frontiers in
COMPUTATIONAL NEUROSCIENCE

ORIGINAL RESEARCH ARTICLE
published: 07 November 2014
doi: 10.3389/fncom.2014.00140

A high-capacity model for one shot association learning in the brain

Hafsteinn Einarsson[1]*, Johannes Lengler[1] and Angelika Steger[1,2]

[1] *Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich, Zürich, Switzerland*
[2] *Collegium Helveticum, Zürich, Switzerland*

Edited by:
Stefano Fusi, Columbia University, USA

Reviewed by:
Andreas Knoblauch, Albstadt-Sigmaringen University, Germany
Marcus K. Benna, Columbia University, USA

****Correspondence:***
Hafsteinn Einarsson, Department of Computer Science, Institute of Theoretical Computer Science, ETH Zürich, Universitätstrasse 6, 8092 Zürich, Switzerland
e-mail: hafsteinn.einarsson@inf.ethz.ch

We present a high-capacity model for one-shot association learning (hetero-associative memory) in sparse networks. We assume that basic patterns are pre-learned in networks and associations between two patterns are presented only once and have to be learned immediately. The model is a combination of an Amit-Fusi like network sparsely connected to a Willshaw type network. The learning procedure is palimpsest and comes from earlier work on one-shot pattern learning. However, in our setup we can enhance the capacity of the network by iterative retrieval. This yields a model for sparse brain-like networks in which populations of a few thousand neurons are capable of learning hundreds of associations even if they are presented only once. The analysis of the model is based on a novel result by Janson et al. on bootstrap percolation in random graphs.

Keywords: one shot learning, hetero-associative memory, relation learning, bootstrap percolation, iterative retrieval, stochastic Hebbian learning, memory capacity

1. INTRODUCTION

In the last decades the problem of fast pattern learning has been intensively studied. Amit and Fusi (1994) introduced a model of auto-associative memory for sparsely coded patterns in fully connected neuronal networks and showed that in this model an ensemble of N neurons can store almost quadratically many patterns before it starts forgetting old ones, even if each pattern is only presented once. In this paper we consider hetero-associative memory instead of auto-associative memory, i.e., relation learning instead of pattern learning. Moreover, we do not only require fast learning, but also fast retrieval of the learned associations. We incorporate this requirement into our model by considering for each retrieval only the first spike of each neuron, ignoring all further spikes. In particular, our model is spike-based rather than rate-based.

Traditionally there have been two main models for hetero-associative memory: the model by Willshaw et al. (1969) based on clipped Hebbian learning, and the networks introduced by Hopfield (1982) (see also Knoblauch et al., 2010 for a review and comparison). Both achieve storage capacities close to the information theoretic upper bound for sparsely coded patterns (Knoblauch et al., 2010). The Hopfield networks are rate-based and aim for convergence to a stable state through auto-feedback, thus they are designed for retrieval in medium or long time scale. The fast learning model in Amit and Fusi (1994) falls in this category, and we compare with it in more detail in Section 2.2. On the other hand, the Willshaw model is both fast-learning and fast-retrieving, but high capacities come at the cost of low retrieval accuracy (Buckingham and Willshaw, 1991). Various ways have been found to overcome this issue, including adaptive thresholds as in Buckingham and Willshaw (1991) and bidirectional iterative retrieval schemes as in Sommer and Palm (1998). Our model is related to the latter approach, except that we are more restrictive in the retrieval procedure so that the model is still fast-retrieving (cf. also Section 4.1): we consider a bipartite graph with partite sets $\mathcal{A}$ and $\mathcal{B}$, where all edges are directed from $\mathcal{A}$ to $\mathcal{B}$ ("afferent edges"), and iterative retrieval is only achieved by the edges in $\mathcal{B}$ ("recurrent edges") (see **Figure 1** for the setup). In this respect, a similar retrieval scheme for the Willshaw model has been studied by Knoblauch and Palm (2001), with the difference that they used inhibition to stop the spread of activity after the pattern is activated, and that they use a global feedback scheme for threshold control. The latter feature allowed for higher fidelity of retrieval and for a threshold that it is independent of the pattern size. While the Willshaw model may also serve as a model of fast learning, we follow the approach in Amit and Fusi (1994) and use binary Hebbian learning with pruning (see below) so that the total number of synapses is unaffected by the number of learned associations. However, in contrast to the model of Amit and Fusi, since we only consider the first spike of each neuron, a neuron can never go from state "active" to "inactive" since it can not retract a spike that it elicited earlier. All these restrictions are biologically motivated, and the biological background can be found in more detail in Section 4.1.

The guiding idea for our model is that a pattern may be stored locally in a cortical column of $N \approx 5000$ neurons, but that it is necessary to associate patterns in different columns or even different regions of the brain. Therefore, the density between different patterns is much lower than the density within a pattern (Binzegger et al., 2004). The combination of low afferent density and a population size of only 5000 neurons makes it impossible to transfer the existing models for fast learning straightforwardly (cf. Section 2.2). However, by making use of recurrent connections (connections between neurons within a

pattern) we are able to show that the resulting iterative retrieval of the pattern allows our model to operate in the range prescribed by biology (see **Figure 2** for an example).

We analyze the model both mathematically and with simulations. The mathematical analysis investigates the limiting case $N \to \infty$. Our main tool is the result of Janson et al. (2012) for bootstrap percolation in a random graph. We extend their result in order to analyze iterative retrieval of a pattern. As a side effect of our calculation we also deduce optimal parameters for a high memory capacity. In particular, we find that the desired plasticity has a non-trivial optimum: it should neither be too small nor too high, cf. **Figure 3**. Similarly, memory capacity depends on the patterns size in a unimodular way, cf. **Figure 4**, that is the pattern size should neither be too small nor too big.

2. METHODS

2.1. MODEL OVERVIEW AND ASSUMPTIONS

2.1.1. Setup and terminology

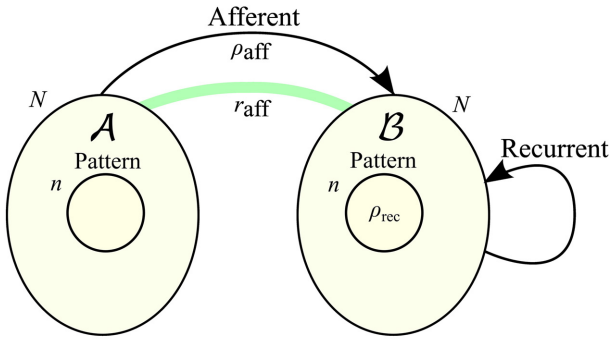
Let G be a directed graph with vertex set $V = \mathcal{A} \cup \mathcal{B}$ where the sets $\mathcal{A}$ and $\mathcal{B}$ are of equal size N (cf. **Figure 1**). Edges between vertices of the same set are called *recurrent*, those from $\mathcal{A}$ to $\mathcal{B}$ *afferent*. All edges between $\mathcal{A}$ and $\mathcal{B}$ are directed toward $\mathcal{B}$. Edges can be either weak or strong. A vertex gets activated if it is connected to at least K active vertices by strong edges, where K is a parameter of the model.

We consider the following learning problem. Let $(A_i)_{i \geq 0}$ and $(B_i)_{i \geq 0}$ be sequences of random subsets (*patterns*) of $\mathcal{A}$ and $\mathcal{B}$, respectively, with sizes $|A_i| = |B_i| = n$ for $i \geq 0$. We sequentially present each pair (A_i, B_i) once. At the presentation of each pair we may change some of the afferent edges from strong to weak or vice versa. In the recall phase we activate all vertices of A_i and let activation propagate. The pair (A_i, B_i) is called *memorized* if activation of the vertices in A_i leads to an activation of the vertices in B_i. More precisely, we want to activate at least $\alpha_{\text{fid}} n$ vertices in B_i (fidelity), and at most $\alpha_{\text{spc}} n$ vertices outside of B_i (specificity). An insertion is considered *successful* if the pair is memorized right after insertion.

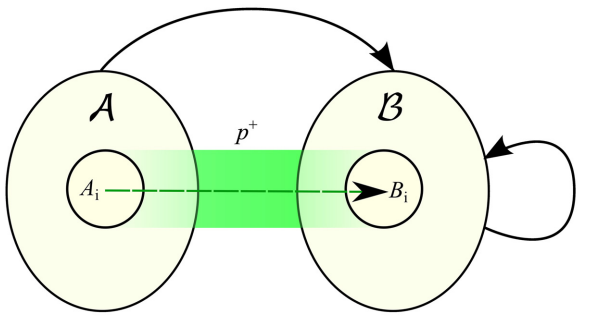
Note that due to the presence of recurrent edges activation can *propagate*: a small set of initially active vertices in B_i (arising from activity in A_i) can eventually activate a much bigger set. More precisely, we start with an active set consisting of the vertices in A_i. In the first round we then activate all vertices that have K neighbors in A_i to which they are connected by strong edges. In the second round all vertices get activated that are connected by K strong edges to vertices in A_i *or* to vertices that were activated in the first round, and so forth. Note that here we tacitly assume that signal propagation is so fast that activation can take place in rounds. Since only strong edges count for activating a neuron, we define the *degree* $\deg(v; S)$ of a vertex v with respect to a set $S \subset V$ to be the number of strong edges between v and S.

Observe that due to our setup the oldest associations have the worst quality. Moreover, we choose a pruning parameter (see below) in such a way that the expected number of strong edges remains constant regardless of the number of shown relations, i.e., the model is a palimpsest (see Nadal et al., 1986). (Note that we take the point of view that the edges within a pattern are fixed, while the afferent edges are plastic; that is, the model is a palimpsest for association learning, not for pattern learning.) We are thus interested in determining the maximum number M (the *capacity* of the model) of additional associations that can be learned so that the set A_0 can still activate its partner B_0.

