# Memory capacities for synaptic and structural plasticity.

## Andreas Knoblauch, Günther Palm, Friedrich Sommer

### 2010

**Preprint:**

*REVIEW* **Communicated by Bruce Graham**

# Memory Capacities for Synaptic and Structural Plasticity

**Andreas Knoblauch**
*andreas.knoblauch@honda-ri.de*
*Honda Research Institute Europe GmbH, D-63073 Offenbach, Germany*

**Günther Palm**
*guenther.palm@uni-ulm.de*
*Institut für Neuroinformatik, Fakultät für Ingenieurwissenschaften und Informatik, Universität Ulm, D-89069 Ulm, Germany*

**Friedrich T. Sommer**
*fsommer@berkeley.edu*
*University of California at Berkeley, Redwood Center for Theoretical Neuroscience, Berkeley, CA 94720-3220, U.S.A.*

**Neural associative networks with plastic synapses have been proposed as computational models of brain functions and also for applications such as pattern recognition and information retrieval. To guide biological models and optimize technical applications, several definitions of memory capacity have been used to measure the efficiency of associative memory. Here we explain why the currently used performance measures bias the comparison between models and cannot serve as a theoretical benchmark. We introduce fair measures for information-theoretic capacity in associative memory that also provide a theoretical benchmark.**

**In neural networks, two types of manipulating synapses can be discerned: *synaptic plasticity*, the change in strength of existing synapses, and *structural plasticity*, the creation and pruning of synapses. One of the new types of memory capacity we introduce permits quantifying how structural plasticity can increase the network efficiency by compressing the network structure, for example, by pruning unused synapses. Specifically, we analyze operating regimes in the Willshaw model in which structural plasticity can compress the network structure and push performance to the theoretical benchmark. The amount $C$ of information stored in each synapse can scale with the logarithm of the network size rather than being constant, as in classical Willshaw and Hopfield nets ($\leq \ln 2 \approx 0.7$). Further, the review contains novel technical material: a capacity analysis of the Willshaw model that rigorously controls for the level of retrieval quality, an analysis for memories with a nonconstant number of active units (where $C \leq 1/e \ln 2 \approx 0.53$), and the analysis of the computational complexity of associative memories with and without network compression.**

 

## 1 Introduction

**1.1 Conventional Versus Associative Memory.** In the classical von Neumann computing architecture, computation and data storage is performed by separate modules, the central processing unit and the random access memory, respectively (Burks, Goldstine, & von Neumann, 1946). A memory address sent to the random access memory gives access to the data content of one particular storage location. *Associative memories* are computing architectures in which computation and data storage are not separated. For example, an associative memory can store a set of associations between pairs of (binary) patterns $\{(\mathbf{u}^\mu \to \mathbf{v}^\mu) : \mu = 1, \ldots, M\}$. Similar to random access memory, a query pattern $\mathbf{u}^\mu$ entered in associative memory can serve as an address for accessing the associated pattern $\mathbf{v}^\mu$. However, the tasks performed by the two types of memory differ fundamentally. Random access is defined only for query patterns that are valid addresses, that is, for the set of $\mathbf{u}$ patterns used during storage. The random access task consists of returning the data record at the addressed location (look-up). In contrast, associative memories accept arbitrary query patterns $\tilde{\mathbf{u}}$, and the computation of any particular output involves all stored data records rather than a single one. Specifically, the associative memory task consists of comparing a query $\tilde{\mathbf{u}}$ with all stored addresses and returning an output pattern equal (or similar) to the pattern $\mathbf{v}^\mu$ associated with the address $\mathbf{u}^\mu$ most similar to the query. Thus, the associative memory task includes the random access task but is not restricted to it. It also includes computations such as pattern completion, denoising, or data retrieval using incomplete cues.

In this review, we compare different implementations of associative memories: First, we study *associative networks*, that is, parallel implementations of associative memory in a network of neurons in which associations are stored in a set of synaptic weights $\mathbf{A}$ between neurons using a local Hebbian learning rule. Associative networks are closely related to Hebbian cell assemblies and play an important role in neuroscience as models of neural computation for various brain structures, for example, neocortex, hippocampus, cerebellum, and mushroom body (Hebb, 1949; Braitenberg, 1978; Palm, 1982; Fransen & Lansner, 1998; Pulvermüller, 2003; Rolls, 1996; Kanerva, 1988; Marr, 1969, 1971; Albus, 1971; Laurent, 2002).

Second, we study compressed associative networks, that is, networks with additional optimal or suboptimal schemes to represent the information contained in the synaptic weight structure efficiently. The analysis of this implementation will enable us to derive a general performance benchmark and understand the difference between structural and synaptic plasticity.

Third, we study sequential implementation of associative memories, that is, computer programs that implement storage (compressed or uncompressed) and memory recall for technical applications and run on an ordinary von Neumann computer.

**1.2 Performance Measures for Associative Memory.** To judge the performance of a computing architecture, one has to relate the size of the achieved computation with the size of required resources. The first popular performance measure for associative memories was the pattern capacity, that is, the ratio between the number of storable association patterns and the number of neurons in the network (Hopfield, 1982). However, in two respects, the pattern capacity is not general enough. First, to compare associative memory with sparse and with dense patterns, the performance measure has to reflect information content of the patterns, not just the count of stored associations. Thus, performance should be measured by the channel capacity of the memory channel, that is, the maximal mutual information (or transinformation) between the stored patterns $\mathbf{v}^\mu$ and the retrieved patterns $\hat{\mathbf{v}}^\mu$ (Cover & Thomas, 1991; Shannon & Weaver, 1949): $T(\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^M; \hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \ldots, \hat{\mathbf{v}}^M)$. Second, the performance measure should take into account the true required storage resources rather than just the number of neurons. The count of neurons in general does not convey the size of the connectivity structure between neurons, which is the substrate where the associations are stored in associative memories. As we will discuss, there is not one universal measure to quantify the storage substrate in associative memories. To reveal theoretical limitations as well as the efficiency of technical and biological implementations of specific models of associative memory, different aspects of the storage substrate are critical. Here we define and compare three performance measures for associative memory models that deviate in how the required storage resources are taken into account.

First, We define (normalized) *network capacity* $C$ as the channel capacity of the associative memory with given network structure, normalized to the number of synaptic contacts between neurons that can accommodate synapses:

$$C = \frac{T(\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^M; \hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \ldots, \hat{\mathbf{v}}^M)}{\#\text{contacts}} \text{ [bit/contact]}. \tag{1.1}$$

In particular, this definition assumes (in contrast to the following two definitions) that the network structure is fixed and independent of the stored data. Definition 1 coincides with the earlier definitions of information-theoretical storage capacity, for example, as employed in Willshaw, Buneman, and Longuet-Higgins (1969), Palm (1980), Amit, Gutfreund, and Sompolinsky (1987b), Nadal (1991), Frolov and Murav'ev (1993), and Palm and Sommer (1996). The network capacity balances computational benefits with the required degree of connectivity between circuit elements. Such a trade-off is important in many contexts, such as chip design and neuroanatomy of the brain. Network capacity quantifies the resources required in a model by just counting contacts between neurons, regardless of the entropy per contact. This property limits the model class for which network capacity defines a

benchmark. Only for associative memories with binary contacts is the network capacity bounded by the value $C = 1$, which marks the achievable optimum as the absolute benchmark. For binary synapses, the normalization constant in the network capacity equals the maximum entropy or Shannon information $I_{\mathbf{A}}$ of the synaptic weight matrix $\mathbf{A}$, assuming statistically independent connections: $C = T/\max[I_{\mathbf{A}}]$. However, in general, network capacity has no benchmark value. Because it does not account for entropy per contact, this measure tends to overestimate the performance of models relying on contacts with high entropy, and conversely, it underestimates models that require contacts with low entropy (cf., Bentz, Hagstroem, & Palm, 1989).

Second, to account for the actual memory requirement of an individual model, we define *information capacity* as the channel capacity normalized by the total entropy in the connections $C^I = T/I(\mathbf{A})$:

$$C^I = \frac{T(\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^M; \hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \ldots, \hat{\mathbf{v}}^M)}{\#\text{ bits of required physical memory}}. \tag{1.2}$$

The information capacity is dimensionless and possesses a model-independent upper bound $C^I_{\text{opt}} = 1$ that defines a general benchmark for associative network models (Knoblauch, 2003a, 2003b, 2005). Note that for efficient implementation of associative memory, large information capacity is necessary but not sufficient. For example, models that achieve large information capacity with low entropy connections rely on additional mechanisms of synaptic compression and decompression to make the implementation efficient. Various compression mechanisms and their neurobiological realizations will be proposed and analyzed in this review. Note further that for models with binary synapses, information capacity is an upper bound of network capacity: $C \leq C^I \leq 1$ (because the memory requirement of the most wasteful model cannot exceed 1 bit per contact).

Third, we define *synaptic capacity* $C^S$ as the channel capacity of the associative memory normalized by the number of nonsilent synapses,

$$C^S = \frac{T(\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^M; \hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \ldots, \hat{\mathbf{v}}^M)}{\#\text{ nonsilent synapses}}\ [\text{bit/synapse}], \tag{1.3}$$

where *nonsilent synapses* are chemical synapses that transmit signals to the postsynaptic cell and have to be metabolically maintained.

Two reasons motivate definition 3. First, the principal cost of neural signaling appears to be restoring and maintaining ionic balances following postsynaptic potentials (Lennie, 2003; Laughlin & Sejnowski, 2003; Attwell & Laughlin, 2001). This suggests that the most critical resource for storing memories in the brain is the physiological maintenance of nonsilent synapses. Thus, our definition of synaptic capacity assesses the number of active synapses commensurate with metabolic energy consumption involved in synaptic transmission.

The second reason is that silent synapses are irrelevant for information retrieval in associative networks (although they are required for storing new information) and could therefore be pruned and replaced by synapses at more useful locations. This idea assumes that the network structure can be adapted to the stored data and has close relations to theoretical considerations about structural plasticity (Stepanyants, Hof, & Chklovskii, 2002; Poirazi & Mel, 2001; Fusi, Drew, & Abbott, 2005). These ideas are also in line with recent neurobiological findings suggesting that structural plasticity (including synaptogenesis and dendritic and axonal growth and remodeling) is a common feature in the physiology of adult brains (Woolley, 1999; Witte, Stier, & Cline, 1996; Engert & Bonhoeffer, 1999; Lamprecht & LeDoux, 2004). Indeed, we have shown in further modeling studies (Knoblauch, 2006, 2009) how ongoing structural plasticity and synaptic consolidation, for example, induced by hippocampal memory replay, can "place" the rare synapses of a sparsely connected network at the most useful locations and thereby greatly increase the information stored per synapse in accordance with our new performance measure $C^S$.

The synaptic capacity is related to the previous definitions of capacity. First, synaptic capacity is an upper bound of the network capacity $C \leq C^S$. Second, for binary synapses with low entropy, the synaptic capacity and the information capacity are proportional $C^S \approx \alpha C^I$: For $r \ll mn$ nonsilent synapses in an $m \times n$-dimensional connectivity matrix $\mathbf{A}$, we have $I_{\mathbf{A}} \approx mnI(r/mn)$ with the single synapse entropy $I(r/mn) \approx r \log(mn)$ (see appendix A) and therefore $\alpha = \log(mn)$. Thus, associative memories with binary low-entropy synapses can be implemented by synaptic pruning, and the upper benchmark is given by $C^S_{\text{opt}} = \log(mn)$.

Finally, we give an example illustrating when and how the three different performance measures are applicable. Consider storing 1 kilo bits of information in a neural network $\mathbf{A}$ of $100 \times 100$ binary synapses, and let 150 of the 10,000 synapses have weight 1. Then the network capacity of the static fully connected net is simply $C = 1000/10{,}000 = 0.1$ bit per binary synapse. However, the synaptic weight matrix $\mathbf{A}$ has only sparsely one-entries with a single synapse entropy of $I(150/10{,}000) = 0.1124$ bit. Then $\mathbf{A}$ can be compressed such that the memory requirements for a computer implementation could decrease to only $I(\mathbf{A}) = 1124$ bit. Thus, the information capacity would be $C^I = 1000/1124 = 0.89$. In a sparsely connected biological network endowed with structural plasticity, it would be possible to prune silent synapses, regenerate new synapses at random locations, and consolidate synapses only at useful positions. Such a network could get along with only 150 nonsilent synapses such that the resulting synaptic capacity is $C^S = 1000/150 = 6.7$ bits per synapse.

**1.3 Associative Memory Models and Their Performance.** How do known associative models perform in terms of the capacities we have introduced? The network capacity was first applied to the Willshaw or Steinbuch

model (Willshaw et al., 1969; Palm, 1980), a feedforward neural associative network with binary neurons and synapses first proposed by Steinbuch (1961; see section 2.2 of this review). The feedforward heteroassociative Willshaw model can achieve a network capacity of $C = \ln 2 \approx 0.7$ bits per contact. The model performs high compared to alternative neural implementations of associative memory with nonbinary synapses and feedback network architectures, which became very popular in the 1980s (Hopfield, 1982, 1984; Hopfield & Tank, 1986; Hertz, Krogh, & Palmer, 1991). The network capacity of the original (nonsparse) Hopfield model stays with 0.14 bits per contact (Amit, Gutfreund, & Sompolinsky, 1987a; Amit et al., 1987b) far below the one for the Willshaw model (see Schwenker, Sommer, & Palm, 1996; Palm, 1991).

The difference in network capacity between the Willshaw model and the Hopfield model turns out to be due to differences in the stored memory patterns. The Willshaw model achieves high network capacity with extremely sparse memory patterns, that is, with a very low ratio between active and nonactive neurons. Conversely, the original Hopfield model is designed for nonsparse patterns with even ratio between active and nonactive neurons. Using sparse patterns in the feedforward Hopfield network with accordingly adjusted synaptic learning rule (Palm, 1991; Dayan & Willshaw, 1991; Palm & Sommer, 1996) increases the network capacity to $1/(2 \ln 2) \approx 0.72$ (Tsodyks & Feigel'man, 1988; Palm & Sommer, 1992). Thus, in terms of network capacity, the sparse Hopfield model outperforms the Willshaw model, but only marginally. The picture is similar in terms of synaptic capacity since the number of nonsilent synapses is the same in both models. However, the comparison between Willshaw and Hopfield model changes significantly when estimating the information capacities. If one assumes a fixed number of bits $h$ assigned to represent each synaptic contact, the network capacity defines a lower bound on the information capacity by $C^I \geq C/h \geq C/\#\{\text{bits per contact}\}$. Thus, for the Willshaw model (with $h = 1$), the information capacity is $C^I \geq 0.69$. In contrast, assuming $h = 2$ in the sparse Hopfield model yields a significantly lower information capacity of $C^I \geq 0.72/2 = 0.36$. In practice, $h > 2$ is used to represent the synapses with sufficient precision, which increases the advantage of the Willshaw model even more.

**1.4 The Willshaw Model and Its Problems.** Since the Willshaw model is not only among the simplest realizations of content-addressable memory but is also promising in terms of information capacity, it is interesting for applications as well as for modeling the brain. However, the original Willshaw model suffers from a number of problems that prevented broader technical application and limited its biological relevance. First, the basic Willshaw model approaches $C = \ln 2$ only for very large (i.e., not practical) numbers $n$ of neurons, and the retrieval accuracy at maximum network capacity is low (Palm, 1980; Buckingham & Willshaw, 1992). Various studies have

shown, however, that modifications of the Willshaw model can overcome this problem: Iterative and bidirectional retrieval schemes (Schwenker et al., 1996; Sommer & Palm, 1999), improved threshold strategies (Buckingham & Willshaw, 1993; Graham & Willshaw, 1995), and retrieval with spiking neurons (Knoblauch & Palm, 2001; Knoblauch, 2003b) can significantly improve network capacity and retrieval accuracy in small memory networks.

But two other problems of the Willshaw model and its derivates remain so far unresolved. The first open question is the sparsity problem that is, the question of whether there is a way to achieve high capacity outside the regime of extreme sparseness in which the number of one-entries $k$ in memory patterns is logarithmic in the pattern size $n$: $k = c \log n$ for a constant $c$ (cf. Figure 3). In the standard model, even small deviations from this sparseness condition reduce network capacity drastically. Although it was possible for some applications to find coding schemes that fulfill the strict requirements for sparseness (Bentz et al., 1989; Rehn & Sommer, 2006), the sparse coding problem cannot be solved in general. The extreme sparsity requirement is problematic not only for applications (e.g., see Rachkovskij & Kussul, 2001) but also for brain modeling because it is questionable whether neural cell assemblies that satisfy the sparseness condition are stable with realistic rates of spontaneous activity (Latham & Nirenberg, 2004). At least for sparsely connected networks realizing only a small given fraction $P$ of the possible synapses, it is possible to achieve nonzero capacities up to $0.53 \leq C \leq 0.69$ for a larger but still logarithmic pattern activity $k = c \log n$, where the optimal $c \to \infty$ increases with decreasing $P \to 0$ (Graham & Willshaw, 1997; Bosch & Kurfess, 1998; Knoblauch, 2006).

The second open question concerning the Willshaw model is the capacity gap problem, that is, the question of why the optimal capacity $C = \ln 2$ is separated by a gap of 0.3 from the theoretical optimum $C = 1$. This question implicitly assumes that the optimal representation of the binary storage matrix is the matrix itself—that is, the distinction between the capacities $C$ and $C^I$ defined here is simply overlooked. For many decades, the capacity gap was considered an empirical fact for distributed storage (Palm, 1991). Although we cannot solve the capacity gap and sparsity problems for the classical definition of $C$, we propose models optimizing $C^I$ (or $C^S$) that can achieve $C^I = 1$ (or $C^S = \log n$) without requiring extremely sparse activity.

**1.5 Organization of the Review.** In section 2 we define the computational task of associative memory, including different levels of retrieval quality. Further, we describe the particular model of associative memory under investigation, the Willshaw model.

Section 3 contains a detailed analysis of the classical Willshaw model, capturing its strengths and weaknesses. We revisit and extend the classical capacity analysis, yielding a simple formula how the optimal network capacity of $C = 0.69$ bits per contact decreases as a function of the noise level in the address pattern. Further, we demonstrate that high values of network

capacity are tightly confined to the regime of extreme sparseness and, in addition, that finite-sized networks cannot achieve high network capacity at a high retrieval quality.

In section 4, the capacity analysis is extended to the new capacity measures we have defined in section 1, to information capacity and synaptic capacity. The analysis of information capacity reveals two efficient regimes that curiously do not coincide with the regime of logarithmic sparseness in which the network capacity is optimal. Interestingly, in the two efficient regimes, the ultrasparse regime ($k < c \log n$) and the regime of moderate sparseness ($k > c \log n$), the information capacity becomes even optimal, that is, $C^I = 1$. Thus, our analysis shows that the capacity gap problem is caused by the bias inherent in the definition of network capacity. Further, the discovery of a regime with optimal information capacity at moderate sparseness points to a solution of the sparsity problem. The analysis of synaptic capacity reveals that if the number of active synapses rather than the total number of synaptic contacts is the critical constraint, the capacity in finite-size associative networks increases from less than 0.5 bit per synaptic contact to about 5 to 10 bits per active synapse.

In section 5 we consider the computational complexity of the retrieval process. We focus on the time complexity for a sequential implementation on a digital computer, but the results can also be interpreted metabolically in terms of energy consumption since retrieval time is dominated by the number of synaptic operations. In particular, we compare two-layer implementations of the Willshaw model to three-layer implementations or look-up tables with an additional hidden grandmother cell layer.

After the discussion in section 6, appendix A gives an overview of binary channels. Appendix B reviews exact formulas for analyzing the Willshaw models with fixed pattern activity that are used to verify the results of this review and compute exact capacities for various finite network sizes (see Table 2). Appendix C points out some fallacies with previous analyses, for example, relying on gaussian approximations of dendritic potential distributions. Finally, appendix D extends our theory to random pattern activity, where it turns out $C \leq 1/(e \ln 2)$.

