0 - j^3p^2\ln p_0) \tag{4.21}$$

$$= 1 - \binom{j}{2}p(1-q)\ln p_0 + \Theta(j^3p^2\ln p_0(1 - j\ln p_0)) \tag{4.22}$$

□

4.3. Variance. Applying theorem 4.2, we can easily compute the second raw and central moments of the Willshaw probability from the well-known second moments of the corresponding binomial probability $p_B(x; z, 1-p_0)$ (see also eqs. A.4-A.5). Thus, replacing p_0^j by $p_0^{(j)}$ in

$$E_{p_{B(x;z,1-p_0)}}X^2 = zp_0(1-p_0) + z^2(1-p_0)^2 = z^2 + p_0(z-2z^2) + p_0^2(z^2-z)$$

gives us immediately the second moment and variance of the Willshaw probability p_{Wh} for *hetero-association*,

$$E_{p_{\mathrm{Wh}}}X^2 = z^2 + p_0(z-2z^2) + p_0^{(2)}(z^2-z) \tag{4.23}$$

$$= z^2(1-2p_0+p_0^{(2)}) + z(p_0 - p_0^{(2)}) \tag{4.24}$$

$$\mathrm{Var}_{p_{\mathrm{Wh}}}X = E_{p_{\mathrm{Wh}}}X^2 - E^2_{p_{\mathrm{Wh}}}X = z^2(p_0^{(2)} - p_0^2) + z(p_0 - p_0^{(2)}) \ . \tag{4.25}$$

where $p_0^{(2)} = \tilde{p}_0^2(1-(kl/mn)(2-k/m))^M$. Note that in eqs. 4.24,4.25 all coefficients of z are positive since with eq. 4.11 we have $p_0 > p_0^{2-k/m} > p_0^{(2)} > p_0^2$ and $1-2p_0+p_0^{(2)} > (1-p_0)^2 > 0$ for $0 < k/m, l/n < 1$. Thus, the variance increases monotonically with z.

Indeed, the variance of the Willshaw probability scales with z^2 while the corresponding binomial variance scales only with z [21]. Applying eq. 4.7 with $d = 2$ we easily obtain the second moments of the Willshaw probability p_{Wa} for *auto-association*,

$$\begin{aligned} E_{p_{\mathrm{Wa}}}(X-\mu)^2 &= (z-\mu)^2 - (2(z-\mu)-1)(zp_0 - \sigma(p_0-\bar{p}_0)) \\ &\quad +(z-1)((z-2\sigma)p_0^{(2)} + 2\sigma\bar{p}_0\tilde{p}_0) \qquad (4.26) \\ \mathrm{Var}_{p_{\mathrm{Wa}}}X &= z^2(p_0^{(2)} - p_0^2) + z(p_0 - p_0^{(2)} - 2\sigma(p_0^{(2)} - p_0^2 - \bar{p}_0(\tilde{p}_0 - p_0))) \\ &\quad -\sigma(p_0 - 2p_0^{(2)} + \sigma(p_0-\bar{p}_0)^2 + \bar{p}_0(2\tilde{p}_0 - 1)) \qquad (4.27) \\ &= \mathrm{Var}_{p_{\mathrm{Wh}}}X - 2\sigma z(p_0^{(2)} - p_0^2 - \bar{p}_0(\tilde{p}_0 - p_0)) \\ &\quad -\sigma(p_0 - 2p_0^{(2)} + \sigma(p_0-\bar{p}_0)^2 + \bar{p}_0(2\tilde{p}_0 - 1)) \ . \qquad (4.28) \end{aligned}$$

Here eq. 4.26 is true for any offset μ. In particular, the second raw moment follows with $\mu = 0$, and the variance eq. 4.27 follows with $\mu = E_{p_{\mathrm{Wa}}}X = z - zp_0 + \sigma(p_0 - \bar{p}_0)$ (see eq. 4.2).

4.4. Auto-association vs. hetero-association. As long as the dendritic potential distributions has a Gaussian shape the variance determines retrieval quality, i.e., the larger the variance the larger the error probabilities (e.g., see eq. 3.42). Thus, in order to answer the question whether retrieval quality is better for auto-association or hetero-association (with $k = l$ and $m = n$), the following lemma investigates the asymptotic behavior of the variances difference $\delta_{\mathrm{Var}_{\mathrm{WaWh}}} := \mathrm{Var}_{p_{\mathrm{Wa}}}X - \mathrm{Var}_{p_{\mathrm{Wh}}}X$. To obtain general results, we fix the memory load $p_1 = p_{1\epsilon}$ to its maximum under quality constraint ϵ, as discussed in section 2.3 (see eq. 2.15).

LEMMA 4.4. *For $p = k/n \to 0$, $\sigma/p \sim 1$, $z \sim k$, $p \ln p \ln k \to 0$, $z \sim k$, fixed $\tilde{p}_1$, and "hifi" memory load $1 - p_0 = p_{1\epsilon} = (\epsilon p)^{1/z}$ as in eq. 2.15 we have*

$$\begin{aligned} \delta_{\mathrm{Var}_{\mathrm{WaWh}}} &\approx 2\sigma z p p_0^2 \ln p_0 - \sigma(p_0 - 2p_0^2) \qquad (4.29) \\ &\approx \begin{cases} +\sigma \ , & p_0 \to 1 \\ -\sigma p_0 \ , & p_0 \to 0 \ . \end{cases} \ . \qquad (4.30) \end{aligned}$$