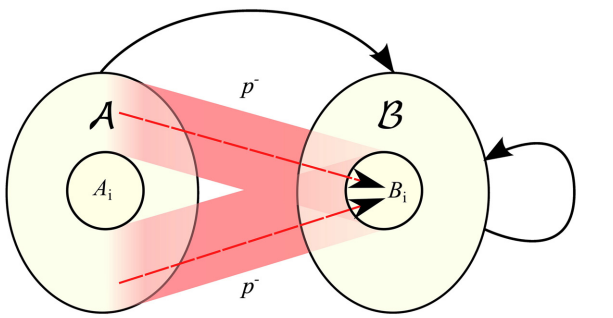
We study learning in a sparse random setting. We assume that afferent edges are present with probability ρ_{aff}, independently. Before learning starts we turn every afferent edge strong with probability r_{aff}, independently. Note that r_{aff} impacts how many edges a vertex outside B_0 receives from A_0 which also depends on n.



Network setup: Each edge between $\mathcal{A}$ and $\mathcal{B}$ is present with probability ρ_{aff} and each such edge is strong with probabililty r_{aff}. Within patterns in $\mathcal{B}$ the recurrent edges are present and strong with probability ρ_{rec}.



Learning step: Each afferent edge between patterns is made strong with probability p^+.



Pruning step: Each afferent edge with source outside pattern is pruned away with probability p^-.

FIGURE 1 | Network setup and learning procedure. See Section 2.1 for a more detailed explanation.

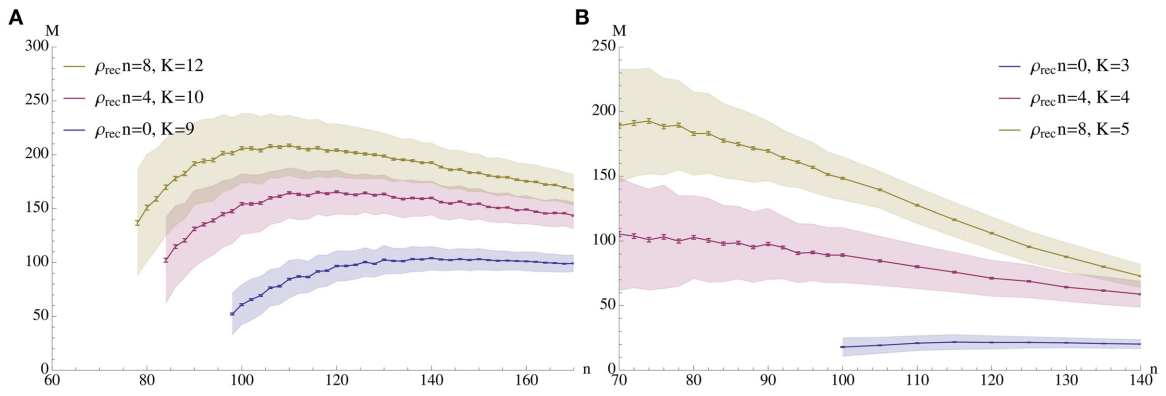


FIGURE 2 | Capacity (*M*) vs. pattern size (*n*). The plots show the mean number of associations that can be inserted until the first association is no longer memorized. The blue curves correspond to a setup without percolation (i.e., recurrent density is 0), while the red and yellow curve correspond to percolation with recurrent degree 4 and 8 respectively. In **(A)** $\rho_{\text{aff}} = 0.2$, $r_{\text{aff}} = 0.1$, $p^+ = 0.6$; the values for K were chosen to give stable results for n in a wide range. **(B)** Shows a sparser setup with $\rho_{\text{aff}} = 0.05$. Here $r_{\text{aff}} = 0.05$, $p^+ = 0.6$ and K is again chosen to optimize capacity. In both experiments $N = 5000$ and the mean is over 500 trials with standard deviation as shown; error bars represent standard mean error. In both cases capacity was only considered if at least 99% of all insertions were successful.

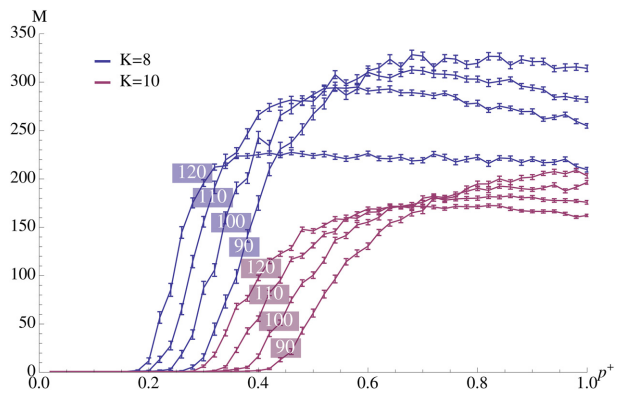


FIGURE 3 | Plasticity (afferent insertion probability, p^+) vs. capacity for different pattern sizes and values of *K*. The relevant parameters are chosen as in **Figure 2A** where $\rho_{\text{rec}} n = 4$. Error bars represent standard mean error and there are 100 trials per data point. We observe that values close to the optimum can be obtained with relatively small insertion probability. Note that the exact value of p^+ is not very important as long as it exceeds a certain threshold.

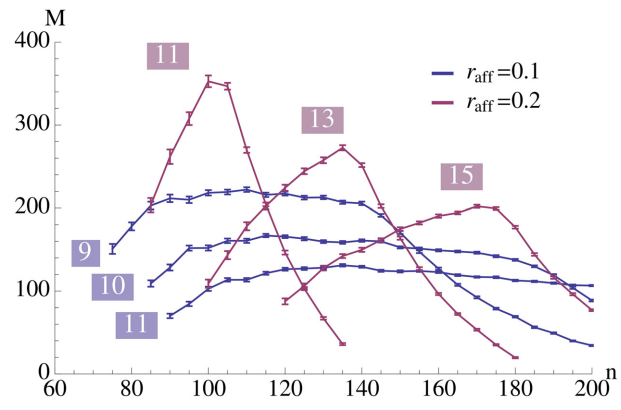


FIGURE 4 | This plot illustrates the same experiment as in Figure 2A, but we vary r_{aff} and provide the data for different values of *K* (as indicated in the plot). It shows that increasing r_{aff} can yield a higher capacity but it comes at the price of only working for a small range of possible pattern sizes. Each data point is the mean of 100 trials where capacity is only considered if at least 99% of the insertions were successful.

As we assume that patterns B_i correspond to "concepts" that are already known, we insert recurrent edges as follows. Each edge in $\mathcal{B}$ is present with probability ρ_{rec} independent of other edges and all of them are initially weak. For each pattern B_i we turn all the edges between pairs of vertices in it strong. In particular, $\mathcal{B}$ corresponds to a sparsely connected Willshaw network.

2.1.2. Learning procedure

In order to learn an association (A_i, B_i) during its presentation we

- Turn each weak afferent edge between A_i and B_i strong ("insert it") with probability p^+,
- Turn each strong afferent edge between $A \setminus A_i$ and B_i weak ("prune it") with probability p^-,

cf. Section 4.2 Note that the first step is a stochastic form of Hebbian learning (Barrows, 1998). The second step is a normalization step. Hence, we choose p^- in such a way that the expected degree for each vertex in B_i stays constant. Observe that the "randomness" assumption means that a vertex $b \in B_i$ is expected to have $\rho_{\text{aff}} n$ edges from vertices in A_i out of which an r_{aff}-fraction are strong and $(N - n)\rho_{\text{aff}}$ edges from vertices in $\mathcal{A} \setminus A_i$ out of which also an r_{aff}-fraction are strong. The learning procedure will thus, in expectation, turn $(1 - r_{\text{aff}})\rho_{\text{aff}} n p^+$ edges strong and $r_{\text{aff}}\rho_{\text{aff}}(N - n)p^-$ edges weak. We thus set

$$p^- = \frac{1 - r_{\text{aff}}}{r_{\text{aff}}} \cdot \frac{n}{N - n} \cdot p^+. \tag{1}$$

2.2. COMPARISON WITH THE MODEL OF AMIT AND FUSI

Our model builds upon the work on the well-studied (Amit and Fusi, 1994) model (*AF model*) and its extensions (cf. Battaglia and Fusi, 1994; Brunel et al., 1998; Romani et al., 2008; Amit and Huang, 2010). In particular, the learning paradigm is identical. The main differences are:

- The AF model studies *auto-associative memory* instead of *hetero-associative memory*. Thus, it considers only one population of neurons (instead of two populations in our model), and in the learning phase patterns are presented instead of pairs of patterns. Consequently, the AF model does not need to distinguish between recurrent and afferent connections. It is well-known that association learning is easier to humans than pattern learning (see Fanselow, 1990).
- All studies on the AF model assumed a complete underlying graph. However, it is straightforward to extend the model to sparse, randomly connected graphs, cf. below. The asymptotics of the capacity remains the same; more precisely, both for the complete graph and the sparse random graph, it is possible to learn $\theta(N^2/\log^2 N)$ patterns before the first pattern is forgotten. However, the density (probability of two neurons being connected by a synapse) will enter via the constant hidden in the θ-notation. Actually, it dramatically reduces the capacity for neuron populations of size, say, $N = 5000$, i.e., for magnitudes of N where neurophysiologically we may assume a constant density (cf. below).
- The AF model investigates whether an activated pattern forms an attractor state in the state space. Consequently, a pattern is remembered in the AF model if every neuron in a pattern A has at least K neighbors in A, and every neuron outside of A has less than K neighbors in A. This view is not suited if the underlying graph is assumed to be a sparse random graph, as there is always a constant probability that a vertex has less than K (strong or weak) neighbors in the pattern. A pattern containing such a vertex can then never be in an attractor state, even if all the edges in the pattern turn strong. We therefore require that only an α_{fid}-fraction of the pattern is activated, where $0 < \alpha_{\text{fid}} \leq 1$ is a parameter that we may choose. For $\alpha_{\text{fid}} = 1$ we are back in the Amit-Fusi model.