## 2 Associative Memory: Computational Task and Network Model

**2.1 The Memory Task.** *Associative memories* store information about a set of memory patterns. For retrieving memories, three different computational tasks have been discussed in the literature. The first task is familiarity discrimination, a binary classification of input patterns into known and unknown patterns (Palm & Sommer, 1992; Bogacz, Brown, & Giraud-Carrier, 2001). The second task is autoassociation or pattern completion, which involves completing a noisy query pattern to the memory pattern that is most similar to the query (Hopfield, 1982). Here we focus on the third task, heteroassociation, which is most similar to the function of a random

access memory. The memorized patterns are organized in association pairs $\{(\mathbf{u}^\mu \mapsto \mathbf{v}^\mu) : \mu = 1, \ldots, M\}$. During retrieval, the memory performs associations within the stored pairs of patterns. If a pattern $\mathbf{u}^\mu$ is entered, the associative memory produces the pattern $\mathbf{v}^\mu$ (Kohonen, 1977). Thus, in analogy to random access memories, the $\mathbf{u}$-patterns are called *address patterns* and the $\mathbf{v}$-patterns are called *content patterns*. However, the associative memory task is more general than a random access task in that arbitrary query patterns are accepted, not just the set of $\mathbf{u}$-patterns. A *query pattern* $\tilde{\mathbf{u}}$ will be compared to all stored $\mathbf{u}$-patterns, and the best match $\mu$ will be determined. The memory will return an *output pattern* $\hat{\mathbf{v}}$ that is equal or similar to the stored content pattern $\mathbf{v}^\mu$. Note that autoassociation is a special case of heteroassociation (for $\mathbf{u}^\mu = \mathbf{v}^\mu$) and that both tasks are variants of the best match problem in Minsky and Papert (1969). Efficient solutions of the best match problem have widespread applications, for example, for cluster analysis, speech and object recognition, or information retrieval in large databases (Kohonen, 1977; Prager & Fallside, 1989; Greene, Parnas, & Yao, 1994; Mu, Artiklar, Watta, & Hassoun, 2006; Rehn & Sommer, 2006).

*2.1.1 Properties of Memory Patterns.* In this review, we focus on the case of binary pattern vectors. The address patterns have dimension $m$, and the content patterns have dimension $n$. The number of one-entries in a pattern is called the *pattern activity*. The mean activity in each address pattern $\mathbf{u}^\mu$ is $k$, which means that, on average, it has $k$ one-entries and $m - k$ zero-entries. Analogously, the mean activity in each content pattern $\mathbf{v}^\mu$ is $l$. Typically the patterns are sparse, which means that the pattern activity is much smaller than the vector size (e.g., $k \ll m$). For the analyses, we assume that the $M$ pattern pairs are generated randomly according to one of the following two methods. First, in the case of *fixed* pattern activity, each pattern has exactly the same activity. For address patterns, for example, this means that each of the $\binom{m}{k}$ binary vectors of size $m$ and activity $k$ has the same chance to be chosen. Second, in the alternative case of *random* pattern activity, pattern components are independently generated. For address patterns, for example, this means that a pattern component $u_i^\mu$ is one with probability $k/m$ and zero otherwise, independent of other components. It turns out that the distinction between constant and random pattern activity is relevant only for address patterns, not for content patterns. Binary memory patterns can be distorted by two distinct types of noise: *add noise* means that false one-entries are added, and *miss noise* means that one-entries are deleted. The rates of these error types in query and output patterns determine two key features of associative memories: noise tolerance and retrieval quality.

*2.1.2 Noise Tolerance.* To assess how much query noise can be tolerated by the memory model, we form query patterns $\tilde{\mathbf{u}}$ by adding random noise to the $\mathbf{u}$-patterns. For our analyses in the main text, we assume that a query pattern $\tilde{\mathbf{u}}$ has exactly $\lambda k$ "correct" and $\kappa k$ "false" one-entries. Thus, query

patterns have fixed pattern activity $(\lambda + \kappa)k$ (see appendix D for random query activity). Query noise and cross-talk between the stored memories can lead to noise in the output of the memory. Output noise is expressed in deviations between retrieval output $\hat{\mathbf{v}}$ and the stored $\mathbf{v}$-patterns.

*2.1.3 Retrieval Quality.* Increasing the number $M$ of stored patterns will eventually increase the output noise introduced by cross-talk. Thus, in terms of the introduced capacity measures, there will be a trade-off between memory load that increases capacity and the level of output noise that decreases capacity. In many situations, a substantial information loss due to output errors can be compensated by the high number of stored memories, and the capacity is maximized at high levels of output errors. For applications, however, this low-fidelity regime is not interesting, and one has to assess capacity at specified low levels of output noise. Based on the expectation $E_\mu$ of errors per output pattern or Hamming distance $h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) := \sum_{j=1}^{n} |v_j^\mu - \hat{v}_j^\mu|$, we define different retrieval qualities (RQ) that will be studied:

- RQ0: $E_\mu h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) = l p_{10} + (n-l) p_{01} \le \rho_0 n$
- RQ1: $E_\mu h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) = l p_{10} + (n-l) p_{01} \le \rho_1 l$
- RQ2: $E_\mu h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) = l p_{10} + (n-l) p_{01} \le \rho_2$
- RQ3: $E_\mu h(\mathbf{v}^\mu, \hat{\mathbf{v}}^\mu) = l p_{10} + (n-l) p_{01} \le \rho_3 / M$,

where $p_{10} := \mathrm{pr}[\hat{\mathbf{v}}_j^\mu = 0 \mid \mathbf{v}_j^\mu = 1]$ and $p_{01} := \mathrm{pr}[\hat{\mathbf{v}}_j^\mu = 1 \mid \mathbf{v}_j^\mu = 0]$ are the component error probabilities and $\rho_0, \rho_1, \rho_2, \rho_3$ are (typically small) constants. Note that the required quality is increasing from RQ0 to RQ3. Asymptotically for $n \to \infty$, RQ0 requires small constant error probabilities, RQ1 requires the expected number of output errors per pattern to be a small fraction of pattern activity $l$, RQ2 requires the expected number of output errors per pattern to be small, and RQ3 requires the total number of errors (summed over the recall of all $M$ stored patterns) to be small. Making these distinctions explicit allows a unified analysis of associative networks and reconciles discrepancies between previous works (cf. Nadal, 1991).

**2.2 The Willshaw Model.** To represent the described associative memory task in a neural network, neurons with binary values are sufficient, although for the computation neurons with continuous values can be beneficial (Anderson, Silverstein, Ritz, & Jones, 1977; Anderson, 1993; Hopfield, 1984; Treves & Rolls, 1991; Sommer & Dayan, 1998). The patterns $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ describe the activity states of two populations of neurons at time $\mu$. In neural associative memories, the associations are stored in the *synaptic matrix* or *memory matrix*.

*2.2.1 Storage.* In the Willshaw or Steinbuch model (Willshaw et al., 1969; Steinbuch, 1961; Palm, 1980, 1991), not only neurons but also synapses

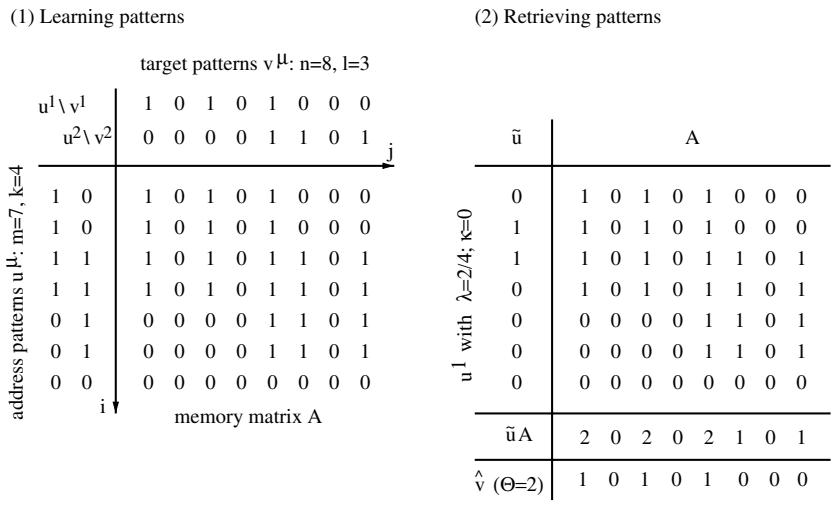


Figure 1: Learning and retrieving patterns in the binary Willshaw model. During learning (left), the associations between a set of address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$ are stored in the synaptic memory matrix $\mathbf{A}$ by clipped Hebbian learning (see equation 2.1). For retrieval (right), a query pattern $\tilde{\mathbf{u}}$ is propagated through the synaptic network by a vector-matrix multiplication followed by a threshold operation (see equation 2.2). In the example, the query pattern contains half of the one-entries of $\mathbf{u}^1$, and the retrieval output $\hat{\mathbf{v}}$ equals $\mathbf{v}^1$ for an optimal threshold $\Theta = |\tilde{\mathbf{u}}| = 2$.

have binary values. The storage and retrieval processes work as follows. The pattern pairs are stored heteroassociatively in a binary memory matrix $\mathbf{A} \in \{0,1\}^{m \times n}$ (see Figure 1), where

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

The network architecture is feedforward; thus, an address population $u$ consists of $m$ neurons projects via the synaptic matrix $\mathbf{A}$ to a content population $v$ consisting of $n$ neurons. Note that the memory matrix is formed by local Hebbian learning, that is, $A_{ij}$ is a (nonlinear) function of the activity values in the pre- and postsynaptic neuron $u_i$ and $v_j$ regardless of other activity in the network. Note further that for the autoassociative case $u = v$ (i.e., if address and content populations are identical), the network can be interpreted as an undirected graph with $m = n$ nodes and edge matrix $\mathbf{A}$, where patterns correspond to cliques of $k = l$ nodes.

*2.2.2 Retrieval.* Stored information can be retrieved by entering a query pattern $\tilde{\mathbf{u}}$. First, a vector-matrix-multiplication yields the dendritic potentials $\mathbf{x} = \tilde{\mathbf{u}} \cdot \mathbf{A}$ in the content neurons. Second, a threshold operation in each content neuron results in the retrieval output $\hat{\mathbf{v}}$,

$$\hat{v}_j = \begin{cases} 1, & x_j = \left( \sum_{i=1}^{m} \tilde{u}_i A_{ij} \right) \geq \Theta \\ 0, & \text{otherwise} \end{cases} . \tag{2.2}$$

A critical prerequisite for high-retrieval quality is the right choice of the threshold value $\Theta$. Values that are too low will lead to high rates of add-errors, whereas values that are too high will result in high rates of miss-errors. A good threshold value is the number of correct one elements in the address pattern because it yields the lowest rate of add errors in the retrieval while still avoiding miss errors entirely. Depending on the types of errors present in the address, this threshold choice can be simple or rather difficult.

For the cases of error-free addresses ($\lambda = 1$ and $\kappa = 0$) and pattern part retrieval, that is, when the address contains miss errors only ($0 < \lambda \leq 1$ and $\kappa = 0$), the optimal threshold value is a simple function of the address pattern $\Theta = |\tilde{\mathbf{u}}| := \sum_{i=0}^{m} \tilde{u}_i$. This threshold value was used in the original Willshaw model, and therefore we will refer to it as the *Willshaw threshold*. This threshold setting can be easily implemented in technical systems and is also biologically very plausible, for example, based on feedforward inhibition via "shadow" interneurons (cf. Knoblauch & Palm, 2001; Knoblauch, 2003b, 2005; Aviel, Horn, & Abeles, 2005).

For the general case of noisy addresses, including miss and add errors ($0 < \lambda \leq 1, \kappa \geq 0$) the optimal threshold is no simple function of the address pattern $\tilde{\mathbf{u}}$. In this case, the number of correct ones is uncertain given the address, and therefore the threshold strategies have to estimate this value based on priori knowledge of $\kappa$ and $\lambda$.

**2.3 Two-Layer Associative Networks and Look-Up Tables.** Essentially the Willshaw model is a neural network with a single layer of neurons $v$ that receive inputs from an address pattern $u$. A number of memory models in the literature can be regarded as an extension of the Willshaw model by adding an intermediate layer of neurons $w$ (see Figure 2). If for each association to be learned, $\mathbf{u}^\mu \to \mathbf{v}^\mu$, one would activate an additional random pattern $\mathbf{w}^\mu$, the two memory matrices $\mathbf{A_1}$ and $\mathbf{A_2}$ would store associations $\mathbf{u}^\mu \to \mathbf{w}^\mu$ and $\mathbf{w}^\mu \to \mathbf{v}^\mu$, respectively. Thus, the two-layer memory would function analogously to the single-layer model (see equation 2.1). However, the two-layer model can be advantageous if address and content patterns are nonrandom or nonsparse because in such cases, the performance of the single-layer model is severely impaired (Knoblauch, 2005; Bogacz and Brown, 2003). The advantage of two-layer models is related to the fact that single-layer perceptrons can learn only linearly separable mappings $\mathbf{u}^\mu \to$

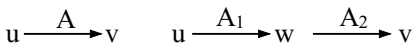


Figure 2: Single-layer Willshaw model (left) and two-layer extension (right) where an additional cell layer $w$ mediates between address layer $u$ and content layer $v$.

$\mathbf{v}^\mu$, while arbitrary mappings require at least a second (hidden) layer. Instead of choosing random patterns $\mathbf{w}^\mu$, one can also try to optimize the intermediary pattern representations. Another interesting model of a two-layer memory is the Kanerva network, where the first memory matrix $\mathbf{A_1}$ is a fixed random projection, and only the second synaptic projection $\mathbf{A_2}$ is learned by Hebbian plasticity (Kanerva, 1988). In addition, two-layer memories are neural implementations of look-up tables if the intermediary layer $w$ has a single active (grandmother) neuron for each association to be stored. In this case, the two memory matrices $\mathbf{A_1}$ and $\mathbf{A_2}$ degenerate to simple look-up tables where the $\mu$th row contains the $\mu$th pattern, respectively. In section 5, we will compare the single-layer model to the two-layer (grandmother cell or look-up table) model. Surprisingly, we will find that the performance of the grandmother cell model is superior to that of the single-layer model in many cases. This is true at least for technical applications, while for biology, the large number of neurons required in the middle layer may be unrealistic, even when it would be possible to select single cells in a WTA-like manner.

## 3 Analysis of Network Capacity

**3.1 Asymptotic Analysis of Network Capacity.** This section summarizes and extends the classical asymptotic analysis of the Willshaw model (Willshaw et al., 1969; Palm, 1980). The fraction of one-entries in the memory matrix $p_1 := \sum_{ij} \mathbf{A}_{ij}/mn$ is a monotonic function of the number of stored patterns and will therefore be referred to as *matrix load* or *memory load*. The probability that a physically present synapse is not activated by the association of one pattern pair is $1 - kl/mn$. Therefore, after learning $M$ patterns, the matrix load is given by

$$p_1 = 1 - \left(1 - \frac{kl}{mn}\right)^M. \tag{3.1}$$

It is often convenient to use equation 3.1 to determine the number of stored patterns,

$$M = \frac{\ln(1 - p_1)}{\ln(1 - kl/mn)} \approx -\frac{mn}{kl} \ln(1 - p_1), \tag{3.2}$$

where the approximation is valid for $kl \ll mn$.

The general analysis of retrieval includes queries $\tilde{\mathbf{u}}$ that contain both noise types, that is, $\lambda \cdot k$ "correct" and $\kappa \cdot k$ "false" one-entries ($0 < \lambda \leq 1$; $0 \leq \kappa$). For clarity, we start with the analysis of pattern part retrieval where the query pattern contains no add noise, that is, $\kappa = 0$ (for investigations of the general case, see section 4.5). For pattern part retrieval with fixed query activity and Willshaw threshold $\Theta = |\tilde{\mathbf{u}}| = \lambda k$, the probability of add noise in the retrieval is

$$p_{01} = p(\hat{v}_i = 1 \mid v_i^\mu = 0) \gtrapprox {p_1}^{\lambda k}. \tag{3.3}$$

(For exact formulas, see equations B.6 to B.8 in appendix B. For random query activity, see appendix D.) The following analysis is based on the binomial approximation equation 3.3, which assumes independently generated one-entries in a subcolumn of the memory matrix. Although this is obviously not true for distributed address patterns with $k > 1$, the approximation is sufficiently exact for most parameter ranges. Knoblauch (2007, 2008) shows that equation 3.3 is generally a lower bound and becomes exact at least for $k = O(n/\log^4 n)$.

With the error probability $p_{01}$, one can compute the mutual information between the memory output and the original content. The mutual information in one pattern component is $T(l/n,\, p_{01}, 0)$ (see equation A.5). When $Mn$ such components are stored, the network capacity $C(k, l, m, n, \lambda, M)$ of equation 1.1 is

$$C = \frac{M}{m} T\left(\frac{l}{n},\, p_{01}, 0\right) \leq \frac{\mathrm{ld}\, p_{01} \ln(1 - p_1)}{k} \tag{3.4}$$

$$\leq \lambda \cdot \mathrm{ld}\, p_1 \cdot \ln(1 - p_1) \leq \lambda \ln 2, \tag{3.5}$$

where we used the bound equation A.6 and the binomial approximation equation 3.3. The first equality is strictly correct only for random activity of address patterns, but still a tight approximation for fixed address pattern activity. The first bound becomes tight at least for $(l/n)/p_{01} \to 0$ (see equation A.6), the second bound for $k \sim O(n/\log^4 n)$ (see references above), and the third bound for $p_1 = 0.5$ and $M \approx 0.69mn/kl$.

Thus, the original Willshaw model can store at most $C = \ln 2 \approx 0.69$ bits per synapse for $\lambda = 1$ (however, for random query activity, we achieve at most $C = 1/(e \ln 2) \approx 0.53$ bits per synapse; see appendix D). The upper bound can be reached for sufficiently sparse patterns, $l \ll n$, $k \ll m$, and balanced memory matrix with an equal number of active and inactive synapses. Strictly speaking, the requirement $(l/n)/p_{01} \ll 1$ implies only low retrieval quality, with the number of false one-entries exceeding the number of correct one-entries $l$. The following section shows that the upper bound can also be reached at higher levels of retrieval quality.

**3.2 Capacity Analysis for Defined Grades of Retrieval Quality.** To ensure a certain retrieval quality, we bound the error probability $p_{01}$ by $p_{01\epsilon}$,

$$p_{01} \leq p_{01\epsilon} := \frac{\epsilon l}{n-l} \Leftrightarrow \lambda \geq \lambda_\epsilon :\approx \frac{\ln \frac{\epsilon l}{n-l}}{k \ln p_1}, \tag{3.6}$$

where we call $\epsilon > 0$ the *fidelity parameter*. For the approximation of minimal address pattern fraction $\lambda_\epsilon$, we again used the binomial approximation equation 3.3. Note that for $p_{10} = 0$ and constant $\epsilon$, this condition ensures retrieval quality of type RQ1 (see section 2.1). More generally, to ensure retrieval quality RQ0-3 at levels $\rho_0 - \rho_3$, the fidelity parameter $\epsilon$ has to fulfill the following conditions:

- RQ0: $\epsilon \leq \rho_0 \frac{n}{l}$
- RQ1: $\epsilon \leq \rho_1$
- RQ2: $\epsilon \leq \rho_2 / l$
- RQ3: $\epsilon \leq \rho_3 \frac{1}{Ml}$.

