Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Neural associative memory with optimal Bayesian learning.

Andreas Knoblauch

2011

Preprint:

This is an accepted article published in Neural Computation. The final authenticated version is available online at: https://doi.org/[DOI not available]

ARTICLE Communicated by Friedrich Sommer

Neural Associative Memory with Optimal Bayesian Learning

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

Neural associative memories are perceptron-like single-layer networks with fast synaptic learning typically storing discrete associations between pairs of neural activity patterns. Previous work optimized the memory capacity for various models of synaptic learning: linear Hopfield-type rules, the Willshaw model employing binary synapses, or the BCPNN rule of Lansner and Ekeberg, for example. Here I show that all of these previous models are limit cases of a general optimal model where synaptic learning is determined by probabilistic Bayesian considerations. Asymptotically, for large networks and very sparse neuron activity, the Bayesian model becomes identical to an inhibitory implementation of the Willshaw and BCPNN-type models. For less sparse patterns, the Bayesian model becomes identical to Hopfield-type networks employing the covariance rule. For intermediate sparseness or finite networks, the optimal Bayesian learning rule differs from the previous models and can significantly improve memory performance. I also provide a unified analytical framework to determine memory capacity at a given output noise level that links approaches based on mutual information, Hamming distance, and signal-to-noise ratio.

1 Introduction

An associative memory is an alternative computing architecture in which, unlike the classical von Neumann machine, computation and data storage are not separated. For example, as illustrated by Figure 1, an associative memory can store a set of associations between pairs of pattern vectors $\{(\mathbf{u}^\mu \to \mathbf{v}^\mu) : \mu = 1, \ldots, M\}$. Similar to random access memory, a query pattern $\mathbf{u}^\mu$ entered in associative memory can serve as an address for accessing the associated content pattern $\mathbf{v}^\mu$. However, unlike random access memory, an associative memory accepts 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

 © 2011 Massachusetts Institute of Technology

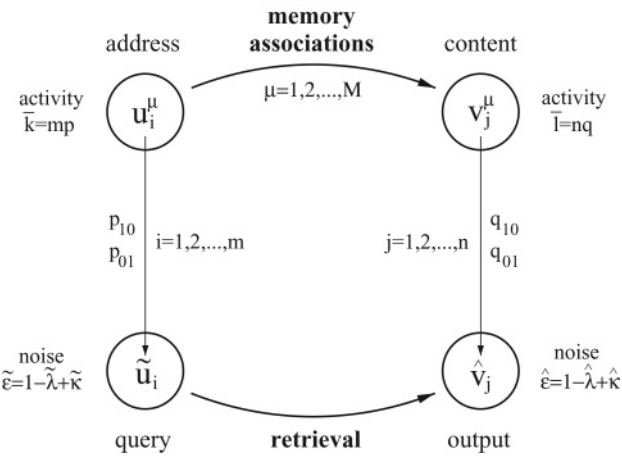

Figure 1: Pattern storage and retrieval in an associative memory. The task is to store M associations between address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$. Address patterns $\mathbf{u}^\mu$ are binary vectors of size m with an average number of $\bar{k} = mp$ active units. Similarly, the content patterns $\mathbf{v}^\mu$ have size n and mean activity $\bar{l} = nq$. During retrieval, the memories are addressed by a query pattern $\tilde{\mathbf{u}}$ being a noisy version of one of the address patterns with component transition probabilities $p_{10} := \mathrm{pr}[\tilde{u}_i = 0 | u_i^\mu = 1]$ and $p_{01} := \mathrm{pr}[\tilde{u}_i = 1 | u_i^\mu = 0]$ corresponding to miss noise and add noise, respectively. Thus, the query contains, on average, a fraction of $\tilde{\lambda} = 1 - p_{10}$ correctly active units and another fraction of $\tilde{\kappa} = p_{01}(1-p)/p$ falsely active units. The total fraction of wrong query components is called query noise $\tilde{\epsilon} = 1 - \tilde{\lambda} + \tilde{\kappa}$. Similarly, the output noise $\hat{\epsilon} = 1 - \hat{\lambda} + \hat{\kappa}$ is the fraction of wrong components in the retrieval output pattern $\hat{\mathbf{v}}$ where $q_{10} := \mathrm{pr}[\hat{v}_j = 0 | v_j^\mu = 1]$ and $q_{01} := \mathrm{pr}[\hat{v}_j = 1 | v_j^\mu = 0]$ are the transition probabilities of the corresponding memory channel.

completion, denoising, or data retrieval using incomplete cues. Moreover, neural implementations of associative memory 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, 1991; Fransen & Lansner, 1998; Pulvermüller, 2003; Johansson & Lansner, 2007; Lansner, 2009; Gardner-Medwin, 1976; Rolls, 1996; Bogacz, Brown, & Giraud-Carrier, 2001; Marr, 1969, 1971; Albus, 1971; Kanerva, 1988; Laurent, 2002).

In its simplest forms, neural associative memories are single-layer perceptrons with fast, typically one-shot, synaptic learning realizing the storage of M discrete associations between binary address and content patterns $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$. The one-shot constraint favors local learning rules where a synaptic

weight w_{ij} depends on only u_i^μ and v_j^μ. Alternative nonlocal learning methods are typically time-consuming and require gradient descent (such as error backpropagation) that is based on global error signals obtained from repeated training of the entire pattern set. Instead, associative memories use simple Hebbian-type learning rules where synaptic weights increase if both the presynaptic and postsynaptic neurons are active during presentation of a pattern pair.

The performance of neural associative memory models can be evaluated by storage capacity, which can be defined, for example, by the number of memories M a network of a given size can store or by the Shannon information C that a synapse can store. More recent work considers also structural compression of synaptic networks and the energy or time requirements per retrieval (Poirazi & Mel, 2001; Stepanyants, Hof, & Chklovskii, 2002; Lennie, 2003; Knoblauch, 2003, 2005, 2009b; Knoblauch, Palm, & Sommer, 2010).

The simplest one-shot learning model is the so-called Steinbuch or Willshaw model with binary synapses and clipped Hebbian learning (Willshaw, Buneman, & Longuet-Higgins, 1969; Steinbuch, 1961; Palm, 1980, 1991; Golomb, Rubin, & Sompolinsky, 1990; Nadal, 1991; Sommer & Dayan, 1998; Sommer & Palm, 1999; Knoblauch et al., 2010). Here a single coincidence of presynaptic and postsynaptic activity is sufficient to increase the synaptic weight from 0 to 1, while further coincidences do not cause further changes.

An alternative model is the linear associative memory, where contributions of different pattern pairs add linearly (Kohonen, 1972; Kohonen & Oja, 1976; Anderson, Silverstein, Ritz, & Jones, 1977; Hopfield, 1982; Palm, 1988a; 1988b; Tsodyks & Feigel'man, 1988; Willshaw & Dayan, 1990; Dayan & Willshaw, 1991; Palm & Sommer, 1992, 1996; Chechik, Meilijson, & Ruppin, 2001; Sterratt & Willshaw, 2008). For example, for binary memory patterns $u_i^\mu, v_j^\mu \in \{0, 1\}$ the general linear learning rule can be described by four values $r_{u_i^\mu v_j^\mu}$ specifying the weight increments for the possible combinations of presynaptic and postsynaptic activity.

Surprisingly, the maximal storage capacity C in bits per synapse is almost identical for the two models: the Willshaw model can achieve up to 0.69 bits per synapse (bps), whereas the linear models achieve only a slightly higher capacity of 0.72 bps in spite of employing real-valued synaptic weights. However, closer investigation reveals that the Willshaw model can achieve nonzero capacity only for extremely sparse activity, where the number of active units per pattern vector scales logarithmic with the vector size. In contrast, the linear model achieves the maximum $C = 0.72$ bps for a much larger range of moderately sparse patterns. Only for a nonvanishing fraction of active units per pattern vector does the performance drop from 0.72 bps to the capacity of the original (nonsparse) Hopfield network (e.g., $C = 0.14$ bps in Hopfield, 1982; Hertz, Krogh, & Palmer, 1991; Palm & Sommer, 1996, or, as we will see below, $C = 0.33$ bps for the heteroassociative feedforward networks considered here). The linear learning

model achieves maximal storage capacity only for the optimal covariance learning rule (e.g., Sejnowski, 1977a, 1977b; Dayan & Willshaw, 1991; Dayan & Sejnowski, 1993; Palm & Sommer, 1996), which becomes equal to the Hebb rule for very sparse patterns and equal to the Hopfield rule for nonsparse patterns. Moreover, simulation experiments show that the capacity of the optimal linear model remains well below the capacity of the Willshaw model for any reasonable finite network size (e.g., $C = 0.2$ bps versus $C = 0.5$ bps for $n = 10^5$ neurons; see Knoblauch, 2009a; Palm & Sommer, 1992). This suggests that the linear covariance rule is not always optimal, in particular not for finite networks and sparse memory representations as found in the brain (Waydo, Kraskov, Quiroga, Fried, & Koch, 2006).

A third model class is based on the Bayesian confidence propagation neural network (BCPNN) rule (Lansner & Ekeberg, 1987, 1989; Kononenko, 1989, 1994; Lansner & Holst, 1996; Sandberg, Lansner, Petersson, & Ekeberg, 2000; Lansner, 2009). This model employs Bayesian maximum-likelihood heuristics for synaptic learning and retrieval (see also a related approach based on maximizing the entropy of synaptic weights: MacKay, 1991). Therefore, it has been suspected that the BCPNN model could achieve optimal performance, or at least exceed the performance of Willshaw and linear models. These conjectures have been supported by some numerical investigations; however, theoretical analyses of the BCPNN model have been lacking so far. As we will see, the BCPNN model becomes optimal only for a limited range of very sparse memory patterns.

This article (see also Knoblauch, 2009a, 2010a) develops the generally optimal associative memory that minimizes output noise and maximizes storage capacity by activating neurons based on Bayesian maximum likelihood decisions. The corresponding neural interpretation of this Bayesian associative memory corresponds in general to a novel nonlinear learning rule resembling the BCPNN rule. Specifically, a theoretical analysis including query noise shows that the previous learning models are only special limit cases of the generally optimal Bayesian model. Asymptotically, for large networks and extremely sparse memory patterns, the Bayesian model becomes essentially identical to the binary Willshaw model (but implemented with inhibitory rather than excitatory synapses; see Knoblauch, 2007). Similarly, the BCPNN model is optimal for a less restricted range of sparse memory patterns where the fraction of active units per memory vector still vanishes. For less sparse and nonsparse patterns, the Bayesian model becomes identical to the linear model employing the covariance rule. For a large range of intermediate sparseness and finite networks, the Bayesian learning rule is shown to perform significantly better than previous models. As a by-product, this work also provides a unified analytical framework to determine memory capacities at a given output noise level that links approaches based on mutual information, Hamming distance, and signal-to-noise ratio.

The organization of the paper is as follows. Section 2 describes the model of neural associative memory with optimal Bayesian learning and analyzes signal-to-noise ratio and storage capacity. Section 3 compares the Bayesian associative memory to previous models in the literature, including inhibitory implementations of the Willshaw network, linear learning models with the covariance rule, and BCPNN-type models, and determines asymptotic conditions when the respective models become equivalent to optimal Bayesian learning. Section 4 presents results from numerical simulation experiments verifying the theoretical results concerning signal-to-noise-ratio, output noise, and storage capacity. Further experiments compare the performance of various learning models for finite network sizes. Section 5 summarizes and discusses the main results of this work. The appendixes include a description for appropriate implementations of Bayesian associative memory (appendix A), an analysis for computing optimal firing thresholds (appendix D), an analysis of the relationship between signal-to-noise ratio and Hamming-distance-based measures for output noise and storage capacity (appendix E), and signal-to-noise ratio analyses for the linear and BCPNN-type models (appendixes G, H).

2 Model of Bayesian Associative Memory

2.1 Memory Storage in Neural and Synaptic Countervariables. The task is to store M associations between address patterns $\mathbf{u}^\mu$ and content patterns $\mathbf{v}^\mu$ where $\mu = 1, \ldots, M$. Here $\mathbf{u}^\mu$ and $\mathbf{v}^\mu$ are binary vectors of size m and n, respectively. Memory associations are stored in first-order (neural) and second-order (synaptic) countervariables. In particular, each address neuron i and each content neuron j can memorize its unit usage:

$$M_1(j) := \#\{\mu : v_j^\mu = 1\}, \tag{2.1}$$

$$M_0(j) := \#\{\mu : v_j^\mu = 0\} = M - M_1(j), \tag{2.2}$$

$$M_1'(i) := \#\{\mu : u_i^\mu = 1\}, \tag{2.3}$$

$$M_0'(i) := \#\{\mu : u_i^\mu = 0\} = M - M_1'(i). \tag{2.4}$$

Similarly, each synapse ij can memorize its synapse usage:

$$M_{11}(ij) := \#\{\mu : u_i^\mu = 1, v_j^\mu = 1\}, \tag{2.5}$$

$$M_{01}(ij) := \#\{\mu : u_i^\mu = 0, v_j^\mu = 1\} = M_1(j) - M_{11}(ij), \tag{2.6}$$

$$M_{00}(ij) := \#\{\mu : u_i^\mu = 0, v_j^\mu = 0\} = M_0'(i) - M_{01}(ij), \tag{2.7}$$

$$M_{10}(ij) := \#\{\mu : u_i^\mu = 1, v_j^\mu = 0\} = M_0(j) - M_{00}(ij), \tag{2.8}$$

where $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Note that it is sufficient to memorize M, M_1, M_1', and M_{11}. Thus, an implementation on a digital computer requires about $(mn + m + n + 1)\mathrm{ld}M$ memory bits. The following analyses consider optimal Bayesian retrieval, assuming that each output unit $j = 1, \ldots, n$ has access to the variables in the set

$$\mathfrak{M}(j) := \{M, M_1(j), M_1'(i), M_{11}(ij) : i = 1, \ldots, m\}. \tag{2.9}$$

The following analyses will show that the mean values of the coincidence counters $\overline{M_{11}} := E(M_{11})$ and unit usages, $\overline{M_1} := E(M_1)$, $\overline{M_1'} := E(M_1')$, have a major role in determining the regime of operation for Bayesian associative memory (see Table 2).

2.2 Neural Formulation of Optimal Bayesian Retrieval. Given a query pattern $\tilde{\mathbf{u}}$ and the countervariables of section 2.1, the memory task is to find the most similar address pattern $\mathbf{u}^\mu$ and return a reconstruction $\hat{\mathbf{v}}$ of the associated content $\mathbf{v}^\mu$. In general, query $\tilde{\mathbf{u}}$ is a noisy version of $\mathbf{u}^\mu$, assuming component transition probabilities given the activity of a content neuron, $v_j^\mu = \mathfrak{a} \in \{0, 1\}$:

$$p_{01|\mathfrak{a}}(ij) := \mathrm{pr}[\tilde{u}_i = 1 | u_i^\mu = 0, v_j^\mu = \mathfrak{a}], \tag{2.10}$$

$$p_{10|\mathfrak{a}}(ij) := \mathrm{pr}[\tilde{u}_i = 0 | u_i^\mu = 1, v_j^\mu = \mathfrak{a}]. \tag{2.11}$$

Now the content neurons j have to decide independently of each other whether to be activated or remain silent. Given the query $\tilde{\mathbf{u}}$, the optimal maximum likelihood decision is based on the odds ratio $\mathfrak{r}_j$,

$$\hat{v}_j = \begin{cases} 1, & \mathfrak{r}_j := \dfrac{\mathrm{pr}[v_j^\mu = 1 | \tilde{\mathbf{u}}, \mathfrak{M}(j)]}{\mathrm{pr}[v_j^\mu = 0 | \tilde{\mathbf{u}}, \mathfrak{M}(j)]} \geq 1 \\ 0, & \text{otherwise} \end{cases}, \tag{2.12}$$

which minimizes the expected Hamming distance $d_H(\mathbf{v}^\mu, \hat{\mathbf{v}}) := \sum_{j=1}^n |v_j^\mu - \hat{v}_j|$ between original and reconstructed content. If the query pattern components are conditional independent given the activity of content neuron j (e.g., assuming independently generated address and query components), we have for $\mathfrak{a} \in \{0, 1\}$

$$\mathrm{pr}[\tilde{\mathbf{u}} | v_j^\mu = \mathfrak{a}, \mathfrak{M}(j)] = \prod_{i=1}^m \mathrm{pr}[\tilde{u}_i | v_j^\mu = \mathfrak{a}, \mathfrak{M}(j)]$$

$$= \prod_{i=1}^{m} \frac{M_{\tilde{u}_i \mathfrak{a}}(ij)(1 - p_{\tilde{u}_i(1-\tilde{u}_i)|\mathfrak{a}}(ij)) + M_{(1-\tilde{u}_i)\mathfrak{a}}(ij)\, p_{(1-\tilde{u}_i)\tilde{u}_i|\mathfrak{a}}(ij)}{M_{\mathfrak{a}}(j)}. \tag{2.13}$$

With the Bayes formula $\mathrm{pr}[v_j^\mu = \mathfrak{a}|\tilde{\mathbf{u}}, \mathfrak{M}(j)] = \mathrm{pr}[\tilde{\mathbf{u}}|v_j^\mu = \mathfrak{a}, \mathfrak{M}(j)]\mathrm{pr}[v_j^\mu = \mathfrak{a}|\mathfrak{M}(j)]/\mathrm{pr}[\tilde{\mathbf{u}}|\mathfrak{M}(j)]$, the odds ratio is

$$\mathfrak{r}_j = \left(\frac{M_0(j)}{M_1(j)}\right)^{m-1} \times \prod_{i=1}^{m} \frac{M_{\tilde{u}_i 1}(ij)(1 - p_{\tilde{u}_i(1-\tilde{u}_i)|1}(ij)) + M_{(1-\tilde{u}_i)1}(ij)\, p_{(1-\tilde{u}_i)\tilde{u}_i|1}(ij)}{M_{\tilde{u}_i 0}(ij)(1 - p_{\tilde{u}_i(1-\tilde{u}_i)|0}(ij)) + M_{(1-\tilde{u}_i)0}(ij)\, p_{(1-\tilde{u}_i)\tilde{u}_i|0}(ij)}. \tag{2.14}$$

For a more plausible neural formulation, we can take logarithms of the probabilities and obtain dendritic potentials $x_j := \log \mathfrak{r}_j$. With $\mathfrak{f}(\tilde{u}_i, ij)$ being the ith factor in the product of equation 2.14, it is

$$\begin{aligned} x_j - (m-1)\log\frac{M_0(j)}{M_1(j)} &= \sum_{i=1}^{m} \log \mathfrak{f}(\tilde{u}_i, ij) \\ &= \sum_{i=1}^{m} (\log \mathfrak{f}(0, ij) + \tilde{u}_i(\log \mathfrak{f}(1, ij) - \log \mathfrak{f}(0, ij))). \end{aligned}$$

Thus, synaptic weights w_{ij}, dendritic potentials x_j, and retrieval output $\hat{v}_j$ are finally

$$w_{ij} = \log \frac{(M_{11}(1 - p_{10|1}) + M_{01}\, p_{01|1})(M_{00}(1 - p_{01|0}) + M_{10}\, p_{10|0})}{(M_{10}(1 - p_{10|0}) + M_{00}\, p_{01|0})(M_{01}(1 - p_{01|1}) + M_{11}\, p_{10|1})}, \tag{2.15}$$

$$x_j = (m-1)\log\frac{M_0}{M_1} + \sum_{i=1}^{m} \log \frac{M_{01}(1 - p_{01|1}) + M_{11}\, p_{10|1}}{M_{00}(1 - p_{01|0}) + M_{10}\, p_{10|0}} + \sum_{i=1}^{m} w_{ij}\tilde{u}_i, \tag{2.16}$$

$$\hat{v}_j = \begin{cases} 1, & x_j \geq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{2.17}$$

such that $\mathrm{pr}[v_j^\mu = 1|\tilde{\mathbf{u}}, \mathfrak{M}(j)] = 1/(1 + e^{-x_j})$ writes as a sigmoid function of x_j, and a content neuron fires, $\hat{v}_j = 1$, iff the dendritic potential is nonnegative. Note that indices of $M_0(j)$, $M_1(j)$, $p_{01|\mathfrak{a}}(ij)$, $p_{10|\mathfrak{a}}(ij)$, $M_{00}(ij)$, $M_{01}(ij)$, $M_{10}(ij)$, and $M_{11}(ij)$ are skipped for readability. Also note that optimal Bayesian learning is nonlinear and, for autoassociation with $\mathbf{u}^\mu = \mathbf{v}^\mu$ and nonzero query noise, asymmetric with $w_{ij} \neq w_{ji}$. Note further that synaptic

weights and dendritic potentials may be infinite, such that accurate implementations require two values per variable for finite and infinite components, respectively (see appendix A).

Nevertheless, evaluating equation 2.16 is much cheaper than equation 2.14,[1] in particular for sparse queries having only a small number of active components with $\tilde{u}_i = 1$. However, the synaptic weights of equation 2.15 may not yet satisfy Dale's law that a neuron is either excitatory or inhibitory. To be more consistent with biology, we may add a sufficiently large constant $w_0 := -\min_{ij} w_{ij}$ to each weight. Then all synapses have nonnegative weights $w'_{ij} := w_{ij} + w_0$ and the dendritic potentials remain unchanged if we replace the last sum in equation 2.16 by

$$\sum_{i=0}^{m} w_{ij}\tilde{u}_i = \sum_{i=0}^{m} w'_{ij}\tilde{u}_i - w_0 \sum_{i=0}^{m} \tilde{u}_i. \tag{2.18}$$

Here the negative sum could be realized, for example, by feedforward inhibition with a strength proportional to the query pattern activity, as suggested by Knoblauch and Palm (2001) and Knoblauch (2005), for example.

