Memory capacities for synaptic and structural plasticity

Andreas Knoblauch[0,1], Günther Palm[2], Friedrich T. Sommer[3]

[1]) Honda Research Institute Europe GmbH, Carl-Legien-Str.30, D-63073 Offenbach, Germany
Tel: ++49 - 69 - 89011-761; Fax: ++49 - 69 - 89011-749
email: andreas.knoblauch@honda-ri.de

[2]) Institut für Neuroinformatik, Fakultät für Ingenieurwissenschaften und Informatik, Universität Ulm,
Oberer Eselsberg, D-89069 Ulm, Germany
Tel: ++49 - 731 - 50-24151; Fax: ++49 - 731 - 50-24156
email: guenther.palm@uni-ulm.de

[3]) University of California at Berkeley, Redwood Center for Theoretical Neuroscience,
156 Stanley Hall, MC# 3220, Berkeley, CA 94720-3220, USA
Tel: ++1 - 510 - 642 - 7251; Fax: ++1 - 510 - 642 - 7206
email: fsommer@berkeley.edu

accepted for publication in Neural Computation
April 28th, 2009

Abstract: 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 to 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 different 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 to quantify 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 paper 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 non-constant 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.

Keywords: associative memory, distributed storage, Willshaw model, look-up-table, best match problem

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

[0]corresponding author

(Burks et al., 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 is not separated. For example, an associative memory can store a set of associations between pairs of (binary) patterns $\{(\mathbf{u}^\mu \to \mathbf{v}^\mu) : \mu = 1, ..., M\}$. Similar as in random access memory, a query pattern $\mathbf{u}^\mu$ entered in the associative memory can serve as address for accessing the associated pattern $\mathbf{v}^\mu$. However, the tasks performed by the two types of memory differ fundamentally. Random access is only defined 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 paper we will compare different implementations of associative memories: First, we will 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, mushroom body (Hebb, 1949; Braitenberg, 1978; Palm, 1982; Fransen and Lansner, 1998; Pulvermüller, 2003; Marr, 1971; Rolls, 1996; Kanerva, 1988; Marr, 1969; Albus, 1971; Laurent, 2002).

Second, we will 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 to understand the difference between structural and synaptic plasticity.

Third, we will 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 v^μ and the retrieved patterns $\hat{v}^\mu$ (Cover and Thomas, 1991; Shannon and 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 does in general 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/biological implementations of specific models of associative memory, different aspects of the storage substrate will be critical. Here we define and compare three different performance measures for associative memory models that deviate in how the required storage resources are taken into account.

1) 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

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

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 et al. (1969); Palm (1980); Amit et al. (1987b); Nadal (1991); Frolov and Murav'ev (1993); Palm and Sommer (1996). The network capacity balances computational benefits with the required degree of connectivity between circuit elements. Such a tradeoff is important in many contexts such as chip design or 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 the network capacity is bounded by the value $C = 1$ which marks the achievable optimum as 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 et al., 1989).

2) 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 $C^I = T/I(\mathbf{A})$.

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

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,b, 2005). Note, that for efficient implementation of associative memory a 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 the following. Note further, that for models with binary synapses, the information capacity is an upper bound of the network capacity: $C \leq C^I \leq 1$ (because the memory requirement of the most wasteful model cannot exceed one bit per contact).

3) We define the *synaptic capacity* C^S as the channel capacity of the associative memory normalized by the number of non-silent synapses

$$C^S \quad = \quad \frac{T(\mathbf{v}^1, \mathbf{v}^2, \ldots, \mathbf{v}^M; \hat{\mathbf{v}}^1, \hat{\mathbf{v}}^2, \ldots, \hat{\mathbf{v}}^M)}{\#\text{ non}-\text{silent synapses}} \text{ [bit/synapse]} , \tag{3}$$

where *non-silent synapses* are chemical synapses that actually transmit signals to the postsynaptic cell and have to be metabolically maintained.

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

Second, 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 et al., 2002; Poirazi and Mel, 2001; Fusi et al., 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 et al., 1996; Engert and Bonhoeffer, 1999; Lamprecht and LeDoux, 2004). Indeed, we have shown in further modeling studies (Knoblauch, 2009, 2006) 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 $r \ll mn$ non-silent synapses in a $m \times n$ 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 100×100 binary synapses and let 150 of the 10000 synapses have weight one. Then the network capacity of the static fully-connected net is simply $C = 1000/10000 = 0.1$ bit per binary synapse. However, the synaptic weight matrix $\mathbf{A}$ has only sparsely one-entries with a single synapse entropy of $I(150/10000) = 0.1124$ 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 non-silent 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 feed forward neural associative network with binary neurons and synapses first proposed by Steinbuch (1961), see section 2.2. The feed-forward "hetero-associative" Willshaw model can achieve a network capacity of $C = \ln 2 \approx 0.7$ bits per contact. The Willshaw model performs high compared to alternative neural implementations of associative memory with non-binary synapses and feedback network architectures which became very popular in the eighties (Hopfield, 1982, 1984; Hopfield and Tank, 1986; Hertz et al., 1991). The network capacity of the original (non-sparse) Hopfield model stays with 0.14 bits/contact (Amit et al., 1987a,b) far below the one for the Willshaw model (see Schwenker et al., 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 non-sparse patterns with even ratio between active and nonactive neurons. Using sparse patterns in the feed-forward Hopfield network with accordingly adjusted synaptic learning rule (Palm, 1991; Dayan and Willshaw, 1991; Palm and Sommer, 1996) increases the network capacity to $1/(2 \ln 2) \approx 0.72$ (Tsodyks and Feigel'man, 1988; Palm and Sommer, 1992). Thus, in terms of network capacity the sparse Hopfield model outperforms the Willshaw model, but only very marginally. The picture is similar in terms of synaptic capacity since the number of non-silent 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 and 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 and Palm, 1999), Improved threshold strategies (Buckingham and Willshaw, 1993; Graham and Willshaw, 1995), and retrieval with spiking neurons (Knoblauch and 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 remained so far unresolved. The first open question is the "sparsity problem" that is, the question 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. Fig. 3). In the standard model even small deviations from this sparseness condition reduce the 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 and Sommer, 2006) the sparse coding problem cannot be solved in general. The extreme sparsity requirement is not only problematic for applications (e.g., see Rachkovskij and 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 and Nirenberg, 2004). At least for sparsely connected networks realizing only a small given fraction P of the possible synapses, it is possible to achieve non-zero 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 and Willshaw, 1997; Bosch and Kurfess, 1998; Knoblauch, 2006).

The second open question concerning the Willshaw model is the "capacity gap" problem, that is, the question 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, i.e., 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 paper

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/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 defined in the introduction, 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 ultra-sparse 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 from less than 0.5 bit per synaptic contact to about 5-10 bit 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 section 6 we give in appendix A an overview on binary channels. Then appendix B reviews exact formulae for the analysis of the Willshaw models with fixed pattern activity which is used to verify the results of this paper and to 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 and Sommer, 1992; Bogacz et al., 2001). The second task is autoassociation or pattern completion, which involves to complete 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 $\{(\mathbf{u}^\mu \mapsto \mathbf{v}^\mu) : \mu = 1, ..., M\}$. 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 μ will be determined. The memory will return an *output pattern* $\hat{v}$ that is equal or similar to the stored content pattern $\mathbf{v}^\mu$. Note that autoassociation is a special case of heteroassociation (for $u^\mu = v^\mu$) 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, e.g. for cluster analysis, speech and object recognition, or information retrieval in large databases (Kohonen, 1977; Prager and Fallside, 1989; Greene et al., 1994; Mu et al., 2006; Rehn and Sommer, 2006).

Properties of memory patterns: In this paper we focus on the case of *binary pattern vectors*.

The address patterns have dimension m, 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. (1) 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. (2) 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^μ is one with probability k/m and zero otherwise, independently of other components. It turns out that the distinction between constant and random pattern activity is relevant only for address patterns, but 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.

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 λk "correct" and κk "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 expresses in deviations between retrieval output $\hat{\mathbf{v}}$ and the stored $\mathbf{v}$-patterns.

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

where $p_{10} := \mathrm{pr}[\hat{\mathbf{v}}_j^\mu = 0 | \mathbf{v}_j^\mu = 1]$ and $p_{01} := \mathrm{pr}[\hat{\mathbf{v}}_j^\mu = 1 | \mathbf{v}_j^\mu = 0]$ 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 et al., 1977; Anderson, 1993; Hopfield, 1984; Treves and Rolls, 1991; Sommer and Dayan, 1998). The patterns $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ describe the activity states of two populations of neurons at time

μ. In neural associative memories the associations are stored in the *synaptic matrix* or *memory matrix*.

Storage: In the Willshaw or Steinbuch model (Willshaw et al., 1969; Steinbuch, 1961; Palm, 1980, 1991) not only neurons but also synapses have binary values. The storage and retrieval processes work as follows. The pattern pairs are stored hetero-associatively in a binary memory matrix $\mathbf{A} \in \{0,1\}^{m \times n}$ (see Fig. 1), where

$$A_{ij} \quad = \quad \min\left(1, \sum_{\mu=1}^{M} u_i^\mu \cdot v_j^\mu\right) \in \{0,1\}\,. \tag{4}$$

The network architecture is feedforward, thus, an address population u consisting 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 auto-associative 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.

Figure 1 about here

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 $\hat{\mathbf{v}}$,

$$\hat{v}_j \quad = \quad \begin{cases} 1\,, & x_j = (\sum_{i=1}^{m} \tilde{u}_i A_{ij}) \geq \Theta \\ 0\,, & \text{otherwise} \end{cases}\,. \tag{5}$$

A critical prerequisite for high retrieval quality is the right choice of the threshold value Θ: Too low values will lead to high rates of add-errors whereas too high values 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 errorfree 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 *Willshaw threshold*. This threshold setting can be easily implemented in technical systems and is also biologically very plausible, for example based on feed-forward inhibition via "shadow" interneurons (cf. Knoblauch and Palm, 2001; Knoblauch, 2003b, 2005; Aviel et al., 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 $\tilde{u}$. 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 κ and λ.

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 extension of the Willshaw model by adding an additional intermediate layer of neurons w (Fig. 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 eq. 4). However, the two-layer model can be advantageous if address and content patterns are non-random or non-sparse 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 $\mathbf{u}^\mu \rightarrow \mathbf{v}^\mu$, 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 μ-th row contains the μ-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 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.