As one stores more and more patterns, the matrix load $p_1$ increases, and the noise level $\lambda_\epsilon$ that can be afforded in the address to achieve the specified retrieval quality drops. Therefore, the maximum number of patterns that can be stored is reached at the point where $\lambda_\epsilon$ reaches the required fault tolerance: $\lambda_\epsilon = \lambda$ (see equation 3.6). Accordingly, the maximum matrix load (and the optimal activity of address patterns) is given by

$$p_{1\epsilon} \approx \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda k}} \qquad \left(\Leftrightarrow k \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{\lambda \mathrm{ld}\, p_{1\epsilon}}\right). \tag{3.7}$$

Thus, with equations 3.2 and 3.4, we obtain the maximal number of stored patterns: the *pattern capacity* $M_\epsilon$ and the *network capacity* $C_\epsilon(k, l, m, n, \lambda, \epsilon) \approx M_\epsilon m^{-1} T(l/n, \epsilon l/(n-l)), 0)$,

$$M_\epsilon \approx -\lambda^2 \cdot (\mathrm{ld}\, p_{1\epsilon})^2 \cdot \ln(1 - p_{1\epsilon}) \cdot \frac{k}{l} \cdot \frac{mn}{\left(\mathrm{ld}\frac{n-l}{\epsilon \cdot l}\right)^2} \tag{3.8}$$

$$C_\epsilon \approx \lambda \cdot \mathrm{ld}\, p_{1\epsilon} \cdot \ln(1 - p_{1\epsilon}) \cdot \eta, \tag{3.9}$$

where

$$\eta := \frac{T\left(\frac{l}{n}, \frac{\epsilon l}{n-l}, 0\right)}{-\frac{l}{n}\mathrm{ld}\frac{\epsilon l}{n-l}} = \frac{T\left(\frac{l}{n}, \frac{\epsilon l}{n-l}, 0\right)}{I(\frac{l}{n})}$$
$$\cdot \left(\frac{1}{1 + \frac{\ln \epsilon}{\ln(l/n)} - \frac{\ln(1-l/n)}{\ln(l/n)}} + \frac{(n-l)\mathrm{ld}(1-l/n)}{l\mathrm{ld}(\epsilon l/n)}\right) \tag{3.10}$$

$$\approx \frac{1}{1+\frac{\ln \epsilon}{\ln(l/n)}}. \tag{3.11}$$

The approximation equation 3.11 is valid for small $\epsilon, l/n \ll 1$. For high-fidelity recall with small $\epsilon \ll 1$, the error $e_I$ of approximating $T$ by $I$ becomes negligible and even $T/I = (1 - e_I) \to 1$ for $l/n \to 0$ (see equation A.9 for details). For sparse content patterns with $l/n \ll 1$, we have $I(l/n) \approx -(l/n)\mathrm{ld}(l/n)$ (see equation A.1), and the right summand in the brackets can be neglected. Finally, the left summand in the brackets of equation 3.10 becomes 1 for $\ln \epsilon / \ln(l/n) \to 0$.

The next two figures illustrate the results of this analysis with an example: a Willshaw network with a square-shaped memory matrix ($m = n$). The address and content patterns have the same activity ($k = l$), and the input is noiseless, that is, $\lambda = 1$, $\kappa = 0$. Figure 3 presents results for a network with $n = 100{,}000$ neurons, a number that corresponds roughly to the number of neurons below 1 square millimeter of cortex surface (Braitenberg & Schüz, 1991; Hellwig, 2000). Figure 3a shows that high network capacity is assumed in a narrow range around the optimum pattern activity $k_{\mathrm{opt}} = 18$ and decreases rapidly for larger or smaller values. For the chosen fidelity level $\epsilon = 0.01$, the maximum network capacity is $C_\epsilon \approx 0.5$, which is significantly below the asymptotic bound. The dashed line shows how the memory load $p_{1\epsilon}$ increases monotonically with $k$ from 0 to 1. The maximum network capacity is assumed near $p_{1\epsilon} = 0.5$, similar to the asymptotic calculation. Note that the number of patterns $M_\epsilon$ becomes maximal at smaller values $p_{1\epsilon} < 0.5$ ($M_\epsilon \approx 29.7 \cdot 10^6$ for $k = 8$ and $p_{1\epsilon} \approx 0.17$).

Figure 3b explores the case where pattern activity is fixed to the value $k = 18$, which was optimal in Figure 3a, for variable levels of fidelity. The most important observation is that not only the maximum number of patterns, but also the maximum network capacity is obtained for low fidelity: $C \approx 0.64$ occurs for $\epsilon \approx 1.4$. This means that in a finite-sized Willshaw network, a high number of stored patterns outbalances the information loss due to the high level of output errors, an observation made also by Nadal and Toulouse (1990) and Buckingham and Willshaw (1992). However, most applications require low levels of output errors and therefore cannot use the maximum network capacity. Technically the pattern capacity $M$ is unbounded since $M_\epsilon \to \infty$ for $\epsilon \to n/l - 1$. However, this transition corresponds to $p_{1\epsilon} \to 1$ and $p_{01\epsilon} \to 1$, which means that the stored patterns cannot be retrieved anymore. The contour plots in Figures 3c to 3e give an overview of how network capacity, memory load, and the maximum number of stored patterns vary with pattern activity and fidelity level. High-quality retrieval with small $\epsilon$ requires generally larger assembly size $k$. For fixed fidelity level $\epsilon$, optimal $k$ for maximal $M$ is generally smaller than optimal $k$ for maximal $C$ (the latter has about double size; cf. Knoblauch, Palm, & Sommer, 2008).

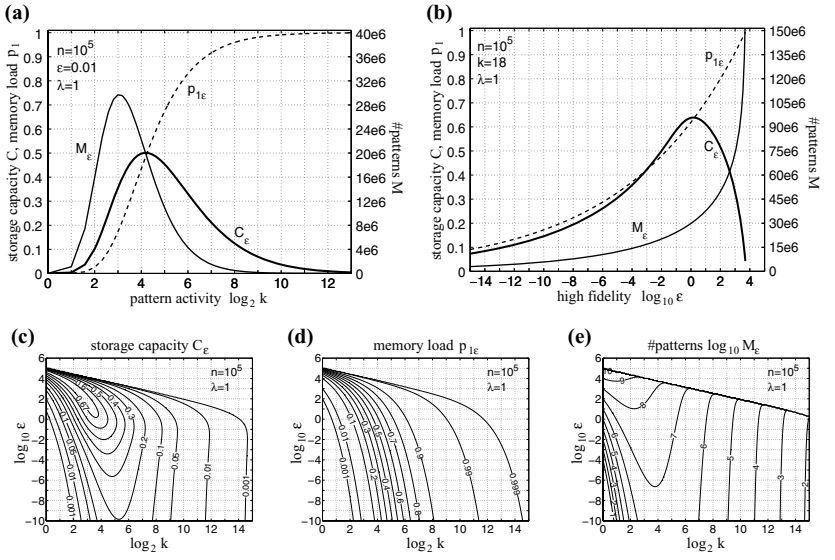


Figure 3: Classical capacity measures $C$ and $M$ for a finite Willshaw network with $m = n = 10^5$ neurons assuming equal pattern activities, $k = l$, and zero input noise, $\lambda = 1$, $\kappa = 0$. (a) Network capacity $C_\epsilon$ (bold line), pattern capacity $M_\epsilon$ (thin line), and memory load $p_{1\epsilon}$ (dashed line) as functions of pattern activity $k$ (log scale). The fidelity level is $\epsilon = 0.01$. The maximum $C_\epsilon \approx 0.49$ is reached for $k = 18$. For larger or smaller $k$, the capacity decreases rapidly. The memory load $p_{1\epsilon}$ increases monotonically with $k$ and is near 0.5 at maximum capacity. (b) Same quantities as in $a$ plotted as functions of $\epsilon$ (log scale) assuming fixed $k = 18$. The maximum $C_\epsilon \approx 0.63$ is reached at low fidelity ($\epsilon \approx 1$) where the retrieval result contains a high level of add noise. (c–e) : Contour plots in the plane spanned by pattern activity $k$ and high-fidelity parameter $\epsilon$ for network capacity $C_\epsilon$ (c), memory load $p_{1\epsilon}$ (d), and pattern capacity $M_\epsilon$ (e).

**3.3 Refined Asymptotic Analysis for Large Networks.** Section 3.2 delineated a theory for the Willshaw associative memory that predicts pattern capacity and network capacity for finite network sizes and defined levels of retrieval quality. Here we use this theory to specify the conditions under which large networks reach the optima of network capacity $C_\epsilon \to \lambda \ln 2$ and pattern capacity $M_\epsilon$. We focus on the case $k \sim l$, which applies to autoassociative memory tasks and heteroassociative memory tasks if the activities of address and content patterns are similar. The results displayed in Figure 4 can be compared to the predictions of the classical analysis recapitulated in section 3.1. Several important observations can be made:

- The upper bound of network capacity can in fact be reached by equation 3.9 for arbitrary small constant $\epsilon$, that is, at retrieval-quality

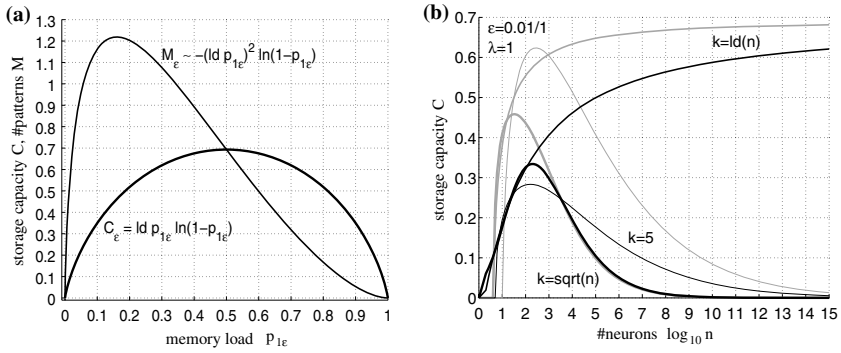


Figure 4: Classical capacity measures $C$ and $M$ for the Willshaw network in the asymptotic limit $n \to \infty$. Other parameter settings are as in Figure 3: $m = n$, $k = l$, $\lambda = 1$, and $\kappa = 0$. (a) Network capacity $C_\epsilon \to \mathrm{ld}\, p_{1\epsilon} \ln(1 - p_{1\epsilon})$ (bold line; see equation 3.9) and pattern capacity $M_\epsilon/(mn/(\mathrm{ld}n)^2) \to -(\mathrm{ld}\, p_{1\epsilon})^2 \ln(1 - p_{1\epsilon})$ (thin line; see equation 3.8) as functions of the matrix load $p_{1\epsilon}$ (see equation 3.7). $C_\epsilon$ is maximal for $p_{1\epsilon} = 0.5$, whereas $M_\epsilon$ is maximal for $p_{1\epsilon} \approx 0.16$. (b) Network capacity $C_\epsilon$ as a function of $n$ for different functions of pattern activity $k(n)$. Black lines correspond to high-fidelity retrieval with $\epsilon = 0.01$, gray lines to low fidelity with $\epsilon = 1$. Bold lines: square root sparseness; solid lines: logarithmic sparseness; thin lines: low, constant activity ($k = 5$).

grade RQ1 at arbitrary high fidelity: $C_\epsilon \to \lambda \ln 2$ for $m, n \to \infty$ and $p_{1\epsilon} \to 0.5$. The latter condition requires logarithmic pattern sparseness $k = \mathrm{ld}n/\lambda$ (see equation 3.7).

- At retrieval quality grade RQ1, network capacity and pattern capacity assume their optima for somewhat different parameter settings. The pattern capacity $M_\epsilon$ (see equation 3.8) peaks at a memory load $p_{1\epsilon} \approx 0.16$, which also requires logarithmic sparseness in the memory patterns, but with a smaller constant than for maximizing network capacity: $k = \mathrm{ld}n/(\lambda \mathrm{ld} 6.25)$. The optimal pattern capacity grows with $mn/(\log n)^2$ (see equation 3.8).
- The optimal bound of network capacity is approached only for logarithmic sparseness $k \sim \log n$, the asymptotically optimal choice of sparseness. For weaker sparseness (e.g., $k \sim \sqrt{n}$) or stronger sparseness (e.g., $k = 5$), the network capacity peaks at some finite network size and vanishes asymptotically. The rate of convergence toward the asymptotic capacity $\ln 2$ depends strongly on the required level of fidelity. For high fidelity (e.g., $\epsilon = 0.01$), this convergence is quite slow, for low fidelity much faster (e.g., $\epsilon = 1$).

With regard to the first statement, it is interesting to ask for what grades of retrieval quality higher than RQ1 the upper bound of network capacity

$C = \lambda \ln 2$ can be achieved. The first statement relies on $\eta \to 1$ (in equation 3.9), that requires $\epsilon > l/n$, a condition that is always fulfilled for the retrieval quality regimes RQ0 and RQ1. It also holds for RQ2 (requiring $\epsilon \sim 1/l$) if $l$ is sufficiently small, for example, $l/n^d \to 0$ for any $d > 0$. In particular, this ansatz describes the usual case of logarithmic sparseness $k \sim l$ and $k \sim \log n$. However, in the strictest "no-error" quality regime RQ3, the upper bound of network capacity is unachievable because it requires $\epsilon \sim 1/(Ml) = k/(mn \ln(1 - p_1)) \sim k/(mn)$, which is incompatible with $\eta \to 1$ or $\ln \epsilon / \ln(l/n) \to 0$. For example, assuming $m \sim n$ yields $\eta \to 1/3$ and therefore the upper bound of network capacity for RQ3 becomes $C = (\lambda \ln 2)/3 \leq 0.23$. Note that this result is consistent with the Gardner bound 0.29 (Gardner & Derrida, 1988) and suggests that previous estimates of RQ3 capacity are wrong or misleading. For example, the result 0.346 computed by Nadal (1991) is correct only for very small-content populations, for example, $n = 1$, where $\epsilon \sim k/m$ and $\eta \to 1/2$.

In summary, the Willshaw model achieves the optimal capacity $\ln 2$ (or $1/e \ln 2$ for random query activity; see appendix D) at surprisingly high grades of retrieval quality. Recall that the Hopfield model achieves nonzero capacity only in the retrieval quality regime RQ0 (Amit et al., 1987a). However, to date no (distributed) associative memory model is known that equals look-up tables in their ability to store an arbitrary large number of patterns without any errors (see section 5). Note that our method of asymptotic analysis is exact, relying only on the binomial approximation, equation 3.3, which has recently been shown to be accurate for virtually any sublinearly sparse patterns (see Knoblauch, 2007, 2008; see also appendix C for linearly sparse and nonsparse patterns). Furthermore, we are able to compute exact capacities even for small networks and thus verify our asymptotic results (see appendixes B and D and Table 2). In contrast, many classical analyses, for example, based on statistical physics (e.g., Tsodyks & Feigel'man, 1988; Golomb, Rubin, & Sompolinsky, 1990), become reliable only for very large networks, assume an infinite relaxation time, and apply only to autoassociation with a recurrent symmetric weight matrix. However, some more recent attempts apply nonequilibrium methods for studying the behavior of recurrent neural networks with symmetric or asymmetric connections far from equilibrium and relaxation (for review, see Coolen, 2001a, 2001b). Alternative approaches based on signal-to-noise theory (e.g., Dayan & Willshaw, 1991; Palm & Sommer, 1996) are better suited for finite feedforward networks with asymmetric weight matrix but require gaussian assumptions on the distribution of dendritic potentials, which may lead to inaccurate results even for very large networks, in particular if patterns are very sparse or nonsparse (see appendix C). Before we proceed to compute synaptic capacity and information capacity for the Willshaw network, we characterize promising working regimes where the synaptic matrix has low entropy, and therefore compression is possible.

**3.4 Regimes of Balanced, Sparse and Dense Potentiation.** Historically, most analyses and model extensions of the Willshaw model have focused on the regime of balanced potentiation with a balanced memory load $0 < p_{1\epsilon} < 1$ in which the network capacity becomes optimal (Willshaw et al., 1969; Palm, 1980; Nadal, 1991; Buckingham & Willshaw, 1992; Sommer & Palm, 1999). Our extended analysis can reveal the optimal values $p_{1\epsilon}$ for arbitrary parameter settings, and it certainly suggests avoiding the regimes $p_{1\epsilon} \to 0$ or $p_{1\epsilon} \to 1$. Equations 3.5 and 3.9 and Figure 4a illustrate that in these regimes, the network capacity drops to zero. It is easy to show that in the limit $n \to \infty$, the following equivalences hold:

$$C_\epsilon > 0 \Leftrightarrow k \sim \log n \Leftrightarrow 0 < p_{1\epsilon} < 1. \tag{3.12}$$

To see this, we can rewrite equation 3.7 as $p_{1\epsilon} = \exp(-d/\lambda c)$ with $c > 0$, logarithmic $k = c \ln n$, and $d := -\ln(\epsilon l/n)/\ln n$. At retrieval quality grades RQ2 and RQ3, $d$ is a constant. Even at RQ1, $d$ remains typically constant for sublinear $l(n)$ (e.g., $d = 1$ if $l$ grows not faster than a polynomial in $\log n$). Then by varying $c$, one can obtain asymptotically for $p_{1\epsilon}$ all possible values in $(0; 1)$, and correspondingly for $C_\epsilon$ all values in $(0; \ln 2]$. Since $p_{1\epsilon}$ is monotonically increasing in $k$, we conclude that in the limit $n \to \infty$, for $d = -\ln p_{01\epsilon}/\ln n \sim 1$ and sublinear $l(n)$ the equivalences 3.12 hold.

Thus, nonzero $C_\epsilon$ is equivalent to logarithmic $k(n) \sim \log n$ and corresponds to the regime of balanced potentiation with $p_{1\epsilon} \in (0; 1)$. For sublogarithmic $k(n)$ the potentiated (1-)synapses in the memory matrix $\mathbf{A}$ are sparse, that is, $p_{1\epsilon} \to 0$, and for supralogarithmic $k(n)$ potentiated synapses are dense, that is, $p_{1\epsilon} \to 1$. Both cases, however, imply $C \to 0$. We will reevaluate these cases of sparse and dense potentiation, which appear inefficient in the light of network capacity, in the following section using the performance measures of information capacity and synaptic capacity that we introduced in section 1.2.

## 4 Analysis of Information Capacity and Synaptic Capacity

**4.1 Information Capacity.** Information capacity (see equation 1.2) relates the stored (retrievable) information to the memory resources required by implementation of an associative memory. Thus, information capacity measures how well a specific implementation exploits its physical substrate. For example, the standard implementation of a Willshaw network allocates one bit of physical memory for each of the $mn$ synapses. Therefore, for a matrix load of $p_1 = 0.5$, the information capacity is identical to the network capacity studied in section 3. However, if the memory load is $p_1 \neq 0.5$, implementations that include a compression of the memory matrix can achieve an information capacity that exceeds the network capacity.

Optimal compression of the memory matrix $\mathbf{A}$ by Huffman (1952) or Golomb (1966) coding (the latter works in cases $p_1 \to 0$ or $p_1 \to 1$) can

decrease the required physical memory by a factor according to the Shannon information $I(p_1) := -p_1 \mathrm{ld}\, p_1 - (1-p_1)\mathrm{ld}(1-p_1)$ of a synaptic weight (see appendix A).[1] Thus, with equation 3.9, the information capacity $C^I$ for optimal compression is written as

$$C^I_\epsilon := \frac{C_\epsilon}{I(p_{1\epsilon})} \approx \lambda \frac{\ln p_{1\epsilon} \ln(1-p_{1\epsilon})}{-p_{1\epsilon} \ln p_{1\epsilon} - (1-p_{1\epsilon}) \ln(1-p_{1\epsilon})} \eta. \tag{4.1}$$

Equation 4.1 reveals the surprising result that in the optimally compressed Willshaw model, the balanced regime is outperformed by the dense and sparse regimes, which both allow approaching the theoretical upper bound of information capacity $C^I \to \lambda\eta$. For small $p_{1\epsilon} \to 0$, we have $I(p_{1\epsilon}) \approx -p_{1\epsilon}\mathrm{ld}\, p_{1\epsilon}$ and $\ln(1-p_{1\epsilon}) \approx -p_{1\epsilon}$, and therefore $C^I \to \lambda\eta$. For large $p_{1\epsilon} \to 1$, we have $I(p_{1\epsilon}) \approx -(1-p_{1\epsilon})\mathrm{ld}(1-p_{1\epsilon})$, and therefore also $C^I \approx (-\ln p_{1\epsilon})/(1-p_{1\epsilon}) \to \lambda\eta$. Thus, a high-fidelity asymptotic information capacity of $\lambda \in (0;1]$ is possible for sparse and dense potentiation, that is, $p_{1\epsilon} \to 0$ or $p_{1\epsilon} \to 1$, for $n \to \infty$ and $\eta \to 1$ (see section 3.4; cf. Knoblauch, 2003a).