Proof. We can apply eq. 4.13 because $p \ln p \ln k \to 0$ implies $p \ln p_0 \to 0$ even for $p_0 \to 0$ with $p_0 \approx -\ln(\epsilon p)/z$ (see eq. 2.20). Thus, eq. 4.29 follows from eq. 4.28 because $p_0 - 2p_0^{(2)} \sim -pp_0^2 \ln p_0$ dominates over $\bar{p}_0 \sim e^{-Mp} = e^{(\ln p_0)/p} = p_0^{n/k}$ even for sparse potentiation with $p_0 \to 1$ and $\bar{p}_0 \sim (1-(\epsilon k/n)^{1/z})^{n/k} \sim \exp(-(\epsilon k/n)^{1/z}(n/k))$, and similarly, $p_0 - 2p_0^{(2)} \approx p_0 - 2p_0^2$ dominates over $\sigma(p_0 - \bar{p}_0)^2 + \bar{p}_0(2\tilde{p}_0 - 1)$.

Eq. 4.30 follows because for sparse potentiation with $p_0 \to 1$ we have $p_0^2 \ln p_0 \to 0$ and $zp \sim k^2/n \to 0$ (see eq. 2.19). Similarly, for dense potentiation with $p_0 \to 0$ we have $-\sigma(p_0 - 2p_0^2) \approx -\sigma p_0$ and for $p_0 \approx -\ln(\epsilon p)/z$ (see eq. 2.20) we have $0 > 2\sigma z p p_0^2 \ln p_0 \approx 2\sigma p_0(-p \ln(\epsilon p) \ln z)$. □

Thus, the autoassociative variance $\mathrm{Var}_{p_{\mathrm{Wa}}}X$ becomes larger than $\mathrm{Var}_{p_{\mathrm{Wh}}}X$ for sparse potentiation with $p_1 \to 0$, but smaller for dense potentiation with $p_1 \to 1$. However, remember from sections 3.4,3.2 that pattern part retrieval with $f = 0$ (i.e., no add errors in the address pattern) implies $\sigma = 0$ and thus identical distributions for auto-association and hetero-association. Also, the following lemma shows that in general the differences between hetero-associative and auto-associative moments vanish asymptotically.

LEMMA 4.5. *For fixed d, $p = k/n \to 0$, $\sigma/p \sim 1$, $z \sim k$, $zp_1 \le \mu \le z$, $p \ln p_0 \to 0$, and "hifi" memory load $1 - p_0 = p_{1\epsilon} = (\epsilon p)^{1/z}$ as in eq. 2.15 we have*

$$E_{p_{\mathrm{Wh}}}(X-\mu)^d - E_{p_{\mathrm{Wa}}}(X-\mu)^d \to 0 \qquad (4.31)$$

Proof. The difference is identical to eq. 4.7 (cf., eq. 4.6) except that the factor $(1-\frac{\sigma j}{z})$ in the first double sum becomes simply $\frac{\sigma j}{z} \sim \frac{j}{n}$. Thus, with eq. A.10 the absolute value of the first double sum becomes zero because

$$|\sum_{j=0}^{d} p_0^{(j)}(-1)^j \frac{\sigma j}{z} z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} b_{di}(z-\mu-j)^{\underline{i-j}}|$$
$$\leq \sum_{j=0}^{d} p_0^{(j)} \frac{\sigma j}{z} z^j \sum_{i=j}^{d} \binom{i}{j} S_{di}(zp_0)^{i-j} \sim \frac{(zp_0)^d}{n} \to 0 \ .$$

The inequality is true for large enough z with $d < z-\mu \leq z - zp_1 = zp_0$. The asymptotic approximations remain true even for small constant z where we used $p_0^{(j)} \sim p_0^j$ (see eq. 4.12). Note here that $z = O(\log n)$ implies sparse or balanced potentiation with $1 \geq p_0 \not\to 0$, while larger z implies dense potentiation with $p_0 \to 0$ and $zp_0 = O(\log n)$ (see eqs. 2.19,2.20). We still have to show that also the second double sum in eq. 4.7 becomes zero:

$$|\sigma \bar{p}_0 \sum_{j=0}^{d} \tilde{p}_0^j (-1)^j \binom{z-1}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} ((z-\mu-i-1)^d - (z-\mu-i)^d))|$$
$$\sim O(\frac{\bar{p}_0 z^{2d+1}}{n}) \to 0 \ .$$

This is obvious for sparse and balanced potentiation when $z \sim k \sim O(\log n)$ (see eq. 2.19), but follows also for dense potentiation ($p_0 = O((\log n)/z) \to 0$; see eq. 2.20) since here $\bar{p}_0 \sim e^{-Mp} = e^{(\ln p_0)/p} = p_0^{n/k}$ quickly approaches zero. □

5. Comparison of Willshaw-Palm to binomial distribution. In this section we compare the Willshaw-Palm probability distribution of the dendritic potentials (see Def. 3.1) to the corresponding binomial approximation $p_B(x; z, p_1)$ which assumes independently generated memory matrix entries (see section 2.3). In particular, we are interested in asymptotic conditions when the two probability distributions, as judged by their moments, become identical for maximal memory load (i.e., $p_1 = p_{1\epsilon}$ and $M = M_\epsilon$ as estimated by eqs. 2.15,2.16). This corresponds to correctness conditions for many previous results that rely on the binomial approximation eq. 2.13 (e.g.,[46, 34, 33, 37, 7, 38, 4, 43, 20]).