Figure 2 about here

3 Analysis of network capacity

3.1 Asymptotic analysis of network capacity

This paragraph 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 $p_1 := \sum_{ij} \mathbf{A}_{ij}/nm$ 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:

$$p_1 = 1 - \left(1 - \frac{kl}{mn}\right)^M, \tag{6}$$

It is often convenient to use (6) to determine the number of stored patterns

$$M = \frac{\ln(1-p_1)}{\ln(1-kl/mn)} \approx -\frac{mn}{kl}\ln(1-p_1), \tag{7}$$

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

The general analysis of retrieval includes queries $\tilde{u}$ 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 purposes of clarity, we will start with the analysis of pattern part retrieval where the address 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 $\Theta = |\tilde{u}| = \lambda k$ the probability of add noise in the retrieval is:

$$p_{01} \quad = \quad p(\hat{v}_i = 1 | v_i^\mu = 0) \gtrapprox {p_1}^{\lambda k}. \tag{8}$$

For exact formulae see eqs. 57-59 in appendix B. For random query activity see appendix D. The following analysis is based on the binomial approximation eq. 8 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 eq. 8 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 eq. 47). Storing Mn such components, the network capacity $C(k, l, m, n, \lambda, M)$ of eq. 1 is

$$C = \frac{M}{m} T\left(\frac{l}{n}, p_{01}, 0\right) \leq \frac{\mathrm{ld} p_{01} \ln(1-p_1)}{k} \tag{9}$$

$$\leq \lambda \cdot \mathrm{ld} p_1 \cdot \ln(1-p_1) \leq \lambda \ln 2 \tag{10}$$

where we used the bound eq. 48 and the binomial approximation eq. 8. 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 eq. 48), 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 actually 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 $p_{01\epsilon}$,

$$p_{01} \leq p_{01\epsilon} := \frac{\epsilon l}{n-l} \quad \Leftrightarrow \quad \lambda \geq \lambda_\epsilon :\approx \frac{\ln \frac{\epsilon l}{n-l}}{k \ln p_{1\epsilon}}, \tag{11}$$

where we call $\epsilon > 0$ the *fidelity parameter*. For the approximation of minimal address pattern fraction λ we used again the binomial approximation eq. 8. Note that for $p_{10} = 0$ and constant ϵ 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 ϵ 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: $\epsilon \leq \rho_3 \frac{1}{Ml}$.

By storing more and more patterns the matrix load $p_{1\epsilon}$ will increase and the noise level λ_ϵ that can be afforded in the address to achieve the specified retrieval quality will drop. Therefore, the maximum number of patterns that can be stored is reached at the point where λ_ϵ reaches the required fault tolerance: $\lambda_\epsilon = \lambda$ (eq. 11). Accordingly, the maximum matrix load (and the optimal activity of address patterns) is given by

$$p_{1\epsilon} \approx \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda \cdot k}} \qquad \left(\Leftrightarrow k \approx \frac{\mathrm{ld} \frac{\epsilon l}{n-l}}{\lambda \mathrm{ld} p_{1\epsilon}}\right), \tag{12}$$

Thus, with eqs. 7,9 we obtain the maximal number of stored patterns, the *pattern capacity* M_ϵ and the *network capacity* $C_\epsilon(k, l, m, n, \lambda, \epsilon) \approx M_\epsilon m^{-1} T(l/n, \epsilon l/(n-l), 0)$,

$$M_\epsilon \approx -\lambda^2 \cdot (\mathrm{ld} p_{1\epsilon})^2 \cdot \ln(1-p_{1\epsilon}) \cdot \frac{k}{l} \cdot \frac{mn}{(\mathrm{ld} \frac{n-l}{\epsilon \cdot l})^2}. \tag{13}$$

$$C_\epsilon \approx \lambda \cdot \mathrm{ld} p_{1\epsilon} \cdot \ln(1-p_{1\epsilon}) \cdot \eta \tag{14}$$

where

$$\eta := \frac{T\left(\frac{l}{n}, \frac{\epsilon l}{n-l}, 0\right)}{-\frac{l}{n}\mathrm{ld}\frac{\epsilon l}{n-l}} = \frac{T\left(\frac{l}{n}, \frac{\epsilon l}{n-l}, 0\right)}{I(\frac{l}{n})} \cdot \left(\frac{1}{1 + \frac{\ln \epsilon}{\ln(l/n)} - \frac{\ln(1-l/n)}{\ln(l/n)}} + \frac{(n-l)\mathrm{ld}(1-l/n)}{l\mathrm{ld}(\epsilon l/n)} \right) \quad (15)$$

$$\approx \frac{1}{1 + \frac{\ln \epsilon}{\ln(l/n)}}. \quad (16)$$

The approximation eq. 16 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 eq. 51 for details). For sparse content patterns with $l/n \ll 1$ we have $I(l/n) \approx -(l/n)\mathrm{ld}(l/n)$ (see eq. 43) and the right summand in the brackets can be neglected. Finally, the left summand in the brackets of eq. 15 becomes 1 for $\ln \epsilon / \ln(l/n) \to 0$.

The next two figures illustrate results of this analysis with an example, a Willshaw network with a square shaped memory matrix ($m = n$). The address and content patterns have same activity ($k = l$) and the input is noiseless, that is, $\lambda = 1$, $\kappa = 0$. Figure 3 presents results for a network with $n = 100000$ neurons, a number that corresponds roughly to the number of neurons below one square millimeter of cortex surface (Braitenberg and Schüz, 1991; Hellwig, 2000). Panel a) 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 as in the asymptotic calculation. Note that the number of patterns M_ϵ becomes maximal at smaller values $p_{1\epsilon} < 0.5$ ($M_\epsilon \approx 29.7 \cdot 10^6$ for $k = 8$ and $p_1 \approx 0.17$).

Figure 3 about here

Fig.3b explores the case where pattern activity is fixed to the value $k = 18$ that was optimal in panel a) 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 spoken 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 Figs. 3c-e give an overview how network capacity, memory load and the maximum number of stored patterns vary with pattern activity and fidelity level. High-quality retrieval with small ϵ require generally larger assembly size k. For fixed fidelity level ϵ, optimal k for maximal M is generally smaller than optimal k for maximal C (the latter has about double size; cf. Knoblauch et al., 2008).

3.3 Refined asymptotic analysis for large networks

The last section has delineated a theory for the Willshaw associative memory that predicts pattern capacity and network capacity for finite network sizes and for 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_ϵ. We focus on the case $k \sim l$ which applies for autoassociative memory tasks and for heteroassociative memory tasks if the activities of address and content patterns are similar. The results displayed in Fig. 4 can be compared to the predictions of the classical analysis recapitulated in section 3.1. Several important observations can be made:

- First, the upper bound of network capacity can in fact be reached by eq. 14 for arbitrary small constant ϵ, that is, at retrieval quality 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 eq. 12).

- Second, at retrieval quality grade RQ1, network capacity and pattern capacity assume their optima for somewhat different parameter settings. The pattern capacity M_ϵ (eq. 13) peaks at a memory load $p_{1\epsilon} \approx 0.16$ which also requires logarithmic sparseness in the memory patterns, however, 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 eq. 13).

- Third, the optimal bound of network capacity is only approached 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 towards 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., $\epsilon = 1$).

Figure 4 about here

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 eq. 14) which requires $\epsilon > l/n$, a condition which 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 and 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 eq. 8, which has recently been shown to be accurate for virtually any sub-linearly sparse patterns (see Knoblauch, 2007, 2008, see also appendix C for linearly sparse and non-sparse patterns). Furthermore, we are able to compute exact capacities even for small networks and, thus, to verify our asymptotic results (see appendices B,D and table 2). In contrast, many classical analyses, for example based on statistical physics (e.g., Tsodyks and Feigel'man, 1988; Golomb et al., 1990), become reliable only for very large networks, assume an infinite relaxation time, and apply only to auto-association with a recurrent symmetric weight matrix. However, some more recent attempts apply non-equilibrium methods for studying the behavior of recurrent neural networks with symmetric or asymmetric connections far from equilibrium and relaxation (for review see Coolen, 2001a,b). Alternative approaches based on signal-to-noise theory (e.g., Dayan and Willshaw, 1991; Palm and Sommer, 1996) are better suited for finite feed-forward 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 non-sparse (see appendix C). Before we proceed to compute synaptic capacity and information capacity for the Willshaw network, in the following 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 and Willshaw, 1992; Sommer and Palm, 1999). Our extended analysis can reveal the optimal values $p_{1\epsilon}$ for arbitrary parameter settings and it certainly suggests to avoid the regimes $p_{1\epsilon} \to 0$ or $p_{1\epsilon} \to 1$: Eqs. 10,14 and Fig. 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 holds:

$$C_\epsilon > 0 \Leftrightarrow k \sim \log n \Leftrightarrow 0 < p_{1\epsilon} < 1. \tag{17}$$

To see this, we can rewrite eq. 12 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 $l(n)$ (for example, $d = 1$ if l grows not faster than a polynomial in $\log n$). Then varying c one can obtain asymptotically for $p_{1\epsilon}$ all possible values in $(0; 1)$, and correspondingly for C_ϵ 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 17 hold.

Thus, non-zero C_ϵ 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 $C \to 0$. These cases of sparse and dense potentiation, that appear inefficient in the light of network capacity, we will reevaluate in the following using the performance measures of information capacity and synaptic capacity that we have introduced in section 1.2.

4 Analysis of information capacity and synaptic capacity

4.1 Information capacity

The information capacity (eq. 2) relates the stored (retrievable) information to the memory resources required by an 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 eq. 14 the information capacity C^I for optimal compression writes

$$C^I_\epsilon \;:=\; \frac{C_\epsilon}{I(p_{1\epsilon})} \;\approx\; \lambda \frac{\ln p_{1\epsilon} \ln(1 - p_{1\epsilon})}{-p_{1\epsilon} \ln p_{1\epsilon} - (1 - p_{1\epsilon}) \ln(1 - p_{1\epsilon})} \eta. \tag{18}$$

Eq. 18 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 to approach 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

[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 $I(p_1)$.

asymptotic information capacity of $\lambda \in (0;1]$ is possible for sparse and dense potentiation, i.e., $p_{1\epsilon} \to 0$ 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 Figs. 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.