This finding is nicely illustrated by the plots of network and information capacity in Figures 5 and 6. The classical maximum of the network capacity $C$ in the balanced regime coincides with the local minimum of the information capacity $C^I$. For all values $p_{1\epsilon} \neq 0.5$, the information capacity surmounts the network capacity and reaches in the sparse and dense regime the theoretical optimum $C^I = 1$. Although networks of reasonable size cannot achieve the theoretical optimum at high retrieval quality, the capacity increases are still considerable, in particular for very sparse activity (e.g., $k = 2$). Moreover, there is a wide range in pattern activity $k$ in which the information capacity $C^I$ exceeds the network capacity $C$ assumed at its narrow optimum. Thus, evaluating the capacity of compressed networks more appropriately by $C^I$ avoids the "sparsity" and "capacity gap" problems of $C$ discussed in section 1.4.

A simple alternative method of synaptic compression would be to form target lists of sparse or dense matrix entries. One can simply store for each address neuron $i$ an index list of postsynaptic targets or nontargets—for $p_1 < 0.5$, the list represents the one-entries in the memory matrix and for $p_1 > 0.5$ the zero-entries. For the latter case, one can adapt the retrieval algorithm in an obvious way such that each 0-synapse decreases the membrane potential of the postsynaptic neuron (see Knoblauch, 2003b, 2006). The target list requires $\min(p_1, 1-p_1)mn\mathrm{ld}n$ bits of physical memory if we neglect

[1] This compression factor is approximate since it assumes independence of the matrix elements, which is not fulfilled for the storage of distributed patterns. Nevertheless, numerical simulations described in Knoblauch et al. (2008) show that the actual compression factor comes very close to $I(p_1)$.

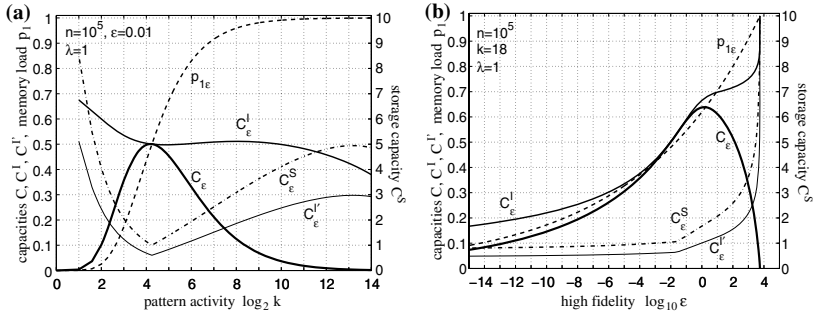


Figure 5: Capacity measures $C^I$ and $C^S$ for a finite Willshaw network with structural compression. Parameters are as in Figure 3 (square weight matrix with $m = n = 10^5$, equal pattern activities $k = l$, zero input noise with $\lambda = 1$, $\kappa = 0$). The plots show information capacity $C^I_\epsilon$ for optimal Huffman-Golomb compression (medium solid), information capacity $C^{I'}_\epsilon$ for simple target lists (thin line), and synaptic capacity $C^S_\epsilon$ (dash-dotted line). For reference, the plots show also network capacity $C_\epsilon$ (thick solid line) and matrix load $p_{1\epsilon}$ (dashed line). Capacities are drawn as either functions of $k$ for fixed fidelity parameter $\epsilon = 0.01$ (a) or functions of $\epsilon$ for fixed $k = 18$ (b).

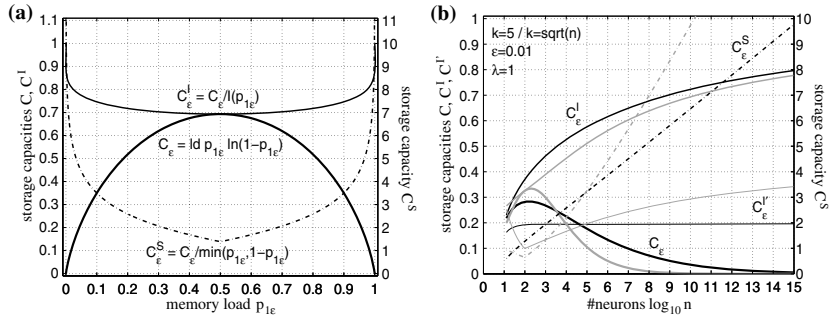


Figure 6: Capacity measures $C^I$ and $C^S$ for the compressed Willshaw model in the asymptotic limit $n \to \infty$. Parameters are as in Figure 4: $m = n$, $k = l$, $\lambda = 1$, $\kappa = 0$. (a) Information capacity $C^I_\epsilon$ (solid line) and synaptic capacity $C^S_\epsilon$ (dash-dotted line) as functions of the matrix load $p_{1\epsilon}$. For reference, the plot also shows network capacity $C_\epsilon$ (bold line). The maximum of $C$ at $p_{1\epsilon} = 0.5$ turns out to be the minimum of $C^I$ and $C^S$. For sparse or dense potentiation with $p_{1\epsilon} \to 0$ or $p_{1\epsilon} \to$, both $C^I_\epsilon \to 1$ and $C^S_\epsilon \sim \ln n \to \infty$ achieve their theoretical bounds. (b) Storage capacities $C_\epsilon$, $C^I_\epsilon$, $C^{I'}_\epsilon$ (thin line), and $C^S_\epsilon$ as functions of the network size $n$ for pattern activities $k(n) = 5$ (black line) and $k(n) = \sqrt{n}$ (gray line) assuming $\epsilon = 0.01$ (cf. Figure 4b). While $C_\epsilon \to 0$ it is $C^I_\epsilon \to 1$ and $C^S_\epsilon \to \infty$ for both functions $k(n)$. $C^{I'}_\epsilon \to 1/k = 0.2$ for $k(n) = 5$. $C^{I'}_\epsilon \to 0.5$ for $k(n) = \sqrt{n}$ (see Table 1).

the additional memory required for $m$ "memory pointers" linking the target lists to the memory matrix.[2] Thus, for large $n$, the resulting compression factor is $\min(p_1, 1-p_1)\mathrm{ld}n$. With equation 3.9, this yields the information capacity for the Willshaw model with the synaptic target list:

$$C_\epsilon^{I'} := \frac{C_\epsilon}{\min(p_{1\epsilon}, 1-p_{1\epsilon})\mathrm{ld}n} \approx \lambda \frac{\mathrm{ld}\, p_{1\epsilon} \cdot \ln(1-p_{1\epsilon})}{\min(p_{1\epsilon}, 1-p_{1\epsilon})\mathrm{ld}n}\eta. \tag{4.2}$$

Figure 5 shows that the information capacity for target list compression $C^{I'}$ stays far below the information capacity for optimal compression $C^I$. As the asymptotic analyses below will show, target list compression achieves the theoretical optimum $C^{I'} = 1$ only for dense potentiation with nearly linear $k(n)$. Nevertheless, target list compression achieves $C^{I'} > C$ for very small or quite large $k$ (e.g., $k \leq 5, k \geq 177$ for $n = 10^5$). The next section shows that $C^{I'}$ has characteristics very similar to synaptic capacity $C^S$, which is more relevant for biological networks.

**4.2 Synaptic Capacity.** Information capacity is clearly important for technical implementations of associative memories on sequential standard computers. But for the brain and also parallel VLSI hardware, it might not be the information content of the required physical memory that really matters. Rather, what matters may be the physiological resources necessary for the physical implementation of the network. For example, the synaptic capacity defined in equation 1.3 measures the mutual information in the memory task per functional synapse. Thus, the physiological resources taken into account are the number of functional synapses, that is, the one-entries in the synaptic matrix, while we assume that silent synapses, the zero-entries, are metabolically cheap and could even be pruned. The synaptic capacity of the Willshaw model can be written as

$$C_\epsilon^S := \frac{C_\epsilon}{\min(p_{1\epsilon}, 1-p_{1\epsilon})} = C_\epsilon^{I'}\mathrm{ld}n \approx \lambda \frac{\mathrm{ld}\, p_{1\epsilon} \cdot \ln(1-p_{1\epsilon})}{\min(p_{1\epsilon}, 1-p_{1\epsilon})}\eta, \tag{4.3}$$

with $\eta$ from equations 3.11 and 3.10. Note that $C^S$ and $C^{I'}$ in equation 4.2 are proportional by a factor of $\mathrm{ld}n$. Another similarity to implementations with target list compression is that in the range of dense connectivity, that is, $p_1 > 0.5$, the synaptic capacity counts the synaptic resources required by an inhibitory network implementation that represents the less frequent $(1-p_1)mn$ zero-entries in the memory matrix with functional synapses (cf. Knoblauch, 2003b, 2006). Such inhibitory implementations of associative memory have been proposed for the cerebellum (Kanerva, 1988; Marr, 1969; Albus, 1971) and might also be relevant for the basal ganglia (Wilson, 2004).

[2] This is negligible for large $n$ if on average a matrix row contains many sparse entries, $\min(p_1, 1-p_1)n \gg 0$, that is, if a neuron has many functional synapses, which is usually true.

Figure 5a shows for $m = n = 10^5$ that the Willshaw model can store up to 8.5 bits per synapse for $k = l = 2$, which exceeds the asymptotic network capacity $C \leq 0.7$ bits per synapse by more than one order of magnitude. As for information capacity, the very steep capacity increase for ultrasparse patterns, $k \to 2$, is remarkable.

For moderately sparse patterns and dense potentiation ($p_{1\epsilon} \to 1$), our analysis (see equation 4.3) suggests synaptic capacities of up to $C^S \approx 4.9$ bits per synapse for $k = 9281$. However, it turns out that the underlying approximation, equation 3.3, of $C^S$ and $C^I$ can become inaccurate for large cell assemblies (see appendixes B and C). Unfortunately, the true values of $C^S$ are significantly smaller, and the maximum occurs for smaller $k$ (see also Table 2 for $\lambda = 0.5$). The reason is that $C^S$ is very sensitive to the compression factor $1 - p_{1\epsilon}$. Thus, even if the true value of $M_\epsilon$ is only a little bit smaller than suggested by equation 3.8, the corresponding value of $1 - p_{1\epsilon}$, and therefore the compressibility of the memory matrix, can be strongly affected for $p_{1\epsilon} \to 1$ (see appendix C for more details; see also section 4.4). In contrast, this effect is not present for ultrasparse patterns with $p_{1\epsilon} \to 0$.

Figures 6a and 5b suggest that $C^S \to \infty$ for $p_{1\epsilon} \to 0$ or $p_{1\epsilon} \to 1$ and very low fidelity $\epsilon \to \infty$, respectively. This means that in principle, it is possible to store an infinite amount of information per synapse. Strictly speaking this is true only for infinitely large networks with $n \to \infty$ because the synaptic capacity $C^S$ is limited by the the number of possible spatial locations, that is, $C^S \leq \mathrm{ld} n$. Note that this is the essential difference between the concepts of synaptic capacity and network capacity: The maximum of network capacity per fixed synapse is determined only by the number of potential synaptic weight states induced by Hebbian plasticity (0 or 1 in the Willshaw model). In contrast, the maximum of synaptic capacity additionally considers the number of potential locations where the synapse can be placed by structural plasticity.

The following two sections derive explicit formulas for storage capacities and memory load for the regimes of sparse and dense potentiation (see section 3.4). Table 1 summarizes all the results for the case $m = n \to \infty$, $k = l$, noiseless addresses $\lambda = 1$ and $\kappa = 0$, and retrieval-quality grade RQ1 with constant $\epsilon \sim 1$.

**4.3 Capacities for Sparse Synaptic Potentiation.** For sparse synaptic potentiation, we have $p_{1\epsilon} \to 0$ and typically sublogarithmic pattern activity $k$ with $k/\mathrm{ld} n \to 0$ (see section 3.4; cf. Table 1). With $-\ln(1 - p_{1\epsilon}) \approx p_{1\epsilon}$ and $I(p_{1\epsilon}) \approx -p_{1\epsilon}\mathrm{ld}\, p_{1\epsilon}$ we obtain from equations 3.7, 3.2, 3.9, 4.1, 4.2, and 4.3 for large $m, n \to \infty$:

$$M_\epsilon \approx \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda k}} \frac{mn}{kl} \approx \epsilon^{\frac{1}{\lambda k}} \frac{m}{k} \left(\frac{n}{l}\right)^{1-\frac{1}{\lambda k}} \tag{4.4}$$

Table 1: Asymptotic Results for High-Fidelity Memory Load $p_{1\epsilon}$, Storable Patterns $M_\epsilon$, and Network Capacity $C_\epsilon$, Information Capacities ${C^I}_\epsilon$ for Optimal Compression and ${C^{I'}}_\epsilon$ for Simple Target Lists, and Synaptic Capacity ${C^S}_\epsilon$.

| $k$ | $p_{1\epsilon}$ | $M_\epsilon$ | $C_\epsilon$ | $C_\epsilon^I$ | $C_\epsilon^{I'}$ | $C_\epsilon^S$ |
|---|---|---|---|---|---|---|
| $c$ | 0 | $\sim n^{2-1/c}$ | 0 | 1 | $1/c$ | $(\mathrm{ld} n)/c \to \infty$ |
| $c(\ln n)^d,\ 0 < d < 1$ | 0 | $\sim n^{2-1/(c(\ln n)^d)}/(\ln n)^{2d}$ | 0 | 1 | 0 | $\sim(\ln n)^{1-d} \to \infty$ |
| $\mathrm{ld} n$ | 0.5 | $(\ln 2) n^2/(\mathrm{ld} n)^2$ | $\ln 2 \approx 0.69$ | $\ln 2$ | 0 | $2 \ln 2$ |
| $c \ln n$ | $\exp(-1/c)$ | $\sim n^2/(\ln n)^2$ | $\in (0; \ln 2)$ | $\in (\ln 2; 1)$ | 0 | $(2 \ln 2; \infty)$ |
| $c(\ln n)^d,\ 1 < d$ | 1 | $\sim n^2 \ln\ln n/(\ln n)^{2d}$ | 0 | 1 | 0 | $\sim \ln\ln n \to \infty$ |
| $\sqrt{n}$ | 1 | $0.5 n \ln n$ | 0 | 1 | 0.5 | $0.5 \mathrm{ld} n \to \infty$ |
| $c n^d,\ 0 < d < 1$ | 1 | $\sim n^{2-2d} \ln n$ | 0 | 1 | $d$ | $d \mathrm{ld} n \to \infty$ |
| $cn,\ 0 < c < 1$ | 1 | $(\ln n)/(-c \ln(1-c))$ | 0 | 0 | 0 | 0 |

Note: Here we consider only the special case of $k = l$, $m = n \to \infty$, noiseless address patterns ($\lambda = 1, \kappa = 0$), and constant fidelity parameter $\epsilon \sim 1$ corresponding to quality regime RQ1.

$$C_\epsilon \approx \frac{\left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda k}} \mathrm{ld}\frac{\epsilon l}{n-l}}{k}\eta \to 0 \tag{4.5}$$

$$C_\epsilon^I \approx \lambda\eta \le \lambda \tag{4.6}$$

$$C_\epsilon^{I'} \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{k\mathrm{ld}n} \cdot \eta \le 1/k \tag{4.7}$$

$$C_\epsilon^S \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{k} \cdot \eta \le \frac{\mathrm{ld}n}{k}. \tag{4.8}$$

The second approximation in equation 4.4 is valid only for $l \ll n$. Thus, for sparse potentiation, we can still store a very large number of ultrasparse patterns where $M$ scales almost with $mn$, for large $k$. However, note that for given $m$, $n$, maximal $M$ is obtained for logarithmic $k$ (cf. Figure 4a). The classical network capacity $C$ vanishes for large $n$, but for optimal compression, we obtain an information capacity with $C^I \to 1$. For simple target lists (see above), the information capacity approaches $C^{I'} \to 1/k$. Thus, $C^{I'}$ is nonzero only for small constant $k$. For constant $k = 1$, we have trivially $C^{I'} \to 1$. However, this result is not very interesting since for $k = 1$, we have no really distributed storage. For $k = 1$, there are only $M = m$ possible patterns to store, and the memory matrix degenerates to a look-up table. Section 5 discusses more closely the relation between the Willshaw model and different implementations of look-up tables.

For the synaptic capacity, we have $C_\epsilon^S \sim \log n \to \infty$ for constant $k \sim 1$, which comes very close to the theoretical optimum $C^S \le \mathrm{ld}n$, the information necessary to determine the target cell of a given synapse among the $n$ potential targets in the content population. Most interestingly, $C^S$ and $C^{I'}$ are independent of the fault tolerance parameter $\lambda$ (and consequently must also be independent of the high-fidelity parameter $\epsilon$). Thus, decreasing $M$ from $M = M_\epsilon$ to $M = 1$ virtually does not affect either $C^S$ or $C^{I'}$. Note that for a single stored pattern, $C^S = (\mathrm{ld}\binom{n}{l})/(kl) \approx (\mathrm{ld}n)/k$ reaches the upper bound of equation 4.8.

**4.4 Capacities for Dense Synaptic Potentiation.** For dense synaptic potentiation, we have $p_{1\epsilon} \to 1$ and typically supralogarithmic pattern activity $k$ with $k/\mathrm{ld}n \to \infty$ (see section 3.4; cf. Table 1). With $I(p_{1\epsilon}) \approx -(1-p_{1\epsilon})\mathrm{ld}(1-p_{1\epsilon})$ and $1 - p_{1\epsilon} \approx -\ln p_{1\epsilon}$ we obtain from equations 3.7, 3.2, 3.9, 4.1, 4.2, and 4.3 for large $n \to \infty$:

$$1 - p_{1\epsilon} \approx \frac{\ln\frac{n-l}{\epsilon l}}{\lambda k} \;\to\; 0 \tag{4.9}$$

$$M_\epsilon \approx \frac{mn}{kl}\left(\ln(\lambda k) - \ln\ln\frac{n-l}{\epsilon l}\right) \tag{4.10}$$

$$C_\epsilon \approx \left(\ln(\lambda k) - \ln\ln\frac{n-l}{\epsilon l}\right)\frac{\mathrm{ld}\frac{n-l}{\epsilon l}}{k}\cdot\eta \;\to\; 0 \tag{4.11}$$

$$C_\epsilon^I \approx \lambda\eta \le \lambda \tag{4.12}$$

$$C_\epsilon^{I'} \approx \lambda\cdot\frac{\ln(\lambda k) - \ln\ln\frac{n-l}{\epsilon l}}{\ln n} \le \lambda\frac{\ln k}{\ln n} \tag{4.13}$$

$$C_\epsilon^S \approx \lambda\cdot\mathrm{ld}(\lambda k) - \mathrm{ld}\ln\frac{n-l}{\epsilon l} \le \lambda\ln n. \tag{4.14}$$

Although the pattern capacity $M_\epsilon$ is much smaller than for balanced and sparse synaptic potentiation, here we can still store many more moderately sparse patterns than there are neurons ($M \gg n$) as long as $k \le \sqrt{n}$ (see equation 4.10; cf. Table 1). The classical network capacity $C$ vanishes for large $n$, but for optimal compression, we obtain a high information capacity $C^I \to 1$. Surprisingly, information capacity can approach the maximum even for nonoptimal compression. For $k = n^d$ and $0 < d < 1$, we obtain $C^{I'} \to \lambda d$ from equation 4.13. Similarly, synaptic capacity achieve its upper bound, $C^S \le \mathrm{ld}n$, for $k = n^d$ with $d \to 1$. Note that here, $C^{I'}$ and $C^S$ achieve factor two larger values than for sparse potentiation and distributed storage with $k \ge 2$ (see equations 4.7 and 4.8). However, the convergence appears to be extremely slow for high fidelity (see appendix B; see also Knoblauch, 2008), and for $d > 0.5$ we obtain asymptotically only $M < n$ (see equation 4.10; cf. Table 1; see also section 5).

For dense synaptic potentiation, both $C^{I'}$ and $C^S$ depend on the fault tolerance requirement $\lambda$ and the high-fidelity parameter $\epsilon$, unlike sparse synaptic potentiation, where these capacities are independent from $\lambda$. Unfortunately, requiring high fidelity and fault tolerance counteracts the compressibility of the memory matrix because $I(p_1)$ increases for decreasing $p_1 > 0.5$. This results in the counterintuitive fact that the amount of necessary physical memory increases with the decreasing number of stored patterns $M$.