The transition probabilities, equations 2.10 and 2.11, can be estimated by maintaining countervariables similar as in section 2.1. For example, if the μth memory v^μ has been queried by $\tilde{M}^\mu$ address queries $\tilde{\mathbf{u}}^{(\mu,\mu')}$ (where $\mu' = 1, 2, \ldots, \tilde{M}^\mu$), then we could estimate for $\mathfrak{a}, \mathfrak{b}, \mathfrak{c} \in \{0, 1\}$,

$$p_{\mathfrak{b}\mathfrak{c}|\mathfrak{a}}(ij) = \frac{\#\{(\mu,\mu') : u_i^\mu = \mathfrak{b}, \tilde{u}_i^{(\mu,\mu')} = \mathfrak{c}, v_j^\mu = \mathfrak{a}\}}{\#\{(\mu,\mu') : u_i^\mu = \mathfrak{b}, v_j^\mu = \mathfrak{a}, 1 \le \mu' \le \tilde{M}^\mu\}}, \tag{2.19}$$

which requires four countervariables per synapse in addition to M_{11}. To reduce storage costs, one may assume

$$\begin{aligned} p_{\mathfrak{b}\mathfrak{c}|\mathfrak{a}}(ij) = p_{\mathfrak{b}\mathfrak{c}}(i) &:= \sum_{\mathfrak{a}\in\{0,1\}} \mathrm{pr}[v_j^\mu = \mathfrak{a}]\, p_{\mathfrak{b}\mathfrak{c}|\mathfrak{a}}(ij) \\ &= \frac{\#\{(\mu,\mu') : u_i^\mu = \mathfrak{b}, \tilde{u}_i^{(\mu,\mu')} = \mathfrak{c}\}}{\#\{(\mu,\mu') : u_i^\mu = \mathfrak{b}, 1 \le \mu' \le \tilde{M}^\mu\}}, \end{aligned} \tag{2.20}$$

independent of j, as do most of the following analyses and experiments for the sake of simplicity, although this assumption may reduce the number of discovered rules (corresponding to infinite w_{ij}) describing deterministic relationships between u_i and v_j.

[1] Evaluating equation 2.14 during retrieval requires about $5m$ multiplications and $2m$ additions even for sparse query activity with $|\tilde{\mathbf{u}}| := \sum_{i=1}^{m} \tilde{u}_i \ll m/2$. By contrast, evaluating equation 2.16 requires only $|\tilde{\mathbf{u}}|$ multiplications and m additions, as the "bias" (first and second summands) of x_j is independent of $\tilde{\mathbf{u}}$ and therefore can be computed in advance.

2.3 Analysis of the Signal-to-Noise Ratio. We would like to build a memory system with high retrieval quality, for example, where the expected Hamming distance,

$$E d_H(\mathbf{v}^\mu, \hat{\mathbf{v}}) = \sum_{j=1}^{n} q(j) q_{10}(j) + (1 - q(j)) q_{01}(j), \tag{2.21}$$

is small. Here, d_H is as defined below equation 2.12, and $q(j) := \mathrm{pr}[v_j^\mu = 1]$ is the prior probability of an active content unit. Thus, retrieval quality is determined by the component output error probabilities,

$$q_{01}(j) := \mathrm{pr}[\hat{v}_j = 1 | v_j^\mu = 0] = \mathrm{pr}[x_j \geq \Theta_j | v_j^\mu = 0], \tag{2.22}$$

$$q_{10}(j) := \mathrm{pr}[\hat{v}_j = 0 | v_j^\mu = 1] = \mathrm{pr}[x_j < \Theta_j | v_j^\mu = 1], \tag{2.23}$$

where the Θ_j are firing thresholds (e.g., $\Theta_j = 0$ for dendritic potentials x_j as in equation 2.16). Intuitively, retrieval quality will be high if the high-potential distribution $\mathrm{pr}[x_j | v_j^\mu = 1]$ and the low-potential distribution $\mathrm{pr}[x_j | v_j^\mu = 0]$ are well separated, that is, if the signal-to-noise ratio (SNR),

$$R(j) := \frac{\mu_{\mathrm{hi}}(j) - \mu_{\mathrm{lo}}(j)}{\max(\sigma_{\mathrm{lo}}(j), \sigma_{\mathrm{hi}}(j))}. \tag{2.24}$$

is large for each content neuron j (Amari, 1977; Palm, 1988a, 1988b; Dayan & Willshaw, 1991; Palm & Sommer, 1996). Here $\mu_{\mathrm{lo}} := E(x_j | v_j^\mu = 0)$ and $\sigma_{\mathrm{lo}}^2 := \mathrm{Var}(x_j | v_j^\mu = 0)$ are the expectation and variance of the low-potential distribution, and $\mu_{\mathrm{hi}} = E(x_j | v_j^\mu = 1)$ and $\sigma_{\mathrm{hi}}^2 := \mathrm{Var}(x_j | v_j^\mu = 1)$ are the expectation and variance of the high-potential distribution. Appendix E shows that under some conditions, the SNR and the Hamming distance are equivalent measures of retrieval quality.

Appendix B computes the SNR $R := R(j)$ for a particular content neuron j with $q := M_1(j)/M$ using the following simplifications:

1. The activation of an address unit i does not depend on other units, and all address units i have the same prior probability $p := p(i) := \mathrm{pr}[u_i^\mu = 1]$ of being active. Thus, on average, an address pattern has $\bar{k} := mp$ active units.
2. Query noise for an address unit i does not depend on other units, and all query components i have the same noise transition probabilities $p_{01} := p_{01}(i) = p_{01|\mathfrak{a}}(ij)$ and $p_{10} := p_{10}(i) = p_{10|\mathfrak{a}}(ij)$. Thus, on average, a query will have $\tilde{\lambda}\bar{k}$ correct and $\tilde{\kappa}\bar{k}$ false one-entries, where $\tilde{\lambda} := 1 - p_{10}$ and $\tilde{\kappa} := (1 - p)p_{01}/p$ define fractions of average miss noise and add noise, respectively, normalized to the mean address pattern activity $\bar{k}$.

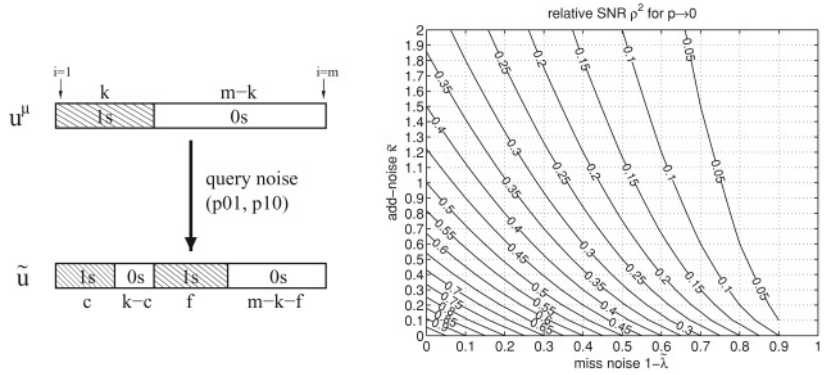

Figure 2: (Left) For the analysis of SNR, we assume that the query pattern $\tilde{\mathbf{u}}$ corresponds to one of the original address patterns $\mathbf{u}^\mu$ that has k one-entries (and $m-k$ zero-entries). Due to query noise, the query $\tilde{\mathbf{u}}$ has only c correct one-entries overlapping with $\mathbf{u}^\mu$ and an addition of f false one-entries. Without loss of generality, the analysis assumes the setting as illustrated. (Right) Contour plot of the relative SNR ρ^2 for sparse address activity with $p \to 0$ (see equation 2.29) as a function of miss noise $1-\tilde{\lambda} \approx (k-c)/k \approx p_{10}$ and add noise $\tilde{\kappa} \approx f/k \approx p_{01}(1-p)/p$ contained in the query $\tilde{\mathbf{u}}$ used for memory retrieval.

3. Retrieval involves a particular query pattern $\tilde{\mathbf{u}}$ being a noisy version of an address pattern $\mathbf{u}^\mu$ that has exactly k one-entries, where the query has c out of k correct one-entries and, additionally, f false one-entries. Without loss of generality, we can assume a setting as illustrated by Figure 2 (left), that is, the address pattern has one-entries $u_i^\mu = 1$ at components $i = 1, 2, \ldots, k$ and zero-entries $u_i^\mu = 0$ at $i = k+1, k+2, \ldots, m$ whereas the query has false entries $\tilde{u}_i = 1 - u_i^\mu$ at $i = c+1, c+2, \ldots, k+f$.
4. The average values of the synaptic coincidence counters diverge: $\overline{M_{11}} = Mpq \to \infty$. Note that this assumption also implies diverging unit usages, $\overline{M_1} = Mq \to \infty$ and $\overline{M_1'} = Mp \to \infty$. For reasons that will become apparent in section 3, the condition $\overline{M_{11}} \to \infty$ is also referred to as the linear learning regime, whereas $\overline{M_{11}} \ll \infty$ will be called the nonlinear learning regime.

From the results of appendix B, we obtain the SNR, equation 2.24 in the asymptotic limit of large $\overline{M_{11}} = Mpq \to \infty$ where all variables will be close to their expectations due to the law of large numbers. In particular, we can assume $k \approx mp$, and, for consistent error estimates, $p_{01} = f/(m-k) = \tilde{\kappa}\, p/(1-p)$, $p_{10} = (k-c)/k = 1-\tilde{\lambda}$. Then we obtain from equation B.6 the mean difference $\Delta\mu := \mu_{\mathrm{hi}} - \mu_{\mathrm{lo}}$ between high potentials and low

potentials:

$$\frac{\Delta\mu}{(\tilde{\lambda}-\frac{p}{1-p}\tilde{\kappa})(1/M_1+1/M_0)} \approx \frac{\tilde{\lambda}p(1-p)m-\tilde{\kappa}p^2m}{p(1+\tilde{\kappa}-1+\tilde{\lambda})} + \frac{mp-mp^2-\tilde{\kappa}mp^2-(mp-\tilde{\lambda}mp)(1-p)}{(1-p)(1-\frac{p}{1-p}\tilde{\kappa}+\frac{p}{1-p}(1-\tilde{\lambda}))} = m\frac{\tilde{\lambda}(1-p)-\tilde{\kappa}p}{(\tilde{\lambda}+\tilde{\kappa})(1-p(\tilde{\lambda}+\tilde{\kappa}))}. \tag{2.25}$$

Similarly, we obtain from equation B.8 for the potential variance:

$$\frac{\sigma^2_{\mathrm{lo/hi}}}{(\tilde{\lambda}-\frac{p}{1-p}\tilde{\kappa})^2(1/M_1+1/M_0)} \approx \frac{pm(\tilde{\lambda}+\tilde{\kappa})(1-p)}{p(1+\tilde{\kappa}-1+\tilde{\lambda})^2} + \frac{mp(1-p(\tilde{\lambda}+\tilde{\kappa}))}{(1-p)(1-\frac{p}{1-p}\tilde{\kappa}+\frac{p}{1-p}(1-\tilde{\lambda}))^2} = m\frac{(1-p)}{(\tilde{\lambda}+\tilde{\kappa})(1-p(\tilde{\lambda}+\tilde{\kappa}))}. \tag{2.26}$$

In order to include randomly diluted networks with connectivity $P \in (0;1]$ where a content neuron v_j receives synapses from only a fraction P of the m address neurons, we can simply replace m by Pm. With $M_1 \approx Mq$ and $M_0 \approx M(1-q)$, the asymptotic SNR $R = \Delta\mu/\sigma$ is

$$R^2 \approx Pm(1/M_1+1/M_0)\frac{(\tilde{\lambda}(1-p)-\tilde{\kappa}p)^2}{(1-p)(\tilde{\lambda}+\tilde{\kappa})(1-p(\tilde{\lambda}+\tilde{\kappa}))} \tag{2.27}$$

$$\approx P\rho^2\frac{m}{Mq(1-q)} \tag{2.28}$$

with

$$\rho^2 \approx \frac{(\tilde{\lambda}(1-p)-\tilde{\kappa}p)^2}{(1-p)(\tilde{\lambda}+\tilde{\kappa})(1-p(\tilde{\lambda}+\tilde{\kappa}))} \approx \begin{cases} \dfrac{\tilde{\lambda}^2}{\tilde{\lambda}+\tilde{\kappa}}, & p \to 0 \\ \dfrac{(\tilde{\lambda}-\tilde{\kappa})^2}{(\tilde{\lambda}+\tilde{\kappa})(2-\tilde{\lambda}-\tilde{\kappa})}, & p = 0.5 \end{cases}. \tag{2.29}$$

Thus, for zero query noise, $\tilde{\lambda}=1, \tilde{\kappa}=0$, the SNR for optimal Bayesian learning is identical to the asymptotic SNR of linear learning with the optimal covariance rule (e.g., see $\rho_3^{\mathrm{Covariance}}$ in Dayan & Willshaw, 1991, p. 259, or

equation 3.28 in Palm & Sommer, 1996, p. 95; see also section 3.2). Nonzero query noise according to $\tilde{\lambda} < 1$ or $\tilde{\kappa} > 0$ decreases the SNR R by a factor $\rho < 1$. Note that ρ characterizes the basin of attraction, defined as the set of queries $\{\tilde{\mathbf{u}} : \hat{\mathbf{v}} \approx \mathbf{v}^\mu\}$ that get mapped to a stored memory $\mathbf{v}^\mu$. For example, we can evaluate which combinations of $\tilde{\lambda}$ and $\tilde{\kappa}$ achieve a fixed desired ρ (and thus R). It turns out that for sparse address patterns, $p < 0.5$, miss noise $\tilde{\lambda} < 1$ impairs network performance more severely than add noise $\tilde{\kappa} > 0$ (see Figure 2, right). As a consequence, the basins of attraction for neural associative memories employing sparse address patterns are not necessarily spheres, but they can be heavily distorted, enlarging toward queries with add noise and shrinking toward queries with miss noise. This implies that the similarity metrics employed by associative networks can strongly deviate from commonly used Hamming or Euclidean metrics. Instead, associative networks appear to follow an information-theoretic metric based on mutual information or transinformation (Cover & Thomas, 1991). This is true at least for random address patterns $\mathbf{u}^\mu$ storing a sufficiently large number of memories such that the synapse usages, in particular M_{11}, are almost never zero. Numerical simulations discussed in section 4 reveal that basins of attraction can behave quite differently if these assumptions are not fulfilled.

2.4 Analysis of Storage Capacity. Let us determine the maximal number of memories that can be stored in an associative network or, equivalently, the maximal amount of information that a synapse can store. To this end, we define storage capacity at a given level of output noise,

$$\hat{\epsilon} := \frac{E d_H(\mathbf{v}^\mu, \hat{\mathbf{v}})}{\bar{l}} = 1 - \hat{\lambda} + \hat{\kappa}, \tag{2.30}$$

being the expected Hamming distance, equation 2.21 normalized to the mean content pattern activity $\bar{l} := \sum_{j=1}^{n} q(j)$. As for query noise, we can write output noise as a sum of miss noise $1 - \hat{\lambda}$ and add noise $\hat{\kappa}$. For ergodic $q := q(j)$, $q_{01} := q_{01}(j)$, $q_{10} := q_{10}(j)$ (or considering only a single output unit j), we have miss noise $1 - \hat{\lambda} = q_{10}$ and add noise $\hat{\kappa} = (1 - q)q_{01}/q$. The weighing between miss noise and add noise can be expressed by the output noise balance,

$$\hat{\xi} := \frac{\hat{\kappa}}{\hat{\epsilon}} = \frac{(1 - q)q_{01}}{q q_{10} + (1 - q)q_{01}} \in [0; 1]. \tag{2.31}$$

For any given distribution of dendritic potentials, there exists a unique optimal firing threshold (see appendix D) and, hence, a corresponding optimal noise balance (see equation E.10) that minimize the output noise $\hat{\epsilon}$. This

minimal output noise $\hat{\epsilon}_{\min} := \min_{\hat{\xi}} \hat{\epsilon}$ is an increasing function of the number M of stored memories (see equation E.6). Therefore, we can define the pattern capacity

$$M_\epsilon := \max\{M : \hat{\epsilon}_{\min} \leq \epsilon\}, \tag{2.32}$$

as the maximal number of memory patterns that can be stored such that the output noise does not exceed a given value ϵ. Assuming that the dendritic potentials follow approximately a gaussian distribution (which is not always true; e.g., see Henkel & Opper, 1990; Knoblauch, 2008), we can apply the results of appendix E and obtain M_ϵ from the SNR, equation 2.24, by solving the equation $R(M_\epsilon) = R_{\min}(\epsilon, q)$ for M_ϵ. Here $R(M_\epsilon)$ is approximately equal to equation 2.28, and $R_{\min}$ is the minimal SNR required for output noise level ϵ and can be computed from solving equation E.6 for R (or, more conveniently, by iterating equations E.9 and E.10). Thus,

$$M_\epsilon = R^{-1}(R_{\min}(\epsilon, q)) \approx P\rho^2 \frac{m}{q(1-q)(R_{\min}(\epsilon, q))^2}. \tag{2.33}$$

where the approximation becomes exact for large networks in the limit $Mpq \to \infty$.

An alternative capacity measure normalizes the stored Shannon information (of the content memories) to the number Pmn of synapses employed in a given network. This is the network capacity

$$C_\epsilon := \frac{M_\epsilon n T(q; q_{01}, q_{10})}{Pmn} \approx \frac{\rho^2}{2\ln 2} \frac{T(q; q_{01}, q_{10})}{q(1-q)(R_{\min}(\epsilon, q))^2} \tag{2.34}$$

where T is the transinformation equation F.4 with error probabilities q_{01}, q_{10} as in equations E.4 and E.5 using $R = R_{\min}$. We can refine these results for two important cases using the results of appendixes E and F.

First, for nonsparse content patterns with $q = 0.5$, it is

$$M_\epsilon \approx P\rho^2 \frac{m}{(G^{c\,-1}(\epsilon/2))^2} \tag{2.35}$$

$$C_\epsilon \approx \frac{\rho^2}{2\ln 2} \frac{1 - I(\epsilon/2)}{(G^{c\,-1}(\epsilon/2))^2} \leq \frac{1}{2\pi(\ln 2)^2} \approx 0.3313. \tag{2.36}$$

As can be seen in Figures 3a and 3b, the upper bound of C_ϵ is achieved for zero query noise ($1 - \tilde{\lambda} = \tilde{\kappa} = 0$) and low fidelity with $\epsilon \to 1$, while $C_\epsilon \to 0$ for high fidelity with $\epsilon \to 0$.

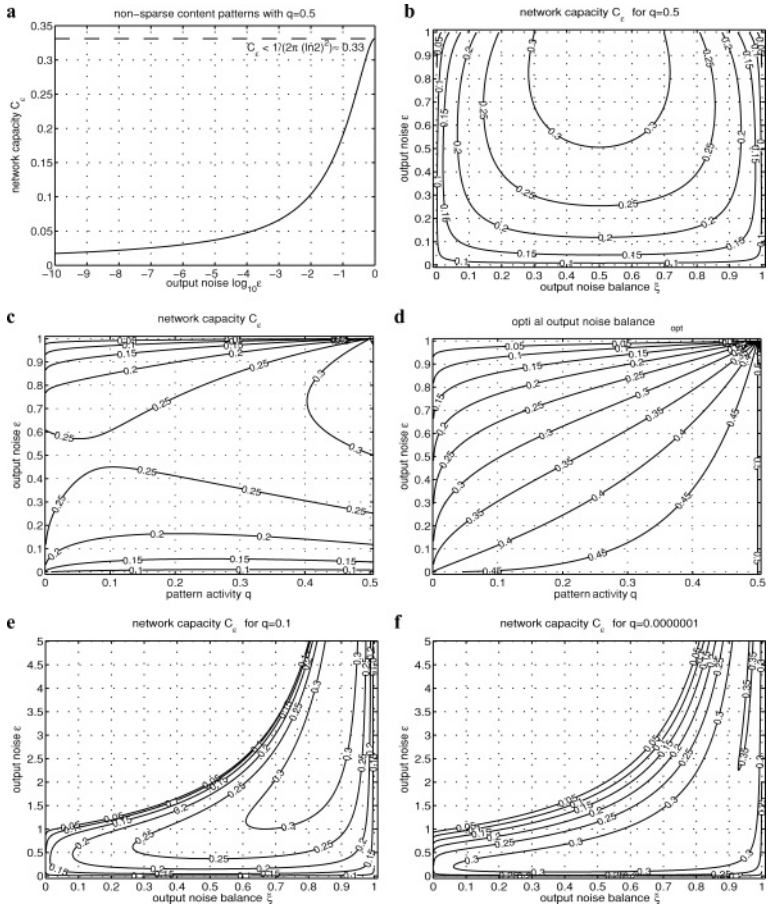

Figure 3: Generalized asymptotic network capacity $C_{\epsilon\xi}$ for zero query noise ($\tilde{\lambda} = 1,\ \tilde{\kappa} = 0,\ \rho = 1$) displayed as a function of output noise parameter ϵ, noise balance parameter ξ, and content pattern activity q (see text below equation 2.38). Here C_ϵ of equation 2.34 is a special case for optimal noise balance $\xi = \hat{\xi}_{\text{opt}}$ as in equation E.10 minimizing output noise $\hat{\epsilon}$ (see equation 2.30). (a) C_ϵ as function of output noise ϵ for nonsparse memories, $q = 0.5$, and optimal noise balance $\xi = \hat{\xi}_{\text{opt}} = 0.5$. (b) Contour plot of general $C_{\epsilon\xi}$ as function of parameters ϵ and ξ for nonsparse $q = 0.5$. (c) Contour plot of C_ϵ as function of q and ϵ for optimal $\xi = \hat{\xi}_{\text{opt}}$. (d) Optimal $\hat{\xi}$ corresponding to panel c. For $q \to 0$ miss noise is dominating with $\hat{\xi} \to 0$ (cf. equation D.10). (e) Contour plot of $C_{\epsilon\xi}$ similar to panel b, but for sparse content memories with $q = 0.1$. (f) Similar to panel e, but for extremely sparse $q = 0.0000001$. Note that $q \to 0$ implies that the maximum of $C_{\epsilon\xi}$ occurs for low fidelity $\epsilon > 1$ and dominating add-noise with $\xi > 0.5$. Thus, although minimizing the Hamming-distance-based output noise, the "optimal" firing threshold Θ_{opt} of appendix E does not necessarily maximize $C_{\epsilon\xi}$ unless $\epsilon \to 0$.