5.1. Difference in moments. For the difference $\Delta_{\text{Wh}}^{(d)}$ between the d-th moments of the *hetero-associative* Willshaw probability p_{Wh} and the corresponding binomial probability $p_B(x; z, 1-p_0)$ (see section 2.3), we obtain from eqs. 4.6,A.5,A.10

$$\Delta_{\text{Wh}}^{(d)}(\mu) := E_{p_{\text{Wh}}}(X-\mu)^d - E_{p_B(x;z,1-p_0)}(X-\mu)^d \tag{5.1}$$
$$= \sum_{j=2}^{d} (p_0^{(j)} - p_0^j)(-1)^j \binom{z}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (z-\mu-i)^d \tag{5.2}$$
$$= \sum_{j=2}^{d} (p_0^{(j)} - p_0^j)(-1)^j z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di}(z-\mu-j)^{\underline{i-j}} \tag{5.3}$$

where μ is again an arbitrary offset (e.g., $\mu = 0$ for the raw moments and $\mu = zp_1$ for the central moments). Thus, $\Delta_{\text{Wh}}^{(d)}$ has the same form as the d-th binomial moment,

written as a polynomial in p_0, but where powers p_0^j have been replaced by $p_0^{(j)} - p_0^j$ (see Theorem 4.2). Also note that $p_0^{(j)} = p_0^j$ for $j = 0, 1$ (see Lemma 4.3). The corresponding difference $\Delta_{\mathrm{Wa}}^{(d)}(\mu) := E_{p_{\mathrm{Wa}}}(X-\mu)^d - E_{p_{\mathrm{B}}(x;z,1-p_0)}(X-\mu)^d$ for the *auto-associative* Willshaw probability p_{Wa} can be obtained in a similar way from eq. 4.7.

5.2. Difference in variance. A particularly interesting case is variance ($d = 2$): As long as the overall distribution of dendritic potentials resembles a Gaussian (which is often true, but see [19, 21, 23]), the retrieval error probabilities are essentially determined by the first two moments, i.e., expectation ($d = 1$) and variance ($d = 2$). Thus, it seems plausible to assume that a necessary condition for convergence of the Willshaw-Palm distribution towards a binomial is that expectation and variance become identical. For hetero-association, the expectations are already identical (eq. 4.1). Thus, it is sufficient to investigate conditions when the difference $\delta_{\mathrm{Wh}B}$ between the two variances vanishes. From eqs. 4.25,4.23 we obtain for hetero-association

$$\delta_{\mathrm{Wh}B} := \mathrm{Var}_{p_{\mathrm{Wh}}} X - \mathrm{Var}_{p_B} X = z^2(p_0^{(2)} - p_0^2) + z(p_0 - p_0^{(2)}) - zp_0(1-p_0) \quad (5.4)$$

$$= (z^2 - z)(p_0^{(2)} - p_0^2) > 0 \quad (5.5)$$

Note that $\delta_{\mathrm{Wh}B}$ is always positive (see eq. 4.11). Thus, the binomial approximation always underestimates the variance of the dendritic potentials. Therefore the binomial approximation generally underestimates the probabilities of retrieval errors and overestimates storage capacity, at least if the Willshaw distribution comes close to a Gaussian which is often true (cf., eq. 3.49 for pattern part retrieval; but see [19, 21, 23]). With eqs. 4.12,4.13 we obtain

$$\delta_{\mathrm{Wh}B} \leq (z^2 - z)p_0^2(p_0^{-k/m} - 1) \approx -(z^2 - z)\frac{k}{m}(1 - \frac{l}{n})p_0^2 \ln p_0, \quad (5.6)$$

where the approximation is true for $(k/m)(1 - \ln p_0) \to 0$ (see also eq. 2.20). Note that eq. 5.6 can become zero for a very large parameter range under maximal memory load (see eqs. 2.19,2.20).

The analysis for auto-association is similar. The difference $\delta_{\mathrm{Wa}B} := \mathrm{Var}_{p_{\mathrm{Wa}}}(X) - \mathrm{Var}_{p_B}(X)$ can be obtained from eqs. 5.5 and eq. 4.28. It is easy to see from eq. 4.31 that in general $\delta_{\mathrm{Wa}B}$ vanishes asymptotically with $\delta_{\mathrm{Wh}B}$. In the following two sections we generalize our asymptotic considerations to higher-order moments.

5.3. Convergence of the raw moments. The following lemma determines asymptotic conditions when the d-th raw moment of the Willshaw-Palm probability p_{Wh} becomes identical to the d-th raw moment of the corresponding binomial probability $p_B(x; z, 1-p_0)$, i.e., conditions when the difference $\Delta_{\mathrm{Wh}}^{(d)}(0) := E_{p_{\mathrm{Wh}}} X^d - E_{p_B(x;z,1-p_0)} X^d$ becomes zero.