  Note that the requirement in the AF model is weakest possible in terms of attractor networks. For example, one might ask what part of the state space is attracted into the pattern state. For such question, the update rule may be important, and it is known that iterative retrieval is superior to one-step-retrieval (Schwenker et al., 1996). However, all such questions break down if the pattern state is not a stable attractor.
- The other, and actually main, difference to the AF model is that we consider hetero-associative memory instead of auto-associative memory, i.e., we do not activate the pattern itself (and require that it stays active), but we activate a pattern A_i and investigate whether this pattern is able to activate its "partner" B_i. Without recurrent edges this boils down to the question whether all (or, cf. above, an α_{fid}-fraction) of the neurons in B_i have at least K neighbors in A_i. This special case is equivalent to the question whether a pattern is memorized in the AF model. With recurrent edges propagation of activity will allow us to show that we actually need only a small fraction of the neurons in B_i to have at least K neighbors in A_i; propagation of activity will then nevertheless ensure that an α_{fid}-fraction of the neurons in B_i is activated (see **Figure 5**). In other words, the AF-model (or rather its hetero-associative equivalent) may be viewed as a starting point of our considerations, as we essentially copied the learning rule and also focus on fast learning. However, as we consider hetero-associative memory we are able to make use of recurrent edges.

Our assumptions are motivated by facts known from neurophysiology. We assume that the two neuronal ensembles are in different areas of the brain. A neuron in the brain is connected to 10–20% of its closest neighbors, and this number drops sharply with distance exceeding 200–300 μ (see Song et al., 2005; Le Bé et al., 2006; Perin et al., 2011; Levy and Reyes, 2012). The size of the input layer of a cortical column contains roughly $N \approx 5000$ neurons (Meyer et al., 2010). This is also roughly the number of neurons within a ball of radius 300 μ (Beaulieu and Colonnier, 1983). The data from Kalisman et al. (2005) suggest that plausible values for the densities within such neuron populations of such a size are of the order of 0.1–0.2, while the afferent density is substantially lower (Binzegger et al., 2004).

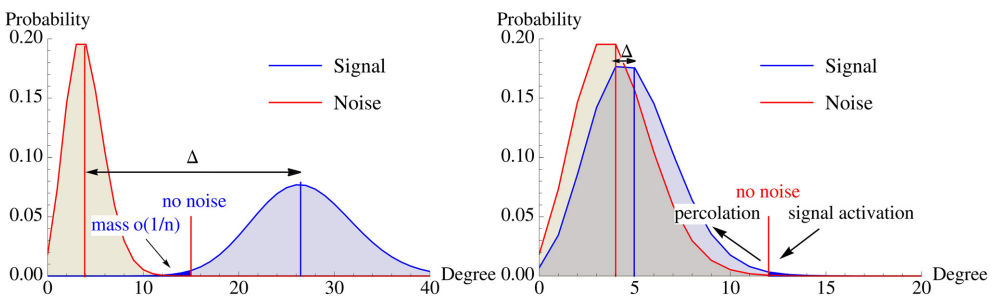


FIGURE 5 | Illustration of the degree distribution of vertices in $\mathcal{B} \setminus B_0$ (noise) and B_0 (signal) w.r.t. A_0. Left: Without recurrent edges, the overlap of both distributions must be extremely small. **Right:** Percolation allows activating the pattern via recurrent edges with fewer strong afferent edges. This dramatically influences the location of the optimal activation threshold K and the signal degree distribution can be much closer to the noise distribution while the pattern is still memorized. The differences between both scenarios are exaggerated to highlight the different mechanisms.

3. RESULTS

3.1. THEORETICAL RESULTS

The effect of learning an association will diminish over time due to later pruning steps. Clearly, this is most critical for the association (A_0, B_0). In this section we thus analyze the recall properties of this association only. Note that we do not aim for precise asymptotics, but rather we give an intuition for the underlying mechanisms of the process. Within the calculations we will therefore make some simplifying assumptions (the Erdős-Rényi assumption in Section 3.1.2 and the Janson assumption in Section 3.1.4). In Section 3.1.5 we then discuss the effect of these assumptions. In order to study whether we can activate pattern B_0 by activating A_0 we need to know the degree distribution of vertices $b \in B_0$ (for fidelity) and $b \in B \setminus B_0$ (for specificity) into A_0. To do so we first consider the probability that a single, fixed edge is strong.

3.1.1. Edge probabilities

Let $a \in A_0$ be arbitrary. For $b \in \mathcal{B}$ we denote by p_{signal} and p_{noise} the probability $\Pr[\{a, b\}\text{strong} \mid \{a, b\}\text{is an edge}]$ in the cases $b \in B_0$ and $b \in \mathcal{B} \setminus B_0$, respectively. First consider $b \in \mathcal{B} \setminus B_0$. After presentation of (A_0, B_0) we have $p_{\text{noise}} = r_{\text{aff}}$ as the learning procedure did not touch the edge $\{a, b\}$. We show by straightforward induction that p_{noise} remains at this value regardless of how many additional pairs (A_i, B_i) are learned, so

$$p_{\text{noise}} = r_{\text{aff}} \tag{2}$$

at any time. Indeed, after presenting one more association (A_i, B_i), $\{a, b\}$ is strong with probability $r_{\text{aff}} + (1 - r_{\text{aff}})p^+$ if $a \in A_i$ (which happens with probability n/N) and with probability $r_{\text{aff}}(1 - p^-)$ if $a \notin A_i$ (which happens with probability $(N - n)/N$). Thus, $\Pr[\{a, b\}\text{strong}] = \frac{n}{N} \cdot (r_{\text{aff}} + (1 - r_{\text{aff}})p^+) + \frac{N-n}{N} \cdot r_{\text{aff}}(1 - p^-) = r_{\text{aff}}$ also in this case, where the last equality follows from Equation (1).

In contrast, p_{signal} changes after each association presentation. Let us denote by $p_{\text{signal}}(i)$ the value after i additional associations were learned. Then $p_{\text{signal}}(0) = r_{\text{aff}} + (1 - r_{\text{aff}})p^+$, and by considering an argument similar as above we see that with each new association the probability of an edge being strong drops as follows:

$$\begin{aligned} p_{\text{signal}}(i+1) &= \frac{N-n}{N} p_{\text{signal}}(i) + \frac{n^2}{N^2}\left(p_{\text{signal}}(i) + (p_{\text{signal}}(i))p^+\right) \\ &\quad + \frac{n(N-n)}{N^2} p_{\text{signal}}(i)(1 - p^-) \\ &= p_{\text{signal}}(i)\left(1 - \frac{n^2 p^+}{N^2 r_{\text{aff}}}\right) + \frac{n^2 p^+}{N^2}, \end{aligned}$$

where the last inequality again follows from Equation (1). In particular, we find that the difference $\Delta(i) := p_{\text{signal}}(i) - r_{\text{aff}}$ decays exponentially with i:

$$\begin{aligned} \Delta(i+1) &= p_{\text{signal}}(i+1) - r_{\text{aff}} \\ &= p_{\text{signal}}(i)\left(1 - \frac{n^2 p^+}{N^2 r_{\text{aff}}}\right) + \frac{n^2 p^+}{N^2} - r_{\text{aff}} \\ &= \left(p_{\text{signal}}(i) - r_{\text{aff}}\right)\left(1 - \frac{n^2 p^+}{N^2 r_{\text{aff}}}\right) \\ &= \Delta(i)\left(1 - \frac{n^2 p^+}{N^2 r_{\text{aff}}}\right). \end{aligned}$$

Consequently, we obtain an explicit formula for $p_{\text{signal}}(i)$ as

$$p_{\text{signal}}(i) = r_{\text{aff}} + \beta^i (1 - r_{\text{aff}})p^+, \tag{3}$$

where $\beta := 1 - (n/N)^2 \cdot p^+/r_{\text{aff}}$. For short reference, we will denote by $p_{\text{signal}} = p_{\text{signal}}(M)$ the probability after M presentations, where M is the capacity of the system cf. below.

3.1.2. Degree distribution

In order to investigate propagation of activity we need to know the degree distribution of vertices $b \in \mathcal{B}$ into A_0. Assuming independence of the probabilities that we computed in the last section, we get

$$\begin{aligned} \deg(b, A_0) &\sim \text{Bin}(n, \rho_{\text{aff}} \cdot p_{\text{noise}}) \\ &= \text{Bin}(n, \rho_{\text{aff}} \cdot r_{\text{aff}}) \qquad \text{for } b \in \mathcal{B} \setminus B_0 \end{aligned} \tag{4}$$

and

$$\begin{aligned} \deg(b, A_0) &\sim \text{Bin}(n, \rho_{\text{aff}} \cdot p_{\text{signal}}(i)) \\ &= \text{Bin}(n, \rho_{\text{aff}} \cdot (r_{\text{aff}} + \beta^i (1 - r_{\text{aff}})p^+)) \text{ for } b \in B_0, \end{aligned} \tag{5}$$

and all these distributions are independent.

For all asymptotic computations we assume that the edges are independent.[1] We call this the "Erdős-Rényi assumption," since it implies that the edges between A_0 and B_0 are given by an Erdős-Rényi random bipartite graph model $B_{n,n;p}$ for some edge probability p. Similarly, we assume that the edges between A_0 and $\mathcal{B} \setminus B_0$ and the edges within B_0 are given by Erdős-Rényi random graphs $G_{n,p'}$ (for some different edge probability p'). Under the "Erdős-Rényi assumption" Equations (4) and (5) are valid. Clearly, we do make some error here; however, one can show that the probability that the assumption is violated tends to zero for N tending to infinity. Similar results are known for the Willshaw model (Knoblauch, 2008). We abstain from estimating the error for finite N, but instead provide some experimental evidence in Section 3.2.