Second, for sparse content patterns with $q \to 0$ and any fixed ϵ, it is

$$M_\epsilon \approx P\rho^2 \frac{m}{-2q(1-q)\ln q}, \tag{2.37}$$

$$C_\epsilon \approx \rho^2 \frac{I(q)(1-\epsilon)}{-2q(1-q)\ln q} \approx \frac{\rho^2(1-\epsilon)}{2\ln 2} \leq \frac{1}{2\ln 2} \approx 0.7214. \tag{2.38}$$

where the upper bound of C_ϵ can be reached for zero query noise and high fidelity with $\epsilon \to 0$. Not surprisingly, this upper bound equals the one found for the linear covariance rule (Palm & Sommer, 1996) as well as the general capacity bound for neural networks (Gardner, 1988). Numerical evaluations (see Figures 3c to 3f) show that a network capacity close to $C_\epsilon \approx 0.72$ requires extremely sparse content memories and very large networks. In fact, finite networks of practical size can reach less than half of the asymptotic value (see Figure 3f). Note that M_ϵ and C_ϵ are defined only for $\epsilon < 1$ assuming optimal firing thresholds to minimize output noise $\hat{\epsilon}$ corresponding to an optimal noise balance $\hat{\xi} := \hat{\kappa}/\hat{\epsilon} \to 0$ as in equation E.10, where output errors are dominated by miss noise (see equation D.10). For generalized definitions of pattern capacity $M_{\epsilon\xi}$ and network capacity $C_{\epsilon\xi}$ at a given output noise balance $\hat{\xi} = \xi$, we can replace $R_{\min}(\epsilon, q)$ by $R_{\min}(\epsilon, q, \xi)$ as given by equation E.9. Here finite networks achieve maximal capacity at low fidelity $\epsilon \gg 1$ and $\xi \to 1$ where output errors are dominated by add noise.

For self-consistency, the analyses so far are valid only for diverging $M_\epsilon pq \sim -mp/\log q \to \infty$. Thus, the results are not reliable for extremely sparse memory patterns, for example, $mp = O(\log n)$, where at least the binomially distributed synaptic countervariables $M_{11} \sim B_{M,pq}$ are small and cannot be approximated by gaussians (where $B_{N,P}$ is defined below equation B.1). In particular for queries without any add noise, $\tilde{\kappa} = p_{01} = 0$, small M_{11} implies very large or even infinite synaptic weights (see equation 2.15) that would also violate the gaussian assumption for the distribution of dendritic potentials. As will be shown below, the Bayesian associative memory becomes equivalent to the Willshaw model with a decreased maximal network capacity $C_\epsilon \leq \ln 2 \approx 0.69$ (or, rather, $C_\epsilon \leq 1/(e \ln 2) \approx 0.53$ for independently generated address pattern components u_i^μ with binomially distributed pattern activities, $k^\mu := \sum_{i=1}^m u_i^\mu \sim B(m, p)$, as assumed here; see Willshaw et al., 1969; Knoblauch et al., 2010, appendix D). The following section investigates more closely the relationships to the Willshaw net, linear Hopfield–type learning rules, and the BCPNN model.

3 Relationships to Previous Models

3.1 Willshaw Model and Inhibitory Networks. The Willshaw or Steinbuch model is one of the simplest models for distributed associative

memory employing synapses with binary weights:

$$w_{ij} = \min(1, M_{11}(ij)) \in \{0, 1\}. \tag{3.1}$$

The dendritic potentials of the content neurons are simply $x_j = \sum_{i=1}^{m} w_{ij}\tilde{u}_i$. Exact potential distributions are well known and can be used to compute optimal firing thresholds (Palm, 1980; Buckingham & Willshaw, 1992, 1993; Knoblauch, 2008; Knoblauch et al., 2010).

The Willshaw model works particularly well for "pattern part retrieval" with zero add noise $\tilde{\kappa} = p_{01} = 0$. Then the active units of a query $\tilde{\mathbf{u}}$ are a subset of an address pattern $\mathbf{u}^\mu$ and the optimal threshold is maximal, that is, equal to the query pattern activity, $\Theta_j = \sum_{i=1}^{m} \tilde{u}_i$. Thus, a single missing query input, $\tilde{u}_i = 1$ but $w_{ij} = 0$, excludes activation of content neuron j. Based on this observation, it has been suggested that the Willshaw model should be interpreted as an essentially inhibitory network where zero weights become negative, positive weights become zero, and the optimal firing threshold becomes zero (Knoblauch, 2007). Such inhibitory implementations of the Willshaw network are very simple and efficient for a wide parameter range of moderately sparse memory patterns with $p \gg \log(n)/n$ where a small number of inhibitory synapses can store a large amount of information, $C^S \sim \log n$ bps, even for diluted networks with low connectivity $P < 1$. Moreover, the inhibitory interpretation offers novel functional hypotheses for strongly inhibitory circuits in the brain, for example, involving basket or chandelier cells (Markram et al., 2004). By contrast, the common excitatory interpretation is efficient only for very sparse memory patterns and cannot implement optimal threshold control in a simple and biologically plausible way (Buckingham & Willshaw, 1993; Graham & Willshaw, 1995).

The following arguments show that the inhibitory Willshaw network is actually a limit case of the optimal Bayesian associative memory in the nonlinear learning limit when synaptic coincidence counters are small, $\overline{M_{11}} = Mpq \ll \infty$, but unit usages are still large, $\overline{M_1'} = Mp \to \infty$, $\overline{M_1} = Mq \to \infty$. For pattern part retrieval with queries containing miss noise only, $p_{01|\mathfrak{a}} = 0$, the optimal Bayesian synaptic weights w_{ij} of equation 2.15 become

$$w_{ij} = \log \frac{M_{11}(1 - p_{10|1})(M_{00} + M_{10}p_{10|0})}{M_{10}(1 - p_{10|0})(M_{01} + M_{11}p_{10|1})} \approx \log \frac{M_{11}M_{00}}{M_{10}M_{01}}, \tag{3.2}$$

where the approximation is valid if query noise is independent of the content, $p_{10|0} = p_{10|1}$, and address patterns have sparse activity, $p \ll 1$, such that $M_{00} \gg M_{10}$ and $M_{01} \gg M_{11}$. In case $p_{10|0} \neq p_{10|1}$, the approximation is still

valid up to an additive offset, $w_0 := \log((1 - p_{10|1})/(p_{10|0}))$, where optimal retrieval can be implemented as described for equation 2.18.[2]

Thus, the optimal Bayesian model has strongly inhibitory weights, $w_{ij} = -\infty$, for $M_{11} = 0$ when the original Willshaw network would have zero weights. For sufficiently small $\overline{M_{11}} = Mpq \ll \infty$, the fraction of synapses with zero coincidence counters will be significant, $p_0 := \mathrm{pr}[M_{11} = 0] = (1 - pq)^M \approx \exp(-Mpq) \gg 0$, and, thus, the dendritic potentials will be dominated by the strongly inhibitory inputs. For still diverging unit usages, $\overline{M_1'} = Mp \to \infty$ and $\overline{M_1} = Mq \to \infty$, the remaining synaptic countervariables will be large and close to their mean values, $M_{00} \approx M(1 - p)(1 - q)$, $M_{01} \approx M(1 - p)q$, $M_{10} \approx Mp(1 - q)$, and therefore approximately equal for all synapses. Thus, up to an additive constant, the synaptic weights become

$$w_{ij} \approx \log M_{11} \tag{3.3}$$

corresponding to a nonlinear incremental Hebbian learning rule. At least for large $p_0 \to 1$, this rule will degenerate to the clipped Hebbian rule of the inhibitory Willshaw model where $w_{ij} = -\infty$ with probability $\mathrm{pr}[M_{11} = 0] = p_0$ and $w_{ij} = 0$ with probability $\mathrm{pr}[M_{11} = 1] \approx 1 - p_0$ whereas $\mathrm{pr}[M_{11} > 1] \approx 0$ becomes negligible. Since $p_0 \to 1$ is equivalent to $\overline{M_{11}} = Mpq \to 0$, this means that the Willshaw model becomes equivalent to Bayesian learning at least for $\max(1/p, 1/q) \ll M \ll 1/(pq)$ (see Figure 6, left panels). Numerical experiments suggest that the Willshaw model may be optimal even for smaller $p_0 \to 0.5$ corresponding to logarithmic pattern activity, $mp \to \tilde{\lambda}^{-1}\mathrm{ld}n$, where the Willshaw capacity becomes maximal, $C \to \tilde{\lambda} \ln 2 \approx 0.69\tilde{\lambda}$ bps, given that individual address pattern activities k^μ are narrowly distributed around mp (see Figure 7b; see also Knoblauch et al., 2010). For even smaller $p_0 < 0.5$ corresponding to $mp > \tilde{\lambda}^{-1}\mathrm{ld}n$ the Willshaw model cannot be optimal because then $C < 0.69\tilde{\lambda}$, whereas the capacity of the optimal Bayesian model increases toward $C \to 0.72\tilde{\lambda}$ bps.[3]

3.2 Linear Learning Models and the Covariance Rule. In general, the synaptic weights of the Bayesian associative network (see equation 2.15) are a nonlinear function of presynaptic and postsynaptic activity. This section shows that in the limit $\overline{M_{11}} = Mpq \to \infty$, the optimal Bayesian rule, equation 2.15, can be approximated by a linear learning rule,

$$w_{ij} = w_0 + r_{00}M_{00} + r_{01}M_{01} + r_{10}M_{10} + r_{11}M_{11}, \tag{3.4}$$

[2]For this, the offset w_0 should not depend on i.

[3]Note that $p_0 \to 0.5$ corresponds to $\overline{M_{11}} = Mpq \to \ln 2 \approx 0.69$. The same argumentation for independently generated address pattern components with binomially distributed $k^\mu \sim B_{m,p}$ would even suggest optimality until $p_0 \to 1/e \approx 0.37$ and $\overline{M_{11}} \to 1$ where the Willshaw model achieves the maximal capacity $C \to \tilde{\lambda}/(e \ln 2) \approx 0.53\tilde{\lambda}$ (see Knoblauch et al., 2010, eq. D.12).

with offset w_0 and learning increments r_{uv} specifying the change of synaptic weight when the presynaptic and postsynaptic neurons have activity $u \in \{0,1\}$ and $v \in \{0,1\}$, respectively. In fact, for diverging unit usages, $M_1, M_0 \to \infty$, the synapse usages will be close to expectation: $M_{00} \approx \overline{M_{00}} = M_0(1-p)$, $M_{01} \approx \overline{M_{01}} = M_1(1-p)$, $M_{10} \approx \overline{M_{10}} = M_0 p$, and $M_{11} \approx \overline{M_{11}} = M_1 p$. These approximations make only a negligible relative error if the standard deviations are small compared to the expectations. The most critical variable is the coincidence counter M_{11} having expectation $M_1 p$ and standard deviation $\sqrt{M_1 p(1-p)}$.[4] Thus, the approximations are valid for large values of the coincidence counter, that is, $\overline{M_{11}} \approx Mpq \to \infty$ for $q := M_1/M$. Then the argument of the logarithm in equation 2.15 will be close to

$$a_0 := \frac{(p(1-p_{10|1})+(1-p)p_{01|1})((1-p)(1-p_{01|0})+pp_{10|0})}{(p(1-p_{10|0})+(1-p)p_{01|0})((1-p)(1-p_{01|1})+pp_{10|1})} = \frac{d_1^* d_2^*}{d_3^* d_4^*}, \tag{3.5}$$

where $d_1^* := p(1-p_{10|1}) + (1-p)p_{01|1}$, $d_2^* := (1-p)(1-p_{01|0}) + pp_{10|0}$, $d_3^* := p(1-p_{10|0}) + (1-p)p_{01|0}$, and $d_4^* := (1-p)(1-p_{01|1}) + pp_{10|1}$. Linearizing the logarithm around a_0 yields

$$\begin{aligned} w_{ij} &\approx f(M_{00}, M_{01}, M_{10}, M_{11}) \\ &:= \log a_0 + \frac{\frac{(M_{11}(1-p_{10|1})+M_{01}p_{01|1})(M_{00}(1-p_{01|0})+M_{10}p_{10|0})}{(M_{10}(1-p_{10|0})+M_{00}p_{01|0})(M_{01}(1-p_{01|1})+M_{11}p_{10|1})} - a_0}{a_0} \\ &= \log a_0 + {a_0}^{-1}\frac{d_1 d_2}{d_3 d_4} - 1, \end{aligned} \tag{3.6}$$

where $d_1 := M_{11}(1-p_{10|1}) + M_{01}p_{01|1}$, $d_2 := M_{00}(1-p_{01|0}) + M_{10}p_{10|0}$, $d_3 := M_{10}(1-p_{10|0}) + M_{00}p_{01|0}$, and $d_4 := M_{01}(1-p_{01|1}) + M_{11}p_{10|1}$ for brevity. Similarly, the resulting function f can be linearized around the expectations of the synapse usages. This gives a learning rule of the form of equation 3.4 with offset $w_0 = \log a_0$ and

$$\begin{aligned} r_{00} &:= \frac{\partial f}{\partial M_{00}}|_{M_{uv}=\overline{M_{uv}}} = \frac{d_1}{a_0 d_4}\frac{(1-p_{01|0})d_3 - d_2 p_{01|0}}{{d_3}^2}|_{M_{uv}=\overline{M_{uv}}} \\ &= \frac{d_1^*}{d_3^* d_4^* a_0 M(1-q)}\left(1 - p_{01|0} - \frac{d_2^*}{d_3^*}p_{01|0}\right) = \eta_{00}pq, \end{aligned} \tag{3.7}$$

[4] Without loss of generality, $p := \mathrm{pr}[u_i^\mu] \le 0.5$ (otherwise, invert the address pattern components).

$$r_{01} := \frac{\partial f}{\partial M_{01}}|_{M_{uv}=\overline{M_{uv}}} = \frac{d_2}{a_0 d_3}\frac{p_{01}d_4 - d_1(1-p_{01})}{{d_4}^2}|_{M_{uv}=\overline{M_{uv}}}$$
$$= \frac{d_2^*}{d_3^* d_4^* a_0 Mq}\left(p_{01|1} - \frac{d_1^*}{d_4^*}(1-p_{01|1})\right) = -\eta_{01}\,p(1-q), \tag{3.8}$$
$$r_{10} := \frac{\partial f}{\partial M_{10}}|_{M_{uv}=\overline{M_{uv}}} = \frac{d_1}{a_0 d_4}\frac{p_{10}d_3 - d_2(1-p_{10})}{{d_3}^2}|_{M_{uv}=\overline{M_{uv}}}$$
$$= \frac{d_1^*}{d_3^* d_4^* a_0 M(1-q)}\left(p_{10|0} - \frac{d_2^*}{d_3^*}(1-p_{10|0})\right) = -\eta_{10}(1-p)q, \tag{3.9}$$
$$r_{11} := \frac{\partial f}{\partial M_{11}}|_{M_{uv}=\overline{M_{uv}}} = \frac{d_2}{a_0 d_3}\frac{p_{01}d_4 - d_1(1-p_{01})}{{d_4}^2}|_{M_{uv}=\overline{M_{uv}}}$$
$$= \frac{d_2^*}{d_3^* d_4^* a_0 Mq}\left(1 - p_{10|1} - \frac{d_1^*}{d_4^*}p_{10|1}\right) = -\eta_{11}(1-p)(1-q), \tag{3.10}$$

where

$$\eta_{00} := \frac{1}{Mpq(1-p)(1-q)} \times\left(\frac{(1-p)(1-p_{01|0})}{(1-p)(1-p_{01|0}) + pp_{10|0}} - \frac{(1-p)p_{01|0}}{p(1-p_{10|0}) + (1-p)p_{01|0}}\right),$$
$$\eta_{01} := \frac{1}{Mpq(1-p)(1-q)} \times\left(\frac{(1-p)(1-p_{01|1})}{(1-p)(1-p_{01|1}) + pp_{10|1}} - \frac{(1-p)p_{01|1}}{p(1-p_{10|1}) + (1-p)p_{01|1}}\right),$$
$$\eta_{10} := \frac{1}{Mpq(1-p)(1-q)} \times\left(\frac{p(1-p_{10|0})}{p(1-p_{10|0}) + (1-p)p_{01|0}} - \frac{pp_{10|0}}{(1-p)(1-p_{01|0}) + pp_{10|0}}\right) = \eta_{00},$$
$$\eta_{11} := \frac{1}{Mpq(1-p)(1-q)} \times\left(\frac{p(1-p_{10|1})}{p(1-p_{10|1}) + (1-p)p_{01|1}} - \frac{pp_{10|1}}{(1-p)(1-p_{01|1}) + pp_{10|1}}\right) = \eta_{01}. \tag{3.11}$$

If the query noise is independent of the contents, $p_{01} = p_{01|0} = p_{01|1}$ and $p_{10} = p_{10|0} = p_{10|1}$, then the four constants become identical, $\eta := \eta_{11} = \eta_{10} = \eta_{01} = \eta_{00}$, the offset becomes zero, $w_0 = 0$, and the

synaptic weight becomes

$$\frac{w_{ij}}{\eta} \approx pq\,M_{00} - p(1-q)M_{01} - (1-p)q\,M_{10} + (1-p)(1-q)M_{11}. \tag{3.12}$$

This is essentially (up to factor η) the linear covariance rule as discussed in much previous work (e.g., Sejnowski, 1977a, 1977b; Hopfield, 1982; Palm, 1988a, 1988b; Tsodyks & Feigel'man, 1988; Willshaw and Dayan, 1990; Dayan & Willshaw, 1991; Palm & Sommer, 1992, 1996; Dayan & Sejnowski, 1993; Chechik et al., 2001; Sterratt & Willshaw, 2008). Thus, together with the results of section 2.3, this shows that, in the asymptotic limit $\overline{M_{11}} = Mpq \to \infty$ with query noise being independent of contents, optimal Bayesian learning becomes equivalent to linear learning models employing the covariance rule. If query noise depends on contents, Bayesian learning differs from the covariance rule, but up to an additive offset, it still follows a linear learning rule.[5]

3.3 BCPNN-Type Models. The BCPNN rule is an early learning model for neural associative memory employing a Bayesian heuristics (Lansner & Ekeberg, 1987, 1989; Kononenko, 1989). The original rule is

$$\Theta_j = -\log \frac{M_1(j)}{M}, \tag{3.13}$$

$$\begin{aligned} w_{ij} &= \log \frac{M_{11}(ij)M}{M_1(j)M_1'(i)} \\ &= \log \frac{M_{11}(ij)(M_{00}(ij) + M_{01}(ij) + M_{10}(ij) + M_{11}(ij))}{(M_{01}(ij) + M_{11}(ij))(M_{10}(ij) + M_{11}(ij))}, \end{aligned} \tag{3.14}$$

where w_{ij} is the synaptic weight and, given a query $\tilde{\mathbf{u}}$, an output neuron will be activated, $\hat{v}_j = 1$, if the dendritic potential $x_j = \sum_{i=1}^{m} w_{ij}\tilde{u}_i$ exceeds the firing threshold Θ_j (see Lansner & Ekeberg, 1989, p. 79).

The following summarizes the main results of a technical report (Knoblauch, 2010a) comparing the BCPNN rule to the optimal Bayesian rule, equation 2.15. Obviously the two rules are not identical. The reason for this discrepancy is that Lansner and Ekeberg derived the BCPNN rule

[5]For example, if address "feature" $u_i = 1$ is positively correlated with content $v_j = 1$, then it typically occurs that $p_{10|1}(ij) < p_{10|0}(ij)$ and $p_{01|1}(ij) > p_{01|0}(ij)$, such that the optimal coincidence increment, $r_{11}(ij)$, is smaller than expected from the covariance rule, $\eta_{11}/\eta_{00} < 1$, whereas the offset is positive, $w_0(ij) > 0$. The deviation from the covariance rule can be significant, for example, $p = q = 0.1$, $\tilde{\lambda} = 0.75$, $\tilde{\kappa} = 0.25$ (corresponding to $p_{10} = 0.25$, $p_{01} = 0.025$), $p_{10|1} = 0.1p_{10}$, $p_{01|1} = 10p_{01}$ yields $\eta_{11}/\eta_{00} \approx 0.3$ and $w_0 \approx 1.8$.

from the following maximum likelihood decision,

$$\hat{v}_j = \begin{cases} 1, & \mathrm{pr}[v_j^\mu = 1 | \mathbf{1}_{\tilde{\mathbf{u}}}, \mathfrak{M}(j)] \\ & \quad = \dfrac{\mathrm{pr}[v_j^\mu = 1 | \mathfrak{M}(j)] \mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}} | v_j^\mu = 1, \mathfrak{M}(j)]}{\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}} | \mathfrak{M}(j)]} \geq 1/2 \,, \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

where $\mathbf{1}_{\tilde{\mathbf{u}}} := \{i : \tilde{u}_i = 1\}$ is the set of active query components. Thus, there are two main differences to the optimal Bayesian decision, equation 2.12. One is that the BCPNN model considers only active query components $i \in \mathbf{1}_{\tilde{\mathbf{u}}}$ and ignores inactive components $i \in \mathbf{0}_{\tilde{\mathbf{u}}} := \{i : \tilde{u}_i = 0\}$. In contrast, the optimal Bayesian model considers both active and inactive query components. Second, the BCPNN model needs to compute $\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}} | \mathfrak{M}(j)]$, which becomes viable only by wrongly assuming that the query components would be independent of each other, that is, by using

$$\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}} | \mathfrak{M}(j)] \approx \prod_{i \in \mathbf{1}_{\tilde{\mathbf{u}}}}^{m} \mathrm{pr}[\tilde{u}_i = 1 | \mathfrak{M}(j)] = \prod_{i \in \mathbf{1}_{\tilde{\mathbf{u}}}} \frac{M_1'(i)}{M}. \tag{3.16}$$

This approximation is inaccurate because the query components given the storage variables depend on each other even for independently generated query components with $\mathrm{pr}[\tilde{\mathbf{u}}] = \prod_i \mathrm{pr}[\tilde{u}_i]$. For example, consider the following simple network motif of two input units, $m = 2$, and a single output unit, $n = 1$. After storing M memories, let

$$M_{10}(1) = 0, \quad M_{11}(1) = 1, \quad M_{10}(2) = 1, \quad M_{11}(2) = 0, \quad M_1 = 1, \tag{3.17}$$

where, for brevity, the indices are skipped for the output unit. Then, for zero query noise, it is $\mathrm{pr}[\tilde{u}_1 = 1 | \mathfrak{M}(j)] > 0$, but $\mathrm{pr}[\tilde{u}_1 = 1 | \tilde{u}_2 = 1, \mathfrak{M}(j)] = 0$. Note that the optimal Bayesian model avoids this problem by computing the odds ratio $\mathrm{pr}[v_j^\mu = 1 | \tilde{\mathbf{u}}, \mathfrak{M}(j)] / \mathrm{pr}[v_j^\mu = 1 | \tilde{\mathbf{u}}, \mathfrak{M}(j)]$ such that $\mathrm{pr}[\tilde{\mathbf{u}} | \mathfrak{M}(j)]$ cancels.