LEMMA 5.1. *For fixed d and $(k/m)\ln p_0 \to 0$ the following bounds become asymptotically true,*

$$|\Delta^{(d)}_{\mathrm{Wh}}(0)| \le \sum_{j=2}^{d}(p_0^{(j)} - p_0^j) z^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di}(z-j)^{\underline{i-j}} \tag{5.7}$$

$$\le -(e-1)\frac{k}{m}\sum_{j=2}^{d}\binom{j}{2} z^{\underline{j}} p_0^j \ln p_0 \sum_{i=j}^{d}\binom{i}{j} S_{di}(z-j)^{\underline{i-j}} \tag{5.8}$$

$$\le -d^2\frac{k}{m} z^d p_0^2 \ln p_0 \sum_{j=2}^{d}\sum_{i=j}^{d}\binom{i}{j} S_{di} \sim \frac{k}{m} z^d p_0^2 \ln p_0 \le \frac{kz^d}{m} \tag{5.9}$$

Proof. The lemma follows from eqs. 5.3,4.12 and $R_j(k/m) < \binom{j}{2}(k/m)$ (see Lemma 4.3). □

Thus, the raw moments of the Willshaw-Palm probability p_{Wh} and the corresponding binomial probability $p_B(x; z, 1-p_0)$ become identical if the address pattern activities $k := |\mathbf{u}^\mu|$ and $z := |\tilde{\mathbf{u}}|$ grow at most polynomial in the logarithm $\log m$ of the address population size m.

In the following section we will see that even for larger $k(m)$ the two probability distributions can still become essentially identical as judged by the difference of the *central* moments. The reason for this effect can be easily explained: Consider two probability distributions p_A and p_B with zero mean values and $\delta(x) := p_A(x) - p_B(x) \to 0$ and also $\delta(x)/p_A(x) \to 0$ and $\delta(x)/p_A(x) \to 0$ for any x. Then assume that the d-th (central) moments converge, i.e., $\sum x^d \epsilon(x) \to 0$. Then it is still possible that the corresponding distributions p'_A and p'_B with mean $\mu > 0$ have diverging moments because $\sum (x+\mu)^d \epsilon(x)$ can grow arbitrarily with μ. This is the motivation to have a closer look at the convergence of the central moments in the following section.

5.4. Convergence of the central moments. Here we determine asymptotic conditions when the d-th *central* moments of the Willshaw-Palm probabilities p_{Wh} and p_{Wa} become identical to the d-th central moment of the corresponding binomial probability $p_B(x; z, 1-p_0)$, i.e., conditions when $\Delta^{(d)}_{\mathrm{Wh}}(\mu)$ and $\Delta^{(d)}_{\mathrm{Wa}}(\mu)$ become zero (see section 5.1).

LEMMA 5.2. *For fixed d, $(k/m)\ln p_0 \to 0$, and $d < z - \mu \le z - zp_1 = zp_0$ the following bounds become asymptotically true,*

$$|\Delta^{(d)}_{\mathrm{Wh}}(\mu)| \le \sum_{j=2}^{d}(p_0^{(j)} - p_0^j) z^{\underline{j}} \sum_{i=j}^{d}\binom{i}{j} S_{di}(z-\mu-j)^{\underline{i-j}} \tag{5.10}$$

$$\le -(e-1)\frac{k}{m}\sum_{j=2}^{d}\binom{j}{2} z^{\underline{j}} p_0^j \ln p_0 \sum_{i=j}^{d}\binom{i}{j} S_{di}(zp_0)^{\underline{i-j}} \tag{5.11}$$

$$\le -(e-1)\frac{k}{m}\sum_{j=2}^{d}\binom{j}{2} \ln p_0 \sum_{i=j}^{d}\binom{i}{j} S_{di}(zp_0)^i \tag{5.12}$$

$$\le -d^2\frac{k}{m}(zp_0)^d \ln p_0 \sum_{j=2}^{d}\sum_{i=j}^{d}\binom{i}{j} S_{di} \sim \frac{-k(zp_0)^d \ln p_0}{m} \tag{5.13}$$

Proof. The lemma follows from eqs. 5.3,4.12 and $R_j(k/m) < \binom{j}{2}(k/m)$ (see Lemma 4.3). ∎

With this we can easily find asymptotic convergence conditions for maximal memory load as approximately analyzed in section 2.3.

THEOREM 5.3. *For maximal memory load as estimated by the binomial approximation 2.13, i.e., for $p_1 = p_{1\epsilon}$ and $M = M_\epsilon$ as estimated by eqs. 2.15,2.16, the d-th central moment of the Willshaw-Palm probability p_{Wh} becomes identical to the d-th central moment of the corresponding binomial probability $p_B(x; z, p_{1\epsilon})$, i.e.,*

$$\Delta^{(d)}_{\mathrm{Wh}}(zp_{1\epsilon}) \to 0, \quad \text{if} \quad \frac{k(\ln \frac{n}{\epsilon l})^d \ln z}{m} \to 0 \ . \tag{5.14}$$

Thus, for n polynomial in m the d-th central moments converge at least for $k = O(m/\log^{d+2} m)$. In particular, the variances converge at least for $k = O(m/\log^4 m)$. Moreover, all central moments converge, and therefore the Willshaw probability p_{Wh} becomes identical to the binomial approximation, at least for $k = O(m^P)$ with fixed $P < 1$.

Proof. For $zp_{0\epsilon} \to \infty$ with $p_{0\epsilon} := 1 - p_{1\epsilon}$, the theorem follows from Lemma 5.2 by using $zp_{0\epsilon} \sim \log(n/(\epsilon l))$ (see eq. 2.20). For smaller (e.g., constant) z the convergence of the central moments follows already from the convergence of the raw moments (see Lemma 5.1). ∎

A particular case are "symmetric" networks with $m = n$ and $k = l$, for example for auto-association. It turns out that for such networks the range of convergence is even larger: Assume $k = n/\ln^P n$. Then the convergence condition in eq. 5.14 becomes $(k/n)(\ln((\ln^P n)/\epsilon))^d \ln n \to 0$. Thus, here the central moments converge at least for $k = O(n/\log^2 n)$. Note that the results for hetero-association apply also to auto-association due to Lemma 4.5. Together these considerations suggest that the theoretical results on neural associative networks apply to a much larger range than assumed previously [34, 38, 19]. This includes large portions of the dense potentiation regime with $k/\log n \to \infty$ (see section 2.4). Here previous analyses relying on the binomial approximation have suggested the potential for very efficient computer implementations and new biological hypotheses about the roles of structural plasticity and inhibitory neurons [22, 24].