3.1.3. Learning without recurrent edges

In order to understand the effect of recurrent edges, we first consider the case of no recurrent edges. This scenario is actually very closely related to the AF model. Recall that the AF model assumes that the input must be able to activate *all* neurons in the pattern ($\alpha_{\text{fid}} = 1.0$). For a sparse setting this seems overly restricive. In this section we thus also consider the case $\alpha_{\text{fid}} = 0.5$ (for which the calculations below are particularly easy). As we will see the

[1] Strictly speaking, edge probabilities are *not* independent. For example, if there exists $1 \le i \le M$, $a_1, a_2 \in A_0 \cap A_i$ and $b \in B_0 \cap B_i$ then the events that $\{a_1, b\}$ is strong respectively that $\{a_2, b\}$ is strong are positively correlated.

benefits (in terms of memory capacity) of a such a seemingly much smaller value is in fact quite moderate.

In the previous section we argued that we may assume the degree distribution to be binomial. In this section we will furthermore assume that for large enough values binomial distributions are well approximated by normal distributions.Recall from Equation (3) that the expected probability for an edge between A_0 and B_0 to be strong is

$$p_{\text{signal}} = p_{\text{signal}}(i) = r_{\text{aff}} + \beta^i p_{\text{signal}}(0),$$

where $p_{\text{signal}}(0)$ is the probability immediately after presenting association (A_0, B_0), and $\beta = \left(1 - \frac{n^2 p^+}{N^2 r_{\text{aff}}}\right)$. Recall also that the difference $\Delta(i) = p_{\text{signal}}(i) - r_{\text{aff}}$ decays with each additional pattern by a factor of β.

The memory capacity M is determined by three variables: the factor β by which the differences $\Delta(i)$ decay, the initial difference $\Delta(0)$, and the minimal difference Δ for which the pattern can still be retrieved. More precisely, the capacity is given by $M = \log_\beta(\Delta(0)/\Delta)$.

As Amit and Fusi noticed in their seminal paper, in the $N \to \infty$ limit it is possible to learn a large number of patterns by making the decay factor β very close to one. More precisely, setting $n = \theta(\log N)$, the quotient $\Delta(0)/\Delta$ turns out to be constant, and

$$\begin{aligned} M = \log_{1/\beta}(\Delta(0)/\Delta) &= \frac{\log(\Delta(0)/\Delta)}{\log(1/\beta)} \\ &= \theta\left(\frac{1}{\log(1/\beta)}\right) = \theta\left(\frac{N^2}{n^2}\right). \end{aligned} \tag{6}$$

Here we will investigate the effect of a smaller activity threshold α_{fid}. The value of α_{fid} obviously does not change β and $\Delta(0)$. It only affects the minimal difference Δ. So we need to estimate Δ for various values of α_{fid}.

The minimal difference Δ is determined by two requirements on the activation threshold K. Firstly, K must be large enough that no noise occurs. This is the case if the probability that a neuron in $\mathcal{B} \setminus B_0$ has degree K is at most $\alpha_{\text{spc}} n/N$. Since we assume the degree distribution of such neurons to be binomially distributed with mean $\mu_{\text{spc}} = n\rho_{\text{aff}} p_{\text{noise}} = n\rho_{\text{aff}} r_{\text{aff}}$ and variance $\sigma^2_{\text{spc}} = n\rho_{\text{aff}} r_{\text{aff}}(1 - \rho_{\text{aff}} r_{\text{aff}})$, we use the normal approximation of the binomial distribution to deduce that we need

$$\frac{\alpha_{\text{spc}} n}{N} > \text{Prob}[\mathcal{N}(\mu_{\text{spc}}, \sigma_{\text{spc}}) \geq K] \approx \frac{1}{\sqrt{2\pi\sigma^2_{\text{spc}}}} e^{-(K-\mu_{\text{spc}})^2/(2\sigma^2_{\text{spc}})},$$

or equivalently

$$K > n\rho_{\text{aff}} r_{\text{aff}} + \sigma_{\text{spc}} \sqrt{2\log\left(\frac{N}{\alpha_{\text{spc}} n\sqrt{2\pi\sigma^2_{\text{spc}}}}\right)}. \tag{7}$$

Secondly, K must be small enough that we can activate an α_{fid} fraction of B_0. Similarly as above, this time using the normal distribution with mean $\mu_{\text{fid}} = n\rho_{\text{aff}} p_{\text{signal}}$ and variance $\sigma^2_{\text{fid}} = n\rho_{\text{aff}} p_{\text{signal}}(1 - \rho_{\text{aff}} p_{\text{signal}})$, we get for $\alpha_{\text{fid}} = 1 - \frac{1}{n}$ that

$$K < n\rho_{\text{aff}} p_{\text{signal}} - \sigma_{\text{fid}} \sqrt{2\log\left(\frac{1}{\alpha_{\text{fid}}\sqrt{2\pi\sigma^2_{\text{fid}}}}\right)}. \tag{8}$$

If, on the other hand, we are satisfied with $\alpha_{\text{fid}} = 0.5$, then we only need the mean of the distribution to be larger than K, so we only need

$$K < n\rho_{\text{aff}} p_{\text{signal}} \tag{9}$$

in this case.

For $\alpha_{\text{fid}} = 0.5$ we may combine inequality Equation (7) and (9) to obtain an explicit formula for the minimal difference $\Delta = p_{\text{signal}} - r_{\text{aff}}$ that is sufficient for recall:

$$\Delta \approx \frac{\sigma_{\text{spc}}}{n\rho_{\text{aff}}} \sqrt{2\log\left(\frac{N}{\alpha_{\text{spc}} n\sqrt{2\pi\sigma^2_{\text{spc}}}}\right)}. \tag{10}$$

Note that we need $\Delta < 1$, as Δ is supposed to be a probability. From this we deduce that n cannot be too small. More precisely, we need $n = \Omega(\log N)$, as already observed by Amit and Fusi (1994).

For $\alpha_{\text{fid}} = 1 - \frac{1}{n}$, we may combine inequality Equation (7) and (8) to get a bound on Δ. In this case, an explicit solution is not possible. However, keeping in mind that Δ remains bounded as $N \to \infty$, we may rewrite $p_{\text{signal}} = r_{\text{aff}} + \Delta$ to deduce

$$\begin{aligned} \Delta > \frac{\sigma_{\text{spc}}}{n\rho_{\text{aff}}} &\sqrt{2\log\left(\frac{N}{\alpha_{\text{spc}} n\sqrt{2\pi\sigma^2_{\text{spc}}}}\right)} \\ &+ \frac{\sigma_{\text{fid}}}{n\rho_{\text{aff}}} \sqrt{2\log\left(\frac{1}{\alpha_{\text{fid}}\sqrt{2\pi\sigma^2_{\text{fid}}}}\right)}. \end{aligned} \tag{11}$$

Since $\sigma_{\text{fid}} = \theta(\sqrt{n})$ the second term tends to 0. On the other hand for $n = \theta(\log N)$ the first term remains constant [and thus $\sigma_{\text{spc}} = \theta(\sqrt{\log N})$]. Therefore, we will get the same asymptotic behavior for the memory capacity from Equations (10) and (11). Thus, in the limit we will not see any difference (not even in the leading constant factor). For small values of n and N, however, both terms in Equation (11) are of the same order of magnitude. So here we do see a difference between 100% activation and 50% activation. Note however that even if both terms are of the same order of magnitude we only gain a factor of ≈2—but we would gain much more if we could replace the plus sign in Equation (11) by a minus sign. Recurrent edges allow essentially that, as we will see in the next section.

3.1.4. Learning with recurrent edges: percolation

In the previous section we derived that the number of patterns that can be learned satisfies $M \approx \frac{N^2 r_{\text{aff}}}{n^2 p^+} \log\left(\frac{\Delta(0)}{\Delta}\right)$, where $\Delta(0) = p_{\text{signal}}(0) - r_{\text{aff}}$ is the difference between signal and noise at start and Δ is the minimal difference for which retrieval is possible.

While this formula is asymptotically very satisfactory, it fails to give good results for realistic values like $N = 5000$ and $r_{\text{aff}} = 0.1$. Working out the numbers one sees that then the fraction $\frac{\Delta(0)}{\Delta}$ will be extremely close to 1 or even less than 1 (in which case no learning is possible at all). We have also seen that decreasing α_{fid} from 1.0 to, say, 0.5 has no dramatic effect as it only increases Δ by a factor of roughly two. Similarly, allowing more noise only increases Δ by a small, constant factor.

In this section we show that using recurrent edges and percolation theory can overcome this problem for small constants. **Figure 5** illustrates the underlying idea. Without recurrent edges one has to ensure that the degree distributions of the signal and the noise are so far apart that one can choose an activation threshold K such that the noise distribution has only a tiny part to the left (as these are the vertices that will get activated outside the pattern), while the signal distribution should have a small part to the right of K (as these are the vertices within the pattern that will not get activated). Using iterative retrieval allows to essentially move the two distribution on top of each other, as the condition for the signal is replaced by "activate a small fraction" instead of "activate almost everything."

Percolation or, more precisely, bootstrap percolation was studied by Janson et al. (2012) for random graphs. Given an Erdős-Rényi graph $G_{n,p}$ and a random subset A of *active* vertices of size $|A| = a$. Percolation studies the question for which sizes of A (as a function of the size of the graph n and the edge density p) activity spreads to all (or at least almost all) vertices. Activity spreads according to a K-threshold rule, i.e., a vertex turns active if it has at least K active neighbors and once it turns active it remains active. Janson et al. (2012) gave a complete characterization of all occurring cases and phenomena. We do not state their result formally, but instead give a sketch of their proof. Subsequently, we then show how it can be transferred to our setting.