Appendix H generalizes the BCPNN rule for noisy queries and describes two improved BCPNN-type rules, each of them fixing one of the two problems described: the BCPNN2 rule (see equation H.9), includes inactive query components but still uses an approximation similar to equation 3.16, and the BCPNN3 rule (see equation H.12) does not employ approximation equation 3.16, but still ignores inactive query components. For the latter, it is possible to compute the SNR in analogy to section 2.3. It turns out that in the linear learning regime, $\overline{M_{11}} = Mpq \to \infty$, the squared SNR R^2 (and thus also storage capacity M_ϵ and C_ϵ) is factor $1 - p(\tilde{\lambda} + \tilde{\kappa})$ below the optimal value equation 2.28. This implies also that the original BCPNN rule performs at least factor $1 - p(\tilde{\lambda} + \tilde{\kappa})$ worse than the optimal Bayesian rule and

thus, at most, is equivalent to the suboptimal linear homosynaptic rule (e.g., see rule R3 in Dayan & Willshaw, 1991). In the complementary nonlinear regime $\overline{M_{11}} \ll \infty$ corresponding to very sparse patterns, similar arguments as in section 3.1 show that the BCPNN model becomes equivalent to optimal Bayesian learning and the Willshaw model.

4 Results from Simulation Experiments

This section has two purposes: to verify the theoretical results and compare the performances of the different learning models. To this end, I have implemented associative memory networks with optimal Bayesian learning (see section 2.2), BCPNN-type learning (see appendix H and section 3.3), linear learning (see appendix G and section 3.2), and Willshaw-type clipped Hebbian learning (see section 3.1). All experiments assume full network connectivity ($P = 1$).

4.1 Verification of SNR R. A first series of experiments illustrated by Figure 4 implemented networks of size $m = n = 1000$ and compared experimental SNR R of dendritic potentials (black curves; see equation 2.24) to the theoretical values (gray curves). Here the theoretical values have been computed from equation 2.28 (Bayes), equations G.7 to G.9 (linear), and equation H.21 (BCPNN3). Data correspond to four experimental conditions testing sparse versus nonsparse memory patterns and queries having miss noise versus add noise. For each condition, the corresponding plot shows SNR R as a function of stored memories M. All experiments assumed ideal conditions where each query pattern $\tilde{\mathbf{u}}$ was generated from an address pattern $\mathbf{u}^\mu$ having $k = pm$ one-entries, where $\tilde{\mathbf{u}}$ contained $c = \tilde{\lambda}k$ correct one-entries and $f = \tilde{\kappa}k$ false one-entries (see Figure 2, left). Furthermore, all tested content neurons had unit usages $M_1 = Mq$.

For most conditions and models, the theoretical predictions match the experimental SNR very well. This is true in particular for the three tested linear models (Hebb rule, homosynaptic rule, and covariance rule), but also for the Bayesian and BCPNN-type rules if the mean value of the coincidence counter is sufficiently large, $\overline{M_{11}} = Mpq \gg 1$, as presumed at the beginning of section 2.3. For example, for nonsparse patterns, the theoretical results become virtually exact for $M > 70$ or $\overline{M_{11}} > 70/4 = 17.5$. For fewer coincidences, $\overline{M_{11}} \not\gg 1$, the SNR curves of the Bayesian and BCPNN-type models are similar as for the Willshaw model. Here the SNR is not a good predictor of retrieval quality and cannot easily be compared to the regime with $\overline{M_{11}} \gg 1$ for the following reasons. First, variances of dendritic potentials between high and low units become significantly different, $\sigma_{\text{hi}} \not\approx \sigma_{\text{lo}}$ (cf. equation 2.26).[6] Second, the distributions of dendritic potentials become

[6] For example, $\sigma_{\text{hi}} = 0$ for pattern part retrieval in the Willshaw model (see section 3.1).

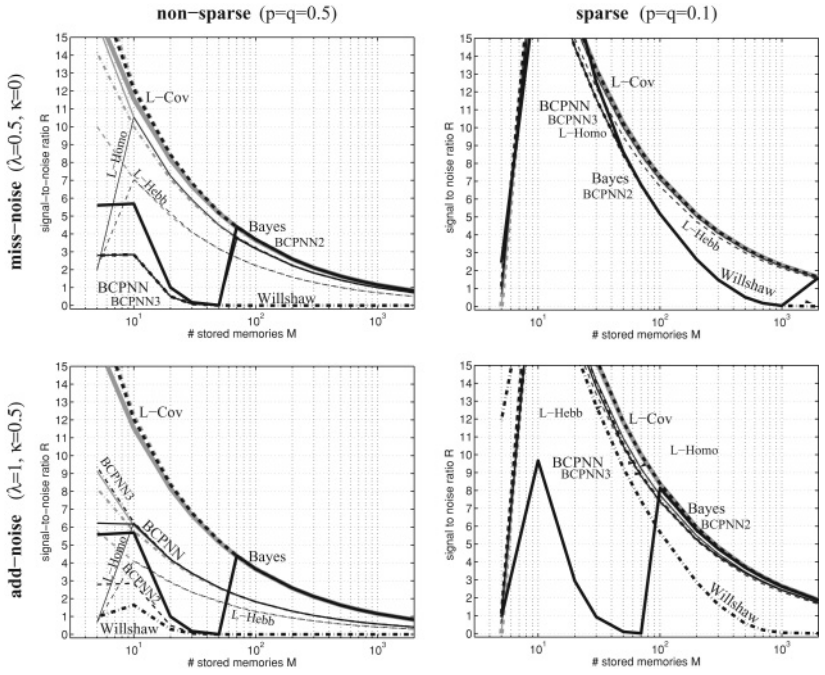

Figure 4: Verification and comparison of SNR R for different learning models (see equation 2.24). Each plot shows SNR R as a function of stored memories M for a network of size $m = n = 1000$, with data from simulation experiments (black) and theory (gray). Individual curves correspond to the optimal Bayesian model (thick solid; see section 2.2, equation 2.28), linear covariance rule (thick dashed; see appendix G), Willshaw model (thick dash-dotted; see section 3.1), BCPNN rule (medium solid; see section H.1), BCPNN2 rule (medium dashed; see section H.2), BCPNN3 rule (medium dash-dotted; see sections H.3 and H.4, equation H.21), linear homosynaptic rule (thin solid; see appendix G), and the linear Hebb rule (thin dashed; see appendix G). Top panels correspond to pattern part retrieval with miss noise only ($\tilde{\lambda} = 0.5$, $\tilde{\kappa} = 0$). Bottom panels correspond to queries including add noise ($\tilde{\lambda} = 1$, $\tilde{\kappa} = 0.5$). Left panels correspond to nonsparse memory patterns with $p = q = 0.5$. Right panels correspond to (moderately) sparse patterns with $p = q = 0.1$. Each data value averages over 10,000 networks, each tested with a single query under ideal theoretical conditions (see text).

nongaussian (Knoblauch, 2008; cf. appendix E). Third, in particular for very small $\overline{M_{11}} \ll 1$, dendritic potentials may be contaminated by infinite synaptic inputs (see equations 2.15, 3.2, and 3.14). This reasoning also explains the nonmonotonicity of the SNR curves visible in Figure 4 for the Bayesian

and BCPNN-type models as a transition from a nonlinear Willshaw-type to a linear covariance-type regime of operation.

4.2 Verification of Output Noise $\hat{\epsilon}$. In a second step, I verified the theory for output noise $\hat{\epsilon}$ (see equation 2.30) as described in appendix E using the same network implementations as described before. In fact, appendix E shows that there is a bijective relation between the SNR R and (minimal) output noise $\hat{\epsilon}$ if the dendritic potentials are gaussian and the high and low potentials have identical variances. Thus, given that the theory of SNR is correct, here it is tested whether these two conditions hold true.

Figure 5 shows output noise $\hat{\epsilon}$ as a function of stored memories M assuming the same conditions as described for Figure 4. As before, for most conditions and models, the theoretical predictions match the experimental $\hat{\epsilon}$ very well. In fact, the match is good even for the Bayesian and BCPNN-type rules when assuming relatively small $\overline{M_{11}}$ where the theoretical estimates of SNR are still inaccurate. Again, the theory is inaccurate only for the Bayesian and BCPNN-type models for the condition of sparse memories and miss noise only. Here the theory basically suggests equivalence to the linear covariance rule, whereas the Bayesian and BCPNN-type models perform much better due to the infinitely negative synaptic weights caused by the $M_{11} = 0$ events, which allow rejecting a neuron activation by a single presynaptic input.

4.3 Verification of Storage Capacity M_ϵ. A further series of experiments illustrated by Figure 6 tested the theory of storage capacity M_ϵ (see equations 2.32 and 2.33) for different network sizes $m = n = 100, 1000, 10{,}000$, a larger range of pattern activities mp ($=nq$), and relaxing the restrictive assumption of having fixed k, c, f, M_1. This means that a query pattern was generated by randomly selecting one of the M address patterns $\mathbf{u}^\mu$ and applying query noise according to parameters $p_{10} = 1 - \tilde{\lambda}$ and $p_{01} = \tilde{\kappa}\, p/(1-p)$. Similarly, all content neurons were included in the analysis. Thus, the previously fixed parameters became binomials, $k \sim B_{m,p}$, $c \sim B_{m,\tilde{\lambda}p}$, $f \sim B_{m,\tilde{\kappa}p}$, $M_1 \sim B_{M,q}$, where $B_{N,P}$ is as explained below equation B.1.

Each plot shows output noise $\hat{\epsilon}$ as a function of mean pattern activity mp. For each value of mp, the number of stored patterns, M_ϵ, was computed from equation 2.33 for the optimal Bayesian rule and a low-output noise level $\epsilon = 0.01$ (see parameter sets 1–6 in Table 1). For small networks ($m = n = 100$; upper panels) the theory is generally inaccurate. For example, for the optimal Bayesian learning rule, the theory strongly overestimates storage capacity for sparse memory patterns and underestimates capacity for nonsparse patterns. For larger networks (middle and bottom panels), there is a large range of mp where the theory precisely predicts storage capacity. Only for very sparse memory patterns (with small $\overline{M_{11}} \ll 1$) does the theory remain inaccurate. For queries containing add noise, the theory

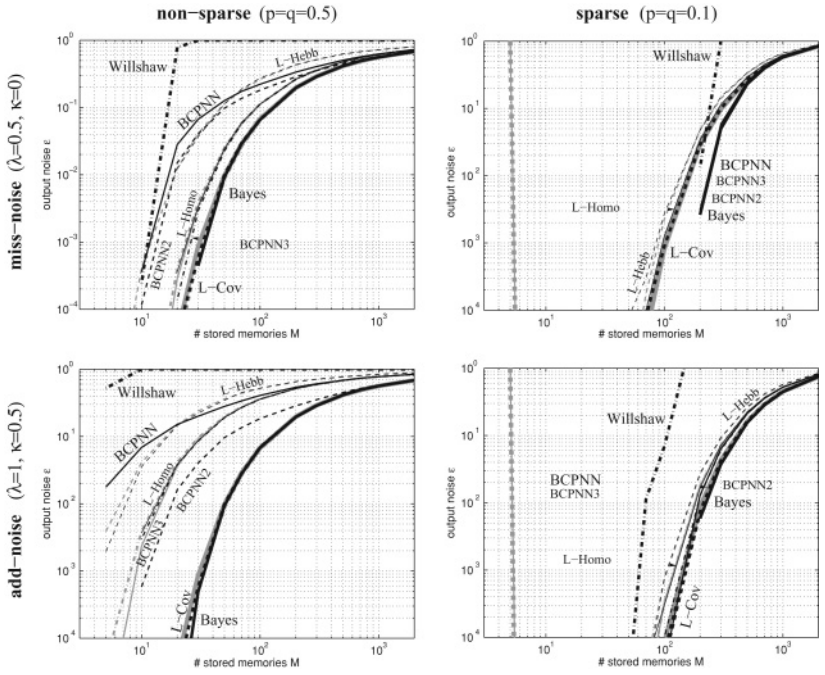

Figure 5: Verification and comparison of output noise $\hat{\epsilon}$ for different learning models (see equation 2.30). Each plot shows $\hat{\epsilon}$ as a function of stored memories M for a network of size $m = n = 1000$, including data from simulation experiments (black) and theory (gray; see equation E.6). Individual curves correspond to the optimal Bayesian model (thick solid; see section 2.2, equation 2.28), linear covariance rule (thick dashed; see appendix G), Willshaw model (thick dash-dotted; see section 3.1), BCPNN rule (medium solid; see section H.1), BCPNN2 rule (medium dashed; see section H.2), BCPNN3 rule (medium dash-dotted; see sections H.3 and H.4 and equation H.21), linear homosynaptic rule (thin solid; see appendix G), and the linear Hebb rule (thin dashed; see appendix G). Top panels correspond to pattern part retrieval with miss noise only ($\tilde{\lambda} = 0.5$, $\tilde{\kappa} = 0$). Bottom panels correspond to queries including add noise ($\tilde{\lambda} = 1$, $\tilde{\kappa} = 0.5$). Left panels correspond to nonsparse memory patterns with $p = q = 0.5$. Right panels correspond to (moderately) sparse patterns with $p = q = 0.1$. Each data value averages over 10,000 networks each tested with a single query under ideal theoretical conditions (see text; same data as in Figure 4).

generally overestimates true capacity. For queries containing only miss noise, the theory overestimates capacity for extremely sparse patterns but underestimates capacity for patterns with intermediate sparseness.

For larger networks and $\overline{M_{11}} \gg 1$, the theory becomes very precise for the optimal Bayes rule, the BCPNN3 rule, and the linear covariance rule.

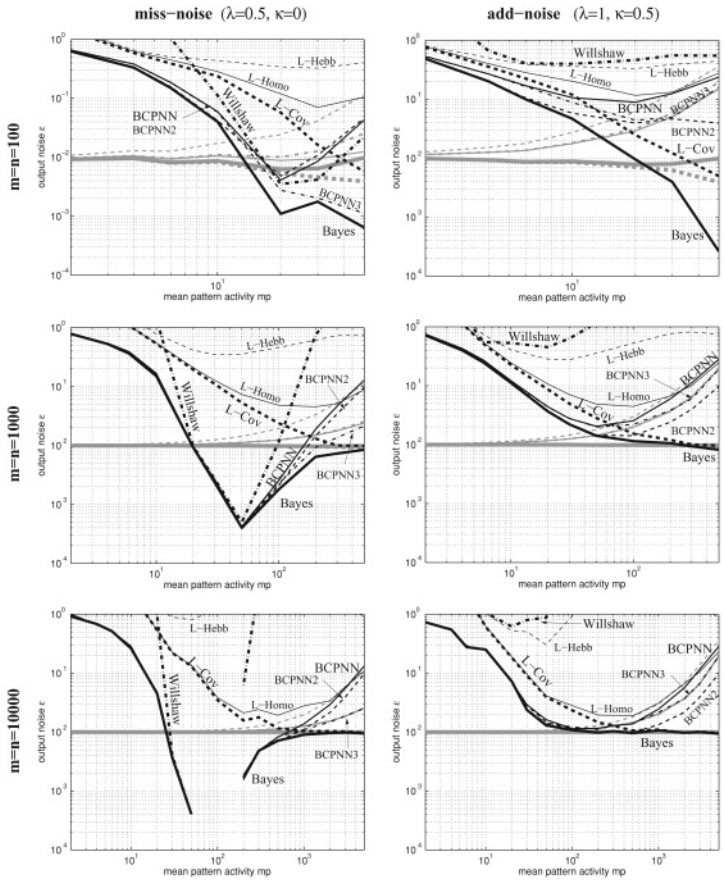

Figure 6: Verification and comparison of pattern capacity M_ϵ for different learning models (see equation 2.32). Each plot shows output noise $\hat{\epsilon}$ as function of mean pattern activity $mp = nq$ when storing memories at the theoretical capacity limit M_ϵ of Bayesian learning for low output noise (equation 2.33, $\epsilon = 0.01$; see parameter sets 1–6 in Table 1). Plots show data from simulation experiments (black; see equation 2.30) and theory (gray; see equation E.6). Individual curves correspond to the optimal Bayesian model (thick solid; see section 2.2, equation 2.28), linear covariance rule (thick dashed; see appendix G), Willshaw model (thick dash-dotted; see section 3.1), BCPNN rule (medium solid; see section H.1), BCPNN2 rule (medium dashed; see section H.2), and BCPNN3 rule (medium dash-dotted; see sections H.3 and H.4), linear homosynaptic rule (thin solid; see appendix G), and the linear Hebb rule (thin dashed; see appendix G). Left panels correspond to pattern part retrieval with miss noise only ($\tilde{\lambda} = 0.5$, $\tilde{\kappa} = 0$). Right panels correspond to queries including add noise ($\tilde{\lambda} = 1$, $\tilde{\kappa} = 0.5$). Top panels correspond to small networks with $m = n = 100$. Middle panels correspond to medium networks with $m = n = 1000$. Bottom panels correspond to larger networks with $m = n = 10{,}000$. Each data value averages over 10,000 retrievals in 100 networks storing random patterns with independent components.

Table 1: Theoretical Pattern Capacities M_ϵ at Output Noise Level $\epsilon = 0.01$ for Optimal Bayesian Learning (Parameter Sets 1–6) and the Willshaw Model (Parameter Set 7) as Employed for the Simulation Experiments Illustrated by Figures 4 to 7.

$mp = nq$	M_ϵ at $\epsilon = 0.01$ parameter set 1 Bayes $m = n = 100$ $\tilde{\lambda} = 0.5, \tilde{\kappa} = 0$	2 $\tilde{\lambda} = 1, \tilde{\kappa} = 0.5$	3 $m = n = 1000$ $\tilde{\lambda} = 0.5, \tilde{\kappa} = 0$	4 $\tilde{\lambda} = 1, \tilde{\kappa} = 0.5$	5 $m = n = 10000$ $\tilde{\lambda} = 0.5, \tilde{\kappa} = 0$	6 $\tilde{\lambda} = 1, \tilde{\kappa} = 0.5$	7 Willshaw $m = n = 1000$ $\tilde{\lambda} = 0.5, \tilde{\kappa} = 0$
2	63	85	5371	7161	468,070	624,093	6
4	34	45	2815	3753	243,308	324,411	315
6	23	31	1932	2577	166,089	221,453	988
10	15	20	1206	1608	102,781	137,042	1578
20	8	11	639	853	53,710	71,613	1252
30	6	8	443	591	36,794	49,059	851
50	5	5	281	374	22,886	30,514	448
100			154	205	12,063	16,084	156
200			88	116	6399	8531	47
300			66	84	4435	5912	22
500			50	50	2813	3749	9
1000					1546	2056	
2000					886	1163	
3000					664	845	
5000					502	502	

Notes: Data assume various network sizes $m = n$, mean pattern activities $mp = nq$, and query noise parameters $\tilde{\lambda}, \tilde{\kappa}$. Capacities for the Bayesian model have been computed from equation 2.33 (assuming independent pattern components). Capacities for the Willshaw model have been computed from Knoblauch et al. (2010, eq. 57) and are exact for fixed pattern activities $k = mp$ (whereas independent memory components would imply $M_\epsilon = 1$ for a large range of sparse memory patterns; cf., Knoblauch et al., 2010, eq. 65).

In contrast, even for $m = n = 10{,}000$ and $pm > 1000$, the theory for the linear homosynaptic rule underestimates output noise $\hat{\epsilon}$ by about a factor of two. The underestimation of $\hat{\epsilon}$ is even worse for the linear Hebbian rule. Here the reasoning is that in contrast to covariance and homosynaptic rule, the mean synaptic weight $\bar{w}_{ij}/M = r_{00}(1-p)(1-q) + r_{01}(1-p)q + r_{10}p(1-q) + r_{11}pq$ is nonzero for the Hebbian rule. Therefore inhomogeneities in c, f, and k can cause a much larger variance in dendritic potentials than predicted by the theory, assuming fixed given values for c, f, and k.

4.4 Comparison of the Different Learning Models. The simulation experiments confirm that the Bayesian learning rule is the general optimum leading to maximal SNR, minimal output noise, and highest storage capacity. Nevertheless, the simulations show also that for particular parameter ranges, some of the previous learning models can also become optimal.

The linear covariance rule becomes optimal in the linear learning regime, $\overline{M_{11}} = Mpq \to \infty$, which, for given output noise level $\hat{\epsilon}$, corresponds to moderately sparse or nonsparse memory patterns with $mp/\ln q \to \infty$ (see equations 2.35 and 2.37). However, for sparse memory patterns of finite size, the linear rules can perform much worse than the optimal Bayesian model—even worse than the Willshaw model.

Similarly, the BCPNN-type models become optimal in the limit of sparse query activity, $p(\tilde{\lambda} + \tilde{\kappa}) \to 0$. For finite size or nonsparse query patterns, the storage capacity can be significantly (factor $1 - p(\tilde{\lambda} + \tilde{\kappa})$) below the optimal value.

Finally, the Willshaw model becomes optimal only for pattern part retrieval ($\tilde{\kappa} = 0$) and few coincidence counts, $\overline{M_{11}} \ll \infty$ corresponding to very sparse memory patterns with $mp = O(\ln q)$. For finite networks, the Willshaw model achieves the performance of the Bayesian model only if the output noise level $\hat{\epsilon}$ is low and the address pattern activities k^μ are constant or narrowly distributed around mp. In all other cases, the Willshaw model performs much worse than the optimal Bayesian rule.