5.5. Numerical evaluations. The results of this section are verified by Figure 5.1 showing data from numerical experiments testing how well the binomial theory approximates exact values. In fact, the reliability of the binomial approximation depends both on the network size (n) and the pattern activity (k). In general, the binomial theory becomes better for larger n and smaller k. The approximations of pattern capacity M_ϵ (eq. 2.16) and network capacity C_ϵ (eq. 2.17) are comparably reliable and, even for linear $k = n/2$ and small n, overestimate the true values by less than factor two (Fig. 5.1a).

However, the derived "compression capacities" C^I and C^S depend on the maximal memory load $p_{1\epsilon}$ (or $1 - p_{1\epsilon}$; see sections 2.2,2.4) which can be strongly overestimated by the binomial theory (Fig. 5.1b). For linear $k \sim n$ the relative error seems to grow without bound implying $C^I \to 0$ and possibly $C^S \to 0$. Nevertheless, for smaller k the binomial approximation is much better already for realistic network sizes. For example, for $n = 10^5$ the information capacity C^I is about 100% of the binomial estimate for constant $k = 4$, 95% for $k = n^{1/2}$, 70% for $k = n^{2/3}$, and still 40% for $k = n^{3/4}$ (similar values for C^S; data not shown). Interestingly, the binomial

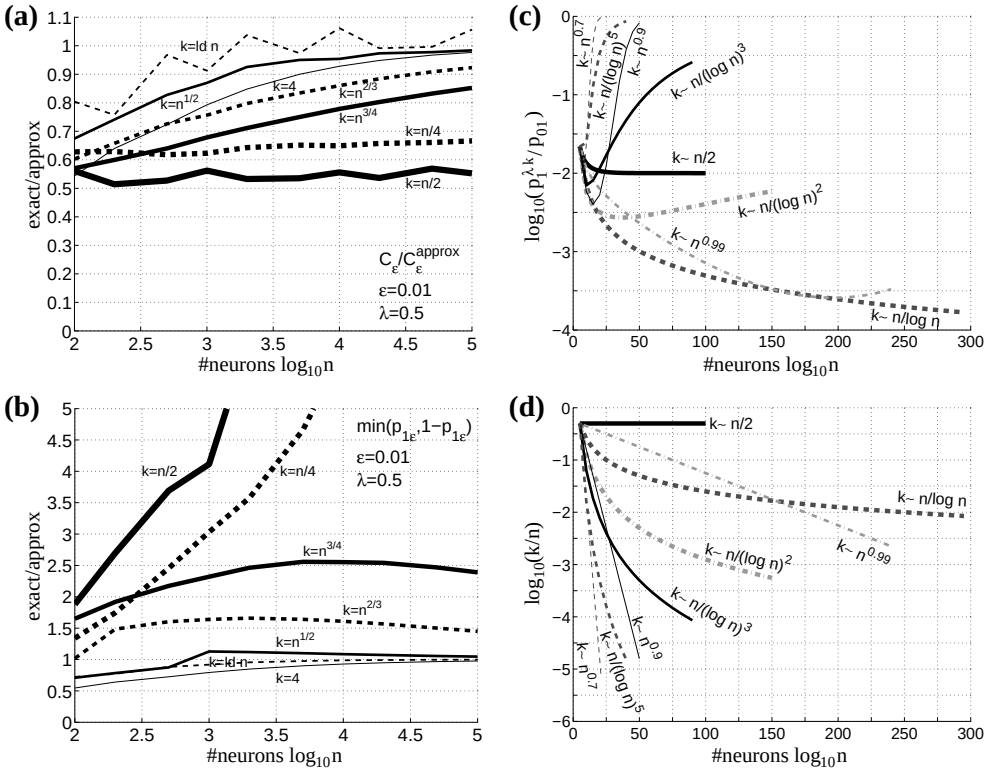

FIG. 5.1. *Numerical experiments comparing the binomial approximative analysis to the exact theory for $m = n$, $k = l$, $\tilde{p}_1 = 0$, $\epsilon = 0.01$, and pattern part retrieval with half addresses ($\lambda = 0.5$).* **a:** *Relative approximation error for network storage capacity, $C_\epsilon/C_\epsilon^{\text{approx}}$ (see eqs. 2.10,2.17). Each curve corresponds to a particular pattern activity function $k(n)$ growing with the neuron number n (log-scale) as indicated in the plots. Relative errors for pattern capacity M_ϵ are virtually identical.* **b:** *Relative approximation error for the memory load $p_{1\epsilon}$ at maximal pattern load M_ϵ (see eqs. 2.9,2.2), similar to panel a. More exactly, the plots show $\min(p_{1\epsilon}, 1 - p_{1\epsilon})/\min(p_{1\epsilon}^{\text{approx}}, 1 - p_{1\epsilon}^{\text{approx}})$ where $p_{1\epsilon}^{\text{approx}}$ is the approximation eq. 2.15. The corresponding approximation errors for the related compression capacities C_ϵ^I and C_ϵ^S (see section 2.4) look qualitatively very similar (cf., [23]).* **c:** *Relative approximation error (log-scale) for the retrieval error probability p_{01} (see eqs. 2.13,3.48) when storing M_ϵ patterns approximated by eq. 2.16. Each curve corresponds to a particular function $k(n)$ with $k(10^5) = 50000$. Each case was evaluated for increasing n until a maximal computation time was reached (about 50h per data point on a 2.4GHz AMD Opteron processor evaluating relevant summands of eq. 3.48 with computing precision 1000bit (see [23] for further details). The plots indicate convergence $p_1^{\lambda k}/p_{01} \to 1$ for $k = O(n/\log^2 n)$, but divergence for $k \sim n/\log n$, thus verifying the theoretical results of section 5.4.* **d:** *Actual pattern activities k/n (log-scale) corresponding to panel c.*