As can be seen in Figure 5a, both information capacities $C^I$ and $C^{I'}$ and synaptic capacity $C^S$ exhibit local maxima at $k^I_{\mathrm{opt}}$ and $k^S_{\mathrm{opt}}$ ($= k^{I'}_{\mathrm{opt}}$) for $k > \mathrm{ld}n$. In Knoblauch (2003b, appendix B.4.2) these maxima are computed (not shown here). The resulting asymptotic optima are approximately

$$k^S_{\mathrm{opt}} \sim n\cdot(e^{\sqrt{-\ln\epsilon}})^{-\sqrt{\ln n}} \tag{4.15}$$

$$k^I_{\mathrm{opt}} \sim n^{1-\frac{-\ln\epsilon-\sqrt{-\ln\epsilon}}{-\ln\epsilon-1}}. \tag{4.16}$$

Note that $k^S_{\mathrm{opt}}$ grows faster than $n^d$ for any $d < 1$, but slower than the upper bound $n/\log^4 n$, where our theory based on the binomial approximation equation 3.3, is valid.

For linear $k = cm$ and $l = dn$, the binomial approximation is invalid, and we have to use alternative methods as described in appendix C. Here the Willshaw model can store only $M \sim \log m$ pattern associations with vanishing storing capacities $C, C^I, C^S \to 0$. There are much better alternative models for this parameter regime. For example, the classical Hopfield model can store a much larger number of $M = 0.14n$ nonsparse patterns resulting in 0.14 bits per (nonbinary) synapse (Hopfield, 1982; Amit et al., 1987a, 1987b). Thus, for nonsparse patterns, synapses with gradual weight states such as employed in the Hopfield model appear to make a big difference to binary clipped Hebbian learning, as in the Willshaw model.

**4.5 Remarks on Fault Tolerance and Attractor Shape.** How does increasing noise $(1 - \lambda, \kappa)$ in the query patterns $\tilde{\mathbf{u}}$ affect the number of storable patterns $M_\epsilon$ and the other capacity measures $(C_\epsilon, C^I_\epsilon, C^S_\epsilon)$ for a given network size and pattern activity?[3] It is particularly simple to answer this question for pattern part retrieval where query patterns contain miss noise only ($\kappa = 0$). Using equations 3.2 and 3.7, we can introduce the fraction of storable patterns as a function of the query noise $\lambda$,

$$m_\lambda := \frac{M_\epsilon(\lambda)}{M_\epsilon(1)} \approx \frac{\ln(1 - p_{1\epsilon}(\lambda))}{\ln(1 - p_{1\epsilon}(1))} \in (0; 1]$$
$$\begin{cases} \approx p_{1\epsilon}(1)^{(1-\lambda)/\lambda} \to 0, & p_{1\epsilon}(1) \to 0 \\ \to 1, & p_{1\epsilon}(1) \to 1 \end{cases}, \tag{4.17}$$

where we used $\ln(1 - p_{1\epsilon}) \approx -p_{1\epsilon}$ for $p_{1\epsilon} \to 0$ and de l'Hôpital's rule for $p_{1\epsilon} \to 1$. The fraction of storable patterns with increasing fault tolerance differs markedly for the regimes of sparse, balanced, and dense synaptic potentiation (cf. sections 3.4, 4.3, and 4.4): Figure 7a shows that the decrease is steep for very sparse memory patterns and $p_{1\epsilon} \to 0$ and shallow for moderately sparse patterns and $p_{1\epsilon} \to 1$. Thus, relatively large cell assemblies with $k \gg \log n$ are much more robust against miss noise than small cell assemblies with $k \leq \log n$ (cf. Table 1). The same conclusion is true for network capacity, $C_\epsilon(\lambda) := m_\lambda \cdot C_\epsilon(1)$ (see equations 3.4 and 3.9).

Increasing fault tolerance or attractor size of a memory will decrease not only $M_\epsilon$ but also $p_{1\epsilon}$. Therefore, the compressibility of the memory matrix also will change. In analogy to $m_\lambda$ for $M_\epsilon$, we can compute the relative compressibility $i_\lambda$ for $C^I_\epsilon$,

$$i_\lambda := \frac{I(p_{1\epsilon}(\lambda))}{I(p_{1\epsilon}(1))} \begin{cases} \approx p_{1\epsilon}(1)^{((1-\lambda)/\lambda)}/\lambda \to 0, & p_{1\epsilon}(1) \to 0 \\ \to 1/\lambda, & p_{1\epsilon}(1) \to 1 \end{cases}, \tag{4.18}$$

[3] Note the difference between assessing fault tolerance for either a given memory load $p_{1\epsilon}$ or given pattern activities $k, l$, since the former is a function of the latter.

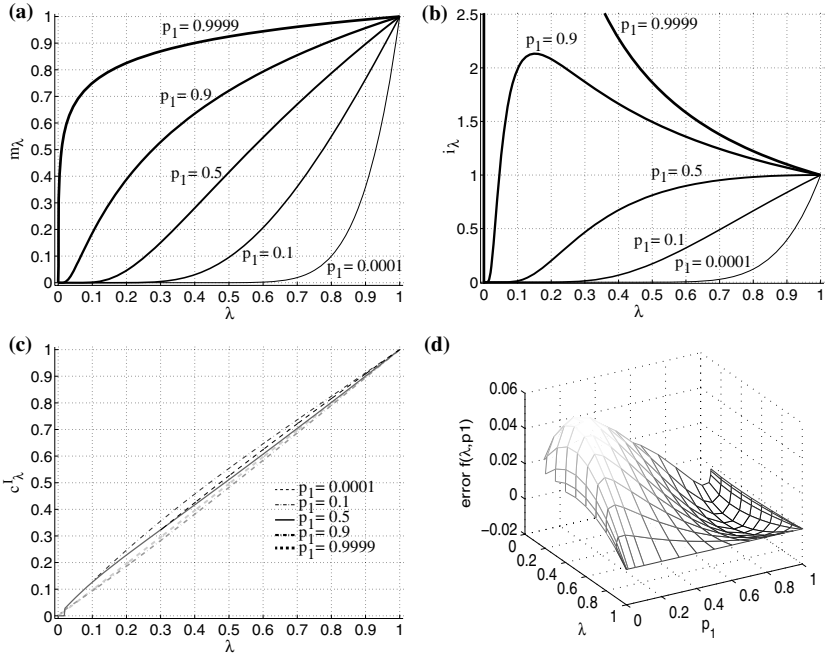


Figure 7: Impact of miss noise on the number of storable patterns and the compressibility of the memory matrix for different $p_1$. Query patterns $\tilde{\mathbf{u}}$ are assumed to contain $\lambda k$ out of the $k$ original ones, but no false ones ($\kappa = 0$). Here $p_1 := p_{1\epsilon}(1)$ is the maximal matrix load for $\lambda = 1$ (see equation 3.7). (a) Fraction of storable patterns $m_\lambda$ versus $\lambda$ (see equation 4.17). (b) Relative compressibility $i_\lambda$ versus $\lambda$ (see equation 4.18). (c) For all values of $p_1$, we have $c_\lambda^I := m_\lambda / i_\lambda \approx \lambda$ (see equation 4.19). (d) The error $f(\lambda, p_1) := c_\lambda^I - \lambda$ of approximating $c_\lambda^I$ by $\lambda$ is small ($-0.02 < f < 0.06$) and even vanishes for $p_1 \to 0$ and $p_1 \to 1$.

where we used $I(p_1) \approx -p_1 \mathrm{ld}\, p_1$ for $p_{1\epsilon}(1) \to 0$ and de l'Hôpital's rule for $p_{1\epsilon}(1) \to 1$ (cf. Knoblauch, 2003b). The relative compressibility is depicted in Figure 7b. Note that always $i_\lambda < 1$ for $p_{1\epsilon}(1) < 0.5$, but usually $i_\lambda > 1$ for $p_{1\epsilon}(1) > 0.5$. The latter occurs for dense potentiation and moderately (e.g., supralogarithmically) sparse address patterns (see Table 1) and implies the counterintuitive fact that although fewer patterns are stored, more physical memory is required. Thus, the dependence of information capacity on miss noise is

$$c_\lambda^I := \frac{C_\epsilon^I(\lambda)}{C_\epsilon^I(1)} = \frac{m_\lambda}{i_\lambda} = \lambda + f(\lambda, p_{1\epsilon}(1)) \approx \lambda, \tag{4.19}$$

for a small error function $f$ with $f \to 0$ for $p_{1\epsilon} \to 0$ and $p_{1\epsilon} \to 1$. The plots of $c^I_\lambda$ in Figure 7c reveal the surprising result that the relative decrease in information capacity is almost linear in $\lambda$ in all the regimes of pattern sparsity. One can verify numerically that $-0.02 < f(\lambda, p_1) < 0.06$ for $\lambda, p_1 \in (0; 1)$ (see Figure 7d).

Similar considerations for the synaptic capacity $C^S$ (that apply also to information capacity $C^{I'}$) reveal that

$$c^S_\lambda := \frac{C^S_\epsilon(\lambda)}{C^S_\epsilon(1)} = \frac{m_\lambda \min(p_{1\epsilon}(1), 1 - p_{1\epsilon}(1))}{\min(p_{1\epsilon}(\lambda), 1 - p_{1\epsilon}(\lambda))}$$
$$\approx \begin{cases} C^S_\epsilon(1), & p_{1\epsilon}(1) \to 0 \\ \lambda C^S_\epsilon(1), & p_{1\epsilon}(1) \to 1 \end{cases}. \tag{4.20}$$

It is remarkable that $C^S$ is independent of $\lambda$ for ultrasparse patterns with $k/\log n \to 0$ and sparse potentiation $p_{1\epsilon} \to 0$. Thus, decreasing $M$ from $M = M_\epsilon(1)$ to $M = M_\epsilon(\lambda)$ affects neither $C^S$ nor $C^{I'}$. Actually, for a single stored pattern, $C^S = (\mathrm{ld}\binom{n}{l})/(kl) \approx (\mathrm{ld} n)/k$ is identical to the upper bound of equation 4.8. Thus, $C^S_\epsilon(\lambda)$ actually increases for $\lambda \to 0$ (or $\epsilon \to 0$).

A theoretical analysis including add noise ($\kappa \geq 0$) is more difficult (cf. Palm & Sommer, 1996; Sommer & Palm, 1999; Knoblauch, 2003b). In numerical experiments, we have investigated retrieval quality as a function of miss noise ($\lambda < 1$) and add-noise ($\kappa > 0$) using exact expressions for retrieval errors $p_{01}$ and $p_{10}$ (see equations B.1 and B.2). For given network size (here $m = n = 1000$) and sparsity level ($k = l = 4, 10, 50, 100, 300$), the number of stored patterns $M$ has been chosen such that for noiseless query patterns ($\lambda = 1, \kappa = 0$), a high-fidelity criterion $\epsilon \leq 0.01$ was fulfilled. Then we computed retrieval quality for noisy query patterns $\tilde{\mathbf{u}}$ with activity $z := |\tilde{\mathbf{u}}|$. For $z \leq k$, queries were pattern parts ($0 < \lambda \leq 1, \kappa = 0$). For $z > k$, queries were supersets of the original address patterns ($\lambda = 1, \kappa \geq 0$). The retrieval quality was measured by minimizing $\epsilon^T := (T(k/n, p_{01}, p_{10}) - I(k/n))/I(k/n)$ with respect to the neuron threshold $\Theta$. Here $\epsilon^T$ corresponds to the normalized information loss between retrieved and originally stored patterns, but using the Hamming distance based measure $\epsilon$ as defined in section 3.2 leads qualitatively to the same results (see Knoblauch et al., 2008). Figure 8a shows for each noise level the retrieval quality and Figure 8b the optimal threshold.

These numerical experiments validate our theoretical results for pattern part retrieval (without add noise). For $\lambda < 1$, ultrasparse patterns (e.g., constant $k = 4$) appear to be very vulnerable to miss noise (i.e., $\epsilon$ increases very steeply with decreasing $\lambda$). In contrast, moderately sparse patterns (e.g., $k = 1000$ for $n = 10{,}000$) are much more robust against miss noise (i.e., the increase of $\epsilon$ is much weaker). On the other hand, our data also show that ultrasparse cell assemblies are very robust against add noise (i.e., the fidelity

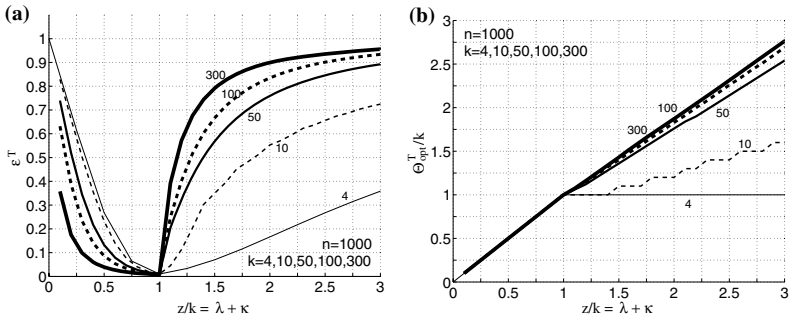


Figure 8: Impact of query noise on the retrieval quality of the Willshaw model for $m = n = 1000$ neurons and different pattern activities $k = l = 4, 10, 50, 100, 300$ (increasing line thickness) storing $M = 4928, 4791, 663, 207, 27$ patterns in each case (corresponding to $\epsilon = 0.01$ for noiseless queries). Data are computed from exact error probabilities (see equations B.1 and B.2). (a) Retrieval quality $\epsilon^T := (T(k/n, p_{01}, p_{10}) - I(k/n))/I(k/n)$ as a function of query pattern activity $z = (\lambda + \kappa)k$. The queries were noiseless for $z/k = 1$, contained only miss noise for $z/k < 1$ (i.e., $\lambda < 1$, $\kappa = 0$), and contained only add noise for $z/k > 1$ (i.e., $\lambda = 1$, $\kappa > 0$). The threshold $\Theta$ is chosen such that $\epsilon^T(\lambda, \kappa)$ is minimized. (b) Optimal threshold $\Theta_{\text{opt}}$ for minimal $\epsilon^T$ shown in *a*. The plots for $\epsilon$ instead of $\epsilon^T$ are qualitatively the same (Knoblauch et al., 2008).

parameter $\epsilon$ increases only relatively slowly with increasing add noise level $\kappa$). In contrast, the large cell assemblies are quite vulnerable to add noise: Here $\epsilon$ increases very steeply with $\kappa$. Our results show that the attractors around memories $\mathbf{u}^\mu$ (i.e., the subspace of query patterns $\tilde{\mathbf{u}}$ that map to $\mathbf{v}^\mu$) have only little similarity to spheres in Hamming space. Rather, for ultrasparse patterns ($k/\log n \to 0$), attractors are elongated toward query patterns with more add noise than miss noise, whereas for moderately sparse patterns ($k/\log n \to \infty$), attractors are elongated toward query patterns with more miss noise than add noise.

Figure 8b illustrates another important difference between sparse and dense synaptic potentiation corresponding to ultrasparse or moderately sparse activity. For ultrasparse patterns, the optimal threshold depends mainly on $\lambda$, but only very weakly on $\kappa$. In contrast, for moderately sparse patterns, the optimal threshold has a strong dependence on both $\lambda$ and $\kappa$. As a consequence, in particular for biological systems, it may be much easier to implement the optimal threshold for retrieving ultrasparse patterns. In a noisy regime with $\kappa \gg 0$, it will be sufficient to simply choose a constant threshold identical to the assembly size, $\Theta = k$, assuming that information processing is usually accomplished with complete patterns, $\lambda = 1$. This bears in particular the possibility of activating superpositions of many different ultrasparse cell assemblies. Actually, a reasonable interpretation of

seemingly random or spontaneous ongoing activity (Arieli, Sterkin, Grinvald, & Aertsen, 1996; Softky & Koch, 1993) would be that a large number of small cell assemblies or synfire chains (Abeles, 1982; Abeles, Bergman, Margalit, & Vaadia, 1993; Diesmann, Gewaltig, & Aertsen, 1999; Wennekers & Palm, 1996) are active at the same time, independent of each other.

## 5 Computational Complexity and Energy Requirements

**5.1 Compressed and Uncompressed Willshaw Network.** So far we have been concerned with the storage capacity and fault tolerance of the Willshaw associative memory. Another important question is how fast the information can be retrieved for implementation on a sequential digital computer. To retrieve a pattern in the Willshaw model, we have to compute potentials $\mathbf{x} = \tilde{\mathbf{u}}\mathbf{A}$ and afterward apply a threshold on each component of $\mathbf{x}$, that is, the retrieval time (or number of retrieval steps) is

$$t^{\mathrm{W}}_{\mathrm{seq}} = z \cdot n + n \approx zn, \tag{5.1}$$

where $z := (\lambda + \kappa)k$ is the query pattern activity. Note that retrieval time is dominated by synaptic operations. Thus, our temporal measure also has an interpretation in terms of energy consumption. However, for this interpretation, it may be more relevant to consider only nonsilent synapses (see section 1.2 and Lennie, 2003; Laughlin & Sejnowski, 2003), which is captured by the following analysis for the "compressed" model.

Matrix compression (or eliminating silent synapses) in the sparse and dense connectivity regimes not only improves storage capacity but generally accelerates retrieval. For sparse connectivity with $p_1 \to 0$, the memory matrix $\mathbf{A}$ contains sparsely one-entries, and computing the potentials $\mathbf{x}$ requires only $p_1 \cdot n$ steps per activated address neuron. Similarly, for dense connectivity with $p_1 \to 1$, we can compute the potentials by $\mathbf{x} = \mathbf{z} - \tilde{\mathbf{u}}\mathbf{A}'$ where $\mathbf{A}' := \mathbf{1} - \mathbf{A}$ contains sparsely one-entries (see also Knoblauch, 2006). Thus, the retrieval time is

$$t^{\mathrm{cW}}_{\mathrm{seq}} = c \cdot z \cdot n \cdot \min(p_1, 1 - p_1), \tag{5.2}$$

where $c$ is a (small) constant accounting for decompression of $\mathbf{A}$ (or $\mathbf{A}'$), keeping track of neurons selected by $\mathbf{A}$ (or $\mathbf{A}'$) in a list, and finally applying the threshold to the neurons in that list (note that $zn \min(p_1, 1 - p_1)$ may be $\ll n$). Obviously, $t^{\mathrm{cW}}_{\mathrm{seq}} / t^{\mathrm{W}}_{\mathrm{seq}} \to 0$, at least for sparse and dense potentiation with $p_1 \to 0$ or $p_1 \to 1$. However, it may be unfair to compare the compressed to the uncompressed Willshaw model since the latter works in an optimal manner for $p_1 = 0.5$, where compression is not possible. Thus, we may want to compare the two models for different pattern sparseness $k, l$. Such an approach has been conducted by Knoblauch (2003b) showing that the compressed model is superior to the uncompressed even if one normalizes the amount of retrieved information to the totally stored information.

**5.2 Comparison to Look-Up Tables and "Grandmother Cell" Networks.** It has been pointed out that Willshaw associative memory can allow much faster access to stored pattern information than a simple look-up table (e.g., see Palm, 1987). A look-up table implementation of associative memory would require an $M \times m$ matrix $\mathbf{U}$ for the address pattern vectors and an $M \times n$ matrix $\mathbf{V}$ for the content patterns such that $\mathbf{U}_\mu = \mathbf{u}^\mu$ and $\mathbf{V}_\mu = \mathbf{v}^\mu$ for $\mu = 1, \ldots, M$ (each matrix row corresponds to a pattern vector). We also refer to the look-up table as a *grandmother cell model* (or briefly grandmother model; cf. Knoblauch, 2005; Barlow, 1972) because its biological interpretation corresponds to a two-layer architecture where an intermediary population contains $M$ neurons, one "grandmother" cell for each stored association (see section 2.3). Thus, grandmother cell $\mu$ receives inputs via synapses corresponding to the $\mu$th row of $\mathbf{U}$. A winner-takes-all dynamics activates only the most excited grandmother cell, which can activate the content population according to the corresponding synaptic row in $\mathbf{V}$.