Figure 5 about here

Figure 6 about here

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, 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 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 eq.14 this yields the information capacity for the Willshaw model with synaptic target list:

$$C_\epsilon^{I'} \quad := \quad \frac{C_\epsilon}{\min(p_{1\epsilon}, 1-p_{1\epsilon})\mathrm{ld}n} \quad \approx \quad \frac{\mathrm{ld}p_{1\epsilon} \cdot \ln(1-p_{1\epsilon})}{\min(p_{1\epsilon}, 1-p_{1\epsilon})\mathrm{ld}n}. \tag{19}$$

Fig. 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 be not 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 eq. 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

$$C_\epsilon^S \quad := \quad \frac{C_\epsilon}{\min(p_{1\epsilon}, 1-p_{1\epsilon})} = C_\epsilon^{I'}\mathrm{ld}n \quad \approx \quad \lambda\frac{\mathrm{ld}p_{1\epsilon} \cdot \ln(1-p_{1\epsilon})}{\min(p_{1\epsilon}, 1-p_{1\epsilon})}\eta \tag{20}$$

with η from eqs. 16,15. Note that C^S and $C^{I'}$ in eq. 19 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

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

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).

Fig. 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 ultra-sparse patterns, $k \to 2$, is remarkable.

For moderately sparse patterns and dense potentiation ($p_{1\epsilon} \to 1$) our analysis eq. 20 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 eq. 8 of C^S and C^I can become inaccurate for large cell assemblies (see appendices B,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 for this is that C^S is very sensitive to the compression factor $1 - p_{1\epsilon}$. Thus, even if the true value of M_ϵ is only a little bit smaller than suggested by eq. 13, then 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 ultra-sparse patterns with $p_{1\epsilon} \to 0$.

Figs. 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 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, i.e., $C^S \leq \mathrm{ld} n$. 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 formulae 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 $\epsilon \sim 1$.

Table 1 about here

4.3 Capacities for sparse synaptic potentiation

For sparse synaptic potentiation we have $p_{1\epsilon} \to 0$ and typically sub-logarithmic 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 eqs. 12, 7, 14, 18, 19, and 20 for large $m, n \to \infty$

$$M_\epsilon \approx \left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda k}} \frac{mn}{kl} \approx \epsilon^{\frac{1}{\lambda k}} \frac{m}{k} \left(\frac{n}{l}\right)^{1-\frac{1}{\lambda k}} \tag{21}$$

$$C_\epsilon \approx \frac{\left(\frac{\epsilon l}{n-l}\right)^{\frac{1}{\lambda k}} \mathrm{ld}\frac{\epsilon l}{n-l}}{k} \eta \to 0 \tag{22}$$

$$C^I_\epsilon \approx \lambda\eta \leq \lambda \tag{23}$$

$$C^{I'}_\epsilon \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{k\mathrm{ld} n} \cdot \eta \leq 1/k \tag{24}$$

$$C^S_\epsilon \approx \frac{\mathrm{ld}\frac{\epsilon l}{n-l}}{k} \cdot \eta \leq \frac{\mathrm{ld} n}{k} \tag{25}$$

The second approximation in eq. 21 is valid only for $l \ll n$. Thus, for sparse potentiation we can still store a very large number of ultra-sparse 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. Fig.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 non-zero 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 not 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 $C^S_\epsilon \sim \log n \to \infty$ for constant $k \sim 1$, which comes very close to the theoretical optimum $C^S \leq \mathrm{ld} n$ which is 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 λ (and consequently must also be independent of the high fidelity parameter ϵ). Thus, decreasing M from $M = M_\epsilon$ to $M = 1$ virtually does not affect neither C^S nor $C^{I'}$. 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 eq. 25 for $M = 1$.

4.4 Capacities for dense synaptic potentiation

For dense synaptic potentiation we have $p_{1\epsilon} \to 1$ and typically supra-logarithmic 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 eqs. 12,7,14, 18, 19, and 20 for large $n \to \infty$

$$1 - p_{1\epsilon} \approx \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k} \to 0 \tag{26}$$

$$M_\epsilon \approx \frac{mn}{kl}\left(\ln(\lambda k) - \ln\ln\frac{n-l}{\epsilon l}\right) \tag{27}$$

$$C_\epsilon \approx \left(\ln(\lambda k) - \ln\ln\frac{n-l}{\epsilon l}\right)\frac{\mathrm{ld}\frac{n-l}{\epsilon l}}{k} \cdot \eta \to 0 \tag{28}$$

$$C^I_\epsilon \approx \lambda\eta \leq \lambda \tag{29}$$

$$C^{I'}_\epsilon \approx \lambda \cdot \frac{\ln(\lambda k) - \ln\ln\frac{n-l}{\epsilon l}}{\ln n} \leq \lambda\frac{\ln k}{\ln n} \tag{30}$$

$$C^S_\epsilon \approx \lambda \cdot \mathrm{ld}(\lambda k) - \mathrm{ld}\ln\frac{n-l}{\epsilon l} \leq \lambda \ln n \tag{31}$$

Although the pattern capacity M_ϵ 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 $k \leq \sqrt{n}$ (see eq.27, 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 non-optimal compression: For $k = n^d$ and $0 < d < 1$ we obtain $C^{I'} \to \lambda d$ from eq.30. Similarly, synaptic capacity achieve its upper bound, $C^S \leq \mathrm{ld} n$, 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 $k \geq 2$ (see eqs. 24,25). 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 eq. 27, cf. table 1; see also section 5).

For dense synaptic potentiation both $C^{I'}$ and C^S depend on the fault tolerance requirement λ and the high fidelity parameter ϵ, unlike to sparse synaptic potentiation where these capacities are independent from λ. 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 counter-intuitive fact that the amount of necessary physical memory increases with decreasing number of stored patterns M.

As can be seen in Fig. 5a, both information capacities C^I and $C^{I'}$ and synaptic capacity C^S exhibit local maxima at k^I_{opt} and k^S_{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

$$k^S_{\rm opt} \sim n \cdot (e^{\sqrt{-\ln \epsilon}})^{-\sqrt{\ln n}} \tag{32}$$

$$k^I_{\rm opt} \sim n^{1-\frac{-\ln\epsilon-\sqrt{-\ln\epsilon}}{-\ln\epsilon-1}}. \tag{33}$$

Note that $k^S_{\rm opt}$ 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 eq. 8 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 $M = 0.14n$ non-sparse patterns resulting in 0.14 bits per (non-binary) synapse (Hopfield, 1982; Amit et al., 1987a,b). Thus, for non-sparse 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 affects increasing noise $(1-\lambda,\kappa)$ in the query patterns $\tilde{\mathbf{u}}$ the number of storable patterns M_ϵ and the other capacity measures (C_ϵ, C^I_ϵ, C^S_ϵ) for 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 eqs. 7 and 12 we can introduce the fraction of storable patterns as a function of the query noise λ,

$$m_\lambda := \frac{M_\epsilon(\lambda)}{M_\epsilon(1)} \approx \frac{\ln(1-p_{1\epsilon}(\lambda))}{\ln(1-p_{1\epsilon}(1))} \in (0;1] \begin{cases} \approx p_{1\epsilon}(1)^{(1-\lambda)/\lambda} \to 0\,, & p_{1\epsilon}(1) \to 0 \\ \to 1 \quad\quad\quad\quad\;\; , & p_{1\epsilon}(1) \to 1 \end{cases} \tag{34}$$

where we used $\ln(1-p_{1\epsilon}) \approx -p_{1\epsilon}$ for $p_{1\epsilon} \to 0$ and de l'Hospital'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 3.4,4.3,4.4): Fig. 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. Tab. 1). The same conclusion is true for network capacity, $C_\epsilon(\lambda) := m_\lambda \cdot C_\epsilon(1)$ (see eqs. 9,14).

Figure 7 about here

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

$$i_\lambda := \frac{I(p_{1\epsilon}(\lambda))}{I(p_{1\epsilon}(1))} \begin{cases} \approx p_{1\epsilon}(1)^{((1-\lambda)/\lambda)}/\lambda \to 0\,, & p_{1\epsilon}(1) \to 0 \\ \to 1/\lambda \quad\quad\quad\;\; , & p_{1\epsilon}(1) \to 1 \end{cases}, \tag{35}$$

where we used $I(p_1) \approx -p_1 \mathrm{ld} p_1$ for $p_{1\epsilon}(1) \to 0$ and de l'Hospital's rule for $p_{1\epsilon}(1) \to 1$ (cf. Knoblauch, 2003b). The relative compressibility is depicted in Fig.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., supra-logarithmically sparse address patterns (see Table 1) and implies the counter-intuitive fact that although fewer patterns are stored *more* physical memory is required. Thus, the dependence of information capacity on miss-noise is

$$c^I_\lambda := \frac{C^I_\epsilon(\lambda)}{C^I_\epsilon(1)} = \frac{m_\lambda}{i_\lambda} = \lambda + f(\lambda, p_{1\epsilon}(1)) \approx \lambda, \tag{36}$$

[3] Note that there is a 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.

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_λ in Fig.7c reveals the surprising result that the relative decrease in information capacity is almost linear in λ 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 Fig.7d).

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

$$c^S_\lambda \;:=\; \frac{C^S_\epsilon(\lambda)}{C^S_\epsilon(1)} = \frac{m_\lambda \min(p_{1\epsilon}(1), 1 - p_{1\epsilon}(1))}{\min(p_{1\epsilon}(\lambda), 1 - p_{1\epsilon}(\lambda))} \approx \begin{cases} C^S_\epsilon(1)\;, & p_{1\epsilon}(1) \to 0 \\ \lambda C^S_\epsilon(1)\,, & p_{1\epsilon}(1) \to 1 \end{cases}\;, \tag{37}$$

It is remarkable that C^S is independent of λ for ultra-sparse patterns with $k/\log n \to 0$ and sparse potentiation $p_{1\epsilon} \to 0$. Thus, decreasing M from $M = M_\epsilon(1)$ to $M = M_\epsilon(\lambda)$ does neither affect 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 eq. 25. 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 and Sommer, 1996; Sommer and 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 eqs. 52,53). 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 $z := |\tilde{u}|$. 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 Θ. Here ϵ^T corresponds to the normalized information loss between retrieved and originally stored patterns, but using the Hamming distance based measure ϵ as defined in section 3.2 leads qualitatively to the same results (see Knoblauch et al., 2008). Figure 8 shows for each noise level the retrieval quality (panel (a)) and the optimal threshold (panel (b)).