approximations first become worse with growing n until a turning point is reached (e.g., $n = 10^4$ for $k = n^{3/4}$), and only then approach finally the exact values.

Figure 5.1cd shows results for very large network size n and comparably large pattern activities $k(n)$. For near linear $k(n)$ the turning points are reached only for n too large to be useful for applications or relevant for biology. Nevertheless, for smaller pattern activities, for example $k = O(n^{0.8})$, the convergence is much faster. Turning points as described above are still visible for $k = O(n/\log^2 n)$, but seem absent for $k \sim n/\log n$. Thus, the numerical experiments are consistent with the theoretical bound derived at the end of section 5.4.

6. Conclusions. Theories on neural associative networks with binary synapses often use a binomial approximation of the dendritic potential distribution to estimate retrieval error probabilities and performance measures such as storage capacity or retrieval speed [46, 34, 37, 33, 4, 43, 20]. However, for finite network size n or patterns with a relatively large number of active units k this approximation can be very inaccurate. So far, the convergence of the binomial approximation to the true potential distribution and thus the asymptotic correctness of the classical theory has been demonstrated only for some special cases involving very sparse activity patterns, where a binary pattern vector of n neurons contains on average only $k = \log n$ or $k \leq n^{1/3}$ active units [34, 38]. This appeared sufficient because it was believed that neural associative networks would be efficient only for extreme sparseness anyway [34, 43]. In contrast, recent applications of the theory to problems requiring less sparse patterns has gained increased attention for a number of reasons described in the introduction. For example, theoretical analyses based on the binomial approximation suggest that associative networks can operate very efficiently for large pattern activities with $k/\log n \to \infty$ (or equivalently "dense potentiation" with memory load $p_1 \to 1$) if the synaptic matrix is adequately compressed [18, 19, 20, 22]. However, the correctness of these results have been doubted because it remained unclear whether the binomial approximation is sufficiently good for large pattern activity $k(n)$.

Here I have solved this problem. For this it was necessary to compute general expressions for the true potential distribution by defining different versions of the Willshaw-Palm probability including hetero-association, auto-association, fixed and random pattern activities (see section 3). I then focused on the characterization of the probability distributions for random pattern activities. This involved computation of the raw and central moments of the Willshaw-Palm probability (section 4) from the corresponding moments of the binomial probability [25]. Finally, I have investigated the convergence of the two probabilities by determining conditions when the moments become identical. The analysis reveals that the moments become identical for almost any sublinear sparseness, for example $k = O(n/(\log n)^2)$ (see section 5.4), and thus verifies the theory on associative networks for large pattern activities.

Appendix A. Lemmas.

The following lemmas are required to prove the claims in this work. Proofs of the lemmas can be found in a technical report [23, 25] or in the standard literature of information theory, combinatorics, analysis, and probability theory (e.g., [9, 39]).

Let $X \in \{0, 1\}$ be a *binary* random variable with $p := \mathrm{pr}[X = 1]$ and information $I(p) := -p\mathrm{ld}p - (1-p)\mathrm{ld}(1-p)$. Further let Y be the result of transmitting X over a binary memoryless channel with transmission error probabilities $p_{01} := \mathrm{pr}[Y = 1|X = 0]$ and $p_{10} := \mathrm{pr}[Y = 0|X = 1]$. Then the *transinformation* between X and Y is

$$T(X;Y) = T(p, p_{01}, p_{10}) := I_Y(p, p_{01}, p_{10}) - I_{Y|X}(p, p_{01}, p_{10}) \tag{A.1}$$

where $I_Y(p, p_{01}, p_{10}) := I\left(p\left(1 - p_{10}\right) + (1-p)\, p_{01}\right)$ is the information (or entropy) of Y and $I_{Y|X}(p, p_{01}, p_{10}) := p \cdot I(p_{10}) + (1-p) \cdot I(p_{01})$ is the information of Y given X.

Now let $X \in \{0, 1, \ldots, N\}$ be a *binomially* distributed random variable with parameters N and P. Then X has expectation $E_{p_B} X = NP$, and the probability and moment generating functions are

$$\mathrm{pr}[X = x] = p_B(x; N, P) := \binom{N}{x} P^x (1-P)^{N-x} \tag{A.2}$$

$$G_{p_B}(t; N, P) := E_{p_B} e^{tX} = (Pe^t + (1-P))^N \ . \tag{A.3}$$

Furthermore, substituting $Q := 1 - P$, it has been proven in [25] that the d-th raw and central moments of X can be written as polynomials in Q,

$$\mathfrak{m}_{r,p_B}(d,N,P) := E_{p_B} X^d = \sum_{j=0}^{d} (-Q)^j \sum_{i=j}^{d} \binom{i}{j} S_{di} N^{\underline{i}} \quad \text{with} \tag{A.4}$$