For naive retrieval using a query pattern $\tilde{\mathbf{u}}$, one would compare $\tilde{\mathbf{u}}$ to each row of $\mathbf{U}$ and select the most similar $\mathbf{u}^\mu$. If each row of $\mathbf{U}$ contains $k \ll m$ one-entries, we may represent each pattern by the (ordered) list of the positions (indices) of its one-entries. Then the retrieval takes only

$$t_{\text{seq}}^{\text{nLUT}} = M \cdot (z + k). \tag{5.3}$$

Then for $M/n \to \infty$, we indeed have $t_{\text{seq}}^{\text{nLUT}} / t_{\text{seq}}^{\text{W}} \geq M/n \to \infty$. Thus, the Willshaw model is more efficient than a naive look-up table if we store more patterns $M$ than we have content neurons $n$.

However, in many cases, compressed look-up tables can be implemented more efficiently than the Willshaw model even for $M \gg n$. So far, by representing lists of one-entries for each pattern in the look-up table, we have essentially compressed the matrix rows. However, it turns out that compressing the columns is always more efficient (Knoblauch, 2005). If we optimally compress the columns of $\mathbf{U}$ (e.g., by Huffman or Golomb coding, similar to the compressed Willshaw model), then information capacity becomes $C^I \to 1$ and a retrieval requires only

$$t_{\text{seq}}^{\text{cLUT}} = c \cdot z \cdot M \cdot k/m \tag{5.4}$$

steps. Compared with the compressed Willshaw model, this yields

$$\begin{aligned} \nu &:= \frac{{t_{\text{seq}}}^{\text{cLUT}}}{{t_{\text{seq}}}^{\text{cW}}} \approx \frac{-\ln(1-p_1)}{l \min(p_1, 1-p_1)} \leq \frac{-\ln(1-p_{1\epsilon})}{l \min(p_{1\epsilon}, 1-p_{1\epsilon})} \\ &\to \begin{cases} 1/l, & p_{1\epsilon} \to 0 \\ \lambda \frac{k}{l} \dfrac{\ln(\lambda k) - \ln\ln \frac{n}{\epsilon l}}{\ln \frac{n}{\epsilon l}}, & p_{1\epsilon} \to 1, \end{cases} \end{aligned} \tag{5.5}$$

where we used $1 - p_{1\epsilon} \approx -\ln p_{1\epsilon}$ for $p_{1\epsilon} \to 1$. Remember from section 3.1 that the memory matrix is sparse ($p_{1\epsilon} \to 0$), balanced ($0 < \delta < p_{1\epsilon} < 1 - \delta$), or dense ($p_{1\epsilon} \to 1$) for sublogarithmic, logarithmic, or supralogarithmic $k(n)$. Thus, the Willshaw model performs worse than the grandmother model for most parameters. The Willshaw model is unequivocally superior only for asymmetric networks with large $k$ and small $l$. If we require $m = n$ and $k = l$ (e.g., for autoassociation), the Willshaw model is superior with $\nu \to \lambda d/(1 - d)$ only for almost linear $k = n^d$ with $1/(1 + \lambda) < d < 1$.

Look-up tables are also superior to distributed associative networks with respect to fault tolerance because they always find the exact nearest neighbor. In order to have a fair comparison with respect to fault tolerance, we can dilute the look-up tables by randomly erasing one-entries in matrix **U**. This will further accelerate retrieval in look-up tables and cut even the remaining parameter range where the Willshaw model is superior (Knoblauch et al., 2008). At least for asymmetric networks, there remains a narrow parameter range where the Willshaw model beats diluted look-up tables. This seems to be the case for large $m$, small $l$, $n$, and relatively small $k$ (but still large enough with supralogarithmic $k/\log n \to \infty$ to obtain dense potentiation).

**5.3 Parallel Implementations.** For full (i.e., synapse-) parallel hardware implementations (like brain tissue or VLSI chips; Chicca et al., 2003; Heittmann & Rückert, 2002), the retrieval time is O(1), and the remaining constant is mainly determined by the hardware properties. Here the limiting resource is the connectivity (e.g., the number of nonsilent synapses), and our analysis so far can be applied again.

However, there are also neuron-parallel computers with reduced hardware connectivity. One big advantage of the Willshaw model is that there are obvious realizations for such architectures (Palm & Palm, 1991; Hammerstrom, 1990; Hammerstrom, Gao, Zhu, & Butts, 2006). For example, on a computer with $n$ processors (one per neuron) and a common data bus shared by all processors, a retrieval takes time ${t_{\text{prl}}}^{\text{W}} = z + 1$. In comparison, a corresponding implementation of the grandmother model or a look-up table will require $M$ processors and time ${t_{\text{prl}}}^{\text{LUT}} = z + \log M$. In particular for $M \gg n$, there is no obvious parallelization of look-up tables that would beat the Willshaw model.

In summary, both the Willshaw and the grandmother model are efficient ($t_{\text{seq}}/M, t_{\text{prl}}/n \to 0$) only for sparse address patterns. Nonsparse patterns require additionally a sparse recoding (or indexing) as is done in multi-index hashing (Greene et al., 1994). Although there are quite efficient computer implementations, it appears that distributed neural associative memories have only minor advantages over compressed look-up tables or multi-index hashing, at least for solving the best match problem on sequential computers. On particular parallel computers, the Willshaw model remains superior.

## 6 Summary and Discussion

Neural associative memories are promising models for computations in the brain (Hebb, 1949; Anderson, 1968; Willshaw et al., 1969; Marr, 1969, 1971; Little, 1974; Gardner-Medwin, 1976; Braitenberg, 1978; Hopfield, 1982; Amari, 1989; Palm, 1990), as well as potentially useful in technical applications such as cluster analysis, speech and object recognition, or information retrieval in large databases (Kohonen, 1977; Bentz et al., 1989; Prager & Fallside, 1989; Greene et al., 1994; Knoblauch, 2005; Mu et al., 2006; Rehn & Sommer, 2006).

In this review, we have raised the question of how to evaluate the efficiency of associative memories, that is, how to quantify the achieved computation and the used resources. The common measure of efficiency is network capacity, that is, the amount of information per synapse that can be stored in a network of fixed structure (Willshaw et al., 1969; Palm, 1980, 1991; Amit et al., 1987a, 1987b; Nadal, 1991; Buckingham & Willshaw, 1992; Sommer & Palm, 1999; Bosch & Kurfess, 1998). Here we have argued that network capacity is biased because it disregards the entropy of the synapses and thus underestimates models with low synaptic entropy and overestimates models with high synaptic entropy. To account for the synaptic entropy, it was necessary to introduce information capacity, a new performance measure. Interestingly, network capacity and information capacity draw radically different pictures in what range associative memories work efficiently. For example, the Willshaw model is known to optimize the network capacity if the distribution of 0-synapses and 1-synapses is even and thus the synaptic entropy is maximal (Willshaw et al., 1969; Palm, 1980). In contrast, the Willshaw model reaches the optimum information capacity in regimes of small synaptic entropy if either almost all synapses remain silent (sparse potentiation with memory load $p_1 \to 0$) or if almost all synapses are active (dense potentiation with memory load $p_1 \to 1$). We have shown that the regimes of optimal information capacity that we discovered have direct practical implications. Specifically, we have constructed models of associative memory using mechanisms like Huffman or Golomb coding for synaptic compression, which can outperform their counterparts without matrix compression.

Further, the discovery of regimes in associative memories with high information capacity could be a key to understanding the computational function of the various types of structural plasticity in the brain. In structural plasticity, functionally irrelevant silent synapses are pruned and replaced by new synapses generated at other locations in the network. This process can lead to a sparsely connected neural network in which each synapse carries a large amount of information about previously learned patterns (Knoblauch, 2009). To quantify the effects of structural plasticity, we have introduced the definition of synaptic capacity, which measures the information stored per functionally necessary synapse (i.e., not counting

silent synapses, which could be pruned). Our model analyses indicate that information capacity and synaptic capacity become optimal in the same regimes of operation. Thus, structural plasticity can be understood as a form of synaptic compression required to optimize information capacity in biological networks.

Although our new definitions of performance measures for associative memories are general, for practical reasons we had to restrict the model analysis to two simple yet interesting examples of associative memories. The simplest possible version is a linear associative memory in which learning corresponds to forming the correlation matrix of the data and retrieval corresponds to a matrix-vector multiplication (Kohonen, 1977). However, the efficiency of linear associative memories is very limited. The cross-talk can be predicted to set in if the stored patterns deviate from the principal components of the data, which will necessarily be the case if the number of stored patterns exceeds the dimension of the patterns. The Willshaw model is a feedforward neural network similar to the linear associative memory but much more efficient by any standards, because nonlinearities in the neural transfer function and in the superposition of memory traces keep the cross-talk small, even if the number of stored patterns scales almost with the square of the dimension of the patterns (Willshaw et al., 1969; Palm, 1980). Thus, we chose to analyze the Willshaw network. In addition, to compare neural associative memories to look-up tables (LUT), the classical structure for content-addressable memory in computer science, we also analyzed a two-layer extension of the Willshaw network with winner-take-all (WTA) activation in the hidden layer, which implements a look-up table.

Previous analyses of the Willshaw network revealed that network capacity is optimized in a regime in which stored patterns are sparse (the number of active units grows only logarithmically in the network size, $k \sim \log n$) and the number of stored patterns grows as $n^2/(\log n)^2$ (Willshaw et al., 1969; Palm, 1980). However, these analyses determined the upper bound of the network capacity with the level of retrieval errors undefined. In practice, computations rely on a specific and guaranteed level of retrieval quality. Therefore, for fair and meaningful comparisons of the three definitions of storage capacity, network, and information and synaptic capacity, we had to develop new analytical procedures to quantify the different capacities at a defined level of retrieval errors.

The new analyses revealed three important new results. First, implicit in classical analyses, a high network capacity $0 < C \leq \ln 2 \approx 0.69$ or $0 < C \leq 1/e \ln 2 \approx 0.53$ is restricted to a very narrow range of logarithmic pattern sparseness (see section 3.4 and appendix D). Second, the information and synaptic capacities assume high values for quite wide ranges of pattern activities (see Figure 5). Third, the optimal regimes of information and synaptic capacities, $C^I \to 1$ and $C^S \sim \log n$, coincide but are distinct from the optimal regime for network capacity. For example, the information capacity has the minimum in the regime of optimal network capacity and assumes

the theoretical optimum $C^I \to 1$ either for ultrasparse patterns $k/\log n \to 0$ or for moderately sparse patterns $k/\log n \to \infty$ (see Perez-Orive et al., 2002; Hahnloser, Kozhevnikov, & Fee, 2002; Quiroga, Reddy, Kreiman, Koch, & Fried, 2005; Waydo, Kraskov, Quiroga, Fried, & Koch, 2006, for experimental evidence supporting sparse representations in the brain).

In addition, the new analyses revealed how the robustness of content-addressable memory against different types of noise in the address patterns varies in the different regimes of operation. While the effects of additional activity (add errors) and missing activity (miss errors) were quite balanced for log-sparse patterns (see Figure 8), the effects strongly varied with error type in the ultrasparse and moderately sparse regime. Specifically, the retrieval of ultrasparse patterns ($k \ll \log n$) was robust against add errors in the address pattern but vulnerable to miss errors. The inverse relation was found for the retrieval of moderately sparse patterns. Thus, the ultrasparse regime could be of particular interest if a memory has to be recognized in superpositions of many patterns, whereas the moderately sparse regime allows completing a memory pattern from a small fragment.

The retrieval speed defined as the time (or number of computation steps) required to retrieve a pattern is another important performance measure for associative memory. Previous work has hypothesized that neural associative memory is an efficient means for information retrieval in the context of the best match problem (Minsky & Papert, 1969), even when implemented on conventional computers. For example, Palm (1987) has argued that distributed neural associative memory would have advantages over local representations such as in the LUT network. While this may hold true for plain (uncompressed) and parallel implementations (Hammerstrom, 1990; Palm & Palm, 1991; Knoblauch, 2003b; Chicca et al., 2003), we showed in section 5 that the compressed LUT network implemented on a sequential architecture outperforms the Willshaw network for almost all parameters (see equation 5.5). Asymptotically, sequential implementations of the single-layer Willshaw model remain superior only for almost nonsparse patterns ($k \sim n^d$ with d near 1) or if content patterns are much sparser than address patterns.

The neurobiological implications of the new efficient regimes we discovered in the Willshaw model (sparse and dense synaptic potentiation corresponding to ultrasparse and moderately sparse patterns) rely on two oversimplifications that need to be addressed in future work.

First, our analyses have assumed that learning starts in a fully connected network and is followed by a pruning phase, where the silent dispensable synapses can be pruned. Since neural networks of the brain have generally low connectivity at any time, this highly simplified model must be refined. Currently we investigate a more realistic model for cortical memory in which a low-capacity memory buffer network (e.g., the hippocampus) interacts with a high-capacity associative projection (e.g., a cortico-cortical synaptic connection), which is subject to structural plasticity. Pattern

associations are temporarily stored in the low-capacity buffer and repeatedly replayed to the high-capacity network. The combination of repetitive training, structural plasticity, and an adequate consolidation of activated synapses emulates a fully connected network equivalent to the model analyzed in this work, although the connectivity level in the cortical module is always low (Knoblauch, 2006, 2009).

Second, it needs to be explained how the regime of moderately sparse patterns with $k/\log n \to \infty$ corresponding to dense synaptic potentiation with $p_1 \to 1$ can be realized in realistic neuronal circuitry. This regime becomes efficient in terms of high synaptic capacity or few synaptic operations per retrieval but only if implemented with inhibitory neurons where the rare silent (0-)synapses are maintained and the large number of active (1-)synapses can be pruned (Knoblauch, 2006). The implementation of this regime is conceivable in brain structures that are dominated by inhibitory neurons (e.g., cerebellum, basal ganglia) and also by using specific types of inhibitory interneurons in cortical microcircuits.

## Appendix A: Binary Channels

The Shannon information $I(X)$ of a binary random variable $X$ on $\Omega = \{0, 1\}$ with $p := \mathrm{pr}[X = 1]$ equals

$$\begin{aligned} I(p) &:= -p \cdot \mathrm{ld}\, p - (1-p) \cdot \mathrm{ld}(1-p) \\ &\approx \begin{cases} -p \cdot \mathrm{ld}\, p, & p \ll 0.5 \\ -(1-p) \cdot \mathrm{ld}(1-p), & 1-p \ll 0.5 \end{cases} \end{aligned} \tag{A.1}$$

(Shannon & Weaver, 1949; Cover & Thomas, 1991). Note the symmetry $I(p) = I(1-p)$, and that $I(p) \to 0$ for $p \to 0$ (and $p \to 1$). A binary memoryless channel is determined by the two error probabilities $p_{01}$ (false one) and $p_{10}$ (false zero). For two binary random variables $X$ and $Y$, where $Y$ is the result of transmitting $X$ over the binary channel, we can write

$$I(Y) = I_Y(p, p_{01}, p_{10}) := I\left(p\,(1-p_{10}) + (1-p)\,p_{01}\right) \tag{A.2}$$

$$I(Y|X) = I_{Y|X}(p, p_{01}, p_{10}) := p \cdot I(p_{10}) + (1-p) \cdot I(p_{01}) \tag{A.3}$$

$$T(X;Y) = T(p, p_{01}, p_{10}) := I_Y(p, p_{01}, p_{10}) - I_{Y|X}(p, p_{01}, p_{10}). \tag{A.4}$$

For the analysis of pattern part retrieval in section 3.1, the case $p_{10} = 0$ is of particular interest:

$$T(p, p_{01}, 0) = I\left(p + p_{01} - p p_{01}\right) - (1-p) \cdot I(p_{01}) \tag{A.5}$$

$$\begin{aligned} &\le I(p_{01}) + I'(p_{01}) \cdot (p(1-p_{01})) - (1-p) \cdot I(p_{01}) \\ &= -p \mathrm{ld}\, p_{01}. \end{aligned} \tag{A.6}$$

For the upper bound, we have linearized $I$ in $p_{01}$ and used the convexity of $I(p)$—$(dI/dp)^2 = -1/(p(1-p)\ln 2) < 0$. The upper bound becomes exact for $p/p_{01} \to 0$. For high fidelity, we are typically interested in $p_{01} \ll p := l/n$ (see section 3.2). Thus, linearization of $I$ in $p$ yields a better upper bound,

$$T(p_1, p_{01}, 0) \leq I(p) + I'(p) \cdot (1-p) \cdot p_{01} - (1-p) \cdot I(p_{01}) \leq I(p), \tag{A.7}$$

where the approximations become exact in the limit $p_{01}/p \to 0$. For the relative error $e_I$ of approximating $T(p, p_{01}, p_{10})$ by $I(p)$, we can write

$$\begin{aligned} e_I &:= \frac{I(p_1) - T(p_1, p_{01}, p_{10})}{I(p_1)} \approx (1-p_1)\frac{I(p_{01}) - I'(p_1) \cdot p_{01}}{I(p_1)} \\ &\approx \frac{I(p_{01})}{I(p_1)} - \frac{p_{01}}{p_1}, \end{aligned} \tag{A.8}$$

where for the last approximation, we additionally assume $p \ll 0.5$ and correspondingly $1 - p \approx 1$, $I(p) \approx -p\mathrm{ld}\,p$, and $I'(p) \approx -\mathrm{ld}\,p$.

Applying these results to our analysis of the Willshaw model in section 3.2, using $p := l/n \ll 0.5$ and $p_{01} := \epsilon p$ for $\epsilon \ll 1$, we obtain

$$e_I \leq \frac{I(\epsilon\frac{l}{n})}{I(\frac{l}{n})} - \epsilon \approx \epsilon \cdot \frac{\mathrm{ld}\epsilon}{\mathrm{ld}(\frac{l}{n})} \approx \frac{I(\epsilon)}{-\mathrm{ld}(\frac{l}{n})} \leq \begin{cases} I(\epsilon), & \text{in any case} \\ \epsilon, & \mathrm{l/n} \leq \epsilon \end{cases}. \tag{A.9}$$

Note that typically sparse patterns with $l/n \ll 1/100$ are used. Thus, requiring for example, $\epsilon = 0.01$ implies that the relative error of approximating $T$ by $I$ in equation 3.10, is smaller than 1%.

## Appendix B: Exact Retrieval Error Probabilities for Fixed Query Activity

Our analysis so far used the binomial approximation, equation 3.3. Here we give the exact expressions for fixed query pattern activity, that is, when the query pattern $\tilde{\mathbf{u}}$ has exactly $c := \lambda k$ correct one-entries from one of the address patterns $\mathbf{u}^\mu$ and, additionally, $f := \kappa k$ false one-entries ($0 < \lambda \leq 1$, $\kappa \geq 0$). Retrieving with threshold $\Theta$, the exact retrieval error probabilities $p_{01} := \mathrm{pr}(\hat{v}_i = 1 | v_i^\mu = 0)$ of a false one-entry and $p_{10} := \mathrm{pr}(\hat{v}_i = 0 | v_i^\mu = 1)$ of a missing one-entry are

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{WP}}(x; k, l, m, n, M-1, c+f) \tag{B.1}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{WP}}(x-c; k, l, m, n, M-1, f)\,, \tag{B.2}$$

where $p_{\mathrm{WP}}(x; k, l, m, n, M, z)$ is the distribution of dendritic potential $x$ when stimulating with a random query pattern having exactly $z$ one-entries and $m - z$ zero entries ($0 \le x \le z$). It is

$$\begin{aligned} & p_{\mathrm{WP}}(x; k, l, m, n, M, z) \\ &= \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} \left(1 - \frac{l}{n}(1 - B(m, k, s+z-x))\right)^M && \text{(B.3)} \\ &\approx \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} \left(1 - \frac{l}{n}\left(1 - \left(1 - \frac{k}{m}\right)^{s+z-x}\right)\right)^M && \text{(B.4)} \\ &= \sum_{i=0}^{M} p_B(i; M, l/n)\, p_B(x; z, 1-(1-k/m)^i), && \text{(B.5)} \end{aligned}$$

where we used $B(a, b, c) := \binom{a-b}{c} / \binom{a}{c} = \prod_{i=0}^{c-1} (a-b-i)/(a-i)$ and the binomial probability $p_B(x; N, P) := \binom{N}{x} P^x (1-P)^{N-x}$. Equation B.3 is exact for fixed address pattern activity, that is, if each address pattern $\mathbf{u}^\mu$ has exactly $k$ one-entries and has been found by Knoblauch (2008), generalizing a previous approach of Palm (1980) for the particular case of zero noise ($c = k$, $f = 0$). The approximations, equations B.4 and B.5, would be exact for random address pattern activity, that is, if $u_i^\mu$ is one with probability $k/m$ (but still fixed $c$, $f$). Equation B.5 averages over the so-called unit-usage (the number of patterns a given content neuron belongs to) and has been found by Buckingham and Willshaw (1992) and Buckingham (1991). The transformation to equation B.4 has been found by Sommer and Palm (1999). Equations B.3 and B.4 are numerically efficient to evaluate for low query pattern activity $c + f$, whereas equation B.5 is efficient for a few stored patterns $M$. The distinction between fixed and random address pattern activity, $|u^\mu|$, is of minor interest for moderately large networks, because then equations B.3 to B.5 yield very similar values (Knoblauch, 2006, 2008). However, the distinction between fixed and random query pattern activity, $|\tilde{\mathbf{u}}|$, remains important even for large networks (see appendix D).