Figure 8 about here

These numerical experiments validate our theoretical results for pattern part retrieval (without add-noise). For $\lambda < 1$ ultrasparse patterns (e.g., constant $k = 4$) appears to be very vulnerable to miss-noise, i.e., ϵ increases very steeply with decreasing λ. In contrast, moderately sparse patterns (e.g., $k = 1000$ for $n = 10000$) are much more robust against miss-noise, i.e., the increase of ϵ is much weaker. On the other hand, our data also show that ultra-sparse cell assemblies are very robust against add-noise, i.e., the fidelity parameter ϵ increases only relatively slowly with increasing add-noise level κ. In contrast, the large cell assemblies are quite vulnerable to add-noise: Here ϵ increases very steeply with κ. Our results show that the attractors around memories $\mathbf{u}^\mu$ (i.e., the subspace of query patterns $\tilde{\mathbf{u}}$ that map to $\mathbf{u}^\mu$) have only little similarity to spheres in Hamming space. Rather, for ultra sparse patterns ($k/\log n \to 0$) attractors are elongated towards query patterns with more add-noise than miss-noise, whereas for moderately sparse patterns ($k/\log n \to \infty$) attractors are elongated towards query patterns with more miss-noise than add-noise.

Figure 8b illustrates another important difference between sparse and dense synaptic potentiation corresponding to ultra-sparse or moderately-sparse activity. For ultra-sparse patterns, the optimal threshold depends mainly on λ, but only very weakly on κ. In contrast, for moderately-sparse patterns, the optimal threshold has a strong dependence both on λ and κ. As a consequence, in particular for biological systems it may be much easier to implement the optimal threshold for retrieving ultra-sparse 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 ultra-sparse cell-assemblies. Actually, a reasonable interpretation of seemingly random or spontaneous ongoing activity (Arieli et al., 1996; Softky and Koch, 1993) would be that a large number of small cell assemblies or synfire chains (Abeles, 1982; Abeles et al., 1993; Diesmann et al., 1999; Wennekers and Palm, 1996) are active at the

same time independently of each other.

5 Computational complexity and energy requirements

5.1 Compressed and uncompressed Willshaw network

So far we were 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 a 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 afterwards apply a threshold on each component of $\mathbf{x}$, i.e., the retrieval time (or number of retrieval steps) is

$$t_{\text{seq}}^{\text{W}} \quad = \quad z \cdot n + n \approx zn \tag{38}$$

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

$$t_{\text{seq}}^{\text{cW}} \quad = \quad c \cdot z \cdot n \cdot \min(p_1, 1-p_1), \tag{39}$$

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, $t_{\text{seq}}^{\text{cW}}/t_{\text{seq}}^{\text{W}} \to 0$ 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 a 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 $\mu = 1, ..., M$, i.e., each matrix row corresponds to a pattern vector. We also refer to the look-up table as *grandmother cell model* (or briefly grandmother model, cf. Knoblauch, 2005; Barlow, 1972) because its biological interpretation actually 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 μ receives inputs via synapses corresponding to the μ-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

$$t_{\text{seq}}^{\text{nLUT}} \quad = \quad M \cdot (z + k). \tag{40}$$

Then for $M/n \to \infty$ we have indeed $t_{\text{seq}}^{\text{nLUT}}/t_{\text{seq}}^{\text{W}} \geq M/n \to \infty$. Thus, the Willshaw model is indeed 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

$$t_{\text{seq}}^{\text{cLUT}} \quad = \quad c \cdot z \cdot M \cdot k/m \tag{41}$$

steps. Comparing with the compressed Willshaw model this yields

$$\nu \quad := \quad \frac{{t_{\text{seq}}}^{\text{cLUT}}}{{t_{\text{seq}}}^{\text{cW}}} \approx \frac{-\ln(1-p_1)}{l \min(p_1, 1-p_1)} \leq \frac{-\ln(1-p_{1\epsilon})}{l \min(p_{1\epsilon}, 1-p_{1\epsilon})} \to \begin{cases} 1/l & , \quad p_{1\epsilon} \to 0 \\ \lambda \frac{k}{l} \frac{\ln(\lambda k) - \ln \ln \frac{n}{\epsilon l}}{\ln \frac{n}{\epsilon l}} & , \quad p_{1\epsilon} \to 1, \end{cases} \tag{42}$$

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 auto-association) 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 network 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 $\mathbf{U}$. 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 and 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 non-silent 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 and Palm, 1991; Hammerstrom, 1990; Hammerstrom et al., 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. Non-sparse 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 they are 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 and Fallside, 1989; Greene et al., 1994; Knoblauch, 2005; Mu et al., 2006; Rehn and Sommer, 2006).

In this paper 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; Amit et al., 1987a,b; Palm, 1991; Nadal, 1991; Buckingham and Willshaw, 1992; Sommer and Palm, 1999; Bosch and 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 understand 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 crosstalk can be predicted to set in if the stored patterns deviate from the principal components of the data which will be necessarily the case if the number of stored patterns exceeds the dimension of the patterns. The Willshaw model is a feed-forward 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 crosstalk 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 had revealed that the 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 between the three definitions of storage capacity, network, 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 already 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 Fig. 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 ultra-sparse patterns $k/\log n \to 0$ or for moderately sparse patterns $k/\log n \to \infty$ (see Perez-Orive et al., 2002; Hahnloser et al., 2002; Quiroga et al., 2005; Waydo et al., 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 Fig. 8) the effects strongly varied with error type in the ultra-sparse and moderately sparse regime. Specifically, the retrieval of ultra-sparse 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 ultra-sparse regime could be of particular interest if a memory has to be recognized in superpositions of many patterns whereas the moderately sparse regime allows to complete a memory pattern already 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 and 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 and Palm, 1991; Knoblauch, 2003b; Chicca et al., 2003), we have shown in section 5 that the compressed LUT network implemented on a sequential architecture outperforms the Willshaw network for almost all parameters (see eq. 42). Asymptotically, sequential implementations of the single layer Willshaw model remain superior only for almost non-sparse 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 ultra-sparse 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.

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

$$I(p) \;:=\; -p\cdot \mathrm{ld}p - (1-p)\cdot \mathrm{ld}(1-p) \approx \begin{cases} -p\cdot \mathrm{ld}p & , \quad p \ll 0.5 \\ -(1-p)\cdot \mathrm{ld}(1-p) & , \quad 1-p \ll 0.5 \end{cases} \tag{43}$$

(Shannon and Weaver, 1949; Cover and 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

$$\begin{aligned} I(Y) &= I_Y(p,p_{01},p_{10}) := I\left(p\,(1-p_{10}) + (1-p)\,p_{01}\right) && (44)\\ I(Y|X) &= I_{Y|X}(p,p_{01},p_{10}) := p\cdot I(p_{10}) + (1-p)\cdot I(p_{01}) && (45)\\ T(X;Y) &= T(p,p_{01},p_{10}) := I_Y(p,p_{01},p_{10}) - I_{Y|X}(p,p_{01},p_{10}). && (46) \end{aligned}$$

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

$$\begin{aligned} T(p,p_{01},0) &= I\left(p + p_{01} - pp_{01}\right) - (1-p)\cdot I(p_{01}) && (47)\\ &\le I(p_{01}) + I'(p_{01})\cdot(p(1-p_{01})) - (1-p)\cdot I(p_{01}) = -p\mathrm{ld}p_{01} && (48) \end{aligned}$$

For the upper bound we have linearized I in p_{01} and used the convexity of $I(p)$, i.e., $(dI/dp)^2 = -1/(p(1-p)\ln 2) < 0$. 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,

$$T(p_1,p_{01},0) \;\le\; I(p) + I'(p)\cdot(1-p)\cdot p_{01} - (1-p)\cdot I(p_{01}) \le I(p), \tag{49}$$

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

$$e_I \;:=\; \frac{I(p_1) - T(p_1,p_{01},p_{10})}{I(p_1)} \approx (1-p_1)\frac{I(p_{01}) - I'(p_1)\cdot p_{01}}{I(p_1)} \approx \frac{I(p_{01})}{I(p_1)} - \frac{p_{01}}{p_1}. \tag{50}$$

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 $I'(p) \approx -\mathrm{ld}p$. Thus we obtain

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

$$e_I \;\le\; \frac{I(\epsilon\frac{l}{n})}{I(\frac{l}{n})} - \epsilon \approx \epsilon\cdot\frac{\mathrm{ld}\epsilon}{\mathrm{ld}(\frac{l}{n})} \approx \frac{I(\epsilon)}{-\mathrm{ld}(\frac{l}{n})} \le \begin{cases} I(\epsilon)\,, & \text{in any case} \\ \epsilon\;, & \mathrm{l/n} \le \epsilon \end{cases}. \tag{51}$$

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 eq. 15 is smaller than one percent.