4.5 Further Results Concerning Memory Statistics and Retrieval Methods. Figure 7 shows additional simulation experiments testing the various learning models for different retrieval methods and different ways of generating random patterns ($m = n = 1000$ and pattern part retrieval with $\tilde{\lambda} = 0.5$, $\tilde{\kappa} = 0$). Since the Bayesian theory can strongly overestimate pattern capacity M_ϵ for very sparse memory patterns (see equation 2.37), memories were stored at the much lower capacity limit of the Willshaw model assuming a fixed pattern activity $k^\mu = mp$ for all memories (see equation 57 in Knoblauch et al., 2010; see parameter set 7 in Table 1). Then testing the networks again with random patterns having independent components (and binomial activity $k^\mu \sim B_{m,p}$) yields qualitatively similar results as before (compare the top left panel of Figure 7 to the middle left panel of

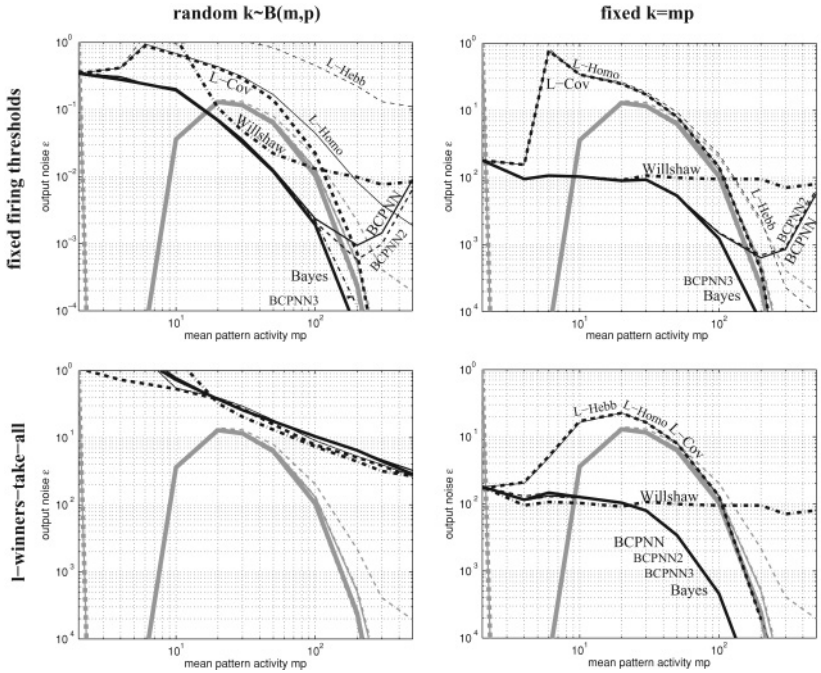

Figure 7: Effect of memory statistics and retrieval method on the performance of different learning models. Each plot shows output noise $\hat{\epsilon}$ as a function of mean pattern activity $mp = nq$ when storing memories at the theoretical capacity limit M_ϵ of the Willshaw model ($\epsilon = 0.01$; see parameter set 7 in Table 1) assuming network size $m = n = 1000$ and queries containing miss noise only ($\tilde{\lambda} = 0.5$, $\tilde{\kappa} = 0$). Plots show data from simulation experiments (black; see equation 2.30) and theory (gray; see equation E.6). Individual curves correspond to the optimal Bayesian model (thick solid; see section 2.2, equation 2.28), linear covariance rule (thick dashed; see appendix G), Willshaw model (thick dash-dotted; see section 3.1), BCPNN rule (medium solid; see section H.1), BCPNN2 rule (medium dashed; see section H.2), BCPNN3 rule (medium dash-dotted; see sections H.3, and H.4), linear homosynaptic rule (thin solid; see appendix G), and the linear Hebb rule (thin dashed; see appendix G). Left panels correspond to random memory patterns with independently generated components, that is, $k^\mu := \sum_{i=1}^{m} u_i^\mu$ follows a binomial distribution, $k^\mu \sim B_{m,p}$. Right panels correspond to random memory patterns with a fixed pattern activity $k^\mu = mp$. Top panels correspond to fixed optimal firing thresholds Θ_j (see appendix D). Bottom panels correspond to l-winners-take-all retrieval activating the $l := nq$ neurons having the largest dendritic potentials x_j. Each data value averages over 10,000 retrievals in 100 networks.

Figure 6). Further simulations suggest that the Bayesian and BCPNN-type models have a high-fidelity capacity for very sparse patterns that is almost as low as reported for the Willshaw model (basically $M_\epsilon = 1$ for $\epsilon \ll 1$ and $k/\log n \to 0$; see appendix D in Knoblauch et al., 2010).

In contrast, for random patterns with fixed activity $k^\mu = mp$, the Bayesian and BCPNN-type models perform equivalent to the Willshaw model for a large range of sparse patterns (see Figure 7, top right panel). Moreover, for less sparse patterns, BCPNN2 becomes equivalent to the BCPNN rule, and BCPNN3 becomes equivalent to optimal Bayesian learning. There is also a strong improvement for the linear homosynaptic and Hebb rules now closely matching the theoretical values (for independent pattern components and binomial k^μ) where the homosynaptic rule becomes equivalent to the covariance rule.

So far, retrieval used fixed firing thresholds to minimize output noise (see appendix D). A simple alternative is $\bar{l}$-winners-take-all (WTA) retrieval activating the $\bar{l} := nq$ neurons with the largest dendritic potentials x_j (as may be implemented in the brain by recurrent inhibition, for example).[7] Figure 7 (bottom left panel) shows simulation results for $\bar{l}$-WTA and memory patterns with independent components and binomial $k^\mu \sim B_{m,p}$. Surprisingly, all of the various learning models show almost identical performance at relatively high levels of output noise $\hat{\epsilon}$. There are two reasons that can partly explain this result. First, $\bar{l}$-WTA cannot achieve high fidelity with $\hat{\epsilon} \to 0$ because the content patterns $\mathbf{v}^\mu$ have a distributed pattern activity $l^\mu \sim B_{n,q}$ which is unknown beforehand. Thus, activating the $\bar{l}$ most excited units causes a positive baseline level of output noise. Second, storing patterns at the relatively low-capacity limit of the Willshaw model implies, for fixed thresholds, low output noise for all models. Therefore, the actual output noise for $\bar{l}$-WTA will be dominated by the baseline errors described. Nevertheless, further simulations confirmed that even for a larger number of stored patterns, the performances of the different models are much more similar than for fixed firing thresholds.

For l-WTA and fixed pattern activity $l^\mu = nq$ the performance generally improves (Figure 7, bottom right panel). As before, l-WTA seems to even out the performance differences of various synaptic learning models: Surprisingly, the linear Hebbian, homosynaptic, and covariance rule now show identical high performance, precisely matching the theoretical values for the covariance rule. Also the Bayesian and BCPNN-type rules show identical performance. Further simulations show that for queries including add noise ($\tilde{\kappa}$, $p_{01} > 0$), l-WTA retrieval becomes identical even between the Bayesian-type and linear model groups. These results support the view that homeostatic mechanisms, such as regulating total activity level, may play an equally important role as tuning the synaptic learning parameters

[7] Although l-WTA retrieval is simple to implement, it is much more difficult to analyze.

(Turrigiano, Leslie, Desai, Rutherford, & Nelson, 1998; Van Welie, Van Hooft, & Wadman, 2004; Chechik et al., 2001; Knoblauch, 2009c).

5 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; Lansner, 2009), 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, Hagstroem, & Palm, 1989; Prager & Fallside, 1989; Greene, Parnas, & Yao, 1994; Huyck & Orengo, 2005; Knoblauch, 2005; Mu, Artiklar, Watta, & Hassoun, 2006; Wichert, 2006; Rehn & Sommer, 2006).

In this paper, I have developed and analyzed the generally optimal neural associative memory that minimizes the Hamming-distance-based output noise $\hat{\epsilon}$ and maximizes pattern capacity M_ϵ and network storage capacity C_ϵ by making Bayesian maximum likelihood considerations. In general, the resulting optimal synaptic learning rule, equation 2.15 is nonlinear and asymmetric, and it differs from previously investigated linear learning models of the Hopfield type, simple nonlinear learning models of the Willshaw type, and BCPNN-type Bayesian learning heuristics. As revealed by detailed theoretical and experimental comparisons, the previous models are rather special cases of Bayesian learning that becomes optimal only in the asymptotic limit of large networks and for particular ranges of pattern activity p, q and query noise $\tilde{\lambda}, \tilde{\kappa}$ (see Table 2).

For example, the Willshaw model becomes optimal only in the limit of small coincidence counters, $\overline{M_{11}} = Mpq \leq 1$, for queries without any add noise, $\tilde{\kappa} = p_{01} = 0$. For maximal $M = M_\epsilon$ at the capacity limit, $\overline{M_{11}} \ll \infty$ can be achieved only for extremely sparse memory patterns where the number of active units per memory vector scales typically logarithmic in the population size, for example, $p, q \sim \log n/n$ (Knoblauch et al., 2010). Nevertheless, one may be surprised how a simple model employing binary synapses can already perform optimal Bayesian retrieval. The reason is that a low value of $\overline{M_{11}} = Mpq$ guarantees that a large fraction $p_0 := (1 - pq)^M$ of synaptic weights remains zero in the Willshaw model or minus infinity in the corresponding Bayesian interpretation (see equation 3.2). Then retrieval gets dominated by rejecting activations of postsynaptic neurons based on single but strongly inhibitory inputs. In particular, for small but nonvanishing p_0, the inhibitory Willshaw network becomes very efficient by storing large amounts of information with a small number of synapses (Knoblauch, 2007). Such an inhibitory interpretation of associative memory may also offer novel functional hypotheses for strongly inhibitory cortical circuits, for example, involving chandelier or basket cells (Markram et al., 2004), and

Table 2: Asymptotic Conditions When the Various Learning Rules Become Optimal (Equivalent to the Bayesian Rule).

Learning Rule	General Conditions for Optimality	Conditions at Capacity Limit $M = M_\epsilon$
Optimal Bayesian	None	None
BCPNN type	$p \to 0$	$p \to 0$
Linear covariance	$\overline{M_{11}} \to \infty$	$(mp)/\log m \to \infty$
Linear homosynaptic	$\overline{M_{11}} \to \infty$ and $p \to 0$	$(mp)/\log m \to \infty$ and $p \to 0$
Linear heterosynaptic	$\overline{M_{11}} \to \infty$ and $q \to 0$	$(mp)/\log m \to \infty$ and $q \to 0$
Linear Hebb	$\overline{M_{11}} \to \infty$ and $p, q \to 0$	$(mp)/\log m \to \infty$ and $p, q \to 0$
Linear Hopfield	$\overline{M_{11}} \to \infty$ and $p, q \to 0.5$	$p, q \to 0.5$
Willshaw	$\overline{M_{11}} \le 1$ and $\overline{M_1}, \overline{M_1'} \to \infty$ and $\tilde{\kappa} \to 0$	$mp \sim \log m$ and $\tilde{\kappa} \to 0$

Notes: The constraints depend on the fraction of active units in an address pattern ($p := \mathrm{pr}[u_i^\mu = 1]$) or content pattern ($q := \mathrm{pr}[v_j^\mu = 1]$), the size of the address population (m), the mean value of the synaptic coincidence counter ($\overline{M_{11}} = Mpq$, where M is the number of stored memories), the mean unit usages ($\overline{M_1} = Mq$, $\overline{M_1'} = Mp$), and the fraction of add noise in the query pattern ($\tilde{\kappa}$). The right column reexpresses the general conditions of the middle column for the case when M equals the pattern capacity M_ϵ.

also for inhibition-dominated brain structures such as cerebellum and basal ganglia (Marr, 1969; Albus, 1971; Kanerva, 1988; Wilson, 2004).

In contrast to the Willshaw model, the linear covariance rule becomes optimal in the linear learning regime where the synaptic coincidence counters diverge, $\overline{M_{11}} = Mpq \to \infty$. Then linearization of the optimal Bayesian rule yields the covariance rule, and the two rules have the same asymptotic SNR. Correspondingly, the fraction of synapses with infinite weights vanishes, $p_0 \to 0$, which, at the capacity limit $M = M_\epsilon$ (see equation 2.33), corresponds to moderately or nonsparse memory patterns with typically $p, q \gg \log n/n$. Numerical experiments indicate that in reasonably large but finite networks, the optimal Bayesian model still performs significantly better than the linear covariance rule for a large range of pattern activities $p \ll 0.5$. Furthermore, the SNR analysis allows a characterization of basins of attraction in terms of miss noise and add noise (see equation 2.29 and Figure 2, right). It turns out that in the linear learning regime, $\overline{M_{11}} \to \infty$, the network is more vulnerable against miss noise ($\tilde{\lambda} < 1$) than add noise ($\tilde{\kappa} > 0$). This contrasts with the nonlinear learning regime, $\overline{M_{11}} \ll \infty$, where the network is more vulnerable against add noise, mainly because add noise destroys the network's ability to reject postsynaptic activations by single strongly inhibitory synaptic inputs. Alternative linear learning models such as the Hebb, homosynaptic, and heterosynaptic rules behave similar to the covariance rule but have a lower signal-to-noise ratio unless $p \to 0$ and/or $q \to 0$ (Dayan & Willshaw, 1991).

The original BCPNN model of Lansner and Ekeberg has a similar formulation as the optimal Bayesian model but neglects inactive query neurons and employs an inaccurate approximation (see equation 3.16). More recent hypercolumnar variants of the BCPNN model for discrete valued memories remedy the first problem by employing extra neurons to represent inactivity (Lansner & Holst, 1996; Johansson, Sandberg, & Lansner, 2002), but require (at least) double the network size of the optimal Bayesian model. For comparison, I have extended the original BCPNN model to include query noise and derived two improved BCPNN-type rules: The BCPNN2 rule also considers the inactive query neurons, whereas the BCPNN3 rule does not make use of the inaccurate approximation. Similar to the Willshaw model, the BCPNN-type rules become optimal at least in the nonlinear learning regime, $\overline{M_{11}} \ll \infty$, corresponding to very sparse patterns where active units dominate the total information contained in a query pattern. Moreover, for the linear learning regime $\overline{M_{11}} = Mpq \to \infty$, I have analyzed the SNR of the BCPNN3 rule being an upper bound for the original BCPNN rule. The analysis revealed that the SNR of the BCPNN3 model is equivalent to the linear homosynaptic rule, that is, factor $1 - p(\tilde{\lambda} + \tilde{\kappa})$ worse than for optimal Bayesian learning (see also Dayan & Willshaw, 1991). Thus, the original BCPNN rule achieves at most the capacity of the homosynaptic rule and becomes optimal only for sparse address patterns with $p \to 0$ or low query activity with small $\tilde{\lambda} + \tilde{\kappa} \to 0$. Even for sparse address patterns with $p \to 0$, the BCPNN-type models have reduced basins of attraction in the sense that they are more vulnerable to add noise with large $\tilde{\kappa} \gg 0$ than the optimal Bayesian model.

MacKay (1991) has suggested a learning model based on maximizing the entropy of synaptic weights that is closely related to optimal Bayesian associative memory. In particular, he arrived at a similar learning rule and also discussed the convergence to the covariance rule as well as the necessity of infinite synaptic weights. The current approach goes beyond these previous results by generalizing the learning rule for query noise and providing an SNR analysis for Bayesian learning. The latter, in connection with the results of appendix E, rigorously proves the equivalence of Bayesian learning and the covariance rule in the limit $\overline{M_{11}} \to \infty$ (whereas Taylor expansion of the BCPNN rule, for example, also leads to the covariance rule in spite of BCPNN being suboptimal; see section H.4). Moreover, this analysis also discusses convergence of the Bayesian learning rule to linear learning rules other than the covariance rule when the query noise is not independent of the stored contents (as can be expected for any real-world data).

As with most previous approaches, the "optimal" Bayesian memory model still makes the naive assumption that address attributes are independent of each other. Although this assumption is almost never fulfilled in real-world data, experiments reveal that naive Bayesian classifiers perform surprisingly well or even optimal in many domains that contain clear attribute dependencies (Zhang, 2004; Domingos & Pazzani, 1997). Moreover,

it may be possible to extend the model by semi-naive approaches including higher-order dependencies, for example, as suggested by Kononenko (1991, 1994).

At least for independent address attributes, the Bayesian neural associative memory presented in this work is, by definition, the optimal local learning model maximizing M_ϵ and C_ϵ. On the other hand, there exist general bounds on the storage capacity of neural networks that do not refer to any particular learning algorithm (Gardner, 1988; Gardner & Derrida, 1988). As the linear covariance rule, the optimal Bayesian model reaches the Gardner bound for sparse memory patterns $p, q \to 0$ in the limit $Mpq \to \infty$ corresponding to moderately sparse patterns with $mp \gg \log(n)$ where the network can store $C_\epsilon = 1/(2 \ln 2) \approx 0.72$ bps (compare equation 37 to equation 40 in Gardner, 1988). However, for logarithmic sparse memory patterns with $mp \sim \log n$, the storage capacity of the optimal Bayesian rule is below the Gardner bound and cannot exceed the maximal capacity of the Willshaw model, which is at $C_\epsilon = \ln 2 \approx 0.69$ bps (or, rather, $C_\epsilon = 1/e \ln 2 \approx 0.53$ bps for distributed pattern activities; see Knoblauch et al., 2010, appendix D). For even sparser memory patterns with $mp/\log n \to 0$, the storage capacity vanishes, $C_\epsilon \to 0$. Also for nonsparse patterns where $p \to 0.5$, the Gardner bound of 2 bps cannot be reached. Here the optimal Bayesian rule achieves at most $C_\epsilon \approx 0.33$ bps for very low-fidelity retrieval with $\epsilon \to 1$, and only $C_\epsilon \to 0$ for high-fidelity retrieval with vanishing output noise $\epsilon \to 0$ (see Figure 3). Thus, as noted by Sommer and Dayan (1998), at least for nonsparse address patterns with $p \to 0.5$, local learning is insufficient, and the optimal synaptic weights must be found by more sophisticated algorithms, including nonlocal information.

Even if the Bayesian associative memory could reach the Gardner bound, the resulting storage capacity of at most 2 bits per synapse would be low compared to the physical memory actually required to represent real-valued synaptic weights (or, alternatively, the countervariables described in section 2.1). Even worse, an accurate neural implementation of the Bayesian associative memory requires two numbers per synaptic weight: a real-valued variable for the finite contributions and an integer variable for the infinite contributions (see appendix A). In fact, if we take into account the computational resources required to represent the resulting network, the Willshaw model outperforms all other models due to the binary weights (Knoblauch et al., 2010): For implementations on digital hardware, the Willshaw model can reach the theoretical maximum of $C^I = 1$ bit per computer bit (Knoblauch, 2003). Correspondingly, parallel hardware implementations of structurally plastic Willshaw networks can reach the theoretical maximum of $C^S = \log n$ bits per synapse (Knoblauch, 2009b). However, these high capacities (per synapse) are achieved only for a relatively low absolute number of stored memories, M, far below the Gardner bound, equation 2.37. Some preliminary work (Knoblauch, 2009c, 2010b) indicates that the Bayesian associative memory can be efficiently discretized such

that structurally compressed network implementations can store $C^I \to 1$ bit per computer bit or $C^S \to \log n$ bits per synapse, whereas M (and C) can still be close to the Gardner bound. Another future direction will be to investigate more closely the biological relevance of Bayesian learning by implementing more realistic network models that include spikes, forgetful synapses, and inhibitory circuits (Sandberg et al., 2000; Fusi, Drew, & Abbott, 2005; Markram et al., 2004).

Appendix A: Implementation of Infinite Weights and Thresholds

As noted in section 2.2, synaptic weights (see equation 2.15) and dendritic potentials (see equation 2.16) may be plus or minus infinity. Naive neural network implementations lead to suboptimal performance if neglecting that positively and negatively infinite contributions may cancel each other. To obtain accurate results, it is necessary to represent synaptic weights and firing thresholds each with two numbers for finite and infinite components. For $d_1 := M_{11}(1 - p_{10|1}) + M_{01}p_{01|1}$, $d_2 := M_{00}(1 - p_{01|0}) + M_{10}p_{10|0}$, $d_3 := M_{10}(1 - p_{10|0}) + M_{00}p_{01|0}$, $d_4 := M_{01}(1 - p_{01|1}) + M_{11}p_{10|1}$, the synaptic weight, equation 2.15, can be expressed by

$$w_{ij} = \log \frac{\mathfrak{F}(d_1)\mathfrak{F}(d_2)}{\mathfrak{F}(d_3)\mathfrak{F}(d_4)}, \tag{A.1}$$

$$w_{ij}^{\infty} = \mathfrak{G}(d_3) + \mathfrak{G}(d_4) - \mathfrak{G}(d_1) - \mathfrak{G}(d_2) \in \{-2, -1, 0, 1, 2\}, \tag{A.2}$$

with the gating functions $\mathfrak{F}(\mathfrak{x}) = \mathfrak{x}$ for $\mathfrak{x} > 0$ and $\mathfrak{F}(\mathfrak{x}) = 1$ for $\mathfrak{x} \leq 0$, and $\mathfrak{G}(\mathfrak{x}) = 0$ for $\mathfrak{x} > 0$ and $\mathfrak{G}(\mathfrak{x}) = 1$ for $\mathfrak{x} \leq 0$. Thus, w_{ij} represents the finite weight-neglecting infinite components, whereas w_{ij}^{∞} counts the number of contributions toward plus and minus infinity. Similarly, the finite and infinite components of firing thresholds (corresponding to the "bias" in equation 2.16) write as

$$\Theta_j = -(m-1)\log \frac{\mathfrak{F}(M_0)}{\mathfrak{F}(M_1)} - \sum_{i=1}^{m} \log \frac{\mathfrak{F}(d_4)}{\mathfrak{F}(d_2)}, \tag{A.3}$$

$$\Theta_j^{\infty} = -(m-1)(\mathfrak{G}(M_1) - \mathfrak{G}(M_0)) - \sum_{i=1}^{m} (\mathfrak{G}(d_2) - \mathfrak{G}(d_4)). \tag{A.4}$$

Then finite and infinite components of dendritic potentials are $x_j = \sum_{i=1}^{m} \tilde{u}_i w_{ij}$ and $x_j^{\infty} = \sum_{i=1}^{m} \tilde{u}_i w_{ij}^{\infty}$, such that a postsynaptic neuron j gets activated if either $x_j^{\infty} > \Theta_j^{\infty}$ or $x_j^{\infty} = \Theta_j^{\infty}$ and $x_j \geq \Theta_j$.