For the particular case of pattern part retrieval, $c = \lambda k$ and $f = 0$, we can use the Willshaw threshold $\Theta = \lambda k$, and the error probabilities are $p_{10} = 0$ and

$$p_{01} = \sum_{s=0}^{\lambda k} (-1)^s \binom{\lambda k}{s} \left[1 - \frac{l}{n}(1 - B(m, k, s))\right]^{M-1} \tag{B.6}$$

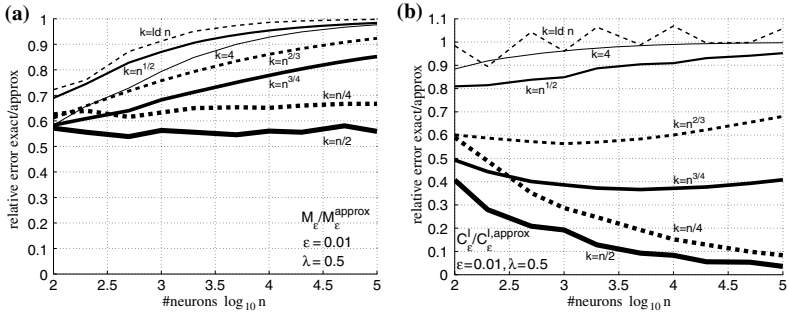


Figure 9: Approximation quality of our analysis in sections 3 and 4 based on equation 3.3 for $m = n, k = l$, high-fidelity parameter $\epsilon = 0.01$, when addressing with half address patterns ($\lambda = 0.5$, $\kappa = 0$). (a) Relative approximation quality of the pattern capacity $M_\epsilon / M_\epsilon^{\text{approx}}$ as a function of neuron number $n$. The exact value $M_\epsilon$ is computed as in Table 2 and the approximation $M_\epsilon^{\text{approx}}$ is computed from equation 3.8. The different lines correspond to different pattern activities $k(n) = 4, \text{ld}n, \sqrt{n}, n^{2/3}, n^{3/4}, n/4, n/2$ (increasing line thickness; alternation of solid and dashed lines). Approximation quality for network capacity $C_\epsilon$ is qualitatively the same. (b) Relative approximation quality similar to $a$, but for the information capacity $C_\epsilon^I$. $C^{I,\text{approx}}$ is computed from equation 4.1. Approximation quality for the synaptic capacity $C_\epsilon^S$ is qualitatively the same.

$$\approx \sum_{s=0}^{\lambda k} (-1)^s \binom{\lambda k}{s} \left[ 1 - \frac{l}{n}(1 - (1 - k/m)^s) \right]^{M-1} \tag{B.7}$$

$$= \sum_{i=0}^{M-1} p_B(i; M-1, l/n)(1 - (1 - k/m)^i)^{\lambda k} \tag{B.8}$$

$$\geq p_1^{\lambda k} \ . \tag{B.9}$$

Here equations B.6 to B.8 correspond to equations B.3 to B.5, and the bound corresponds to the binomial approximation, equation 3.3. Knoblauch (2007, 2008) shows that this lower bound becomes tight at least for $k \sim O(n/\log^4 n)$ or, for $m = n, k = l$, already for $k \sim O(n/\log^2 n)$. Thus, our theory based on the binomial approximation, equation 3.3, becomes exact for virtually any sublinear $k(n)$.

We have validated these results in extensive numerical experiments, which can be found in Knoblauch (2006, 2008) and Knoblauch et al. (2008). Table 2 shows some exact results when addressing with half patterns ($\lambda = 0.5$, $\kappa = 0$). Figure 9 plots the quality of the binomial approximation, equation 3.3, for pattern capacity $M$ and information capacity $C^I$ for different sparsity levels and increasing network size $n \to \infty$.

Table 2: Exact Capacities of the Willshaw Model Computed from Equation B.6 for $m = n$, $k = l$, High Fidelity $\epsilon = 0.01$ When Addressing with Half Address Patterns ($\lambda = 0.5$, $\kappa = 0$).

| $n =$ | 100 | 200 | 500 | 1000 | 2000 | 5000 | 10,000 | 20,000 | 50,000 | 100,000 |
|---|---|---|---|---|---|---|---|---|---|---|
| $k = 4$ | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |
| $M_\epsilon$ | 7 | 23 | 102 | 315 | 951 | 3985 | 11,614 | 33,561 | 135,216 | 386,157 |
| $C_\epsilon$ | 0.016734 | 0.016080 | 0.013581 | 0.011749 | 0.009820 | 0.007427 | 0.005876 | 0.004581 | 0.003239 | 0.002467 |
| $C_\epsilon^I$ | 0.189510 | 0.213911 | 0.239855 | 0.257522 | 0.272803 | 0.289919 | 0.301034 | 0.310883 | 0.322324 | 0.330003 |
| $C_\epsilon^S$ | 1.501279 | 1.755475 | 2.087170 | 2.337024 | 2.586433 | 2.915938 | 3.165234 | 3.414622 | 3.744455 | 3.994076 |
| $k = \mathrm{ld} n$ | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 16 | 17 |
| $M_\epsilon$ | 26 | 73 | 530 | 1578 | 6825 | 31481 | 130517 | 410162 | 2239454 | 8958499 |
| $C_\epsilon$ | 0.093667 | 0.087255 | 0.136318 | 0.126214 | 0.166369 | 0.152057 | 0.185759 | 0.169994 | 0.185909 | 0.211443 |
| $C_\epsilon^I$ | 0.177045 | 0.174203 | 0.216708 | 0.210461 | 0.239663 | 0.234620 | 0.258792 | 0.248317 | 0.254089 | 0.272940 |
| $C_\epsilon^S$ | 0.781248 | 0.790925 | 0.863820 | 0.864564 | 0.891863 | 0.916887 | 0.938451 | 0.933671 | 0.907202 | 0.926973 |
| $k = \sqrt{n}$ | 10 | 14 | 22 | 32 | 45 | 71 | 100 | 141 | 224 | 316 |
| $M_\epsilon$ | 20 | 67 | 294 | 791 | 2122 | 7082 | 17013 | 40294 | 119800 | 271628 |
| $C_\epsilon$ | 0.092180 | 0.120686 | 0.150986 | 0.159572 | 0.162795 | 0.150634 | 0.136076 | 0.120799 | 0.098333 | 0.082962 |
| $C_\epsilon^I$ | 0.134642 | 0.140982 | 0.152895 | 0.160997 | 0.175765 | 0.189566 | 0.198546 | 0.211598 | 0.224751 | 0.235512 |
| $C_\epsilon^S$ | 0.506227 | 0.430356 | 0.347634 | 0.358847 | 0.476765 | 0.628296 | 0.745907 | 0.895062 | 1.088783 | 1.249831 |
| $k = n^{2/3}$ | 22 | 34 | 63 | 100 | 159 | 292 | 464 | 737 | 1357 | 2154 |
| $M_\epsilon$ | 11 | 27 | 76 | 156 | 310 | 736 | 1371 | 2509 | 5454 | 9662 |
| $C_\epsilon$ | 0.081660 | 0.086933 | 0.081497 | 0.071901 | 0.061067 | 0.046564 | 0.036626 | 0.028186 | 0.019377 | 0.014325 |
| $C_\epsilon^I$ | 0.083180 | 0.087490 | 0.092952 | 0.097348 | 0.104483 | 0.114870 | 0.124077 | 0.134520 | 0.149156 | 0.160552 |
| $C_\epsilon^S$ | 0.194163 | 0.191892 | 0.275016 | 0.344860 | 0.435923 | 0.575533 | 0.703203 | 0.852483 | 1.078028 | 1.268922 |
| $k = n/4$ | 25 | 50 | 125 | 250 | 500 | 1250 | 2500 | 5000 | 12500 | 25000 |
| $M_\epsilon$ | 10 | 16 | 24 | 31 | 39 | 49 | 56 | 64 | 74 | 82 |
| $C_\epsilon$ | 0.079104 | 0.063284 | 0.037970 | 0.024522 | 0.015425 | 0.007752 | 0.004430 | 0.002531 | 0.001171 | 0.000649 |
| $C_\epsilon^I$ | 0.079241 | 0.067368 | 0.050885 | 0.042898 | 0.038121 | 0.030659 | 0.024775 | 0.021308 | 0.016677 | 0.014209 |
| $C_\epsilon^S$ | 0.166347 | 0.177726 | 0.178703 | 0.181323 | 0.191142 | 0.183161 | 0.164436 | 0.157467 | 0.138863 | 0.128935 |

Note: Table entries correspond to network size $n$, pattern activity $k$, pattern capacity $M_\epsilon$, network capacity $C_\epsilon$, information capacity $C_\epsilon^I$, and synaptic capacity $C_\epsilon^S$.

## Appendix C: Fallacies for Extremely Sparse and Nonsparse Activity

As discussed in section 3.3, our analysis method is exact for both small and very large networks, whereas alternative methods are inaccurate for finite networks and, for some parameter ranges, even in the asymptotic limit. For example, previous analyses of feedforward associative networks with linear learning, such as the covariance rule, often compute capacity as a function of the so-called signal-to-noise ratio $\mathrm{SNR} = (\mu_{\mathrm{hi}} - \mu_{\mathrm{lo}})^2/\sigma^2$, defined as the mean potential difference between "high units" (which should be active in the retrieval result $\hat{\mathbf{v}}$) and "low units" (which should be inactive) divided by the potential variance (Dayan & Willshaw, 1991; Palm, 1991; Palm & Sommer, 1996). Assuming gaussian dendritic potentials, such analyses propose an asymptotic network capacity $C = 0.72$ for linear associative networks with covariance learning and $k/m \to 0$, which seems to be better than the binary Willshaw model. However, numerical evaluations prove that even for moderate sparseness in large, finite networks, the Willshaw model performs better (data not shown). To analyze the reason for this discrepancy, we compute the SNR for the Willshaw model,

$$\mathrm{SNR}_{\mathrm{Willshaw}} \approx \frac{(\lambda k(1-p_1))^2}{\lambda k p_1(1-p_1)} = \frac{\lambda k(1-p_1)}{p_1}. \tag{C.1}$$

The SNR for the network with linear learning and the optimal covariance rule has been found to be $m/(M(l/n)(1-l/n))$ for zero query noise (Dayan & Willshaw, 1991; Palm & Sommer, 1996). Using $M$ as in equation 3.2 and assuming small $p_1 \to 0$, this becomes

$$\mathrm{SNR}_{\mathrm{Cov}} \approx \frac{mkl}{-mn(l/n)\ln(1-p_1)} = \frac{k}{-\ln(1-p_1)}. \tag{C.2}$$

Thus, for small $p_1 \to 0$, the SNR will be $k/p_1$ for both models, which falsely suggests, assuming gaussian dendritic potentials, that the Willshaw model could also store 0.72 bits per synapse, which is, of course, wrong. In fact, for $k/\log n \to 0$ (which is equivalent to $p_{1\epsilon} \to 0$), equation 3.12 proves zero capacity for the Willshaw model and strongly suggests the same result for the covariance rule in the linear associative memory. Further numerical experiments and theoretical considerations show that even for $k \sim \log n$, the Willshaw model performs better than linear covariance learning, although it cannot exceed $C = 0.69$ or $C = 0.53$. This shows that the SNR method and the underlying gaussian approximation become reliable only for dense potentiation with $p_{1\epsilon} \to 1$ and $k/\log n \to \infty$ (see also Knoblauch, 2008; Henkel & Opper, 1990).

But even for dense potentiation, the gaussian assumption is inaccurate for linear pattern activities $k = cn$ and $l = dn$ with constant $c$ and $d$,

falsely suggesting constant pattern capacity $M_\epsilon \sim 1$ for $m, n \to \infty$ (note that dense potentiation may imply highly asymmetric potential distributions; see Knoblauch, 2003b). In fact, $M_\epsilon \to \infty$ diverges for RQ1 as can be seen in equation B.8. Moreover, we can compute upper and lower bounds for equation B.8 by assuming that all content neurons have a unit usage $i$ larger or smaller than $Md + \xi\sqrt{Md(1-d)}$ (note that $p_{01}$ given $i$ increases with $i$),

$$p_{01} \approx (1-(1-c)^{Md+\xi\sqrt{Md(1-d)}})^{\lambda cm}. \tag{C.3}$$

For sufficiently large positive (but for RQ1 still constant) $\xi$, this approximation is an upper bound. For example, we can choose $\xi := G^{c-1}(\epsilon_1 d)$ with $\epsilon_1 \ll \epsilon$ such that only a few content neurons have a unit usage more than $\xi$ standard deviations larger than the mean unit usage (here, $G^c(x) := 0.5\mathrm{erfc}(x/\sqrt{2})$ is the gaussian tail integral). Similarly, for large negative $\xi$ we obtain a lower bound. Requiring $p_{01} \le \epsilon d/(1-d)$, we obtain for the pattern capacity

$$M_\epsilon + \xi\sqrt{M_\epsilon(1-d)/d} \approx \frac{\ln(1-(\epsilon d/(1-d))^{1/(\lambda cm)})}{d\ln(1-c)} \approx \frac{\ln m}{-d\ln(1-c)}. \tag{C.4}$$

Thus, the pattern capacity is essentially independent of $\xi$. However, compared to equations 3.8 and 4.10, the asymptotic pattern capacity is reduced by a factor $f := (-\ln(1-c))/c < 1$. This turns out to be the reason that the Willshaw network has zero information capacity $C^I \to 0$ and zero synaptic capacity $C^S \to 0$ for linear address pattern activity $k = cm$. With $\tilde{p}_{0\epsilon} := (1-cd)^{fM_\epsilon} \to 0$ (see equation 3.1), it is $p_{0\epsilon} := 1 - p_{1\epsilon} = \tilde{p}_{0\epsilon}^f$ (see equation 3.7). Therefore, equation 4.1 becomes $C_\epsilon^I \sim \mathrm{ld}(1-\tilde{p}_{0\epsilon})(\ln \tilde{p}_{0\epsilon})/(\tilde{p}_{0\epsilon}^f \mathrm{ld}\, \tilde{p}_{0\epsilon}^f) \sim \tilde{p}_{0\epsilon}^{1-f} \to 0$. Similarly, equation 4.3 becomes $C_\epsilon^S \sim \mathrm{ld}(1-\tilde{p}_{0\epsilon})(\ln \tilde{p}_{0\epsilon})/\tilde{p}_{0\epsilon}^f \approx \tilde{p}_{0\epsilon}^{1-f} \ln \tilde{p}_{0\epsilon} \to 0$.

## Appendix D: Corrections for Random Query Activity

So far, our exact theory in appendix B as well as the approximative theory in sections 3 to 5 assume that the query pattern $\tilde{\mathbf{u}}$ has exactly $\lambda k$ correct one-entries (and $\kappa k$ false one-entries). This is sufficient for many applications where specifications assume a minimal quality of query patterns in terms of a lower bound for the number of correct one-entries. However, in particular for small $k$ or large $\lambda$ near 1, we may want to include the case of random query pattern activity. In the following, we assume that the address patterns have random activity—each pattern component $u_i^\mu$ is one with probability $k/m$ independent of other components. Similarly, in a query pattern $\tilde{\mathbf{u}}$, a one-entry is erased with probability $1-\lambda$. For simplicity,

we assume no add noise (i.e., $\kappa = 0$). Thus, a component in the query pattern, $\tilde{u}_i$, is one with probability $\lambda k/m$. Then the query pattern activity $Z$ is a binomially distributed random variable, $\mathrm{pr}[Z = z] = p_B(z; m, \lambda k/m)$ (for $p_B$, see equation B.5). For a particular $Z = z$, the exact error probability $p_{01}$ is given by equation B.7 (or equation B.8), replacing $\lambda k$ by $z$. Averaging over all possible $z$ yields

$$\begin{aligned} p_{01}^* &= \sum_{z=0}^{m} p_B(z; m, \lambda k/m) \sum_{s=0}^{z} (-1)^s \binom{z}{s} \left[1 - \frac{l}{n}(1 - (1 - k/m)^s)\right]^{M-1} \\ &= \sum_{s=0}^{m} \left(-\frac{\lambda k}{m}\right)^s \binom{m}{s} \left[1 - \frac{l}{n}(1 - (1 - k/m)^s)\right]^{M-1} \qquad \text{(D.1)} \\ &= \sum_{i=0}^{M-1} p_B(i; M-1, l/n) \sum_{z=0}^{m} p_B(z; m, \lambda k/m)(1 - (1 - k/m)^i)^z \\ &= \sum_{i=0}^{M-1} p_B(i; M-1, l/n) \left(1 - \frac{\lambda k}{m}(1 - k/m)^i\right)^m. \qquad \text{(D.2)} \end{aligned}$$

The first equation is numerically efficient for small $k$, the last equation for small $M$. For the binomial approximative analyses, we can rewrite equation 3.3 as

$$p_{01}^* \approx \sum_{z=0}^{m} p_B(z; m, \lambda k/m) p_1^z = \left(1 - \lambda \frac{k}{m}(1 - p_1)\right)^m. \qquad \text{(D.3)}$$

Controlling for retrieval quality, $p_{01}^* \leq \epsilon l/(n - l)$, the maximal memory load, equation 3.7, becomes

$$p_{1\epsilon}^* \approx 1 - \frac{1 - (\frac{\epsilon l}{n-l})^{1/m}}{\lambda k/m}. \qquad \text{(D.4)}$$

Note that positive $p_{1\epsilon}^* \geq 0$ requires $\epsilon \geq e^{-\lambda k}(n - l)/l$ or, equivalently, $k \geq \ln((n - l)/(\epsilon l))/\lambda$. Consequently, even for logarithmic $k, l = O(\log n)$, it may be impossible to achieve retrieval quality levels RQ1 or higher (see section 2.1). For example, $k \leq c \log n$ with $c < 1$ implies diverging noise $\epsilon \geq n^{1-c}/l$, while RQ1 would require constant $\epsilon \sim 1$ and RQ2 or RQ3 even vanishing $\epsilon \to 0$. This is a major difference to the model with fixed query pattern activity.