Let us recall the setup: we are given a random graph $G_{n,p}$ and we start with a (random) subset A of size $|A| = a$ of active vertices. Instead of immediately activating all vertices with enough active neighbors, we expose the random graph $G_{n,p}$ step by step by the following, equivalent process.

Consider a set U of unexposed vertices and a set E of exposed vertices. At the beginning we initialize U with the vertices from A and let E start empty. Every time we expose a vertex from U (by removing it from U, adding it to E and exposing the edges from it to $V \setminus E$) we add newly active vertices to U. If U gets empty at some point in time we add a random (unexposed) vertex to it.

In order for the process to percolate one needs that at every time t there are still unexposed vertices, i.e., the set U is non-empty. Observe that at time t (that is, when $|E| = t$) every vertex in $V \setminus E$ has revealed exactly t (potential) edges. That is, it is active at time t with probability $p' = \Pr[\text{Bin}(t, p) \geq K]$. Let $S(t)$ denote the set of vertices in $V \setminus A$ that are active at time t and let $s(t) = |S(t)|$. Then $s(t)$ is distributed as $\text{Bin}(n - a, p')$. Since we assume the process to percolate, the t exposed vertices are all active. Hence, the size of U at time t is $s(t) + a - t$.

So we percolate if and only if for all t we have $\text{Bin}(n - a, p') > t - a$. In Janson et al. (2012), the authors proved that for large n we may replace the binomial distribution by its expectation (we call this the Janson assumption). Thus, we percolate if and only if we have

$$(n - a)\Pr[\text{Bin}(t, p) \geq K] > t - a \qquad \text{for all } t \geq 0. \quad (12)$$

Essentially, one can read off the conditions for percolation from Equation (12), cf. Janson et al. (2012) for the formal derivation. The key point is that whenever the edge probability is sufficiently high (e.g., $p \geq (1 + \delta)\log n/n$, for any $\delta > 0$) then we only need to activate a set A of size $\gg (np^K)^{-1/(K-1)} = o(n)$ in order to have (almost) full percolation with high probability.

We now transfer these results to the learning scenario studied in this paper. If we assume that within B_i the edges form a random graph with density ρ_{rec} then the above percolation result tells us that we only have to activate a tiny portion of B_i directly in order to achieve full activation of B_i. **Figure 6** illustrates this effect. As

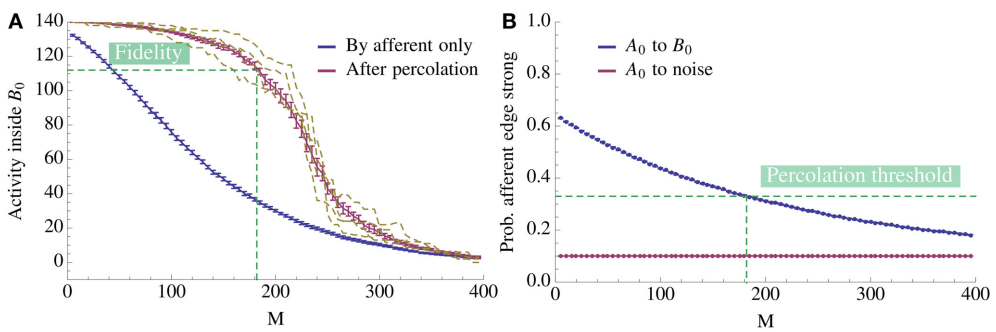


FIGURE 6 | Change of activity inside the pattern as more competing associations are learned. (A) The *x*-axis denotes the number of associations exposed to the model and the y-axis gives the mean number of active vertices over 20 trials. We compare the activity due to afferent edges only (blue) and after percolation via recurrent edges (red). The dashed yellow curves give examples of single instances of the process after percolation. Note that since fidelity is 80% we count the pattern as being activated if we percolate to at least 112 vertices. Thus, in expectation we can learn 182 patterns for this choice of parameters. **(B)** Mean ratio of strong edges to all edges over 20 trials. We compare the density of strong edges amongst present edges between A_0 and B_0 and between A_0 and $\mathcal{B} \setminus B_0$. As expected, due to normalization of strong edges the density toward $\mathcal{B} \setminus B_0$ remains fixed. Note also that the fidelity threshold from **(A)** induces a density threshold, dashed green line. Above threshold percolation works but below it does not. Here the value of the threshold is ≈ 0.33. For both plots $N = 5000$, $n = 140$, $K = 12$, $p^+ = 0.6$, $r_{\text{aff}} = 0.1$, $\rho_{\text{aff}} = 0.2$, and the recurrent degree is 8. For both figures the error bars represent standard mean error.

Figure 6A shows, it suffices to activate even a small bootstrap afferently in order to activate the whole pattern by percolation. Moreover, observe the threshold effect: while the afferent density stays above some threshold value, percolation activates almost the complete pattern; below this threshold, activity does not spread. This is the basis for our analysis: once we know the threshold, we can compute how the afferent density evolves over time to determine when it hits the threshold (**Figure 6B**).

It remains to determine the threshold. Actually, our situation is even better than the one studied in Janson et al. (2012): every vertex in B_i has an afferent degree into A_i distributed as $\mathrm{Bin}(n, \rho_{\mathrm{aff}} p_{\mathrm{signal}})$. For some vertices this degree will be at least K and they thus get activated immediately. Other may have degree almost K, but the recurrent edges to the vertices that were activated immediately will bring the degree above K, etc. For a formal study we proceed similarly as above: we consider a set U of unexposed vertices that at $t = 0$ contains all vertices whose afferent degree is at least K. While percolation runs we again add active vertices to U. Observe that in this scenario a vertex of B_i is active at time t with probability

$$\begin{aligned} p' &= \Pr[\mathrm{Bin}(n, \rho_{\mathrm{aff}} p_{\mathrm{signal}}) + \mathrm{Bin}(t, \rho_{\mathrm{rec}}) \geq K] \\ &= \Pr[\mathrm{Bin}(n, \rho_{\mathrm{aff}} p_{\mathrm{signal}}) \geq K] \\ &\quad + \sum_{i=0}^{K-1} \Pr[\mathrm{Bin}(n, \rho_{\mathrm{aff}} p_{\mathrm{signal}}) = i] \cdot \Pr[\mathrm{Bin}(t, \rho_{\mathrm{rec}}) \geq K - i]. \end{aligned}$$

Again we denote by $S(t)$ the set of vertices in B_i that are active at time t and let $s(t) = |S(t)|$. Then $s(t)$ is distributed as $\mathrm{Bin}(n, p')$. In order to percolate we need $s(t) > t$ for all $0 \leq t < n$. Replacing the binomial distribution by its expectation (as we may do under the Erdős-Rényi assumption by Janson et al., 2012) we obtain that we percolate if and only if

$$n \cdot \Pr[\mathrm{Bin}(n, \rho_{\mathrm{aff}} p_{\mathrm{signal}}) + \mathrm{Bin}(t, \rho_{\mathrm{rec}}) \geq K] > t \quad \text{for all } t \geq 0. \tag{13}$$

For a fixed value of ρ_{rec} Equation (11) thus allows us to determine the probability of edges being strong afferently p_{signal} that we need in order to achieve percolation.

We close this section with the remark that while percolation has a dramatic effect for finite values, it does not change the asymptotics of the memory capacity. To see this observe that we need to be able to activate at least one vertex in B_i due to the afferent edges alone. By a similar argument as for Equation (11), we thus get

$$\Delta_{\mathrm{percolation}} > \frac{\sigma_{\mathrm{spc}}}{n\rho_{\mathrm{aff}}} \sqrt{2 \log\left(\frac{N}{\alpha_{\mathrm{spc}} n \sqrt{2\pi\sigma_{\mathrm{spc}}^2}}\right)} - \frac{\sigma_{\mathrm{fid}}}{n\rho_{\mathrm{aff}}} \sqrt{2 \log\left(\frac{n}{\sqrt{2\pi\sigma_{\mathrm{fid}}^2}}\right)}. \tag{14}$$

Note that the main change compared to Equation (11) is the sign of the second term. As before, for $N \to \infty$, the first term will remain constant [for $n = \theta(\log N)$], while the second term will tend to 0. Hence, we will not see any difference in the asymptotic capacity. The influence of percolation is limited to finite values of N; but, as we saw in **Figure 4**, the differences of the two models are substantial for values of N and ρ_{aff} as they occur in the brain.

3.1.5. Error estimates

The calculations in the previous sections rely on some approximations that are all valid in the $N \to \infty$ limit. There are three sources of errors that need consideration:

1. The Erdős-Rényi assumption (independence) may not hold.
2. The Janson assumption [cf. Equations (12) and (13)] may not hold.
3. There is an error term that comes from replacing the binomial distribution by a normal approximation.

In the previous section we handled (1) and (2) by arguing that in the limit the probability that at least one these properties does not hold tends to zero. By then analyzing the situation under the condition that the Janson assumption and the Erdős-Rényi assumption do hold, cf. Equation (12), we get an estimate for what happens in the "typical" case. Unfortunately, to actually quantify the errors seems very hard, as for example, the paper Janson et al. (2012) does not provide precise bounds for the probability that (2) is violated.

In this section we thus show experimentally that the errors induced by the approximations (1)–(3) are indeed small for the chosen parameters. For each K, **Figure 9** contains four curves:

a) The simulation result;
b) We use simulations to estimate the threshold for percolation p_{signal} in an Erdős-Rényi random graph, and computed the capacity by Equation (6);
c) We use the Janson assumption in Equation (13) for $t = 0, \ldots, \alpha_{\mathrm{fid}} n$ to estimate p_{signal}, and compute the capacity by Equation (6);
d) We use Equation (13) to estimate p_{signal} as in (b), but with the binomial distributions replaced by normal approximations. Then we compute again the capacity by Equation (6).