Appendix B: Analysis of the SNR for Optimal Bayesian Retrieval

The following computes the SNR (see equation 2.24) for neural associative memory with optimal Bayesian learning (section 2.2) making the same definitions and simplifications as detailed at the beginning of section 2.3. Section B.1 computes the mean difference $\Delta\mu := \mu_{\rm hi} - \mu_{\rm lo}$ between the dendritic potential of a high and a low unit, and section B.2 computes the variances $\sigma^2_{\rm hi}$ and $\sigma^2_{\rm lo}$ for the corresponding distributions of dendritic potentials.

B.1 Mean Values of Dendritic Potentials. Equivalent to equation 2.16 (but replacing $m-1$ by m and skipping indices i, j for brevity), a content neuron j will be activated if the dendritic potential x_j exceeds the threshold $\Theta_j := \log(M_0/M_1)$ (instead of $\Theta_j = 0$), where

$$\begin{aligned}
x_j &= m \log \frac{M_0}{M_1} + \sum_{i=1}^{c} \log \frac{M_{11}(1-p_{10}) + M_{01}p_{01}}{M_{10}(1-p_{10}) + M_{00}p_{01}} \\
&\quad + \sum_{i=c+1}^{k} \log \frac{M_{01}(1-p_{01}) + M_{11}p_{10}}{M_{00}(1-p_{01}) + M_{10}p_{10}} \\
&\quad + \sum_{i=k+1}^{k+f} \log \frac{M_{11}(1-p_{10}) + M_{01}p_{01}}{M_{10}(1-p_{10}) + M_{00}p_{01}} \\
&\quad + \sum_{i=k+f+1}^{m} \log \frac{M_{01}(1-p_{01}) + M_{11}p_{10}}{M_{00}(1-p_{01}) + M_{10}p_{10}} \\
&= m \log \frac{M_0}{M_1} + \sum_{i=1}^{c} \log \frac{M_1 p_{01} + M_{11}(1-p_{01}-p_{10})}{M_0(1-p_{10}) - M_{00}(1-p_{01}-p_{10})} \\
&\quad + \sum_{i=c+1}^{k} \log \frac{M_1(1-p_{01}) - M_{11}(1-p_{01}-p_{10})}{M_0 p_{10} + M_{00}(1-p_{01}-p_{10})} \\
&\quad + \sum_{i=k+1}^{k+f} \log \frac{M_1 p_{01} + M_{11}(1-p_{01}-p_{10})}{M_0(1-p_{10}) - M_{00}(1-p_{01}-p_{10})} \\
&\quad + \sum_{i=k+f+1}^{m} \log \frac{M_1(1-p_{01}) - M_{11}(1-p_{01}-p_{10})}{M_0 p_{10} + M_{00}(1-p_{01}-p_{10})}. \qquad \text{(B.1)}
\end{aligned}$$

Given M_1, M_0, the remaining variables are binomially distributed—$M_{00} \sim B_{M_0,1-p}$ and $M_{11} \sim B_{M_1,p}$, where $\mathrm{pr}[B_{N,P} = z] = \binom{N}{z} P^z (1-P)^{N-z}$. For large $NP(1-P)$ the binomial $B_{N,P}$ can be approximated by a gaussian $G_{\mu,\sigma}$ with mean $\mu = NP$ and variance $\sigma^2 = NP(1-P)$. Given u^μ_i and v^μ_j, we then

have

$$M_{11}(i,j) \sim \begin{cases} B_{M_1,p} \sim G_{M_1 p,\sqrt{M_1 p(1-p)}}, & (u_i^\mu, v_j^\mu) = (0,0) \\ B_{M_1,p} \sim G_{M_1 p,\sqrt{M_1 p(1-p)}}, & (u_i^\mu, v_j^\mu) = (1,0) \\ B_{M_1-1,p} \sim G_{(M_1-1)p,\sqrt{(M_1-1)p(1-p)}}, & (u_i^\mu, v_j^\mu) = (0,1) \\ 1 + B_{M_1-1,p} \sim G_{1+(M_1-1)p,\sqrt{(M_1-1)p(1-p)}}, & (u_i^\mu, v_j^\mu) = (1,1) \end{cases}, \tag{B.2}$$

$$M_{00}(i,j) \sim \begin{cases} 1 + B_{M_0-1,1-p} \sim G_{1+(M_0-1)(1-p),\sqrt{(M_0-1)p(1-p)}}, & (u_i^\mu, v_j^\mu) = (0,0) \\ B_{M_0-1,1-p} \sim G_{(M_0-1)(1-p),\sqrt{(M_0-1)p(1-p)}}, & (u_i^\mu, v_j^\mu) = (1,0) \\ B_{M_0,1-p} \sim G_{M_0(1-p),\sqrt{M_0 p(1-p)}}, & (u_i^\mu, v_j^\mu) = (0,1) \\ B_{M_0,1-p} \sim G_{M_0(1-p),\sqrt{M_0 p(1-p)}}, & (u_i^\mu, v_j^\mu) = (1,1) \end{cases}, \tag{B.3}$$

From this, we can approximate the distribution of the dendritic potential x_j for low units and high units, respectively. For large k and $m-k$, the sums of logarithms in equation B.1 are approximately gaussian distributed. In principle, the mean potentials μ_{lo} and μ_{hi} for low units and high units can be computed exactly from equation B.12. Fortunately, it turns out that the mean potential difference $\Delta\mu := \mu_{\text{hi}} - \mu_{\text{lo}}$ required for the SNR can be well approximated by using only the first-order term in equation B.12 (while all higher-order terms become virtually identical for μ_{hi} and μ_{lo}; for more details, see Knoblauch, 2009a, appendixes D, F). These first-order approximations $\mu'_{\text{lo}}, \mu'_{\text{hi}}$ of $\mu_{\text{lo}}, \mu_{\text{hi}}$ are

$$\begin{aligned} \mu'_{\text{lo}} = {} & m \log \frac{M_0}{M_1} + c \log \frac{M_1 p_{01} + M_1 p(1 - p_{01} - p_{10})}{M_0(1 - p_{10}) - (M_0 - 1)(1-p)(1 - p_{01} - p_{10})} \\ & + (k - c) \log \frac{M_1(1 - p_{01}) - M_1 p(1 - p_{01} - p_{10})}{M_0 p_{10} + (M_0 - 1)(1-p)(1 - p_{01} - p_{10})} \\ & + f \log \frac{M_1 p_{01} + M_1 p(1 - p_{01} - p_{10})}{M_0(1 - p_{10}) - (1 + (M_0 - 1)(1-p))(1 - p_{01} - p_{10})} \\ & + (m - k - f) \log \frac{M_1(1 - p_{01}) - M_1 p(1 - p_{01} - p_{10})}{M_0 p_{10} + (1 + (M_0 - 1)(1-p))(1 - p_{01} - p_{10})} \\ = {} & c \log \frac{M_0(p_{01} + p(1 - p_{01} - p_{10}))}{M_0(p_{01} + p(1 - p_{01} - p_{10})) + (1-p)(1 - p_{01} - p_{10})} + (k - c) \\ & \times \log \frac{M_0(p_{10} + (1-p)(1 - p_{01} - p_{10}))}{M_0(p_{10} + (1-p)(1 - p_{01} - p_{10})) - (1-p)(1 - p_{01} - p_{10})} \\ & + f \log \frac{M_0(p_{01} + p(1 - p_{01} - p_{10}))}{M_0(p_{01} + p(1 - p_{01} - p_{10})) - p(1 - p_{01} - p_{10})} + (m - k - f) \end{aligned}$$

$$
\begin{aligned}
&\times \log \frac{M_0(p_{10} + (1-p)(1-p_{01}-p_{10}))}{M_0(p_{10} + (1-p)(1-p_{01}-p_{10})) + p(1-p_{01}-p_{10})} \\
&\approx -c \frac{(1-p)(1-p_{01}-p_{10})}{M_0(p_{01} + p(1-p_{01}-p_{10}))} \\
&\quad + (k-c) \frac{(1-p)(1-p_{01}-p_{10})}{M_0(p_{10} + (1-p)(1-p_{01}-p_{10}))} \\
&\quad + f \frac{p(1-p_{01}-p_{10})}{M_0(p_{01} + p(1-p_{01}-p_{10}))} \\
&\quad - (m-k-f) \frac{p(1-p_{01}-p_{10})}{M_0(p_{10} + (1-p)(1-p_{01}-p_{10}))} \qquad \text{(B.4)} \\
\mu'_{\mathrm{hi}} &= m \log \frac{M_0}{M_1} + c \log \frac{M_1 p_{01} + (1 + (M_1-1)p)(1-p_{01}-p_{10})}{M_0(1-p_{10}) - M_0(1-p)(1-p_{01}-p_{10})} \\
&\quad + (k-c) \log \frac{M_1(1-p_{01}) - (1 + (M_1-1)p)(1-p_{01}-p_{10})}{M_0 p_{10} + M_0(1-p)(1-p_{01}-p_{10})} \\
&\quad + f \log \frac{M_1 p_{01} + (M_1-1)p(1-p_{01}-p_{10})}{M_0(1-p_{10}) - M_0(1-p)(1-p_{01}-p_{10})} \\
&\quad + (m-k-f) \log \frac{M_1(1-p_{01}) - (M_1-1)p(1-p_{01}-p_{10})}{M_0 p_{10} + M_0(1-p)(1-p_{01}-p_{10})} \\
&= c \log \frac{M_1(p_{01} + p(1-p_{01}-p_{10})) + (1-p)(1-p_{01}-p_{10})}{M_1(p_{01} + p(1-p_{01}-p_{10}))} + (k-c) \\
&\quad \times \log \frac{M_1(p_{10} + (1-p)(1-p_{01}-p_{10})) - (1-p)(1-p_{01}-p_{10})}{M_1(p_{10} + (1-p)(1-p_{01}-p_{10}))} \\
&\quad + f \log \frac{M_1(p_{01} + p(1-p_{01}-p_{10})) - p(1-p_{01}-p_{10})}{M_1(p_{01} + p(1-p_{01}-p_{10}))} + (m-k-f) \\
&\quad \times \log \frac{M_1(p_{10} + (1-p)(1-p_{01}-p_{10})) + p(1-p_{01}-p_{10})}{M_1(p_{10} + (1-p)(1-p_{01}-p_{10}))} \\
&\approx c \frac{(1-p)(1-p_{01}-p_{10})}{M_1(p_{01} + p(1-p_{01}-p_{10}))} \\
&\quad - (k-c) \frac{(1-p)(1-p_{01}-p_{10})}{M_1(p_{10} + (1-p)(1-p_{01}-p_{10}))} \\
&\quad - f \frac{p(1-p_{01}-p_{10})}{M_1(p_{01} + p(1-p_{01}-p_{10}))} \\
&\quad + (m-k-f) \frac{p(1-p_{01}-p_{10})}{M_1(p_{10} + (1-p)(1-p_{01}-p_{10}))}. \qquad \text{(B.5)}
\end{aligned}
$$

where the approximations are valid for large $M_0 p, M_1 p \to \infty$ and sufficiently small p_{01}, p_{10}. Therefore, the mean difference $\Delta\mu := \mu_{\text{hi}} - \mu_{\text{lo}}$ between the high and low distributions is

$$\begin{aligned}
&\frac{\Delta\mu}{1 - p_{01} - p_{10}} \\
&\quad \approx \frac{\mu'_{\text{hi}} - \mu'_{\text{lo}}}{1 - p_{01} - p_{10}} \approx \frac{c(1-p)}{p_{01} + p(1 - p_{01} - p_{10})} \left(\frac{1}{M_1} + \frac{1}{M_0}\right) \\
&\qquad - \frac{(k-c)(1-p)}{p_{10} + (1-p)(1 - p_{01} - p_{10})} \left(\frac{1}{M_1} + \frac{1}{M_0}\right) \\
&\qquad - \frac{fp}{p_{01} + p(1 - p_{01} - p_{10})} \left(\frac{1}{M_1} + \frac{1}{M_0}\right) \\
&\qquad + \frac{(m-k-f)p}{p_{10} + (1-p)(1 - p_{01} - p_{10})} \left(\frac{1}{M_1} + \frac{1}{M_0}\right) \\
&\quad = \left(\frac{c(1-p) - fp}{p(1 + \frac{1-p}{p} p_{01} - p_{10})} + \frac{(m-k-f)p - (k-c)(1-p)}{(1-p)(1 - p_{01} + \frac{p}{1-p} p_{10})} \right) \\
&\qquad \times \left(\frac{1}{M_1} + \frac{1}{M_0}\right). \qquad \text{(B.6)}
\end{aligned}$$

B.2 Variance of Dendritic Potentials. In order to get the SNR, equation 2.24, we have to compute the variances σ^2_{lo} and σ^2_{hi} for x_j in equation B.1. Given the unit usages $M_1(j)$, the random variables $M_{00}(i, j)$ and $M_{11}(i, j)$ are independent, and thus the variances simply add. Because each variance summand is positive, for large $M_1 p, M_0 p \to \infty$, we can simply assume $M_{11} \sim G_{M_1 p, \sqrt{M_1 p(1-p)}}$ and $M_{00} \sim G_{M_0(1-p), \sqrt{M_0 p(1-p)}}$ in all cases (cf. equations B.2 and B.3). With equation B.13 we get

$$\begin{aligned}
&\operatorname{Var}(\log(M_1 p_{01} + M_{11}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_1 p(1-p)}{(M_1 p_{01} + M_1 p(1 - p_{01} - p_{10}))^2}, \\
&\operatorname{Var}(\log(M_0(1 - p_{10}) - M_{00}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_0 p(1-p)}{(M_0(1 - p_{10}) - M_0(1-p)(1 - p_{01} - p_{10}))^2}, \\
&\operatorname{Var}(\log(M_1(1 - p_{01}) - M_{11}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_1 p(1-p)}{(M_1(1 - p_{01}) - M_1 p(1 - p_{01} - p_{10}))^2},
\end{aligned}$$

$$
\begin{aligned}
&\mathrm{Var}(\log(M_0 p_{10} + M_{00}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_0 p(1-p)}{(M_0 p_{10} + M_0(1-p)(1 - p_{01} - p_{10}))^2}, \\
&\mathrm{Var}(\log(M_1 p_{01} + M_{11}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_1 p(1-p)}{(M_1 p_{01} + M_1 p(1 - p_{01} - p_{10}))^2}, \\
&\mathrm{Var}(\log(M_0(1 - p_{10}) - M_{00}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_0 p(1-p)}{(M_0(1 - p_{10}) - M_0(1-p)(1 - p_{01} - p_{10}))^2}, \\
&\mathrm{Var}(\log(M_1(1 - p_{01}) - M_{11}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_1 p(1-p)}{(M_1(1 - p_{01}) - M_1 p(1 - p_{01} - p_{10}))^2}, \\
&\mathrm{Var}(\log(M_0 p_{10} + M_{00}(1 - p_{01} - p_{10}))) \\
&\quad \approx \frac{(1 - p_{01} - p_{10})^2 M_0 p(1-p)}{(M_0 p_{10} + M_0(1-p)(1 - p_{01} - p_{10}))^2}.
\end{aligned} \tag{B.7}
$$

Thus, the variances $\mathrm{Var}(x_j)$ for the potentials of both low units and high units are approximately

$$
\begin{aligned}
&\frac{\sigma_{\mathrm{lo}}^2}{(1 - p_{01} - p_{10})^2} \\
&\quad \approx \frac{\sigma_{\mathrm{hi}}^2}{(1 - p_{01} - p_{10})^2} \approx c \frac{1-p}{M_1 p(1 + \frac{1-p}{p} p_{01} - p_{10})^2} \\
&\qquad + c \frac{1-p}{M_0 p(1 + \frac{1-p}{p} p_{01} - p_{10})^2} + (k-c) \frac{p}{M_1(1-p)(1 - p_{01} + \frac{p}{1-p} p_{10})^2} \\
&\qquad + (k-c) \frac{p}{M_0(1-p)(1 - p_{01} + \frac{p}{1-p} p_{10})^2} \\
&\qquad + f \frac{1-p}{M_1 p(1 + \frac{1-p}{p} p_{01} - p_{10})^2} + f \frac{1-p}{M_0 p(1 + \frac{1-p}{p} p_{01} - p_{10})^2} \\
&\qquad + \frac{(m-k-f)p}{M_1(1-p)(1 - p_{01} + \frac{p}{1-p} p_{10})^2} + \frac{(m-k-f)p}{M_0(1-p)(1 - p_{01} + \frac{p}{1-p} p_{10})^2} \\
&\quad = \frac{(c+f)(1-p)(1/M_1 + 1/M_0)}{p(1 + \frac{1-p}{p} p_{01} - p_{10})^2} + \frac{(m-c-f)p(1/M_1 + 1/M_0)}{(1-p)(1 - p_{01} + \frac{p}{1-p} p_{10})^2}
\end{aligned}
$$

$$= \left(\frac{(c+f)(1-p)}{p(1+\frac{1-p}{p}p_{01}-p_{10})^2} + \frac{(m-c-f)p}{(1-p)(1-p_{01}+\frac{p}{1-p}p_{10})^2} \right) \times \left(\frac{1}{M_1} + \frac{1}{M_0} \right). \tag{B.8}$$

B.3 Lemmas for Computing Dendritic Potential Distributions. Let X be a random variable with normal distribution, $X \sim G_{0,\sigma}$, that is, X is a gaussian with zero mean and variance σ^2. Then the dth moment is

$$E(X^d) = \begin{cases} 0, & d = 2i+1 \\ 1 \cdot 3 \cdots (d-1)\sigma^d, & d = 2i \end{cases}. \tag{B.9}$$

Proofs can be found in standard textbooks of statistics and probability theory (e.g., see equation 5.44 in Papoulis, 1991).

Then the Taylor expansion of $\log(x)$ around μ (also called the Newton-Mercator series) is

$$\log(\mu+\Delta) = \log\mu + \log(1+\Delta/\mu)$$

$$= \log\mu + \frac{\Delta}{\mu} - \frac{1}{2}(\Delta/\mu)^2 + \frac{1}{3}(\Delta/\mu)^3 + \ldots \tag{B.10}$$

$$= \log\mu + \sum_{d=1}^{\infty} (-1)^{d+1} \frac{(\Delta/\mu)^d}{d} \tag{B.11}$$

for $-1 < \Delta/\mu \leq 1$. Proofs can be found in standard textbooks of analysis (e.g., see Borwein & Bailey, 2003; Weisstein, 1999; Abramowitz & Stegun, 1972).

Now let X be a gaussian random variable, $X \sim G_{\mu,\sigma}$, with mean μ and variance σ^2. Then for $\sigma \ll \mu$, we have

$$E(\log X) = \log\mu - \sum_{i=1}^{\infty} \frac{1 \cdot 3 \cdots (2i-1)}{2i} (\sigma/\mu)^{2i} \approx \log\mu \tag{B.12}$$

$$\mathrm{Var}(\log X) \approx \left(\frac{\sigma}{\mu}\right)^2, \tag{B.13}$$

where the approximations are tight for $\sigma/\mu \to 0$ if $\mu \not\to 1$.

Proof. We can write $X = \mu + \Delta$ where Δ is normal with variance σ^2. Then equation B.12 follows from eqs. B.9 and B.11. Similarly, the variance

$\mathrm{Var}(\log X) = E((\log X)^2) - (E(\log X))^2$ follows from

$$
\begin{aligned}
(\log X)^2 &= \left(\log\mu + \sum_{d=1}^{\infty}(-1)^{d+1}\frac{(\Delta/\mu)^d}{d}\right)^2 \\
&= (\log\mu)^2 + 2(\log\mu)\sum_{d=1}^{\infty}(-1)^{d+1}\frac{(\Delta/\mu)^d}{d} \\
&\quad + \sum_{d_1=1}^{\infty}\sum_{d_2=1}^{\infty}(-1)^{d_1+d_2}\frac{(\Delta/\mu)^{d_1+d_2}}{d_1 d_2}, \\
E((\log X)^2) &= (\log\mu)^2 + 2(\log\mu)\sum_{d=1}^{\infty}(-1)^{d+1}\frac{E(\Delta^d)}{d\,\mu^d} \\
&\quad + \sum_{d_1=1}^{\infty}\sum_{d_2=1}^{\infty}(-1)^{d_1+d_2}\frac{E(\Delta^{d_1+d_2})}{d_1 d_2 \mu^{d_1+d_2}}, \\
(E(\log X))^2 &= (\log\mu)^2 + 2(\log\mu)\sum_{d=1}^{\infty}(-1)^{d+1}\frac{E(\Delta^d)}{d\,\mu^d} \\
&\quad + \sum_{d_1=1}^{\infty}\sum_{d_2=1}^{\infty}(-1)^{d_1+d_2}\frac{E(\Delta^{d_1})E(\Delta^{d_2})}{d_1 d_2 \mu^{d_1+d_2}}, \\
\mathrm{Var}(\log X) &= \sum_{d_1=1}^{\infty}\sum_{d_2=1}^{\infty}(-1)^{d_1+d_2}\frac{E(\Delta^{d_1+d_2}) - E(\Delta^{d_1})E(\Delta^{d_2})}{d_1 d_2 \mu^{d_1+d_2}},
\end{aligned}
$$

where in the last equation for $\sigma/\mu \to 0$, the first summand ($d_1 = d_2 = 1$) dominates.

Appendix C: Gaussian Tail Integrals

Let $g(x)$ be the gaussian probability density:

$$g(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}. \tag{C.1}$$

Then the complementary gaussian distribution function is the right tail integral:

$$G^c(x) := \int_x^{\infty} g(t)dt = \frac{1-\mathrm{erf}(x/\sqrt{2})}{2} = \frac{\mathrm{erfc}(x/\sqrt{2})}{2} \lessapprox \frac{e^{-x^2/2}}{\sqrt{2\pi}\,x} < \frac{e^{-x^2/2}}{2}. \tag{C.2}$$

The first bound is true for any $x > 0$, and the corresponding approximation error becomes smaller than 1% for $x > 10$. The second bound is true for any $x > 0$. Inverting G^c yields

$$G^{c\,-1}(x) := y|_{G^c(y)=x} = \sqrt{2}\mathrm{erf}^{-1}(1-2x) = \sqrt{2}\mathrm{erfc}^{-1}(2x)$$
$$\lessapprox \sqrt{-2\ln(\sqrt{2\pi}\,x\,G^{c\,-1}(x))} < \sqrt{-2\ln(2x)}. \tag{C.3}$$

The two approximations correspond to those of equation C.2. In the first approximation, the term $G^{c\,-1}(x)$ can be replaced, for example, by the second approximation $\sqrt{-2\ln(2x)}$.