B Exact retrieval error probabilities for fixed query activity

Our analysis so far used the binomial approximation eq. 8. Here we give the exact expressions for *fixed* query pattern activity, i.e., when the query pattern $\tilde{u}$ has exactly $c := \lambda k$ correct one-entries from one of the address patterns u^μ and additionally $f := \kappa k$ false one-entries ($0 < \lambda \leq 1$, $\kappa \geq 0$). Retrieving with threshold Θ, 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

$$p_{01}(\Theta) = \sum_{x=\Theta}^{c+f} p_{\mathrm{WP}}(x; k, l, m, n, M-1, c+f) \tag{52}$$

$$p_{10}(\Theta) = \sum_{x=c}^{\Theta-1} p_{\mathrm{WP}}(x-c; k, l, m, n, M-1, f) \; . \tag{53}$$

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 ($0 \leq x \leq z$). It is

$$p_{\mathrm{WP}}(x; k, l, m, n, M, z) = \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} (1 - \frac{l}{n}(1 - B(m, k, s+z-x)))^M \tag{54}$$

$$\approx \binom{z}{x} \sum_{s=0}^{x} (-1)^s \binom{x}{s} (1 - \frac{l}{n}(1 - (1 - \frac{k}{m})^{s+z-x}))^M \tag{55}$$

$$= \sum_{i=0}^{M} p_B(i; M, l/n) p_B(x; z, 1 - (1-k/m)^i) \tag{56}$$

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 $p_B(x; N, P) := \binom{N}{x} P^x (1-P)^{N-x}$. Eq. 54 is exact for *fixed* address pattern activity, i.e., if each address pattern u^μ 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 eq. 55,56 would be exact for *random* address pattern activity, i.e., if u_i^μ is one with probability k/m (but still fixed c, f). Eq. 56 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); Buckingham (1991). The transformation to eq. 55 has been found by Sommer and Palm (1999). Eqs. 54,55 are numerically efficient to evaluate for low query pattern activity $c + f$, whereas eq. 56 is efficient for few stored patterns M. The distinction between fixed and random *address* pattern activity, $|u^\mu|$, is of minor interest already for moderately large networks, because then eqs. 54-56 yield very similar values Knoblauch (2006, 2008). However, the distinction between fixed and random query pattern activity, $|\tilde{u}|$, 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

$$p_{01} = \sum_{s=0}^{\lambda k} (-1)^s \binom{\lambda k}{s} [1 - \frac{l}{n}(1 - B(m, k, s))]^{M-1} \tag{57}$$

$$\approx \sum_{s=0}^{\lambda k} (-1)^s \binom{\lambda k}{s} [1 - \frac{l}{n}(1 - (1-k/m)^s)]^{M-1} \tag{58}$$

$$= \sum_{i=0}^{M-1} p_B(i; M-1, l/n)(1 - (1-k/m)^i)^{\lambda k} \tag{59}$$

$$\geq p_1^{\lambda k} \; . \tag{60}$$

Here eqs. 57-59 correspond to eqs. 54-56, and the bound corresponds to the binomial approximation eq. 8. Knoblauch (2007, 2008) show 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 eq. 8 becomes exact for virtually any sub-linear $k(n)$.

We have validated these results in extensive numerical experiments which can be found in Knoblauch (2006, 2008); Knoblauch et al. (2008). Table 2 shows some exact results when addressing with half patterns ($\lambda = 0.5$, $\kappa = 0$). Fig. 9 plots the quality of the binomial approximation eq. 8 for pattern capacity M and information capacity C^I for different sparsity levels and increasing network size $n \to \infty$.

Table 2 about here

Figure 9 about here

C Fallacies for extremely sparse and non-sparse activity

As discussed in section 3.3 our analysis method is exact, both for 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 feed-forward 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 $\hat{v}$) and "low-units" (which should be inactive) divided by the potential variance (Dayan and Willshaw, 1991; Palm, 1991; Palm and 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,

$$\mathrm{SNR}_{\mathrm{Willshaw}} \approx \frac{(\lambda k(1-p_1))^2}{\lambda k p_1(1-p_1)} = \frac{\lambda k(1-p_1)}{p_1} \tag{61}$$

The SNR for the network with linear learning and the optimal covariance rule has been found to be $m/(M(l/n)(1-l/n))$ (Dayan and Willshaw, 1991; Palm and Sommer, 1996). Using M as in eq. 7 and assuming small $p_1 \to 0$ this becomes

$$\mathrm{SNR}_{\mathrm{Cov}} \approx \frac{mkl}{-mn(l/n)\ln(1-p_1)} = \frac{k}{-\ln(1-p_1)} \tag{62}$$

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 $p_{1\epsilon} \to 0$) eq. 17 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 and 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 eq. 59. Moreover, we can compute upper and lower bounds for eq. 59 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),

$$p_{01} \approx (1-(1-c)^{Md+\xi\sqrt{Md(1-d)}})^{\lambda cm} \tag{63}$$

For sufficiently large positive (but for RQ1 still constant) ξ 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 few content neurons have a unit usage more than ξ standard deviations larger than the mean unit usage (here $G^c(x) := 0.5\text{erfc}(x/\sqrt{2})$ is the Gaussian tail integral). Similarly, for large negative ξ we obtain a lower bound. Requiring $p_{01} \leq \epsilon d/(1-d)$ we obtain for the pattern capacity

$$M_\epsilon + \xi\sqrt{M_\epsilon(1-d)/d} \approx \frac{\ln(1-(\epsilon d/(1-d))^{1/(\lambda cm)})}{d\ln(1-c)} \approx \frac{\ln m}{-d\ln(1-c)} \ . \tag{64}$$

Thus, the pattern capacity is essentially independent of ξ. However, compared to eqs. 13,27 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 eq. 6) it is $p_{0\epsilon} := 1-p_{1\epsilon} = \tilde{p}_{0\epsilon}^f$ (see eq. 12). Therefore eq. 18 becomes $C_\epsilon^I \sim \text{ld}(1-\tilde{p}_{0\epsilon})(\ln \tilde{p}_{0\epsilon})/(\tilde{p}_{0\epsilon}^f \text{ld}\tilde{p}_{0\epsilon}^f) \sim \tilde{p}_{0\epsilon}^{1-f} \to 0$. Similarly, eq. 20 becomes $C_\epsilon^S \sim \text{ld}(1-\tilde{p}_{0\epsilon})(\ln \tilde{p}_{0\epsilon})/\tilde{p}_{0\epsilon}^f \approx \tilde{p}_{0\epsilon}^{1-f} \ln \tilde{p}_{0\epsilon} \to 0$.

D Corrections for random query activity

So far, our exact theory in appendix B as well as the approximative theory in sections 3-5 assume that the query pattern $\tilde{u}$ has exactly λk correct one-entries (and κk 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 λ 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, i.e., each pattern component u_i^μ is one with probability k/m independent of other components. Similarly, in a query pattern $\tilde{u}$ 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, $\text{pr}[Z = z] = p_B(z; m, \lambda k/m)$ (for p_B see below eq. 56). For a particular $Z = z$ the exact error probability p_{01} is given by eq. 58 (or eq. 59) replacing λk by z. Averaging over all possible z yields

$$\begin{aligned}
p_{01}^* &= \sum_{z=0}^{m} p_B(z; m, \lambda k/m) \sum_{s=0}^{z} (-1)^s \binom{z}{s} [1 - \frac{l}{n}(1-(1-k/m)^s)]^{M-1} \\
&= \sum_{s=0}^{m} (-\frac{\lambda k}{m})^s \binom{m}{s} [1 - \frac{l}{n}(1-(1-k/m)^s)]^{M-1} \qquad (65) \\
&= \sum_{i=0}^{M-1} p_B(i; M-1, l/n) \sum_{z=0}^{m} p_B(z; m, \lambda k/m)(1-(1-k/m)^i)^z \\
&= \sum_{i=0}^{M-1} p_B(i; M-1, l/n)(1 - \frac{\lambda k}{m}(1-k/m)^i)^m \qquad (66)
\end{aligned}$$

The first equation is numerically efficient for small k, the last equation for small M. For the binomial approximative analyses we can rewrite eq. 8 as

$$p_{01}^* \approx \sum_{z=0}^{m} p_B(z; m, \lambda k/m) p_1^z = (1 - \lambda \frac{k}{m}(1-p_1))^m \tag{67}$$

Controlling for retrieval quality, $p_{01}^* \leq \epsilon l/(n-l)$, the maximal memory load eq. 12 becomes

$$p_{1\epsilon}^* \approx 1 - \frac{1 - (\frac{\epsilon l}{n-l})^{1/m}}{\lambda k/m} \ . \tag{68}$$

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 eq. 12 and eq. 68,

$$\begin{aligned}
\Delta p_{1\epsilon} &\approx e^{(\ln x)/(\lambda k)} - (1 + \frac{e^{(\ln x)/m} - 1}{\lambda k/m}) && (69)\\
&= \sum_{i=1}^{\infty} \frac{(\ln x)^i}{i!(\lambda k)^i} - \frac{(\ln x)^i}{i!\lambda k m^{i-1}} && (70)\\
&= \sum_{i=2}^{\infty} \frac{(\ln x)^i}{i!(\lambda k)^i}(1 - (\lambda k/m)^{i-1}) && (71)\\
&\approx p_{1\epsilon} - 1 - \ln p_{1\epsilon} && (72)
\end{aligned}$$

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, i.e., $\Delta p_{1\epsilon}/p_{1\epsilon} \to 0$ 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

$$\begin{aligned}
p^*_{1\epsilon} &\approx 1 + \ln p_{1\epsilon} \approx 1 - \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k} && (73)\\
p^*_{0\epsilon} &:= 1 - p^*_{1\epsilon} = \frac{\ln \frac{n-l}{\epsilon l}}{\lambda k} \qquad \left(\Leftrightarrow k \approx \frac{\ln \frac{n-l}{\epsilon l}}{\lambda p^*_{0\epsilon}} \right) && (74)\\
M^*_\epsilon &= -\frac{mn}{kl} \ln p^*_{0\epsilon} \approx -\lambda^2 p^2_{0\epsilon} \ln p_{0\epsilon} \frac{k}{l} \frac{mn}{(\ln \frac{n-l}{\epsilon l})^2} && (75)\\
C^*_\epsilon &= M_\epsilon m^{-1} T(l/n, \epsilon l/(n-l), 0) \approx -\lambda p^*_{0\epsilon} \mathrm{ld} p^*_{0\epsilon} \eta && (76)
\end{aligned}$$

The asymptotic bound of network capacity is thus only $C^*_\epsilon \leq 1/(e \ln 2) \approx 0.53$ 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 Fig. 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 $C^* \leq 0.53$ contradicts previous analyses. For example, Nadal (1991) estimates $C^* \leq 0.236$ for $p_1 = 0.389$. We believe that our results are correct and that the discrepancies are due to inaccurate approximations employed by previous works. In fact, we have verified the accuracy of our theory in two steps (see Knoblauch, 2006, 2008; Knoblauch et al., 2008): First, we have verified all our formulae for the exact error probabilities of the different model variants (eqs. 52-59,65-66) by extensive simulations of small networks. Second, we have proven the asymptotic correctness of our binomial approximative theory (see eqs. 8,12-14,73-76) by theoretical considerations and numerical experiments (see also Fig. 10).

Figure 10 about here

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. FTS was supported by NSF grant IIS-0713657 and a Google research award.

References

Abeles, M. (1982). *Local cortical circuits.* Springer, Berlin Heidelberg New York.

Abeles, M., Bergman, H., Margalit, E., and 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., and Sompolinsky, H. (1987a). Information storage in neural networks with low levels of activity. *Phys. Rev. A*, 35:2293–2303.