The four curves quantify the errors 1–3 in the following way:

- In (b) we use the Erdős-Rényi assumption, but nothing else. So the difference between a and b quantifies the error of type 1.
- In (c) we use the Erdős-Rényi assumption and the Janson assumption. So the difference between b and c quantifies the error of type 2.
- In (d) we use the Erdős-Rényi assumption, the Janson assumption, and the normal approximations. So the difference between c and d quantifies the error of type 3. Finally, the difference between a and d quantifies the overall contribution of all three error sources.

To compare the errors to the second order terms in Equations (11) and (14), recall that these terms are at least the difference between the capacities with and without recurrent edges, up to error terms

of type 1, 2, and 3. Therefore, we also plotted the capacity without recurrent edges for different K (including the K that maximizes the capacity). It is clearly visible that the difference between the capacity with recurrent edges (highest blue curve) and the capacity without recurrent edges (highest violett curve) is much larger than the error terms. Thus, the errors of type 1, 2, and 3 are small for plausible parameter values.

3.1.6. The optimal plasticity constant

From our consideration, we can derive the optimal value for the plasticity p^+. Note first that the minimal difference $\Delta = p_{\text{signal}} - r_{\text{aff}}$ for which we can still recall is independent of p^+, regardless of α and regardless of whether we use percolation. Since the capacity is

$$M = \frac{\log(\Delta(0)/\Delta)}{\log(1/\beta)} = \log\left(\frac{(1-r_{\text{aff}})p^+}{\Delta}\right)\frac{1}{\log(1/\beta)} \approx \log\left(\frac{(1-r_{\text{aff}})p^+}{\Delta}\right)\left(\frac{N^2 r_{\text{aff}}}{n^2 p^+}\right), \quad (15)$$

we essentially need to maximize a function of the form $\log(c_1 p^+)\cdot(c_2/p^+)$. Such a function takes its maximum at $p^+ = e/c_1$, where $e = 2.718\ldots$. Hence, the optimal p^+ is

$$p^+ = \frac{e\Delta}{1-r_{\text{aff}}}.$$

For the case without recurrent edges this resembles the findings in Romani et al. (2008). For the maximal capacity we hence get

$$M \approx \log\left(\frac{(1-r_{\text{aff}})p^+}{\Delta}\right)\left(\frac{N^2 r_{\text{aff}}}{n^2 p^+}\right) = \frac{N^2 r_{\text{aff}}(1-r_{\text{aff}})}{n^2 e\Delta}. \quad (16)$$

Note that M is not independent of ρ_{aff} since $\Delta \sim 1/\sqrt{\rho_{\text{aff}}}$.

3.1.7. Noise tolerance

We study two types of noise tolerance, so called *query noise*, where the activation of A_0 is imperfect and *recurrent noise*, where we start the recall with active vertices in $\mathcal{B} \setminus B_0$.

In the case of query noise we activate A_0 with λ precision, $\lambda \in [0,1]$, meaning that we activate λn vertices chosen u.a.r. from A_0 and $(1-\lambda)n$ vertices chosen u.a.r. from $\mathcal{A} \setminus A_0$. Note that since there are n vertices active in $\mathcal{A}$ at the start of percolation every vertex in $\mathcal{B} \setminus B_0$ expects the same amount of inputs as if A_0 was activated with precision $\lambda = 1$ so the specificity constraint is unaffected. However, for vertices in B_0 they now expect to receive $\lambda n p_{\text{signal}} + (1-\lambda) n r_{\text{aff}}$ signals from $\mathcal{A}$. We thus have that we can still recall B_0 after i insertions of competing associations if

$$\lambda(p_{\text{signal}}(i) - r_{\text{aff}}) > \Delta.$$

One easily checks that the difference in capacity between precision 1 and precision λ is

$$\log_{1/\beta}(1/\lambda) = \theta\left(\frac{N^2 \log(1/\lambda)}{n^2}\right).$$

For recurrent noise with m noisy vertices the bootstrap consists of A_0 and m vertices chosen u.a.r. from $\mathcal{B} \setminus B_0$. In this case the activation of B_0 w.r.t. the fidelity requirement is not affected but we run the risk of percolation within $\mathcal{B}$. Note that the edges within $\mathcal{B}$ are not independently strong so we cannot directly apply the percolation theory for Erdős-Rényi graphs. However, empirical observations (see **Figure 7B**) indicate that there is still a threshold phenomenon occurring for percolation which depends on the number of patterns stored in $\mathcal{B}$. Moreover, the same figure shows that the capacity of the system is extremely stable against recurrent noise.

3.2. EXPERIMENTAL RESULTS

The theoretical results obtained in Section 3.1 are for the limiting case $N \to \infty$. It is not possible to obtain explicit error terms since the error terms for the threshold density p_{signal} in the bootstrap percolation are not known explicitly. For this reason we test our results in a bioplausible range with $N = 5000$ neurons in $\mathcal{A}$ and $\mathcal{B}$ each (cf. Section 2.2).

For all the relevant figures we perform one shot learning as described in Section 2.1.2. In order to realize a recurrent density of ρ_{rec} within the patterns we proceed as follows: we initialize the set $\mathcal{B}$ as a random graph with edge probability ρ_{rec} with all the edges weak. When we insert a pattern in $\mathcal{B}$ we turn all the edges inside the pattern strong. In that way we inherit the density of ρ_{rec} for each pattern from the global density within $\mathcal{B}$.

Figure 2 demonstrates how memory capacity depends on the pattern size n when all parameters of the process are fixed and chosen in some optimal way, as argued below. The capacity is the expected number of associations which can be inserted until the first association cannot be recalled any more (due to pruning and/or noise). Throughout we chose $\alpha_{\text{fid}} = 0.8$ and $\alpha_{\text{spc}} = 1.0$ as parameters for fidelity and specificity.

In general, a fixed set of parameters ρ_{aff}, r_{aff}, ρ_{rec}, p^+, and K will only work for a finite range of values for n: if n is too large, then noise is too large and the specificity criterion is violated. On the other hand, if n is too small, then we will not be able to satisfy the fidelity condition even immediately after learning.

In the following figures we illustrate the connections between the various parameters for the case $\rho_{\text{aff}} = 0.2$. In the figures we only show data points for which reliability was at least 99%, meaning that in 99% of the cases the first association could be recalled before competing associations were inserted.

Figure 8 demonstrates the effect of varying the probability of afferent edges being strong, i.e., r_{aff}, for a fixed value of n (here $n = 100$). As it turns out, for each K the curve is unimodular and the maximal values of these curves are also unimodular. The figure shows the best K for recurrent degree 0 respectively 4. It is worthwhile to note that for fixed K the curves drop sharply if r_{aff} exceeds a certain value (as then noise takes over). However, for $\rho_{\text{rec}} > 0$ this drop is less dramatic, making the setup more stable.

Figure 8 seems to indicate that a value of $r_{\text{aff}} \approx 0.25$ is a good choice. In order to test that we compared in **Figure 4** the effect of r_{aff}. We found that while for larger r_{aff} the maximum value that we can achieve is indeed higher, this comes at the price of robustness. More precisely, for larger values of r_{aff} the curves (for a fixed K), tend to be very pointed, while for smaller r_{aff} we can have plateaus

with almost the same value. This is the reason for our choice of $r_{\text{aff}} = 0.1$ in **Figure 2**.

Figure 3 illustrates our choice of $p^+ = 0.6$. We see that when the remaining parameters are fixed, we essentially get a threshold phenomenon: p^+ needs to be sufficiently large, but a further increase does not have a positive effect any more (but may even decrease performance). Intuitively, this phenomenon occurs because percolation within $\mathcal{B}$ becomes possible with a bootstrap of size n before the association is forgotten afferently. A further increase of p^+ thus only increases this effect and therefore does not increase the learning capacity.

Now we are ready to explain our choice of parameters for **Figure 2A**: we chose $r_{\text{aff}} = 0.1$ and $p^+ = 0.6$, as suggested by **Figures 3**, **8**. The figure on the right side of **Figure 2** shows a similar plot for $\rho_{\text{aff}} = 0.05$. Here it turned out that $r_{\text{aff}} = 0.05$ yields better results (due to the smaller memory capacity), so we chose this value, and the learning probability is again $p^+ = 0.6$. For both cases, and each expected recurrent degree, K was chosen so that we obtain stable results for n in a wide range. In the case of zero recurrent degree, the sparsity enforces a small value of K to allow learning at all; in turn, this means that no value of K works for a large interval, so we chose $K = 3$ which yields the best (even though still quite small) capacity for large n's.

Figure 7 demonstrates the two types of noise tolerance we study. In the case of query noise, **Figure 7A**, we choose our parameters as in **Figure 2A** with $n = 100$, $K = 12$, and $\rho_{\text{rec}} n = 8$. In this setting the model is able to satisfy the fidelity requirement with $\lambda = 0.7$ and even after 100 insertions of competing associations the relation (A_0, B_0) can still tolerate $\lambda = 0.8$. For the recurrent noise we observe that with only a few patterns stored recurrently in $\mathcal{B}$ the model does not react to recurrent noise at all. This happens because either the necessary bootstrap size for percolation is too large or percolation within $\mathcal{B}$ is simply impossible due to the density of strong recurrent edges being too low. However, once sufficiently many patterns have been inserted in $\mathcal{B}$ percolation becomes possible and we observe a threshold behavior, see **Figure 7B**.

Figure 9 gives an example of quadratic growth with three theoretical predictions for comparison that quantify the different approximations made in the theoretical predictions, cf. Section 3.1.5.