Appendix D: Optimal Firing Thresholds

Given a query pattern $\tilde{\mathbf{u}}$ resembling one of the original address patterns $\mathbf{u}^\mu$, our goal is to minimize the expected Hamming distance $d_H(\mathbf{v}^\mu, \hat{\mathbf{v}})$ between the corresponding content $\mathbf{v}^\mu$ and the retrieval output $\hat{\mathbf{v}}$ (see equation 2.21). To this end, each content neuron v_j has to adjust its firing threshold Θ in order to minimize

$$H(\Theta) := q\,q_{10} + (1-q)q_{01}, \tag{D.1}$$

where $q := \mathrm{pr}[v_j^\mu = 1]$ is the prior and

$$q_{01} := \int_\Theta^\infty g_{\mathrm{lo}}(x)dx \quad \text{and} \quad q_{10} := \int_{-\infty}^\Theta g_{\mathrm{hi}}(x)dx \tag{D.2}$$

are the probabilities of making an output error (e.g., equations 2.22 and 2.23) assuming a given low distribution $g_{\mathrm{lo}}(x) := \mathrm{pr}[x_j = x|v_j^\mu = 0]$ and high distribution $g_{\mathrm{hi}}(x) := \mathrm{pr}[x_j = x|v_j^\mu = 1]$ for the dendritic potential x_j (e.g., see equation 2.16). Minimizing $H(\Theta)$ requires $dH/d\Theta = 0$ or, equivalently,

$$(1-q)g_{\mathrm{lo}}(\Theta) = q\,g_{\mathrm{hi}}(\Theta), \tag{D.3}$$

as illustrated by Figure 8 (left). The optimal threshold Θ_{opt} can be obtained by solving equation D.3, which is easy if the distributions of dendritic potentials are gaussians. Then equation D.3 is rewritten as

$$(1-q)g\left(\frac{\Theta-\mu_{\mathrm{lo}}}{\sigma_{\mathrm{lo}}}\right) = q\,g\left(\frac{\Theta-\mu_{\mathrm{hi}}}{\sigma_{\mathrm{hi}}}\right), \tag{D.4}$$

where g is the Gaussian density, equation C.1, and $\mu_{\mathrm{lo}}, \mu_{\mathrm{hi}}, \sigma_{\mathrm{lo}}, \sigma_{\mathrm{hi}}$ are means and standard deviations of the low and high dendritic potentials similar as

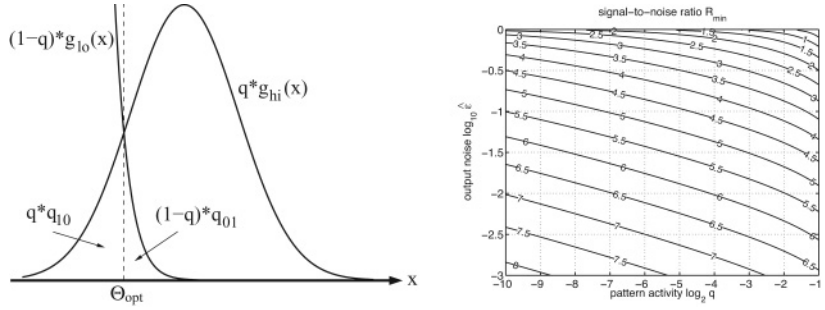

Figure 8: Optimal firing threshold and minimal SNR. (Left) Expected normalized distributions $(1-q)g_{\mathrm{lo}}(x)$ and $qg_{\mathrm{hi}}(x)$ of dendritic potential x for low-units (with $v_j^\mu = 0$) and high-units (with $v_j^\mu = 1$), respectively. The optimal firing threshold is at dendritic potential $x = \Theta_{\mathrm{opt}}$ where the two distributions are equal. For sparse content patterns with $q < 0.5$ the resulting miss noise qq_{10} is larger than the add noise $(1-q)q_{01}$. In fact, for $q \to 0$ and constant ϵ the add noise becomes negligible (see equation D.10; see also Knoblauch, 2009a). (Right) Contour plot showing the minimal SNR $R_{\min}(\hat{\epsilon}, q)$ (see appendix E) required to obtain output noise $\hat{\epsilon}$ for content pattern activity q and optimal firing threshold equation D.9.

defined below equation 2.24. Taking logarithms yields a quadratic equation in Θ with the solution

$$\Theta_{\mathrm{opt},1/2} = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A}, \tag{D.5}$$

$$A = \left(\frac{1}{2\sigma_{\mathrm{hi}}^2} - \frac{1}{2\sigma_{\mathrm{lo}}^2}\right), \tag{D.6}$$

$$B = -\left(\frac{\mu_{\mathrm{hi}}}{\sigma_{\mathrm{hi}}^2} - \frac{\mu_{\mathrm{lo}}}{\sigma_{\mathrm{lo}}^2}\right), \tag{D.7}$$

$$C = \log\left(\frac{1-q}{q}\frac{\sigma_{\mathrm{hi}}}{\sigma_{\mathrm{lo}}}\right) - \frac{1}{2}\left(\frac{\mu_{\mathrm{lo}}}{\sigma_{\mathrm{lo}}}\right)^2 + \frac{1}{2}\left(\frac{\mu_{\mathrm{hi}}}{\sigma_{\mathrm{hi}}}\right)^2, \tag{D.8}$$

where the optimal threshold is either Θ_1 or Θ_2. If the standard deviations are equal, $\sigma_{\mathrm{lo}} = \sigma_{\mathrm{hi}}$, then $A = 0$, and equation D.4 has the unique solution

$$\Theta_{\mathrm{opt}} = -C/B. \tag{D.9}$$

The following lemma characterizes the weighing of add noise ($v_j^\mu = 0$ but $\hat{v}_j = 1$) versus miss noise ($v_j^\mu = 1$ but $\hat{v}_j = 0$) in the retrieval result $\hat{\mathbf{v}}$ when

choosing the optimal firing threshold: If we assume a given constant output noise $\hat{\epsilon} := H(\Theta_{\text{opt}})/q$ (cf. equation 2.30), gaussian potentials with equal standard deviations $\sigma_{\text{lo}} = \sigma_{\text{hi}}$ and optimal firing threshold $\Theta = \Theta_{\text{opt}}$ as in equation D.9, then

$$q \to 0 \quad \text{implies} \quad \frac{(1-q)q_{01}}{q\,q_{10}} \to 0 \quad \text{and thus} \quad \hat{\xi} \to 0. \tag{D.10}$$

that is, for sparse content patterns, the output errors are dominated by miss noise (see equation 2.31). A formal proof of the lemma can be found in Knoblauch (2009a, appendix A, equation 74). Figure 8 (left) gives an intuition as to why the lemma is true. Here $H(\Theta_{\text{opt}})$ is the intersection area of high and low distribution, where the left and right parts of the area correspond to miss noise $q\,q_{10}$ and add noise $(1-q)q_{01}$, respectively (see the arrows). Requiring constant $H(\Theta_{\text{opt}})/q$ implies that the intersection area $H(\Theta_{\text{opt}})$ must be a constant fraction of the area below $q\,g_{\text{hi}}(x)$. Thus, $q \to 0$ implies for $\sigma_{\text{lo}} = \sigma_{\text{hi}}$ that the decrease of $(1-q)g_{\text{lo}}(x)$ with x becomes very steep compared to the increase of $q\,g_{\text{hi}}(x)$ and finally approaches the dashed line corresponding to Θ_{opt}.

Appendix E: The Relation Between SNR R and Output Noise $\hat{\epsilon}$

We can use two different measures to evaluate retrieval quality: section 2.3 uses the SNR R (see equation 2.24), whereas section 2.4 uses output noise $\hat{\epsilon}$, which is based on the Hamming distance (see equation 2.30). This appendix shows that the two measures are actually equivalent if we assume that (1) all content neurons j have the same priors $q := \text{pr}[v_j^\mu = 1]$ and the same distributions for high and low dendritic potentials; (2) all dendritic potentials follow a gaussian distribution; (3) each content neuron optimally adjusts the firing threshold in order to minimize output noise $\hat{\epsilon}$ (see appendix D); and (4) the distributions of high and low dendritic potentials have the same standard deviation, $\sigma := \sigma_{\text{lo}} = \sigma_{\text{hi}}$. Note that all assumptions are fulfilled at least in the limit $Mpq \to \infty$ for reasons discussed in section 2.3.

We first write the output noise $\hat{\epsilon}$ as a function of the SNR R: Due to assumption 1, we can write the output noise, equation 2.30 in terms of the output error probabilities, equations 2.22 and 2.23:

$$\hat{\epsilon} := (1/q - 1)q_{01} + q_{10}. \tag{E.1}$$

Due to assumption 2, the output error probabilities write

$$q_{01} = G^c\left(\frac{\Theta_{\text{opt}} - \mu_{\text{lo}}}{\sigma_{\text{lo}}}\right) \quad \text{and} \quad \mathrm{q}_{10} = \mathrm{G}^{\mathrm{c}}\left(\frac{\mu_{\text{hi}} - \Theta_{\text{opt}}}{\sigma_{\text{hi}}}\right), \tag{E.2}$$

where $G^c(x)$ is the tail integral of a gaussian (see equation C.2), and, due to assumption 3, Θ_{opt} is the optimal firing threshold as explained in appendix D. Due to assumption 4, Θ_{opt} is as in equation D.9:

$$\Theta_{\text{opt}} = \frac{\ln \frac{1-q}{q} + \frac{1}{2}\frac{{\mu_{\text{hi}}}^2 - {\mu_{\text{lo}}}^2}{\sigma^2}}{\frac{\mu_{\text{hi}} - \mu_{\text{lo}}}{\sigma^2}} = \frac{\sigma}{R}\ln\frac{1-q}{q} + \frac{1}{2}\sigma R + \mu_{\text{lo}} \geq \frac{\mu_{\text{lo}} + \mu_{\text{hi}}}{2}. \tag{E.3}$$

The last bound implies that the optimal threshold shifts toward the high potentials for sparse patterns with $q < 0.5$ and centers only for $q = 0.5$. Thus, the error probabilities at optimal threshold are

$$q_{01} = G^c\left(\frac{\Theta_{\text{opt}} - \mu_{\text{lo}}}{\sigma}\right) = G^c\left(R/2 + \frac{\ln(1/q-1)}{R}\right), \tag{E.4}$$

$$\begin{aligned} q_{10} &= G^c\left(\frac{\mu_{\text{hi}} - \Theta_{\text{opt}}}{\sigma}\right) = G^c\left(R - R/2 - \frac{\ln(1/q-1)}{R}\right) \\ &= G^c\left(R/2 - \frac{\ln(1/q-1)}{R}\right), \end{aligned} \tag{E.5}$$

and thus the minimal output noise level $\hat{\epsilon}$ that can be achieved with SNR R equals

$$\hat{\epsilon}_{\min}(R,q) = (1/q-1)G^c\left(R/2 + \tfrac{\ln(1/q-1)}{R}\right) + G^c\left(R/2 - \tfrac{\ln(1/q-1)}{R}\right) \tag{E.6}$$

where G^c can be evaluated with equation C.2.

Vice versa, we obtain the minimal SNR $R_{\min}(\hat{\epsilon}, q)$ required for an output noise level $\hat{\epsilon}$ by resolving equation E.6 for R. We can do this easily for two special cases. First, for nonsparse content patterns with $q = 0.5$, we have $\hat{\epsilon}_{\min} = 2G^c(r/2)$ and thus

$$R_{\min}(\hat{\epsilon}, 0.5) = 2G^{c-1}(\hat{\epsilon}/2), \tag{E.7}$$

where G^{c-1} is as in equation C.3. Second, for sparse content patterns with $q \to 0$, miss-noise will dominate output errors according to equation D.10. Correspondingly, the output noise, equation E.6, is dominated by the second summand. Therefore, $q \to 0$ implies

$$R_{\min}(\hat{\epsilon}, q) \approx G^{c-1}(\hat{\epsilon}) + \sqrt{(G^{c-1}(\hat{\epsilon}))^2 + 2\ln(1/q-1)} \approx \sqrt{-2\ln q}. \tag{E.8}$$

Alternatively, and in particular for $0 \not\approx q \neq 0.5$, we can compute $R_{\min}$ by iteratively applying the following two equations:

$$R_{\min}(\hat{\epsilon}, q, \hat{\xi}) = G^{c-1}\left(\frac{\hat{\xi}\hat{\epsilon}q}{1-q}\right) + G^{c-1}((1-\hat{\xi})\hat{\epsilon}), \tag{E.9}$$

$$\begin{aligned}\hat{\xi}_{\text{opt}} &= 1 - \frac{G^c(R/2 - \frac{\ln(1/q-1)}{R})}{\hat{\epsilon}} \\ &= \frac{1-q}{q\hat{\epsilon}} G^c\left(R/2 + \frac{\ln(1/q-1)}{R}\right),\end{aligned} \tag{E.10}$$

starting with $\hat{\xi} = 0.5$, for example. In the first step, equation E.9 computes the minimal SNR required to obtain output noise $\hat{\epsilon}$ where, in contrast to assumption 3, firing thresholds are chosen such that a given fraction $\hat{\xi} \in [0; 1]$ of the expected output errors is add noise, and the remaining fraction $1 - \hat{\xi}$ is miss noise (here, $\hat{\xi}$ is the output noise balance, equation 2.31; see also equation E.1 and Figure 8, left). In the second step, we insert $R = R_{\min}$ from the first step into equation E.10 and compute the optimal noise balance $\hat{\xi} = \hat{\xi}_{\text{opt}}$ such that output noise $\hat{\epsilon}$ is minimal and assumption 3 is fulfilled again. In practice, few iterations of this procedure (e.g., fewer than 10) are sufficient to obtain an accurate estimate of $R_{\min}(\hat{\epsilon}, q)$, which may be further verified by insertion into equation E.6. For more details, see Knoblauch (2009a, appendix A). Figure 8 (right) computes $R_{\min}(\hat{\epsilon}, q)$ for relevant parameters $\hat{\epsilon}$ and q.

Appendix F: Binary Channels

For a random variable $X \in \{0, 1\}$ with $q := \text{pr}[X = 1]$ the information I(X) equals (Shannon & Weaver, 1949)

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

It is $I(q) = I(1-q)$ and $I(q) \to 0$ for $q \to 0$. A binary memory-less channel is determined by the two error probabilities q_{01} for add noise and q_{10} for miss noise. For two binary random variables X and Y, where Y is the result of transmitting X over the binary channel, we can write

$$I(Y) = I_Y(q, q_{01}, q_{10}) := I\left(q\,(1-q_{10}) + (1-q)\,q_{01}\right), \tag{F.2}$$

$$I(Y|X) = I_{Y|X}(q, q_{01}, q_{10}) := q \cdot I(q_{10}) + (1-q) \cdot I(q_{01}), \tag{F.3}$$

$$T(X;Y) = T(q, q_{01}, q_{10}) := I_Y(q, q_{01}, q_{10}) - I_{Y|X}(q, q_{01}, q_{10}). \tag{F.4}$$

For the analysis of storage capacity of associative networks at noise level ϵ (see section 2.4), we are interested in fulfilling the high-fidelity criterion, equation E.1, with a "noise balance" parameter ξ weighing between add noise and miss noise,

$$q_{01} = \frac{\xi\epsilon q}{1-q} \quad \text{and} \quad \mathrm{q}_{10} = (1-\xi)\epsilon, \tag{F.5}$$

such that

$$T\left(q, \frac{\xi\epsilon q}{1-q}, (1-\xi)\epsilon\right) = I(q - \epsilon q(1-2\xi)) - q\, I\left((1-\xi)\epsilon\right) - (1-q) I\left(\frac{\xi\epsilon q}{1-q}\right). \tag{F.6}$$

Thus, we can compute the component transinformation for several interesting cases:

$$T\left(q, \frac{\xi\epsilon q}{1-q}, (1-\xi)\epsilon\right) \begin{cases} = I\left(\frac{1-\epsilon(1-2\xi)}{2}\right) - 0.5(I((1-\xi)\epsilon) + I(\xi\epsilon)), & q = 0.5 \\ \approx I(q), & \epsilon/q \to 0 \\ \approx I(q)\left(1 - \epsilon(1-\xi+\frac{\mathrm{ld}\epsilon}{\mathrm{ld}q})\right), & q/\epsilon \to 0 \\ \approx I(q), & q, \epsilon \to 0 \end{cases} . \tag{F.7}$$

For details see Knoblauch (2009a, appendix E). Three approximations are of particular interest. For $q = 0.5$ and $\xi = 0.5$, we have $T \approx 1 - I(\epsilon/2)$. For $q \to 0$, constant ϵ, and dominating miss noise with $\xi \to 0$, we have $T \approx I(q)(1-\epsilon)$. For $q \to 0$, constant ϵ, and dominating add noise with $\xi \to 1$, we have $T \approx I(q)$.

Appendix G: Analysis of the SNR for Linear Learning Rules

Here we analyze the SNR for the linear learning rule, equation 3.6, in analogy to the analysis in section 2.3. Without loss of generality, we assume that the query pattern $\tilde{\mathbf{u}} \approx \mathbf{u}^M$ resembles the Mth address pattern and, similarly as illustrated by Figure 2 (left), contains c correct one-entries and f false one-entries. The synaptic weight writes as the linear sum of learning increments r_{uv} due to individual memory associations with presynaptic

activity $u \in \{0, 1\}$ and postsynaptic activity $v \in \{0, 1\}$,

$$w_{ij} = r_{00}M_{00} + r_{01}M_{01} + r_{10}M_{10} + r_{11}M_{11} = \sum_{\mu=1}^{M} r_{u_i^\mu v_j^\mu} \tag{G.1}$$

$$= \begin{cases} r_{u_i^M 1} + \sum_{\mu=1}^{M_0} r_{u_i^\mu 0} + \sum_{\mu=M_0+1}^{M-1} r_{u_i^\mu 1}, & v_j^M = 1 \\ r_{u_i^M 0} + \sum_{\mu=1}^{M_1} r_{u_i^\mu 1} + \sum_{\mu=M_1+1}^{M-1} r_{u_i^\mu 0}, & v_j^M = 0 \end{cases}, \tag{G.2}$$

where, without loss of generality, for a high unit ($v_j^M = 1$), we assume that $v_j^\mu = 1$ for $\mu = M_0 + 1, \ldots, M$; and for a low unit ($v_j^M = 0$), we assume that $v_j^\mu = 1$ for $\mu = 1, \ldots, M_1$. Then the dendritic potential $x_j = \sum_{i=1}^{m} w_{ij} F(\tilde{u}_i)$ with $F(1) = 1$ and $F(0) = a$ is

$$x_j = \sum_{i=1}^{c} w_{ij}|_{u_i^M=1} + a \sum_{i=c+1}^{k} w_{ij}|_{u_i^M=1} + \sum_{i=k+1}^{k+f} w_{ij}|_{u_i^M=0} + a \sum_{i=k+f+1}^{m} w_{ij}|_{u_i^M=0}. \tag{G.3}$$

Thus, the mean dendritic potentials for high and low units are

$$\begin{aligned} \mu_{\text{hi}} &= (c + (k-c)a)[r_{11} + M_0 E(r_{u_i^\mu 0}) + (M_1 - 1)E(r_{u_i^\mu 1})] \\ &\quad + (f + (m-k-f)a)[r_{01} + M_0 E(r_{u_i^\mu 0}) + (M_1 - 1)E(r_{u_i^\mu 1})] \\ &= (c + (k-c)a)r_{11} + (f + (m-k-f)a)r_{01} \\ &\quad + (c + f + (m-c-f)a)[M_0((1-p)r_{00} + pr_{10}) \\ &\quad + (M_1 - 1)((1-p)r_{01} + pr_{11})], \end{aligned} \tag{G.4}$$

$$\begin{aligned} \mu_{\text{lo}} &= (c + (k-c)a)[r_{10} + M_1 E(r_{u_i^\mu 1}) + (M_0 - 1)E(r_{u_i^\mu 0})] \\ &\quad + (f + (m-k-f)a)[r_{00} + M_1 E(r_{u_i^\mu 1}) + (M_0 - 1)E(r_{u_i^\mu 0})] \\ &= (c + (k-c)a)r_{10} + (f + (m-k-f)a)r_{00} \\ &\quad + (c + f + (m-c-f)a)[M_1((1-p)r_{01} + pr_{11}) \\ &\quad + (M_0 - 1)((1-p)r_{00} + pr_{10})], \end{aligned} \tag{G.5}$$

using $E(r_{u_i^\mu 0}) = (1-p)r_{00} + pr_{10}$ and $E(r_{u_i^\mu 1}) = (1-p)r_{01} + pr_{11}$. Similarly, we can compute the variances of dendritic potentials by replacing a by a^2

and E by Var and leaving out constant terms,

$$\begin{aligned}\sigma_{\mathrm{hi}}^2 &= (c + (k-c)a^2)[M_0 \mathrm{Var}(r_{u_i^\mu 0}) + (M_1 - 1)\mathrm{Var}(R_{u_i^\mu 1})] \\ &\quad + (f + (m-k-f)a^2)[M_0 \mathrm{Var}(r_{u_i^\mu 0}) + (M_1 - 1)\mathrm{Var}(r_{u_i^\mu 1})] \\ &= (c + f + (m-c-f)a^2)p(1-p)[M_0(r_{10} - r_{00})^2 \\ &\quad + (M_1 - 1)(r_{11} - r_{01})^2], \qquad \text{(G.6)} \\ \sigma_{\mathrm{lo}}^2 &= (c + (k-c)a^2)[M_1 \mathrm{Var}(r_{u_i^\mu 1}) + (M_0 - 1)\mathrm{Var}(r_{u_i^\mu 0})] \\ &\quad + (f + (m-k-f)a^2)[M_1 \mathrm{Var}(r_{u_i^\mu 1}) + (M_0 - 1)\mathrm{Var}(r_{u_i^\mu 0})] \\ &= (c + f + (m-c-f)a^2)p(1-p)[M_1(r_{11} - r_{01})^2 \\ &\quad + (M_0 - 1)(r_{10} - r_{00})^2], \qquad \text{(G.7)}\end{aligned}$$

using $\mathrm{Var}(r_{u_i^\mu 0}) = p(1-p)(r_{10} - r_{00})^2$ and $\mathrm{Var}(r_{u_i^\mu 1}) = p(1-p)(r_{11} - r_{01})^2$. Then the mean potential difference $\Delta\mu := \mu_{\mathrm{hi}} - \mu_{\mathrm{lo}}$ is

$$\begin{aligned}\Delta\mu &= (c + (k-c)a)(r_{11} - r_{10}) + (f + (m-k-f)a)(r_{01} - r_{00}) \\ &\quad + (c + f + (m-c-f)a)[(1-p)r_{00} + pr_{10} - (1-p)r_{01} - pr_{11}], \qquad \text{(G.8)} \\ &= (c + (k-c)a)(r_{11} - r_{10}) - (f + (m-k-f)a)(r_{00} - r_{01}) \\ &\quad - (c + f + (m-c-f)a)[p(r_{11} - r_{10}) - (1-p)(r_{00} - r_{01})]. \qquad \text{(G.9)}\end{aligned}$$

With this, we can compute the SNR $R := \Delta\mu / \max(\sigma_{\mathrm{hi}}, \sigma_{\mathrm{lo}})$ (see equation 2.24), optimal firing thresholds (see appendix D), and storage capacity (see section 2.4). It is well known that the optimal linear rule (maximizing R) is the so-called covariance rule $r_{00} = pq$, $r_{01} = -p(1-q)$, $r_{10} = -(1-p)q$, $r_{11} = (1-p)(1-q)$, and $a = -(\tilde{\lambda} + \tilde{\kappa})p/(1 - (\tilde{\lambda} + \tilde{\kappa})p)$ where $p := \mathrm{pr}[u_i^\mu = 1]$ and $q := \mathrm{pr}[v_j^\mu = 1]$ (see Dayan & Willshaw, 1991; Palm & Sommer, 1996). Further rules of interest are, for example, the Hebbian rule $r_{11} = 1, r_{00} = r_{01} = r_{10} = a = 0$; the homosynaptic rule $r_{11} = 1 - q, r_{10} = -q$, $r_{00} = r_{01} = a = 0$; and the heterosynaptic rule $r_{11} = 1 - p$, $r_{01} = -p$, $r_{00} = r_{10} = a = 0$.