Amit, D., Gutfreund, H., and 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 Hassoun, M., editor, *Associative Neural Memories.* Oxford University Press, New York.

Anderson, J., Silverstein, J., Ritz, S., and 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., and Aertsen, A. (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science*, 273:1868–1871.

Attwell, D. and 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., and 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., and Palm, G. (1989). Information storage and effective data retrieval in sparse matrices. *Neural Networks*, 2:289–293.

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

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

Bosch, H. and 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 Heim, R. and Palm, G., editors, *Lecture notes in biomathematics (21). Theoretical approaches to complex systems.*, pages 171–188. Springer-Verlag, Berlin Heidelberg New York.

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

Buckingham, J. (1991). Delicate nets, faint recollections: a study of partially connected associative network memories. *PhD thesis, University of Edinburgh.*

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

Buckingham, J. and 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., and von Neumann, J. (1946). Preliminary discussion of the logical design of an electronic computing instrument. Report 1946, U.S. Army Ordonance Department.

Chicca, E., Badoni, D., Dante, V., D'Andreagiovanni, M., Salina, G., Carota, L., Fusi, S., and Del Giudice, P. (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 Moss, F. and Gielen, S., editors, *Handbook of Biological Physics*, volume 4, pages 531–596. Elsevier Science, Amsterdam, Netherlands.

Coolen, A. (2001b). Statistical mechanics of recurrent neural networks II: Dynamics. In Moss, F. and Gielen, S., editors, *Handbook of Biological Physics*, volume 4, pages 597–662. Elsevier Science, Amsterdam, Netherlands.

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

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

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

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

Fransen, E. and 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. and 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., and Abbott, L. (2005). Cascade models of synaptically stored memories. *Neuron*, 45:599–611.

Gardner, E. and 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., and 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. and Willshaw, D. (1995). Improving recall from an associative memory. *Biological Cybernetics*, 72:337–346.

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

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

Hahnloser, R., Kozhevnikov, A., and 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*, pages II:537–543. IEEE Press.

Hammerstrom, D., Gao, C., Zhu, S., and Butts, M. (2006). FPGA implementation of very large associative memories. In Omondi, A. and Rajapakse, J., editors, *FPGA implementations of neural networks*, pages 167–195. Springer US.

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

Heittmann, A. and 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. and Opper, M. (1990). Distribution of internal fields and dynamics of neural networks. *Europhysics Letters*, 11(5):403–408.

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

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. and 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.* MIT Press, Cambridge, MA.

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

Knoblauch, A. (2003b). Synchronization and pattern separation in spiking associative memory and visual cortical areas. *PhD thesis, Department of Neural Information Processing, 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 Report 06-02, Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany.

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

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 Mayor, J., Ruh, N., and Plunkett, K., editors, *Connectionist Models of Behavior and Cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop*. World Scientific Publishing.

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

Knoblauch, A., Palm, G., and Sommer, F. (2008). Performance characteristics of sparsely and densely potentiated associative networks. HRI-EU Report 08-02, Honda Research Institute Europe GmbH, D-63073 Offenbach/Main, Germany.

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

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

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

Laughlin, S. and 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. and Papert, S. (1969). *Perceptrons: An introduction to computational geometry.* MIT Press, Cambridge, MA.

Mu, X., Artiklar, M., Watta, P., and 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. and 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.* Springer, Berlin.

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

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

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

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

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

Prager, R. and 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 University Press, Cambridge, UK.

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

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

Rehn, M. and 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., and Palm, G. (1996). Iterative retrieval of sparsely coded associative memory patterns. *Neural Networks*, 9:445–455.

Shannon, C. and Weaver, W. (1949). *The mathematical theory of communication.* University of Illinois Press, Urbana/Chicago.

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

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

Sommer, F. and 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., and Chklovskii, D. (2002). Geometry and structural plasticity of synaptic connectivity. *Neuron*, 34:275–288.

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

Tsodyks, M. and 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., and Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience*, 26(40):10232–10234.

Wennekers, T. and Palm, G. (1996). Controlling the speed of synfire chains. In Malsburg, C., Seelen, W., Vorbrüggen, J., and Sendhoff, B., editors, *Proceedings of the ICANN 1996*, pages 451–456, Berlin, Heidelberg, New York. Springer Verlag.

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

Wilson, C. (2004). Basal ganglia. In Shepherd, G., editor, *The Synaptic Organization of the Brain (5th edition)*, chapter 9, pages 361–413. Oxford University Press, New York.

Witte, S., Stier, H., and 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 Stuart, G., Spruston, N., and Häusser, M., editors, *Dendrites.*, pages 339–364. Oxford University Press, Oxford, UK.

k	$p_{1\epsilon}$	M_ϵ	C_ϵ	C_ϵ^I	$C_\epsilon^{I'}$	C_ϵ^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	$0.5n \ln n$	0	1	0.5	$0.5\mathrm{ld}n \to \infty$
$cn^d \ , 0 < d < 1$	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

Table 1: Asymptotic results for hifi memory load $p_{1\epsilon}$, storable patterns M_ϵ, and network capacity C_ϵ, information capacities ${C^I}_\epsilon$ for optimal compression and ${C^{I'}}_\epsilon$ for simple target lists, and synaptic capacity ${C^S}_\epsilon$. 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.

$n =$	100	200	500	1000	2000	5000	10000	20000	50000	100000
$k = 4$	4	4	4	4	4	4	4	4	4	4
M_ϵ	7	23	102	315	951	3985	11614	33561	135216	386157
C	0.016734	0.016080	0.013581	0.011749	0.009820	0.007427	0.005876	0.004581	0.003239	0.002467
C_ϵ^I	0.189510	0.213911	0.239855	0.257522	0.272803	0.289919	0.301034	0.310883	0.322324	0.330003
C_ϵ^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_ϵ	26	73	530	1578	6825	31481	130517	410162	2239454	8958499
C_ϵ	0.093667	0.087255	0.136318	0.126214	0.166369	0.152057	0.185759	0.169994	0.185909	0.211443
C_ϵ^I	0.177045	0.174203	0.216708	0.210461	0.239663	0.234620	0.258792	0.248317	0.254089	0.272940
C_ϵ^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_ϵ	20	67	294	791	2122	7082	17013	40294	119800	271628
C_ϵ	0.092180	0.120686	0.150986	0.159572	0.162795	0.150634	0.136076	0.120799	0.098333	0.082962
C_ϵ^I	0.134642	0.140982	0.152895	0.160997	0.175765	0.189566	0.198546	0.211598	0.224751	0.235512
C_ϵ^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_ϵ	11	27	76	156	310	736	1371	2509	5454	9662
C_ϵ	0.081660	0.086933	0.081497	0.071901	0.061067	0.046564	0.036626	0.028186	0.019377	0.014325
C_ϵ^I	0.083180	0.087490	0.092952	0.097348	0.104483	0.114870	0.124077	0.134520	0.149156	0.160552
C_ϵ^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_ϵ	10	16	24	31	39	49	56	64	74	82
C_ϵ	0.079104	0.063284	0.037970	0.024522	0.015425	0.007752	0.004430	0.002531	0.001171	0.000649
C_ϵ^I	0.079241	0.067368	0.050885	0.042898	0.038121	0.030659	0.024775	0.021308	0.016677	0.014209
C_ϵ^S	0.166347	0.177726	0.178703	0.181323	0.191142	0.183161	0.164436	0.157467	0.138863	0.128935

Table 2: Exact capacities of the Willshaw model computed from eq. 57 for $m = n$, $k = l$, high fidelity $\epsilon = 0.01$ when addressing with half address patterns ($\lambda = 0.5$, $\kappa = 0$). Table entries correspond to network size n, pattern activity k, pattern capacity M_ϵ, network capacity C_ϵ, information capacity C_ϵ^I, and synaptic capacity C_ϵ^S.

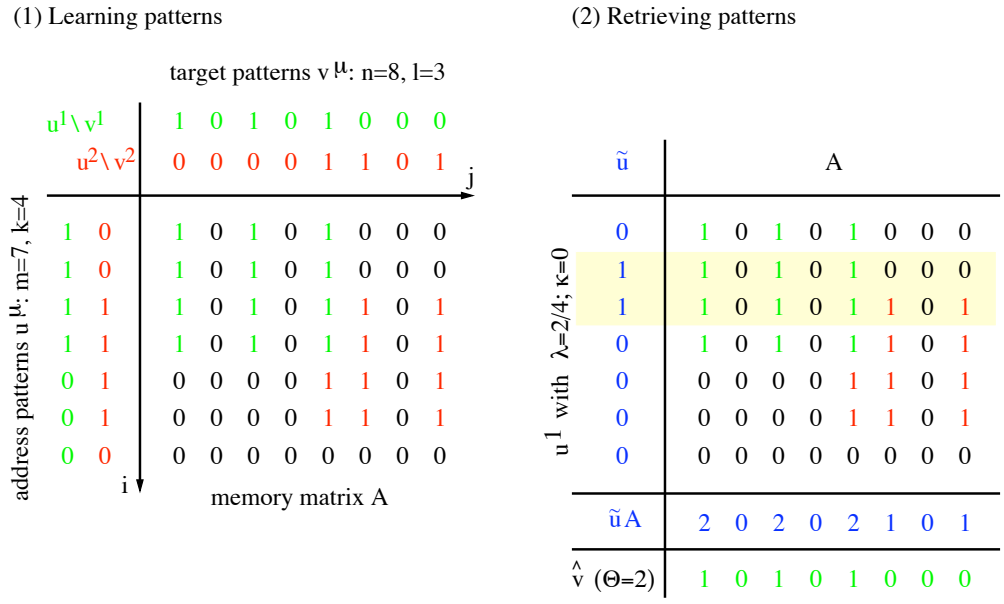

Figure 1: Learning and retrieving patterns in the binary Willshaw model. During learning (left) the associations between a set of address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$ are stored in the synaptic memory matrix $\mathbf{A}$ by clipped Hebbian learning (eq. 4). For retrieval (right) an address pattern $\tilde{\mathbf{u}}$ is propagated through the synaptic network by a vector-matrix multiplication followed by a threshold operation (eq. 5). In the example the address pattern contains half of the one-entries of $\mathbf{u}^1$ and the retrieval result equals $\mathbf{v}^1$ for an optimal threshold $\Theta = |\tilde{\mathbf{u}}| = 2$.