$$E_{p_B}(X-\mu)^d = \sum_{j=0}^{d} (-Q)^j \binom{N}{j} \sum_{k=0}^{j} (-1)^k \binom{j}{k} (N-\mu-k)^d \tag{A.5}$$

where $N^{\underline{i}} := N(N-1)\cdots(N-i+1)$ denotes a falling factorial, $S_{di} \geq 0$ are Stirling numbers of the second kind, and μ is an arbitrary offset. For $\mu = NP$ eq. A.5 yields the d-th central moment $\mathfrak{m}_{c,p_B}(d,N,P)$ of the binomial probability. For $\mu = 0$ eq. A.5 becomes identical to the raw moment eq. A.4 [25]. The following lemma is the sieve formula of Sylvester-Poincaré,

$$\operatorname{pr}\left(\bigcup_{k=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < \ldots < i_k \leq n} \operatorname{pr}\left(\bigcap_{h=1}^{k} A_{i_h}\right) \tag{A.6}$$

The following combinatorial equations are true:

$$\binom{Y}{s-N} = \sum_{t=0}^{N} (-1)^{N+t} \binom{Y+t}{s} \binom{N}{t} \tag{A.7}$$

$$\binom{n}{m}\binom{m}{p} = \binom{n}{p}\binom{n-p}{m-p} = \binom{n}{m-p}\binom{n-m+p}{p} \tag{A.8}$$

$$\sum_{i=0}^{M} p_B(i;M,Q) \cdot (1-P)^{Ji} = (1 - Q(1-(1-P)^J))^M \ . \tag{A.9}$$

$$\binom{N}{j} \sum_{i=0}^{j} (-1)^i \binom{j}{i} (n-\mu-i)^d = N^{\underline{j}} \sum_{i=j}^{d} \binom{i}{j} S_{di} (N-\mu-j)^{\underline{i-j}} \ . \tag{A.10}$$

Eq. A.8 implies $B(a,b,c) = B(a,c,b)$ or $\binom{a}{b}\binom{a-b}{c} = \binom{a}{c}\binom{a-c}{b}$. Eq. A.9 is a variant of the binomial sum $(A+B)^M = \sum_{i=0}^{M} \binom{M}{i} A^i B^{M-i}$. For a proof of eq. A.10 see lemma 3.1. in [25]. Here $N^{\underline{i}}$ denotes again a falling factorial, and $S_{di} \geq 0$ Stirling numbers of the second kind.

Then we have used the following bounds,

$$(1-pq) \gtrless (1-p)^q \quad \text{for} \quad p \in (0;1) \quad \text{and} \quad q \overset{\in}{\not\in} (0;1) \tag{A.11}$$

$$1 + x \leq e^x \leq 1 + (e-1)x \quad \text{for} \quad 0 \leq x \leq 1 \tag{A.12}$$

where the first bound in Lemma A.12 is true for any x. Finally, the following asymptotic equations are true for $n \to \infty$ with $|x(n)|, |y(n)| \to 0$,

$$e^x = 1 + x + \Theta(x^2), \qquad \ln(1+x) = x + \Theta(x^2) \ , \tag{A.13}$$

$$e^{x+\Theta(y)} = 1 + x + \Theta(x^2) + \Theta(y), \qquad \frac{1}{1+x} = 1 - x + \Theta(x^2) \ , \tag{A.14}$$

where for a function $f(n)$ we write $f(n) = \Theta(g(n))$ iff there are constants c_1, c_2, n_0 such that for any $n > n_0$ we have $c_1 g(n) < f(n) < c_2 g(n)$.

Acknowledgments. The author is grateful to Edgar Körner and Marc-Oliver Gewaltig for providing the opportunity to do this work at the Honda Research Institute, and also to Friedrich Sommer and Günther Palm for helpful discussions and comments.

REFERENCES

[1] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Information storage in neural networks with low levels of activity. *Phys. Rev. A*, 35:2293–2303, 1987.

[2] D.J. Amit, H. Gutfreund, and H. Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173:30–67, 1987.

[3] H.J. Bentz, M. Hagstroem, and G. Palm. Information storage and effective data retrieval in sparse matrices. *Neural Networks*, 2:289–293, 1989.

[4] H. Bosch and F. Kurfess. Information storage capacity of incompletely connected associative memories. *Neural Networks*, 11(5):869–876, 1998.

[5] V. Braitenberg. Cell assemblies in the cerebral cortex. In R. Heim and G. Palm, editors, *Lecture notes in biomathematics (21). Theoretical approaches to complex systems.*, pages 171–188. Springer-Verlag, Berlin Heidelberg New York, 1978.

[6] J.T. Buckingham. Delicate nets, faint recollections: a study of partially connected associative network memories. *PhD thesis, University of Edinburgh*, 1991.

[7] J.T. Buckingham and D.J. Willshaw. Performance characteristics of associative nets. *Network: Computation in Neural Systems*, 3:407–414, 1992.

[8] J.T. Buckingham and D.J. Willshaw. On setting unit thresholds in an incompletely connected associative net. *Network: Computation in Neural Systems*, 4:441–459, 1993.

[9] T.M. Cover and J.A. Thomas. *Elements of information theory.* Wiley, New York, 1991.

[10] R. Fay, U. Kaufmann, A. Knoblauch, H. Markert, and G. Palm. Integrating object recognition, visual attention, language and action processing on a robot using a neurobiologically motivated associative architecture. In *Proceedings of the NeuroRobotics Workshop at the 27th German GI Conference on Artificial Intelligence*. University of Ulm, 2004.