4. DISCUSSION

4.1. MODEL ASSUMPTIONS

4.1.1. Synapses and learning

The synapses in our model only have two states: they are either weak or strong. The learning rule follows the Hebbian paradigm "fire together, wire together," followed by a normalization step. Learning mechanisms in the brain are more complicated. In particular, for spike-timing dependent plasticity (STDP) the timing of pre- and post-synaptic spike is crucial. However, it has been shown by Abbott and Nelson (2000); Gerstner and Kistler (2002a,b) that STDP resembles Hebbian learning when there is

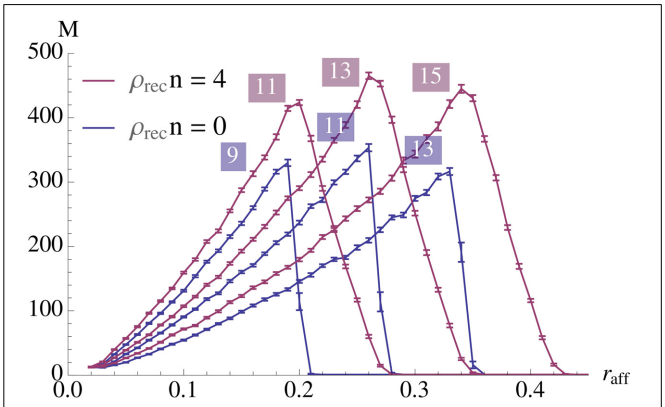


FIGURE 8 | Strong edge density (probability of afferent edges being strong, r_{aff}) vs. capacity for $n = 100$. The labels on the curves denote which value of K was used to generate it. The blue curve for $K = 11$ and red curve for $K = 13$ maximize capacity. Here $N = 5000$, $\rho_{\text{aff}} = 0.2$, and $p^+ = 1$. Each data point is the mean of 100 samples and the error bars represent standard mean error.

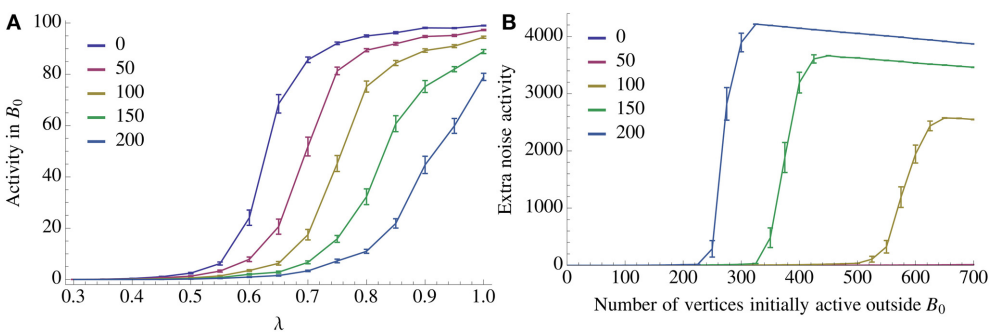


FIGURE 7 | Resilience of recall against noise. We look at two types of noise, an imperfect activation of A_0 (query noise, **A**), and interference within $\mathcal{B} \setminus B_0$ (recurrent noise, **B**). In the case of query noise we u.a.r. activate λn vertices in A_0 and $(1 - \lambda)n$ vertices in $\mathcal{A} \setminus A_0$. We observe a threshold in λ for recall which depends on the number of inserted competing associations. A further discussion of the query noise can be found in Section 3.1.7. For recurrent noise we activate together with A_0 a fixed amount of vertices in $\mathcal{B} \setminus B_0$. For recurrent noise we activate together with A_0 a fixed amount of vertices chosen u.a.r. in $\mathcal{B} \setminus B_0$. We also observe a threshold in how many such vertices the model can tolerate which depends on the number of patterns stored recurrently within $\mathcal{B}$. For both figures the parameters of the process are the same as in **Figure 2A** with $n = 100$, $K = 12$, and $\rho_{\text{rec}} n = 8$, and error bars represent standard mean error.

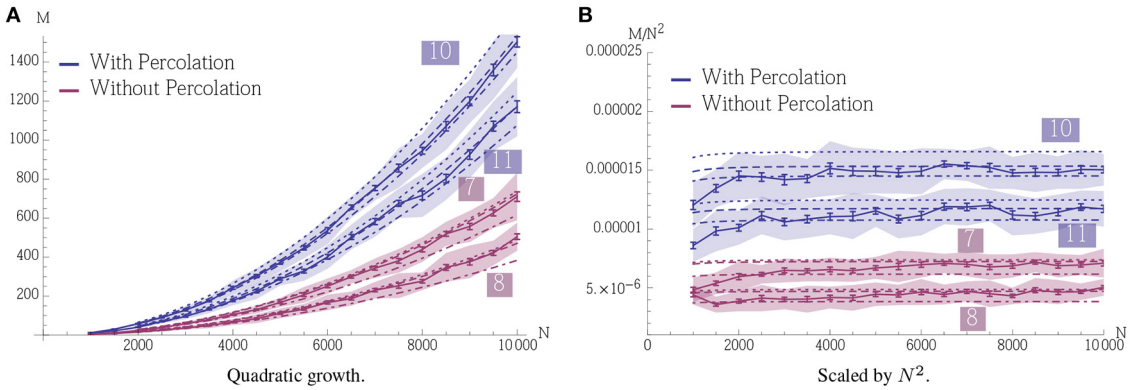


FIGURE 9 | Demonstration of quadratic growth and comparison between simulation and theoretical predictions (A). (B) displays the same data but scaled by N^2 to estimate the constant of the leading term. In case of the simulation, whole curve, the error bars represent standard mean error for 25 trials and the envelope represents the standard deviation. The dashed curve, (b), is calculated using a theoretical prediction where first $\Delta_{\text{percolation}}$ is estimated using a simulation. The simulation determines the smallest p_{signal} for which percolation is still possible for a pattern pair in an Erdős-Rényi random graph and the estimator is an average over 500 such trials. The dashed curve is then obtained using Equation (6). The dotted curve, (c), is also calculated using a theoretical prediction. There we estimate $\Delta_{\text{percolation}}$ using Equation (12) such that it is satisfied for $t = 0, \ldots, \alpha_{\text{fid}} n$ and use Equation (6) again to obtain the curve. Finally the dot-dashed curve, (d), is the prediction where $\Delta_{\text{percolation}}$ is obtained by replacing the binomial distributions in Equation (12) by normal distributions. For a further discussion on these predictions, b, c, and d, and their importance we refer the reader to Section 3.1.5. All parameters in this figure are fixed except for N, they are the same as in **Figure 2A** with $n = 100$ and $\rho_{\text{rec}} n = 8$. Note that for a fixed K the growth stops when the specificity constraint becomes violated due to afferent noise. However, this can be accounted for by altering the parameters, we refer to Section 3.1.3 for a further discussion.

no systematic time shift between different inputs, and at the same time it normalizes the input of each neuron (see Kempter et al., 1999; Abbott and Nelson, 2000; Song et al., 2000; Abbott and Gerstner, 2004; Gilson et al., 2010).

The question whether synapses are binary is unsettled and vividly disputed in Graupner and Brunel (2010); Barbour et al. (2007); Satel et al. (2009). However, some STDP experiments indicate that synapses in the hippocampus are indeed binary: synapses that have been potentiated by an STDP protocol can not be potentiated a second time, but can be depressed again, and vice versa as in Petersen et al. (1998); O'Connor et al. (2005). Also, while such experiments last for minutes, the change is sudden and strong (a factor of 2–3, see Petersen et al., 1998; O'Connor et al., 2005). These findings are compatible with our assumptions of stochastic Hebbian learning.

4.1.2. Activity and dynamics

It is well-known that the brain encodes some information in the firing rate of neurons, and many computational papers take this point of view (e.g., Amit and Fusi, 1994). However, there are also other ways the brain encodes and processes information. E.g., when humans are asked to discriminate between pictures of animals and non-animals, then task-related eye-saccades can be observed after 120 ms (Kirchner and Thorpe, 2006). This amazing speed indicates that feedback loops or rate based encoding do not play a role for these ultra-fast processes, since each region in the brain has only 10–20 ms to process and transmit the signal. Thus, it seems that at least some type of hypothesis forming is done in a single feed-forward sweep of information, based on one or only very few spikes per neuron. Various other physiologic and psychologic experiments came to similar conclusions (Thorpe and Imbert, 1989; Allison et al., 1999; Liu et al., 2002; Crouzet et al., 2010; 't Hart et al., 2013, see also Johnson and Olshausen, 2002 for a review). We designed our model to fit a sweep of activity as described above, and thus we only count whether a neuron emits at least one spike, ignoring any further spikes of this neuron. Janson et al. (2012) proved that such a sweep is extremely fast: For a pattern with n vertices it takes at most time $O(\log\log n)$ if the transmission delays of all edges is 1. In our context $n = O(\log N)$, so percolation only needs time $O(\log\log\log N)$. If the transmission delays are drawn from an exponential distribution with mean 1, then Einarsson et al. (2014) showed that the sweep is even faster: it takes at most constant time, independent of n.

4.2. PATTERN SIZES AND PLASTICITY

Our simulations show that stochastic Hebbian learning enables sparsely connected neuronal ensembles to perform one shot association learning. There is a tradeoff between reliability and capacity. For smaller pattern sizes the successfully inserted patterns can be memorized for a long time yielding a large expected capacity. However, a large portion of the insertions for small patterns are not successful, even with their optimal plasticity parameter $p^+ = 1$. For larger patterns the optimum p^+ is <1 and every pattern is stored successfully but the capacity drops proportional to $\frac{N^2}{n^2}$. By keeping n fixed and varying the plasticity parameter we have a similar tradeoff: if plasticity is too small associations are poorly stored in the first place but if it is too large the ongoing activity in the network will rapidly overwrite older associations. For a fixed population size N the optimum plasticity parameter decays proportional to $\frac{1}{\sqrt{n}}$. Since

the growth rate of the capacity is quadratic we have that eventually every neuron will take part in multiple associations. This turns out to be the case even for $N = 5000$ in a sparsely connected network.