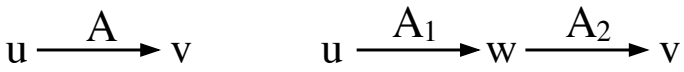

Figure 2: Single layer Willshaw model (left) and two layer extension (right) where an additional cell layer w mediates between address layer u and content layer v.

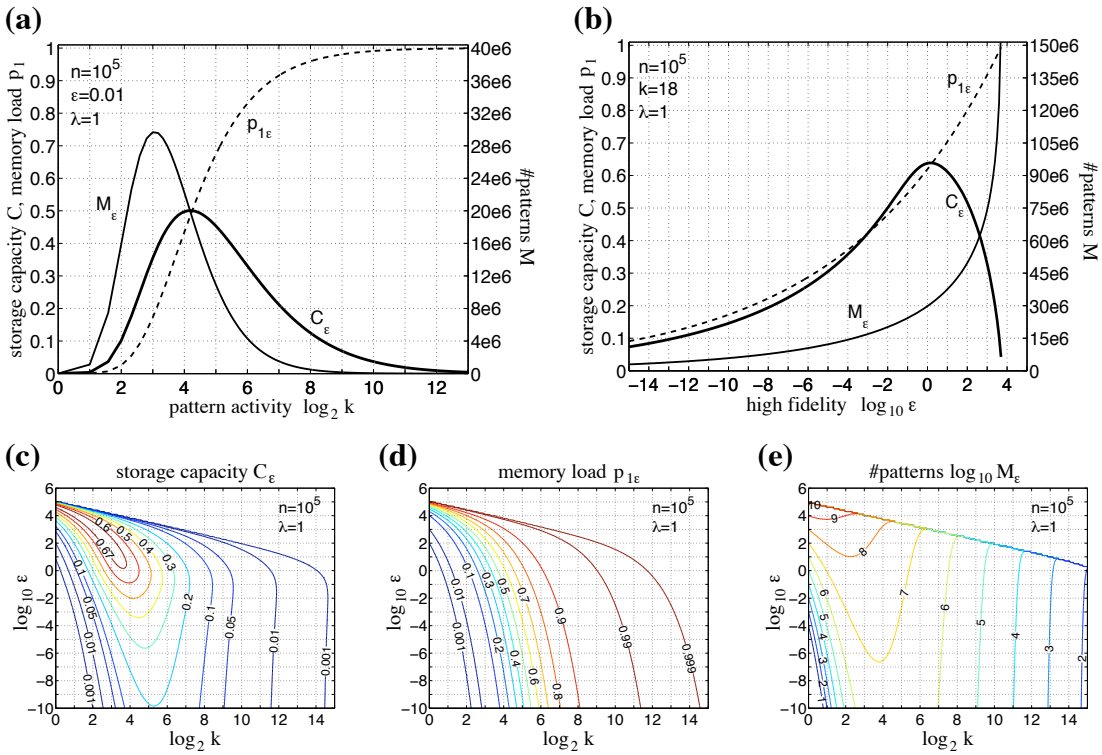

Figure 3: Classical capacity measures C and M for a finite Willshaw network with $m = n = 10^5$ neurons assuming assuming equal pattern activities, $k = l$, and zero input noise, $\lambda = 1$, $\kappa = 0$. **a** : Network capacity C_ϵ (bold line), pattern capacity M_ϵ (thin line) and memory load $p_{1\epsilon}$ (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 ϵ (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 plain spanned by pattern activity k and high fidelity parameter ϵ for network capacity C_ϵ (**c**), memory load $p_{1\epsilon}$ (**d**), and pattern capacity M_ϵ (**e**).

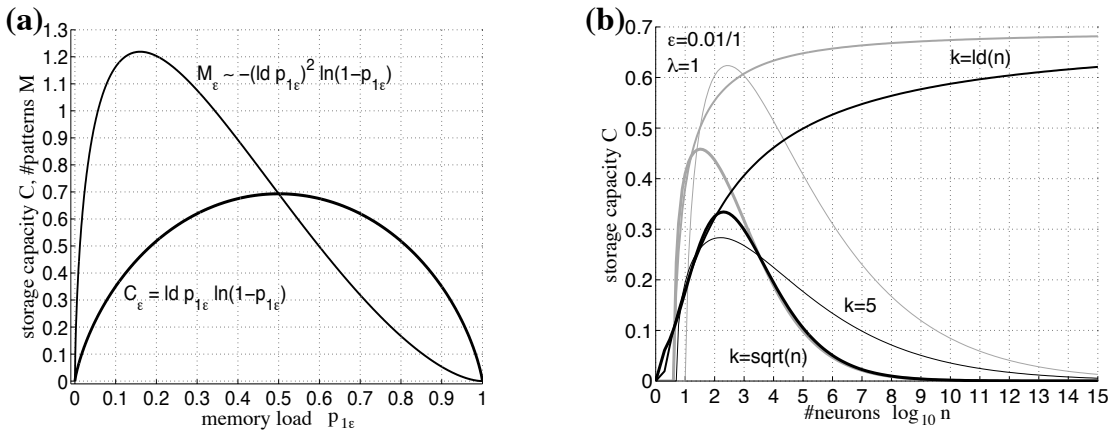

Figure 4: Classical capacity measures C and M for the Willshaw network in the asymptotic limit $n \to \infty$. Other parameter settings are as in figure 3: $m = n$, $k = l$, $\lambda = 1$ and $\kappa = 0$. **a** : Network capacity $C_\epsilon \to \mathrm{ld}p_{1\epsilon} \ln(1 - p_{1\epsilon})$ (bold line, see eq.14) 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 eq.13) as functions of the matrix load $p_{1\epsilon}$ (see eq. 12). C_ϵ is maximal for $p_{1\epsilon} = 0.5$, whereas M_ϵ is maximal for $p_{1\epsilon} \approx 0.16$. **b** : Network capacity C_ϵ as 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: squareroot sparseness; solid lines: logarithmic sparseness; thin lines: low constant activity ($k = 5$).

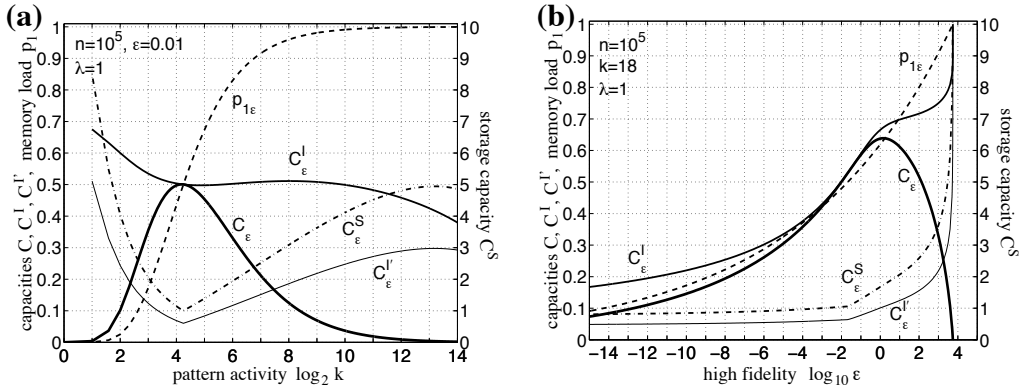

Figure 5: Capacity measures C^I and C^S for a finite Willshaw network with structural compression. Parameters are as in Fig. 3 (square weight matrix with $m = n = 10^5$, equal pattern activities $k = l$, zero input noise with $\lambda = 1$, $\kappa = 0$). The plots show information capacity C^I_ϵ for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

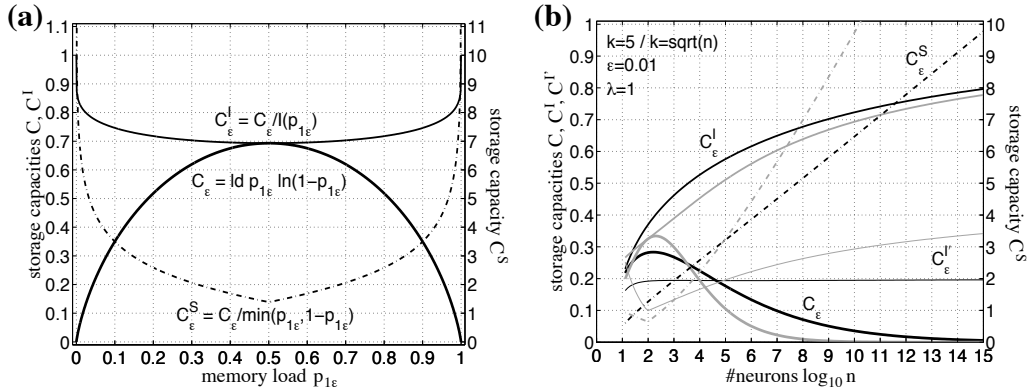

Figure 6: Capacity measures C^I and C^S for the compressed Willshaw model in the asymptotic limit $n \to \infty$. Parameters are as in Fig. 4: $m = n$, $k = l$, $\lambda = 1$, $\kappa = 0$. **a** : Information capacity C^I_ϵ (solid line) and synaptic capacity C^S_ϵ (dash-dotted) as functions of the matrix load $p_{1\epsilon}$. For reference, the plot shows also network capacity C_ϵ (bold). 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_ϵ, C^I_ϵ, $C^{I'}_\epsilon$ (thin), and C^S_ϵ as functions of the network size n for pattern activities $k(n) = 5$ (black) and $k(n) = \sqrt{n}$ (gray) assuming $\epsilon = 0.01$, cf. Fig. 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).