[11] W. Gerstner and J.L. van Hemmen. Associative memory in a network of 'spiking' neurons. *Network*, 3:139–164, 1992.

[12] B. Graham and D. Willshaw. Improving recall from an associative memory. *Biological Cybernetics*, 72:337–346, 1995.

[13] D. Greene, M. Parnas, and F. Yao. Multi-index hashing for information retrieval. *Proceedings of the 35th Annual Symposium on Foundations of Computer Science*, pages 722–731, 1994.

[14] D.O. Hebb. *The organization of behavior. A neuropsychological theory.* Wiley, New York, 1949.

[15] R. Hecht-Nielsen. *Confabulation theory.* Springer-Verlag, Heidelberg, 2007.

[16] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science, USA*, 79:2554–2558, 1982.

[17] P. Kanerva. *Sparse Distributed Memory.* MIT Press, Cambridge, MA, 1988.

[18] A. Knoblauch. Optimal matrix compression yields storage capacity 1 for binary Willshaw associative memory. In O. Kaynak, E. Alpaydin, E. Oja, and L. Xu, editors, *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003.*, LNCS 2714, pages 325–332. Springer Verlag, Berlin, 2003.

[19] A. Knoblauch. Synchronization and pattern separation in spiking associative memory and visual cortical areas. *PhD thesis, Department of Neural Information Processing, University of Ulm, Germany*, 2003.

[20] A. Knoblauch. Neural associative memory for brain modeling and information retrieval. *Information Processing Letters*, 95:537–544, 2005.

[21] A. Knoblauch. Statistical implications of clipped Hebbian learning of cell assemblies. *Neurocomputing*, 65–66:647–652, 2005.

[22] A. Knoblauch. On compressing the memory structures of binary neural associative networks. Internal Report HRI-EU 06-02, Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany, April 2006.

[23] A. Knoblauch. Asymptotic conditions for high-capacity neural associative networks. Internal Report HRI-EU 07-02, Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany, February 2007.

[24] A. Knoblauch. On the computational benefits of inhibitory neural associative networks. Internal Report HRI-EU 07-05, Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany, May 2007.

[25] A. Knoblauch. Closed-form expressions for the moments of the binomial probability distribution. *SIAM Journal on Applied Mathematics*, 69(1):197–204, 2008.

[26] A. Knoblauch and G. Palm. Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks*, 14:763–780, 2001.

[27] A. Knoblauch and G. Palm. Scene segmentation by spike synchronization in reciprocally connected visual areas. II. Global assemblies and synchronization on larger space and time scales. *Biological Cybernetics*, 87(3):168–184, 2002.

[28] T. Kohonen. *Associative memory: a system theoretic approach.* Springer, Berlin, 1977.

[29] P.E. Latham and S. Nirenberg. Computing and stability in cortical networks. *Neural Computation*, 16(7):1385–1412, 2004.

[30] D. Marr. Simple memory: a theory for archicortex. *Philosophical Transactions of the Royal Society of London, Series B*, 262:24–81, 1971.

[31] M.L. Minsky and S. Papert. *Perceptrons: An introduction to computational geometry.* MIT Press, Cambridge, MA, 1969.

[32] X. Mu, M. Artiklar, P. Watta, and M.H. Hassoun. An RCE-based associative memory with application to human face recognition. *Neural Processing Letters*, 23:257–271, 2006.

[33] J.-P. Nadal. Associative memory: on the (puzzling) sparse coding limit. *J.Phys. A: Math. Gen.*, 24:1093–1101, 1991.

[34] G. Palm. On associative memories. *Biological Cybernetics*, 36:19–31, 1980.

[35] G. Palm. *Neural Assemblies. An Alternative Approach to Artificial Intelligence.* Springer, Berlin, 1982.

[36] G. Palm. Computing with neural networks. *Science*, 235:1227–1228, 1987.

[37] G. Palm. Memory capacities of local rules for synaptic modification. A comparative review. *Concepts in Neuroscience*, 2:97–128, 1991.

[38] G. Palm and F. Sommer. Associative data storage and retrieval in neural nets. In E. Domany, J.L. van Hemmen, and K. Schulten, editors, *Models of Neural Networks III*, pages 79–118. Springer-Verlag, New York, 1996.

[39] A. Papoulis. *Probability, Random Variables, and Stochastic Processes. Third edition.* McGraw-Hill, New York, 1991.

[40] R.W. Prager and F. Fallside. The modified Kanerva model for automatic speech recognition. *Computer Speech and Language*, 3:61–81, 1989.

[41] D.A. Rachkovskij and E.M. Kussul. Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Computation*, 13:411–452, 2001.

[42] M. Rehn and F.T. Sommer. Storing and restoring visual input with collaborative rank coding and associative memory. *Neurocomputing*, 69:1219–1223, 2006.

[43] F.T. Sommer and G. Palm. Improved bidirectional retrieval of sparse patterns stored by Hebbian learning. *Neural Networks*, 12:281–297, 1999.

[44] K. Steinbuch. Die Lernmatrix. *Kybernetik*, 1:36–45, 1961.

[45] H. Wersing and E. Körner. Learning optimized features for hierarchical models of invariant object recognition. *Neural Computation*, 15:1559–1588, 2003.

[46] D.J. Willshaw, O.P. Buneman, and H.C. Longuet-Higgins. Non-holographic associative memory. *Nature*, 222:960–962, 1969.