Writing $x := \epsilon l/(n-l)$ and using $e^x = \sum_{i=0}^{\infty} x^i/i!$, we obtain for the difference $\Delta p_{1\epsilon} := p_{1\epsilon} - p^*_{1\epsilon}$ between equation 3.7 and equation D.4:

$$\Delta p_{1\epsilon} \approx e^{(\ln x)/(\lambda k)} - \left(1 + \frac{e^{(\ln x)/m} - 1}{\lambda k/m}\right) \tag{D.5}$$

$$= \sum_{i=1}^{\infty} \frac{(\ln x)^i}{i!(\lambda k)^i} - \frac{(\ln x)^i}{i!\lambda k m^{i-1}} \tag{D.6}$$

$$= \sum_{i=2}^{\infty} \frac{(\ln x)^i}{i!(\lambda k)^i}(1 - (\lambda k/m)^{i-1}) \tag{D.7}$$

$$\approx p_{1\epsilon} - 1 - \ln p_{1\epsilon}, \tag{D.8}$$

where the last approximation is true for balanced potentiation with fixed $p_{1\epsilon}$ and $\lambda k/m \to 0$. Note that for sparse potentiation with $p_{1\epsilon} \to 0$ and $k/\log n \to 0$ we have diverging $\Delta p_{1\epsilon}$. At least for dense potentiation with $p_{1\epsilon} \to 1$ and $k/\log n \to \infty$, the relative differences vanish: $\Delta p_{1\epsilon}/p_{1\epsilon} \to 0$ and even $\Delta p_{1\epsilon}/(1 - p_{1\epsilon}) \to 0$. Thus, at least for dense potentiation, the models with fixed and random query pattern activity become equivalent, including all results on information capacity $C^I$ and synaptic capacity $C^S$ (see sections 3–5). Proceeding as in section 3.2, we obtain

$$p^*_{1\epsilon} \approx 1 + \ln p_{1\epsilon} \approx 1 - \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k} \tag{D.9}$$

$$p^*_{0\epsilon} := 1 - p^*_{1\epsilon} = \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k} \qquad \left(\Leftrightarrow k \approx \frac{\ln \frac{n-l}{\epsilon l}}{\lambda p^*_{0\epsilon}}\right) \tag{D.10}$$

$$M^*_\epsilon = -\frac{mn}{kl} \ln p^*_{0\epsilon} \approx -\lambda^2 p^{*2}_{0\epsilon} \ln p^*_{0\epsilon} \frac{k}{l} \frac{mn}{(\ln \frac{n-l}{\epsilon l})^2} \tag{D.11}$$

$$C^*_\epsilon = M_\epsilon m^{-1} T(l/n, \epsilon l/(n-l), 0) \approx -\lambda p^*_{0\epsilon} \mathrm{ld}\, p^*_{0\epsilon} \eta. \tag{D.12}$$

The asymptotic bound of network capacity is thus only $C^*_\epsilon \le 1/(e \ln 2) \approx 0.53$ for $p^*_{0\epsilon} = 1/e \approx 0.368$ and retrieval quality levels RQ0-RQ2 (for RQ3, the bound decreases by factor 1/3 as discussed in section 3.3). Figure 10 illustrates asymptotic capacities in analogy to Figure 6. For dense potentiation, $p^*_{0\epsilon} \to 0$, results are identical to the model with fixed query pattern activity. For sparse potentiation, $p^*_{0\epsilon} \to 1$, we have $C^{I*}_\epsilon := C^*_\epsilon / I(p^*_{0\epsilon}) \to 0$ and still $C^{S*}_\epsilon := C^*_\epsilon / \min(p^*_{0\epsilon}, 1 - p^*_{0\epsilon}) \to 1/\ln 2 \approx 1.44$. For $k = l$ maximal pattern capacity is $0.18\lambda^2 mn/(\mathrm{ld} n)^2$ for $p^*_{0\epsilon} = 1/\sqrt{e} \approx 0.61$.

Note that our result $C^* \le 0.53$ contradicts previous analyses. For example, Nadal (1991) estimates $C^* \le 0.236$ for $p^*_1 = 0.389$. We believe that our results are correct and that the discrepancies are due to inaccurate approximations employed by previous work. In fact, we have verified the accuracy of our theory in two steps (see Knoblauch, 2006, 2008; Knoblauch et al.,

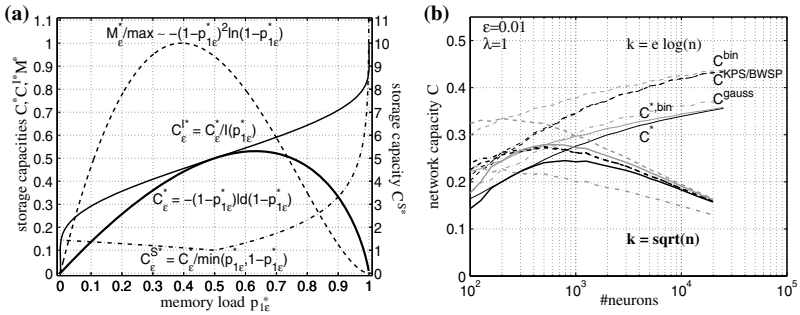


Figure 10: (a) Asymptotic network capacity $C^*_\epsilon$, information capacity $C^{I*}_\epsilon$, synaptic capacity $C^{S*}$, and pattern capacity $M^*_\epsilon$ as functions of memory load $p^*_{1\epsilon}$ for the model variant with random query pattern activity. Compare to Figure 6a. (b) Exact and approximative network capacity for finite network sizes $m = n$ and mean pattern activity $k = l = e \ln n$ (thin lines) or $k = l = \sqrt{n}$ (bold lines). For random query pattern activity, the plot shows results computed with exact equation D.1 ($C^*$; black solid) and binomial approximation equation D.9 ($C^{*,\mathrm{bin}}$; gray solid). For fixed query pattern activity, the plot shows results computed with exact equation B.7 ($C^{\mathrm{BWSP}}$; black dashed) and equation B.6 ($C^{\mathrm{KPS}}$; black dash-dotted), the binomial approximation equation 3.7 ($C^{\mathrm{bin}}$; gray dashed), and a gaussian approximation of dendritic potentials ($C^{\mathrm{gauss}}$; gray dash-dotted; see, Knoblauch, 2008). Note that the binomial approximations closely approximate the exact values already for relatively small networks. In contrast, the gaussian approximation significantly underestimates capacity even for large networks.

2008). First, we have verified all our formulas for the exact error probabilities of the different model variants (equations B.1 to B.8 and D.1 and D.6) by extensive simulations of small networks. Second, we have proven the asymptotic correctness of our binomial approximative theory (see equations 3.3, 3.7–3.9, and D.9–D.12) by theoretical considerations and numerical experiments (see also Figure 10).

## Acknowledgments

We thank Sen Cheng, Marc-Oliver Gewaltig, Edgar Körner, Ursula Körner, Bartlett Mel, and Xundong Wu for helpful discussions, as well as Pentti Kanerva for his comments to an earlier version of the manuscript. F.T.S. was supported by NSF grant IIS-0713657 and a Google research award.

## References

Abeles, M. (1982). *Local cortical circuits.* Berlin: Springer.

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). Spatio-temporal firing patterns in frontal cortex of behaving monkeys. *Journal of Neurophysiology*, *70*, 1629–1643.

Albus, J. (1971). A theory of cerebellar function. *Mathematical Biosciences*, *10*, 25–61.

Amari, S.-I. (1989). Characteristics of sparsely encoded associative memory. *Neural Networks*, *2*, 451–457.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1987a). Information storage in neural networks with low levels of activity. *Phys. Rev. A*, *35*, 2293–2303.

Amit, D., Gutfreund, H., & Sompolinsky, H. (1987b). Statistical mechanics of neural networks near saturation. *Annals of Physics*, *173*, 30–67.

Anderson, J. (1968). A memory storage model utilizing spatial correlation functions. *Kybernetik*, *5*, 113–119.

Anderson, J. (1993). The BSB model: A simple nonlinear autoassociative neural network. In M. Hassoun (Ed.), *Associative neural memories.* New York: Oxford University Press.

Anderson, J., Silverstein, J., Ritz, S., & Jones, R. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413–451.

Arieli, A., Sterkin, A., Grinvald, A., & Aertsen, A. (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science*, *273*, 1868–1871.

Attwell, D., & Laughlin, S. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow and Metabolism*, *21*, 1133–1145.

Aviel, Y., Horn, D., & Abeles, M. (2005). Memory capacity of balanced networks. *Neural Computation*, *17*, 691–713.

Barlow, H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology. *Perception*, *1*, 371–394.

Bentz, H., Hagstroem, M., & Palm, G. (1989). Information storage and effective data retrieval in sparse matrices. *Neural Networks*, *2*, 289–293.

Bogacz, R., & Brown, M. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, *13*, 494–524.

Bogacz, R., Brown, M., & Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience*, *10*, 5–23.

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

Braitenberg, V. (1978). Cell assemblies in the cerebral cortex. In R. Heim & G. Palm (Eds.), *Theoretical approaches to complex systems* (pp. 171–188). Berlin: Springer-Verlag.

Braitenberg, V., & Schüz, A. (1991). *Anatomy of the cortex: Statistics and geometry.* Berlin: Springer-Verlag.

Buckingham, J. (1991). *Delicate nets, faint recollections: A study of partially connected associative network memories.* Unpublished doctoral dissertation, University of Edinburgh.

Buckingham, J., & Willshaw, D. (1992). Performance characteristics of associative nets. *Network: Computation in Neural Systems*, *3*, 407–414.

Buckingham, J., & Willshaw, D. (1993). On setting unit thresholds in an incompletely connected associative net. *Network: Computation in Neural Systems*, *4*, 441–459.

Burks, A., Goldstine, H., & von Neumann, J. (1946). *Preliminary discussion of the logical design of an electronic computing instrument* (Rep. 1946). U.S. Army Ordnance Department.

Chicca, E., Badoni, D., Dante, V., D'Andreagiovanni, M., Salina, G., Carota, L., et al. (2003). A VLSI recurrent network of integrate-and-fire neurons connected by plastic synapses with long-term memory. *IEEE Transactions on Neural Networks*, *14*, 1297–1307.

Coolen, A. (2001a). Statistical mechanics of recurrent neural networks I: Statics. In F. Moss & S. Gielen (Eds.), *Handbook of biological physics* (Vol. 4, pp. 531–596). Amsterdam: Elsevier Science.

Coolen, A. (2001b). Statistical mechanics of recurrent neural networks II: Dynamics. In F. Moss & S. Gielen (Eds.), *Handbook of biological physics* (Vol. 4, pp. 597–662). Amsterdam: Elsevier Science.

Cover, T., & Thomas, J. (1991). *Elements of information theory.* New York: Wiley.

Dayan, P., & Willshaw, D. (1991). Optimising synaptic learning rules in linear associative memory. *Biological Cybernetics*, *65*, 253–265.

Diesmann, M., Gewaltig, M., & Aertsen, A. (1999). Stable propagation of synchronous spiking in cortical neural networks. *Nature*, 402(6761), 529–533.

Engert, F., & Bonhoeffer, T. (1999). Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature*, *399*, 66–70.

Fransen, E., & Lansner, A. (1998). A model of cortical associative memory based on a horizontal network of connected columns. *Network: Computation in Neural Systems*, *9*, 235–264.

Frolov, A., & Murav'ev, I. (1993). Informational characteristics of neural networks capable of associative learning based on Hebbian plasticity. *Network: Computation in Neural Systems*, *4*, 495–536.

Fusi, S., Drew, P., & Abbott, L. (2005). Cascade models of synaptically stored memories. *Neuron*, *45*, 599–611.

Gardner, E., & Derrida, B. (1988). Optimal storage properties of neural network models. *J. Phys. A: Math. Gen.*, *21*, 271–284.

Gardner-Medwin, A. (1976). The recall of events through the learning of associations between their parts. *Proceedings of the Royal Society of London Series B*, *194*, 375–402.

Golomb, D., Rubin, N., & Sompolinsky, H. (1990). Willshaw model: Associative memory with sparse coding and low firing rates. *Phys. Rev. A*, *41*, 1843–1854.

Golomb, S. (1966). Run-length encodings. *IEEE Transactions on Information Theory*, *12*, 399–401.

Graham, B., & Willshaw, D. (1995). Improving recall from an associative memory. *Biological Cybernetics*, *72*, 337–346.

Graham, B., & Willshaw, D. (1997). Capacity and information efficiency of the associative net. *Network: Computation in Neural Systems*, *8*(1), 35–54.

Greene, D., Parnas, M., & Yao, F. (1994). Multi-index hashing for information retrieval. In *Proceedings of the 35th Annual Symposium on Foundations of Computer Science* (pp. 722–731). Piscataway, NJ: IEEE.

Hahnloser, R., Kozhevnikov, A., & Fee, M. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature*, *419*, 65–70.

Hammerstrom, D. (1990). A VLSI architecture for high-performance, low-cost, on-chip learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks 1990* (pp. II:537–543). Piscataway, NJ: IEEE.

Hammerstrom, D., Gao, C., Zhu, S., & Butts, M. (2006). FPGA implementation of very large associative memories. In A. Omondi & J. Rajapakse (Eds.), *FPGA implementations of neural networks* (pp. 167–195). New York: Springer.

Hebb, D. (1949). *The organization of behavior: A neuropsychological theory.* New York: Wiley.

Heittmann, A., & Rückert, U. (2002). Mixed mode VLSI implementation of a neural associative memory. *Analog Integrated Circuits and Signal Processing*, *30*, 159–172.

Hellwig, B. (2000). A quantitative analysis of the local connectivity between pyramidal neurons in layers 2/3 of the rat visual cortex. *Biological Cybernetics*, *82*, 111–121.

Henkel, R., & Opper, M. (1990). Distribution of internal fields and dynamics of neural networks. *Europhysics Letters*, *11*(5), 403–408.

Hertz, J., Krogh, A., & Palmer, R. (1991). *Introduction to the theory of neural computation.* Redwood City, CA: Addison-Wesley.

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

Hopfield, J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science, USA*, *81*(10), 3088–3092.

Hopfield, J., & Tank, D. (1986). Computing with neural circuits. *Science*, *233*, 625–633.

Huffman, D. (1952). A method for the construction of minimum redundancy codes. *Proceedings of the Institute of Radio Engineers*, *40*, 1098–1101.

Kanerva, P. (1988). *Sparse distributed memory.* Cambridge, MA: MIT Press.

Knoblauch, A. (2003a). Optimal matrix compression yields storage capacity 1 for binary Willshaw associative memory. In O. Kaynak, E. Alpaydin, E. Oja, & L. Xu (Eds.), *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003* (LNCS 2714, pp. 325–332). Berlin: Springer-Verlag.

Knoblauch, A. (2003b). *Synchronization and pattern separation in spiking associative memory and visual cortical areas.* Unpublished doctoral dissertation, University of Ulm, Germany.

Knoblauch, A. (2005). Neural associative memory for brain modeling and information retrieval. *Information Processing Letters*, *95*, 537–544.

Knoblauch, A. (2006). *On compressing the memory structures of binary neural associative networks* (HRI-EU Rep. 06-02). Offenbach/Main, Germany: Honda Research Institute Europe.

Knoblauch, A. (2007). *Asymptotic conditions for high-capacity neural associative networks* (HRI-EU Rep. 07-02). Offenbach/Main, Germany: Honda Research Institute Europe.

Knoblauch, A. (2008). Neural associative memory and the Willshaw-Palm probability distribution. *SIAM Journal on Applied Mathematics*, *69*(1), 169–196.

Knoblauch, A. (2009). The role of structural plasticity and synaptic consolidation for memory and amnesia in a model of cortico-hippocampal interplay. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist Models of Behavior and Cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop*. Singapore: World Scientific.

Knoblauch, A., & Palm, G. (2001). Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Networks*, *14*, 763–780.

Knoblauch, A., Palm, G., & Sommer, F. (2008). *Performance characteristics of sparsely and densely potentiated associative networks* (HRI-EU Rep. 08-02). Offenbach/Main, Germany: Honda Research Institute Europe.

Kohonen, T. (1977). *Associative memory: A system theoretic approach.* Berlin: Springer.

Lamprecht, R., & LeDoux, J. (2004). Structural plasticity and memory. *Nature Reviews Neuroscience*, *5*, 45–54.

Latham, P., & Nirenberg, S. (2004). Computing and stability in cortical networks. *Neural Computation*, *16*(7), 1385–1412.

Laughlin, S., & Sejnowski, T. (2003). Communication in neuronal networks. *Science*, *301*, 1870–1874.

Laurent, G. (2002). Olfactory network dynamics and the coding of multidimensional signals. *Nature Reviews Neuroscience*, *3*, 884–895.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, *13*, 493–497.

Little, W. (1974). The existence of persistent states in the brain. *Mathematical Biosciences*, *19*, 101–120.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology*, *202*(2), 437–470.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London, Series B*, *262*, 24–81.

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* Cambridge, MA: MIT Press.

Mu, X., Artiklar, M., Watta, P., & Hassoun, M. (2006). An RCE-based associative memory with application to human face recognition. *Neural Processing Letters*, *23*, 257–271.

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

Nadal, J.-P., & Toulouse, G. (1990). Information storage in sparsely coded memory nets. *Network: Computation in Neural Systems*, *1*, 61–74.

Palm, G. (1980). On associative memories. *Biological Cybernetics*, *36*, 19–31.

Palm, G. (1982). *Neural assemblies: An alternative approach to artificial intelligence.* Berlin: Springer.

Palm, G. (1987). Computing with neural networks. *Science*, *235*, 1227–1228.

Palm, G. (1990). Cell assemblies as a guideline for brain research. *Concepts in Neuroscience*, *1*, 133–148.

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

Palm, G., & Palm, M. (1991). Parallel associative networks: The PAN-system and the Bacchus-chip. In U. Ramacher, U. Rückert, & J. Nossek (Eds.), *Proceedings of the 2nd International Conference on Microelectronics for Neural Networks*. Munich: Kyrill & Method Verlag.

Palm, G., & Sommer, F. (1992). Information capacity in recurrent McCulloch-Pitts networks with sparsely coded memory states. *Network*, *3*, 177–186.

Palm, G., & Sommer, F. (1996). Associative data storage and retrieval in neural nets. In E. Domany, J. van Hemmen, & K. Schulten (Eds.), *Models of neural networks III* (pp. 79–118). New York: Springer-Verlag.

Perez-Orive, J., Mazor, O., Turner, G., Cassenaer, S., Wilson, R., & Laurent, G. (2002). Oscillations and sparsening of odor representations in the mushroom body. *Science*, *297*, 359–365.

Poirazi, P., & Mel, B. (2001). Impact of active dendrites and structural plasticity on the memory capacity of neural tissue. *Neuron*, *29*, 779–796.

Prager, R., & Fallside, F. (1989). The modified Kanerva model for automatic speech recognition. *Computer Speech and Language*, *3*, 61–81.

Pulvermüller, F. (2003). *The neuroscience of language: On brain circuits of words and serial order.* Cambridge: Cambridge University Press.

Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107.

Rachkovskij, D., & Kussul, E. (2001). Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Computation*, *13*, 411–452.

Rehn, M., & Sommer, F. (2006). Storing and restoring visual input with collaborative rank coding and associative memory. *Neurocomputing*, *69*, 1219–1223.

Rolls, E. (1996). A theory of hippocampal function in memory. *Hippocampus*, *6*, 601–620.

Schwenker, F., Sommer, F., & Palm, G. (1996). Iterative retrieval of sparsely coded associative memory patterns. *Neural Networks*, *9*, 445–455.

Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Softky, W., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *Journal of Neuroscience*, *13*(1), 334–350.

Sommer, F., & Dayan, P. (1998). Bayesian retrieval in associative memories with storage errors. *IEEE Transactions on Neural Networks*, *9*, 705–713.

Sommer, F., & Palm, G. (1999). Improved bidirectional retrieval of sparse patterns stored by Hebbian learning. *Neural Networks*, *12*, 281–297.

Steinbuch, K. (1961). Die Lernmatrix. *Kybernetik*, *1*, 36–45.

Stepanyants, A., Hof, P., & Chklovskii, D. (2002). Geometry and structural plasticity of synaptic connectivity. *Neuron*, *34*, 275–288.

Treves, A., & Rolls, E. (1991). What determines the capacity of autoassociative memories in the brain? *Network*, *2*, 371–397.

Tsodyks, M., & Feigel'man, M. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, *6*, 101–105.

Waydo, S., Kraskov, A., Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, *26*(40), 10232–10234.

Wennekers, T., & Palm, G. (1996). Controlling the speed of synfire chains. In C. Malsburg, W. Seelen, J. Vorbrüggen, & B. Sendhoff (Eds.), *Proceedings of the ICANN 1996* (pp. 451–456). Berlin: Springer-Verlag.

Willshaw, D., Buneman, O., & Longuet-Higgins, H. (1969). Non-holographic associative memory. *Nature*, *222*, 960–962.

Wilson, C. (2004). Basal ganglia. In G. Shepherd (Ed.), *The synaptic organization of the brain* (5th ed., pp. 361–413). New York: Oxford University Press.

Witte, S., Stier, H., & Cline, H. (1996). In vivo observations of timecourse and distribution of morphological dynamics in Xenopus retinotectal axon arbors. *Journal of Neurobiology*, *31*, 219–234.

Woolley, C. (1999). Structural plasticity of dendrites. In G. Stuart, N. Spruston, & M. Häusser (Eds.), *Dendrites* (pp. 339–364). New York: Oxford University Press.

---

Received August 9, 2007; accepted April 28, 2009.