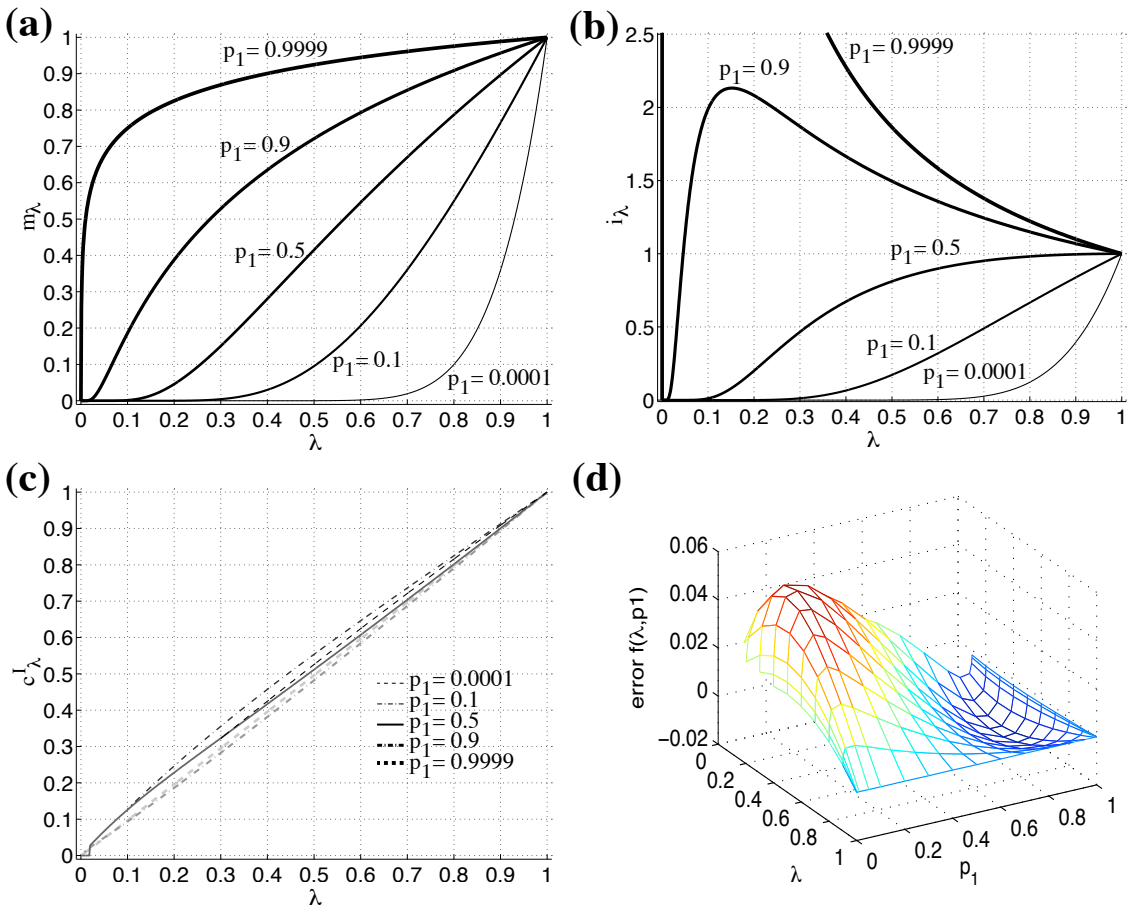

Figure 7: Impact of miss noise on the number of storable patterns and the compressibility of the memory matrix for different p_1. Query patterns $\tilde{\mathbf{u}}$ are assumed to contain λk out of the k original ones, but no false ones ($\kappa = 0$). Here $p_1 := p_{1\epsilon}(1)$ is the maximal matrix load for $\lambda = 1$ (see eq. 12). **a** : Fraction of storable patterns m_λ vs. λ (see eq. 34). **b** : Relative compressibility i_λ vs. λ (see eq. 35). **c** : For all values of p_1 we have $c^I_\lambda := m_\lambda / i_\lambda \approx \lambda$ (see eq. 36). **d** : The error $f(\lambda, p_1) := c^I_\lambda - \lambda$ of approximating c^I_λ by λ is small ($-0.02 < f < 0.06$) and even vanishes for $p_1 \to 0$ and $p_1 \to 1$.

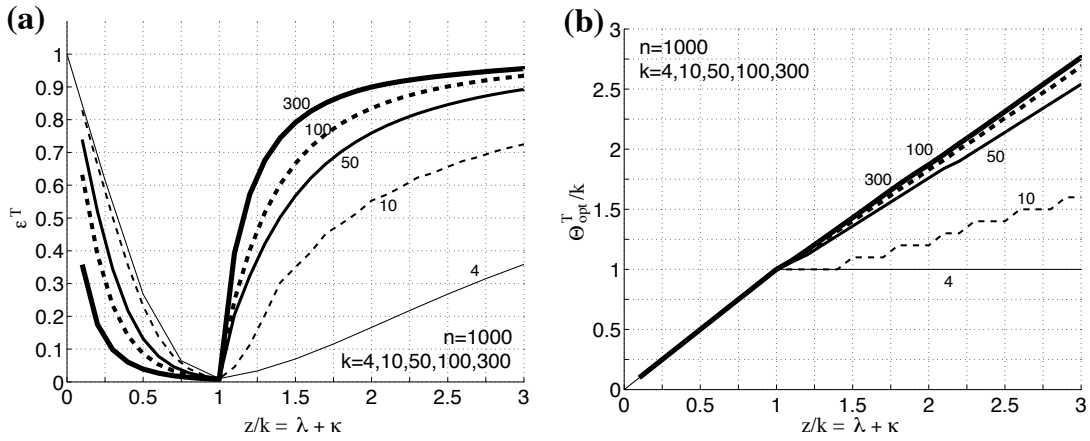

Figure 8: Impact of query noise on the retrieval quality of the Willshaw model for $m = n = 1000$ neurons and different pattern activities $k = l = 4, 10, 50, 100, 300$ (increasing line thickness) storing $M = 4928, 4791, 663, 207, 27$ patterns in each case (corresponding to $\epsilon = 0.01$ for noiseless queries). Data are computed from exact error probabilities eqs. 52,53. **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 Θ is chosen such that $\epsilon^T(\lambda, \kappa)$ is minimized. **b** : Optimal threshold Θ_{opt} for minimal ϵ^T shown in (a). The plots for ϵ instead of ϵ^T are qualitatively the same (Knoblauch et al., 2008).

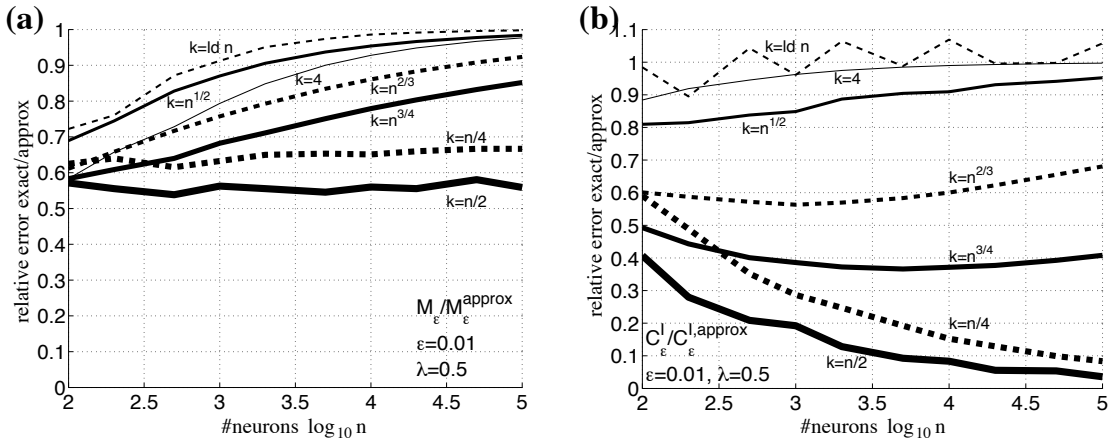

Figure 9: Approximation quality of our analysis in sections 3 and 4 based on eq. 8 for $m = n$, $k = l$, high fidelity parameter $\epsilon = 0.01$, when addressing with half address patterns ($\lambda = 0.5$,$\kappa = 0$). **a** : Relative approximation quality of the pattern capacity $M_\epsilon/M_\epsilon^{\text{approx}}$ as a function of neuron number n. The exact value M_ϵ is computed as in table 2 and the approximation $M_\epsilon^{\text{approx}}$ is computed from eq. 13. 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_ϵ is qualitatively the same. **b** : Relative approximation quality similar to (a), but for the information capacity C_ϵ^I. $C^{I,\text{approx}}$ is computed from eq. 18. Approximation quality for the synaptic capacity C_ϵ^S is qualitatively the same.

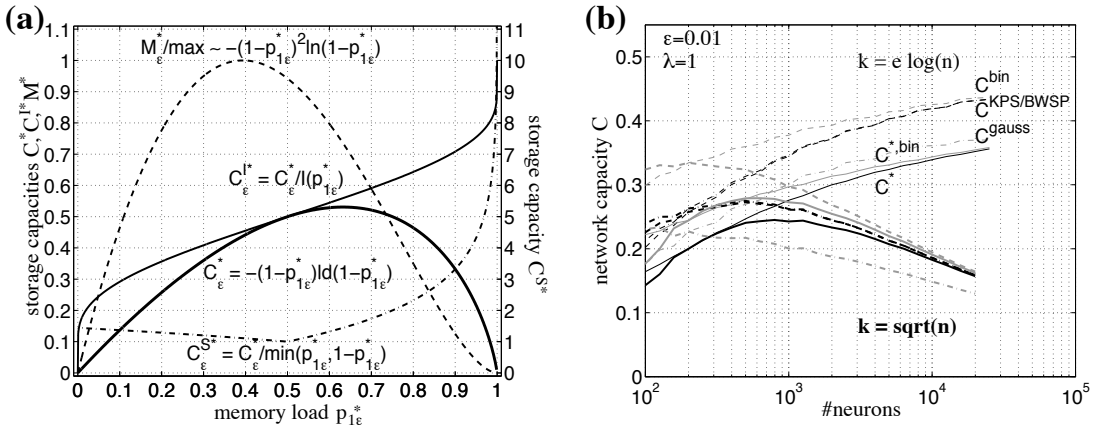

Figure 10: **a** : Asymptotic network capacity C^*_ϵ, information capacity C^{I*}_ϵ, synaptic capacity C^{S*}, and pattern capacity M^*_ϵ as functions of memory load $p^*_{1\epsilon}$ for the model variant with random query pattern activity. Compare to Fig. 6a. **b** : Exact and approximative network 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 eq. 65 (C^*; black solid) and binomial approximation eq. 73 ($C^{*,\mathrm{bin}}$; gray solid). For fixed query pattern activity the plot shows results computed with exact eq. 58 (C^{BWSP}; black dashed) and eq. 57 (C^{KPS}; black dash-dotted), the binomial approximation eq. 12 (C^{bin}; gray dashed), and a Gaussian approximation of dendritic potentials (C^{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.