Appendix H: Generalized BCPNN-Type Learning Rules

H.1 Generalizing the BCPNN Rule for Query Noise. Section 3.3 discusses the original BCPNN rule of Lansner & Ekeberg (1989). The original

BCPNN rule, equation 3.14, does not consider query noise. We can generalize the BCPNN rule including query noise as we have done for the optimal Bayesian rule in section 2.2. Defining $p_{\mathfrak{bc}}(i) := \mathrm{pr}[v_j^\mu = 0]p_{\mathfrak{bc}|0} + \mathrm{pr}[v_j^\mu = 1]p_{\mathfrak{bc}|1}$ (for any j), it is

$$\mathrm{pr}[\tilde{u}_i = 1|\mathfrak{M}(j)] = \frac{M_1'(i)}{M}(1 - p_{10}(i)) + \frac{M_0'(i)}{M}p_{01}(i), \tag{H.1}$$

$$\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}}|\mathfrak{M}(j)] \approx \prod_{i\in\mathbf{1}_{\tilde{\mathbf{u}}}} \mathrm{pr}[\tilde{u}_i = 1|\mathfrak{M}(j)], \tag{H.2}$$

$$\begin{aligned}\mathrm{pr}[v_j^\mu = 1|\mathbf{1}_{\tilde{\mathbf{u}}}, \mathfrak{M}(j)] &= \frac{\mathrm{pr}[v_j^\mu = 1|\mathfrak{M}(j)]\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}}|v_j^\mu = 1, \mathfrak{M}(j)]}{\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}}|\mathfrak{M}(j)]} \\ &= \frac{M_1}{M}\prod_{i\in\mathbf{1}_{\tilde{\mathbf{u}}}} \frac{M_{11}(1 - p_{10|1}) + M_{01}p_{01|1}}{M_1\frac{M_1'(1-p_{10})+M_0'p_{01}}{M}} \\ &= \left(\frac{M}{M_1}\right)^{z-1}\prod_{i\in\mathbf{1}_{\tilde{\mathbf{u}}}} \frac{M_{11}(1 - p_{10|1}) + M_{01}p_{01|1}}{M_1'(1 - p_{10}) + M_0'p_{01}},\end{aligned} \tag{H.3}$$

where $z := |\mathbf{1}_{\tilde{\mathbf{u}}}| = \sum_{i=1}^m \tilde{u}_i$ denotes the number of one-entries in the query vector. Thus, taking logarithms yields synaptic weights w_{ij} and firing thresholds Θ_j,

$$-\Theta_j = \log 2 + \log\frac{M_1}{M}, \tag{H.4}$$

$$w_{ij} := \log\frac{(M_{11}(1 - p_{10|1}) + M_{01}p_{01|1})M}{(M_1'(1 - p_{10}) + M_0'p_{01})M_1}, \tag{H.5}$$

where we have again skipped indices i, j for brevity. Transition probabilities can again be estimated as in equations 2.19 and 2.20.

H.2 The BCPNN2 Rule: Including Inactive Query Components. As discussed in section 3.3, we can improve the BCPNN rule by also considering the zero-entries in a query pattern, that is, by computing

$$\mathrm{pr}[v_j^\mu = 1|\tilde{\mathbf{u}}, \mathfrak{M}(j)] = \frac{M_1}{M}\prod_{i\in\mathbf{1}_{\tilde{\mathbf{u}}}} \frac{\frac{M_{11}(1-p_{10|1})+M_{01}p_{01|1}}{M_1}}{\frac{M_1'(1-p_{10})+M_0'p_{01}}{M}}\prod_{i\in\mathbf{0}_{\tilde{\mathbf{u}}}} \frac{\frac{M_{01}(1-p_{01|1})+M_{11}p_{10|1}}{M_1}}{\frac{M_0'(1-p_{01})+M_1'p_{10}}{M}}, \tag{H.6}$$

$$= \left(\frac{M}{M_1}\right)^{m-1} \prod_{i=1}^{m} \frac{M_{01}(1-p_{01|1}) + M_{11}p_{10|1}}{M'_0(1-p_{01}) + M'_1 p_{10}} \times \prod_{i \in \mathbf{1}_{\tilde{\mathbf{u}}}} \frac{(M_{11}(1-p_{10|1}) + M_{01}p_{01|1})(M'_0(1-p_{01}) + M'_1 p_{10})}{(M_{01}(1-p_{01|1}) + M_{11}p_{10|1})(M'_1(1-p_{10}) + M'_0 p_{01})}, \tag{H.7}$$

and thus

$$-\Theta_j = \log 2 + (m-1)\log\frac{M}{M_1} + \sum_{i=1}^{m} \log \frac{M_{01}(1-p_{01|1}) + M_{11}p_{10|1}}{M'_0(1-p_{01}) + M'_1 p_{10}}, \tag{H.8}$$

$$w_{ij} = \log \frac{(M_{11}(1-p_{10|1}) + M_{01}p_{01|1})(M'_0(1-p_{01}) + M'_1 p_{10})}{(M_{01}(1-p_{01|1}) + M_{11}p_{10|1})(M'_1(1-p_{10}) + M'_0 p_{01})}. \tag{H.9}$$

H.3 The BCPNN3 Rule: Eliminating pr[$\tilde{\mathbf{u}}$]. As discussed in section 3.3, we can improve the BCPNN rule by computing the odds ratio:

$$\begin{aligned} \frac{\mathrm{pr}[v_j^\mu = 1 | \mathbf{1}_{\tilde{\mathbf{u}}}, \mathfrak{M}(j)]}{\mathrm{pr}[v_j^\mu = 0 | \mathbf{1}_{\tilde{\mathbf{u}}}, \mathfrak{M}(j)]} &= \frac{\mathrm{pr}[v_j^\mu = 1 | \mathfrak{M}(j)]\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}} | v_j^\mu = 1, \mathfrak{M}(j)]}{\mathrm{pr}[v_j^\mu = 0 | \mathfrak{M}(j)]\mathrm{pr}[\mathbf{1}_{\tilde{\mathbf{u}}} | v_j^\mu = 0, \mathfrak{M}(j)]} \\ &= \frac{\frac{M_1}{M}\prod_{i \in \mathbf{1}_{\tilde{\mathbf{u}}}} \frac{M_{11}(1-p_{10|1}) + M_{01}p_{01|1}}{M_1}}{\frac{M_0}{M}\prod_{i \in \mathbf{1}_{\tilde{\mathbf{u}}}} \frac{M_{10}(1-p_{10|0}) + M_{00}p_{01|0}}{M_0}} \\ &= \left(\frac{M_0}{M_1}\right)^{z-1} \prod_{i \in \mathbf{1}_{\tilde{\mathbf{u}}}} \frac{M_{11}(1-p_{10|1}) + M_{01}p_{01|1}}{M_{10}(1-p_{10|0}) + M_{00}p_{01|0}}, \end{aligned} \tag{H.10}$$

and thus

$$-\Theta_j = \log \frac{M_1}{M_0}, \tag{H.11}$$

$$w_{ij} = \log \frac{(M_{11}(1-p_{10|1}) + M_{01}p_{01|1})M_0}{(M_{10}(1-p_{10|0}) + M_{00}p_{01|0})M_1}. \tag{H.12}$$

H.4 The SNR of the BCPNN3 Rule. One can show that linearizing the BCPNN-type rules also yields the covariance rule, as shown in section 3.2 for the optimal Bayesian rule (Knoblauch, 2010a). By this, one may be tempted to believe that the BCPNN model would also be optimal in the limit $Mpq \to \infty$. However, asymptotically identical first-order terms of single synaptic weights is not a sufficient condition for identical network

performance since $Mpq \to \infty$ implies a diverging synapse number. In fact, the following analysis shows that the BCPNN3 rule has a lower SNR than the optimal Bayes rule, which also excludes the optimality of the BCPNN model. We can easily adapt the SNR analysis of section 2.3 to the BCPNN3 rule simply by skipping all terms relating to inactive query components $\tilde{u}_i = 0$. Equivalently to equations H.11 and H.12, the biological formulation of the BCPNN3 model as

$$\begin{aligned}\hat{v}_j = 1 \Leftrightarrow (c+f)\log\frac{M_0}{M_1} + \sum_{i=1}^{m} \tilde{u}_i \log\frac{M_{11}(1-p_{10}) + M_{01}p_{01}}{M_{10}(1-p_{10}) + M_{00}p_{01}} \\ \geq \Theta_j := \log\frac{M_0}{M_1}.\end{aligned} \tag{H.13}$$

In analogy to equation B.1, the potential x_j of content neuron j writes as

$$\begin{aligned}x_j = (c+f)\log\frac{M_0}{M_1} + \sum_{i=1}^{c} \log\frac{M_1 p_{01} + M_{11}(1-p_{01}-p_{10})}{M_0(1-p_{10}) - M_{00}(1-p_{01}-p_{10})} \\ + \sum_{i=k+1}^{k+f} \log\frac{M_1 p_{01} + M_{11}(1-p_{01}-p_{10})}{M_0(1-p_{10}) - M_{00}(1-p_{01}-p_{10})}.\end{aligned} \tag{H.14}$$

In analogy to equations B.4 and B.5, the first-order approximations of mean low and high potentials are

$$\mu'_{\mathrm{lo}} \approx -c\frac{(1-p)(1-p_{01}-p_{10})}{M_0(p_{01}+p(1-p_{01}-p_{10}))} + f\frac{p(1-p_{01}-p_{10})}{M_0(p_{01}+p(1-p_{01}-p_{10}))}, \tag{H.15}$$

$$\mu'_{\mathrm{hi}} \approx c\frac{(1-p)(1-p_{01}-p_{10})}{M_1(p_{01}+p(1-p_{01}-p_{10}))} - f\frac{p(1-p_{01}-p_{10})}{M_1(p_{01}+p(1-p_{01}-p_{10}))}. \tag{H.16}$$

In analogy to equation B.6 the mean difference $\Delta\mu := \mu_{\mathrm{hi}} - \mu_{\mathrm{lo}}$ between the high and low distributions is

$$\frac{\Delta\mu}{1-p_{01}-p_{10}} \approx \frac{c(1-p)-fp}{p(1+\frac{1-p}{p}p_{01}-p_{10})}\left(\frac{1}{M_1}+\frac{1}{M_0}\right). \tag{H.17}$$

In analogy to equation B.8, the variances of dendritic potentials are

$$\frac{\sigma_{\text{lo}}^2}{(1-p_{01}-p_{10})^2} \approx \frac{\sigma_{\text{hi}}^2}{(1-p_{01}-p_{10})^2} \approx \frac{(c+f)(1-p)}{p(1+\frac{1-p}{p}p_{01}-p_{10})^2}\left(\frac{1}{M_1}+\frac{1}{M_0}\right). \tag{H.18}$$

Thus, asymptotically, for $c=\tilde{\lambda}k$ and $f=\tilde{\kappa}k$ and assuming large networks and consistent error estimation such that $k=pm$, $p_{01}=f/(m-k)=\tilde{\kappa}\,p/(1-p)$, $p_{10}=(k-c)/k=1-\tilde{\lambda}$, we obtain in analogy to equations 2.25 and 2.26,

$$\frac{\Delta\mu}{(\tilde{\lambda}-\frac{p}{1-p}\tilde{\kappa})(1/M_1+1/M_0)} \approx m\frac{\tilde{\lambda}(1-p)-\tilde{\kappa}\,p}{\tilde{\lambda}+\tilde{\kappa}} \times \frac{\sigma_{\text{lo/hi}}^2}{(\tilde{\lambda}-\frac{p}{1-p}\tilde{\kappa})^2(1/M_1+1/M_0)} \tag{H.19}$$

$$\approx \frac{m(1-p)}{\tilde{\lambda}+\tilde{\kappa}}. \tag{H.20}$$

Therefore, similar to equation 2.28, for large $M_1 \approx Mq$ and including network connectivity P, the SNR $R=\Delta\mu/\sigma$ can be obtained from

$$R^2 \approx \frac{Pm}{Mq(1-q)}\frac{(1-p)(\tilde{\lambda}-\frac{p}{1-p}\tilde{\kappa})^2}{\tilde{\lambda}+\tilde{\kappa}}. \tag{H.21}$$

Thus, asymptotically for $Mpq \to \infty$, the squared SNR for the BCPNN3 rule is factor $1-p(\tilde{\lambda}+\tilde{\kappa})$ worse than for the optimal Bayesian model.

Acknowledgments

I am grateful to Julian Eggert, Marc-Oliver Gewaltig, Helmut Glünder, Edgar Körner, Ursula Körner, Anders Lansner, Günther Palm, Friedrich Sommer, and the two anonymous reviewers for helpful discussions and comments.

References

Abramowitz, M., & Stegun, I. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* New York: Dover.

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

Amari, S.-I. (1977). Neural theory of association and concept-formation. *Biological Cybernetics, 26*, 175–185.

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

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

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

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

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

Borwein, J., & Bailey, D. (2003). *Mathematics by experiment: Plausible reasoning in the 21st century.* Wellesley, MA: AK Peters.

Braitenberg, V. (1978). Cell assemblies in the cerebral cortex. In R. Heim & G. Palm (Eds.), *Lecture notes in biomathematics (21). Theoretical approaches to complex systems* (pp. 171–188). Berlin: Springer-Verlag.

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

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

Chechik, G., Meilijson, I., & Ruppin, E. (2001). Effective neuronal learning with ineffective Hebbian learning rules. *Neural Computation, 13*, 817–840.

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

Dayan, P., & Sejnowski, T. (1993). The variance of covariance rules for associative matrix memories and reinforcement learning. *Neural Computation, 5*, 205–209.

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

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103–130.

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

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

Gardner, E. (1988). The space of interactions in neural network models. *J. Phys. A: Math. Gen., 21*, 257–270.

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

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

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

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

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

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

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

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

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

Huyck, C., & Orengo, V. (2005). Information retrieval and categorization using a cell assembly network. *Neural Computing and Applications, 14*(4), 282–289.

Johansson, C., & Lansner, A. (2007). Imposing biological constraints onto an abstract neocortical attractor network model. *Neural Computation, 19*(7), 1871–1896.

Johansson, C., Sandberg, A., & Lansner, A. (2002). A neural network with hypercolumns. In J. Dorronsoro (Ed.), *Proceedings of the International Conference on Artificial Neural Networks (ICANN)* (pp. 192–197). Berlin: Springer-Verlag.

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

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

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

Knoblauch, A. (2007). On the computational benefits of inhibitory neural associative networks (HRI-EU Rep. 07-05). Offenbach/Main, Germany: Honda Research Institute Europe.

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

Knoblauch, A. (2009a). *Neural associative networks with optimal Bayesian learning* (HRI-EU Rep. 09-02). Offenbach/Main, Germany: Honda Research Institute Europe.

Knoblauch, A. (2009b). The role of structural plasticity and synaptic consolidation for memory and amnesia in a model of cortico-hippocampal interplay. In J. Mayor, N. Ruh, & K. Plunkett (Eds.), *Connectionist Models of Behavior and Cognition II: Proceedings of the 11th Neural Computation and Psychology Workshop* (pp. 79–90). Singapore: World Scientific.

Knoblauch, A. (2009c). *Zip nets: Neural associative networks with non-linear learning* (HRI-EU Rep. 09-03). Offenbach/Main, Germany: Honda Research Institute Europe.

Knoblauch, A. (2010a). *Comparison of the Lansner/Ekeberg rule to optimal Bayesian learning in neural associative memory* (HRI-EU Rep. 10-06). Offenbach/Main, Germany: Honda Research Institute Europe.

Knoblauch, A. (2010b). Zip nets: Efficient associative computation with binary synapses. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (pp. 4271–4278). Piscataway, NJ: IEEE Press.

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

Knoblauch, A., Palm, G., & Sommer, F. (2010). Memory capacities for synaptic and structural plasticity. *Neural Computation, 22*(2), 289–341.

Kohonen, T. (1972). Correlation matrix memories. *IEEE Transactions on Computers, C-21*, 353–359.

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

Kohonen, T., & Oja, E. (1976). Fast adaptive formation of orthogonalizing filters and associative memory in recurrent networks of neuron-like elements. *Biological Cybernetics, 21*(2), 85–95.

Kononenko, I. (1989). Bayesian neural networks. *Biological Cybernetics, 61*(5), 361–370.

Kononenko, I. (1991). Semi-naive Bayesian classifier. In *Proceedings of the 6th European Working Session on Learning* (pp. 206–219). Berlin: Springer-Verlag.

Kononenko, I. (1994). On Bayesian neural networks. *Informatica (Slovenia), 18*(2), 183–195.

Lansner, A. (2009). Associative memory models: From the cell-assembly theory to biophysically detailed cortex simulations. *Trends in Neurosciences, 32*(3), 178–186.

Lansner, A., & Ekeberg, O. (1987). An associative network solving the "4-bit adder problem." In M. Caudill & C. Butler (Eds.), *Proceedings of the IEEE First International Conference on Neural Networks* (pp. II–549). Piscataway, NJ: IEEE.

Lansner, A., & Ekeberg, O. (1989). A one-layer feedback artificial neural network with a Bayesian learning rule. *International Journal of Neural Systems, 1*(1), 77–87.

Lansner, A., & Holst, A. (1996). A higher order Bayesian neural network with spiking units. *International Journal of Neural Systems, 7*(2), 115–128.

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.

MacKay, D. (1991). Maximum entropy connections: Neural networks. In *Proceedings of the 10th International Workshop on Maximum Entropy and Bayesian Methods (MaxEnt 1990).* Dordrecht: Kluwer.

Markram, H., Toledo-Rodriguez, M., Wang, Y., Gupta, A., Silberberg, G., & Wu, C. (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience, 5*, 793–807.

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.

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

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

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

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

Palm, G. (1988a). Local synaptic rules with maximal information storage capacity. In H. Haken (Ed.), *Neural and synergetic computers* (pp. 100–110). Berlin: Springer-Verlag.

Palm, G. (1988b). On the asymptotic information storage capacity of neural networks. In R. Eckmiller & C. von der Malsburg (Eds.), *Neural computers* (pp. 271–280). Berlin: Springer-Verlag.

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

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

Papoulis, A. (1991). *Probability, random variables, and stochastic processes* (3rd ed.). New York: McGraw-Hill.

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

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

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

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

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

Sandberg, A., Lansner, A., Petersson, K., & Ekeberg, O. (2000). A palimpsest memory based on an incremental Bayesian learning rule. *Neurocomputing, 32–33*, 987–994.

Sejnowski, T. (1977a). Statistical constraints on synaptic plasticity. *Journal of Theoretical Biology, 69*, 385–389.

Sejnowski, T. (1977b). Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology, 4*, 303–321.

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

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

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

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

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

Sterratt, D., & Willshaw, D. (2008). Inhomogeneities in heteroassociative memories with linear learning rules. *Neural Computation, 20*, 311–344.

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

Turrigiano, G., Leslie, K., Desai, N., Rutherford, L., & Nelson, S. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature, 391*, 892–896.

Van Welie, I., Van Hooft, J., & Wadman, W. (2004). Homeostatic scaling of neuronal excitability by synaptic modulation of somatic hyperpolarization-activated IH channels. *Proceedings of the National Academy of Sciences, USA, 101*(14), 5123–5128.

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

Weisstein, E. (1999). *Mercator series.* Available online at http://mathworld.wolfram.com/MercatorSeries.html.

Wichert, A. (2006). Cell assemblies for diagnostic problem-solving. *Neurocomputing, 69*, 810–824.

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

Willshaw, D., & Dayan, P. (1990). Optimal plasticity in matrix memories: What goes up must come down. *Neural Computation, 2*, 85–93.

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

Zhang, H. (2004). The optimality of naive bayes. In V. Barr & Z. Markov (Eds.), *Proceedings of the 17th Florida Artificial Intelligence Research Society Conference* (pp. 562–567). Menlo Park, CA: AAAI Press.

Received April 28, 2010; accepted October 27, 2010.