ACKNOWLEDGMENTS

We thank Thomas Rast and Nemanja Skoric for helpful discussions.

REFERENCES

Abbott, L., and Gerstner, W. (2004). Homeostasis and learning through spike-timing dependent plasticity. *Methods Models Neurophys.* Available online at: http://infoscience.epfl.ch/record/114304/files/Abbott04.pdf

Abbott, L. F., and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nat. Neurosci.* 3(Suppl.), 1178–1183. doi: 10.1038/81453

Allison, T., Puce, A., Spencer, D. D., and McCarthy, G. (1999). Electrophysiological studies of human face perception. I: potentials generated in occipitotemporal cortex by face and non-face stimuli. *Cereb. Cortex* 9, 415–430. doi: 10.1093/cercor/9.5.415

Amit, D. J., and Fusi, S. (1994). Learning in neural networks with material synapses. *Neural Comput.* 6, 957–982. doi: 10.1162/neco.1994.6.5.957

Amit, Y., and Huang, Y. (2010). Precise capacity analysis in binary networks with multiple coding level inputs. *Neural Comput.* 22, 660–688. doi: 10.1162/neco.2009.02-09-967

Barbour, B., Brunel, N., Hakim, V., and Nadal, J. (2007). What can we learn from synaptic weight distributions? *Trends Neurosci.* 30, 622–629. doi: 10.1016/j.tins.2007.09.005

Barrows, G. L. (1998). "Stochastic hebbian learning with binary synapses," in *Neural Networks Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Joint Conference on,* Volume 1 (Anchorage, AK), 525–530.

Battaglia, F. P., and Fusi, S. (1994). Learning in neural networks with partially structured synaptic transitions. *Network* 6, 261–270. doi: 10.1088/0954-898X/6/2/007

Beaulieu, C., and Colonnier, M. (1983). The number of neurons in the different laminae of the binocular and monocular regions of area 17 in the cat. *J. Comp. Neurol.* 217, 337–344. doi: 10.1002/cne.902170308

Binzegger, T., Douglas, R., and Martin, K. (2004). A quantitative map of the circuit of cat primary visual cortex. *J. Neurosci.* 24, 8441–8453. doi: 10.1523/JNEUROSCI.1400-04.2004

Brunel, N., Carusi, F., and Fusi, S. (1998). Slow stochastic Hebbian learning of classes of stimuli in a recurrent neural network. *Network* 9, 123–152. doi: 10.1088/0954-898X/9/1/007

Buckingham, J., and Willshaw, D. (1991). Performance characteristics of the associative net. *Network* 3, 407–414. doi: 10.1088/0954-898X/3/4/005

Crouzet, S. M., Kirchner, H., and Thorpe, S. J. (2010). Fast saccades toward faces: face detection in just 100 ms. *J. Vis.* 10, 16.1–17. doi: 10.1167/10.4.16

Einarsson, H., Mousset, F., Lengler, J., Panagiotou, K., and Steger, A. (2014). Bootstrap percolation with inhibition. arXiv:1410.3291.

Fanselow, M. S. (1990). Factors governing one-trial contextual conditioning. *Anim. Learn. Behav.* 18, 264–270. doi: 10.3758/BF03205285

Gerstner, W., and Kistler, W. M. (2002a). Mathematical formulations of Hebbian learning. *Biol. Cybernet.* 87, 404–415. doi: 10.1007/s00422-002-0353-y

Gerstner, W., and Kistler, W. M. (2002b). *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511815706

Gilson, M., Burkitt, A. N., Grayden, D. B., Thomas, D. A., and van Hemmen, J. L. (2010). Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks v: self-organization schemes and weight dependence. *Biol. Cybernet.* 103, 365–386. doi: 10.1007/s00422-010-0405-7

Graupner, M., and Brunel, N. (2010). Mechanisms of induction and maintenance of spike-timing dependent plasticity in biophysical synapse models. *Front. Comput. Neurosci.* 4:136. doi: 10.3389/fncom.2010.00136

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* 79, 2554–2558. doi: 10.1073/pnas.79.8.2554

Janson, S., Łuczak, T., Turova, T., and Vallier, T. (2012). Bootstrap percolation on the random graph $G_{n,p}$. *Ann. Appl. Probab.* 22, 1989–2047. doi: 10.1214/11-AAP822

Johnson, J. S., and Olshausen, B. A. (2002). Timecourse of neural signatures of object recognition. *J. Vis.* 3, 499–512. doi: 10.1167/3.7.4

Kalisman, N., Silberberg, G., and Markram, H. (2005). The neocortical microcircuit as a tabula rasa. *Proc. Natl. Acad. Sci. U.S.A.* 102, 880–885. doi: 10.1073/pnas.0407088102

Kempter, R., Gerstner, W., and Hemmen, J. V. (1999). Hebbian learning and spiking neurons. *Phys. Rev. E* 59:4498. doi: 10.1103/PhysRevE.59.4498

Kirchner, H., and Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. *Vis. Res.* 46, 1762–1776. doi: 10.1016/j.visres.2005.10.002

Knoblauch, A. (2008). Neural associative memory and the Willshaw–Palm probability distribution. *SIAM J. Appl. Math.* 69, 169–196. doi: 10.1137/070700012

Knoblauch, A., and Palm, G. (2001). Pattern separation and synchronization in spiking associative memories and visual areas. *Neural Netw.* 14, 763–780. doi: 10.1016/S0893-6080(01)00084-3

Knoblauch, A., Palm, G., and Sommer, F. T. (2010). Memory capacities for synaptic and structural plasticity. *Neural Comput.* 22, 289–341. doi: 10.1162/neco.2009.08-07-588

Le Bé, J. V., Silberberg, G., Wang, Y., and Markram, H. (2006). Morphological, electrophysiological, and synaptic properties of corticocallosal pyramidal cells in the neonatal rat neocortex. *Cereb. Cortex* 17, 2204–2213. doi: 10.1093/cercor/bhl127

Levy, R. B., and Reyes, A. D. (2012). Spatial profile of excitatory and inhibitory synaptic connectivity in mouse primary auditory cortex. *J. Neurosci.* 32, 5609–5619. doi: 10.1523/JNEUROSCI.5158-11.2012

Liu, J., Harris, A., and Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nat. Neurosci.* 5, 910–916. doi: 10.1038/nn909

Meyer, H. S., Wimmer, V. C., Hemberger, M., Bruno, R. M., de Kock, C. P. J., Frick, A., et al. (2010). Cell type-specific thalamic innervation in a column of rat vibrissal cortex. *Cereb. Cortex* 20, 2287–2303. doi: 10.1093/cercor/bhq069

Nadal, J. P., Toulouse, G., Changeux, J. P., and Dehaene, S. (1986). Networks of formal neurons and memory palimpsests. *Europhys. Lett.* 1, 535–542. doi: 10.1209/0295-5075/1/10/008

O'Connor, D., Wittenberg, G., and Wang, S. (2005). Graded bidirectional synaptic plasticity is composed of switch-like unitary events. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9679. doi: 10.1073/pnas.0502332102

Perin, R., Berger, T. K., and Markram, H. (2011). A synaptic organizing principle for cortical neuronal groups. *Proc. Natl. Acad. Sci. U.S.A.* 108, 5419–5424. doi: 10.1073/pnas.1016051108

Petersen, C. C., Malenka, R. C., Nicoll, R. A., and Hopfield, J. J. (1998). All-or-none potentiation at CA3-CA1 synapses. *Proc. Natl. Acad. Sci. U.S.A.* 95, 4732–4737. doi: 10.1073/pnas.95.8.4732

Romani, S., Amit, D. J., and Amit, Y. (2008). Optimizing one-shot learning with binary synapses. *Neural Comput.* 20, 1928–1950. doi: 10.1162/neco.2008.10-07-618

Satel, J., Trappenberg, T., and Fine, A. (2009). Are binary synapses superior to graded weight representations in stochastic attractor networks? *Cogn. Neurodynam.* 3, 243–250. doi: 10.1007/s11571-009-9083-3

Schwenker, F., Sommer, F. T., and Palm, G. (1996). Iterative retrieval of sparsely coded associative memory patterns. *Neural Netw.* 9, 445–455. doi: 10.1016/0893-6080(95)00112-3

Sommer, F. T., and Palm, G. (1998). Improved bidirectional retrieval of sparse patterns stored by Hebbian learning. *Neural Netw.* 12, 281–297. doi: 10.1016/S0893-6080(98)00125-7

Song, S., Miller, K. D., and Abbott, L. F. (2000). Competitive hebbian learning through spike-timing-dependent synaptic plasticity. *Nat. Neurosci.* 3, 919–926. doi: 10.1038/78829

Song, S., Sjöström, P., Reigl, M., Nelson, S., and Chklovskii, D. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol.* 3:e68. doi: 10.1371/journal.pbio.0030068

't Hart, B. M., Schmidt, H. C. E. F., Klein-Harmeyer, I., and Einhauser, W. (2013). Attention in natural scenes: contrast affects rapid visual processing and fixations alike. *Philos. Trans. Biol. Sci.* 368, 20130067–20130067. doi: 10.1098/rstb.2013.0067

Thorpe, S. J. and Imbert, M. (1989). "Biological constraints on connectionist modelling," in *Connectionism in Perspective* (Elsevier), 63–92. Available online at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.96.6484

Willshaw, D. J., Buneman, O. P., and Longuet-Higgins, H. C. (1969). Non-holographic associative memory. *Nature* 222, 960–962. doi: 10.1038/222960a0

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 20 July 2014; accepted: 16 October 2014; published online: 07 November 2014.
Citation: Einarsson H, Lengler J and Steger A (2014) A high-capacity model for one shot association learning in the brain. Front. Comput. Neurosci. ***8****:140. doi: 10.3389/fncom.2014.00140*
This article was submitted to the journal Frontiers in Computational Neuroscience.