Biologically inspired model of associative memory storage with noisy neurons and synapses

Chi Zhang, Danke Zhang, and Armen Stepanyants

Department of Physics and Center for Interdisciplinary Research on Complex Systems, Northeastern University, Boston

ABSTRACT

Neural networks in the brain function reliably despite various sources of errors and noise present at every step of signal transmission. Here, we examine the effects of such fluctuations during associative memory storage on properties of biologically constrained McCulloch and Pitts networks. Our results show that networks, loaded with associative memories to capacity, display many structural and dynamical features observed in local cortical circuits. We predict that noisy inhibitory and excitatory connections in the cortex are depressed or completely eliminated during learning, and that neuron classes that operate with low firing rates have low connection probabilities and strengths.

INTRODUCTION

Brain networks can reliably store and retrieve memories despite various sources of errors and noise in signal transmission. To explore the effects of reliability in memory storage on brain connectivity, we examined a model of associative memory storage in a biologically constrained network of McCulloch and Pitts neurons [1]. The problem of associative learning [2-4] has received much attention over the years due in part to its theoretical tractability (see e.g. [5-8]) and the facility to incorporate biologically inspired elements, such as sign-constrained postsynaptic connections (inhibitory and excitatory) (see e.g. [9-11]), homeostatically constrained presynaptic connections (see e.g. [12]), and robustness to noise [6,7]. Yet, various sources of noise that accompany signal

transmission in the brain [13-16] are not fully accounted for by the associative learning models. At a high-level description, errors in the input to a neuron combine with fluctuations in synaptic transmission and the neuron's intrinsic sources of noise, producing spiking errors in the neuron's output. The latter are, in turn, injected into the network, completing the error propagation cycle (Figure 1A). The associative learning models generally make two unrealistic assumptions, which preclude the analysis of this cycle. First, they assume that intrinsic noise has a fixed range, and second, they allow no errors during the recall of stored memories (but see Supplementary Material of Brunel et al. [10]). The two assumptions are tightly linked and removing the first necessitates eliminating the second. For example, if the distribution of intrinsic noise is Gaussian, associative memories cannot be retrieved without errors. These assumptions are both inconsistent with experimental evidence and unnecessary. Therefore, we developed a network model of associative memory storage which incorporates the complete error propagation cycle.

RESULTS

Network model of associative memory storage in the presence of errors and noise

We consider an all-to-all potentially (structurally) connected network [17] of N_{inh} inhibitory and ($N - N_{inh}$) excitatory McCulloch and Pitts neurons (Figure 1A). The network is faced with a task of learning a sequence of consecutive states, $X^1 \rightarrow X^2 \rightarrow ...X^{m+1}$, in which X^{μ} is a binary vector representing target activities of all neurons at a time step μ. In this process, individual neurons (e.g. neuron i) must independently learn to associate inputs they receive from the network, X^{μ}, with the corresponding outputs derived from the associative memory sequence, $X_i^{\mu+1}$. The neurons must learn these input-output associations by adjusting the weights of their input connections, J_{ij} (weight of connection from neuron j to neuron i). Similar to [12], two biologically inspired constraints are imposed on the learning process. First, the l_1-norm of input connection weights of each neuron is fixed during learning, $\frac{1}{N}\sum_{j=1}^{N}\left|J_{ij}\right| = w$. Here, parameter w is referred to as the average absolute connection weight. Second, the signs of output connection weights of each neuron

(inhibitory or excitatory) do not change during learning, $J_{ij}g_j \geq 0$. Here, parameter $g_j = 1$ if neuron j is excitatory and -1 if it is inhibitory. Biological motivations for these constraints have been previously discussed [12].

In contrast to previous studies [6,7,10-12,18,19], we account for the fact that learning in the brain is accompanied with multiple sources of errors and noise. Within the associative memory storage model, these sources can be divided into three categories (orange lightning signs in Figure 1A): (1) spiking errors, or errors in X^{μ}, (2) synaptic noise, or noise in J_{ij}, and (3) intrinsic noise, which combines all other sources of noise affecting the neurons' postsynaptic potentials. The last category includes background synaptic activity and stochasticity of ion channel states, and, in the model, it is equivalent to noise in the neurons' thresholds of firing, h. In the following, we use an asterisk to denote the quantities containing errors or noise (e.g. $X^{*\mu}$), whereas variables without asterisks represent the mean (for h and J_{ij}) or target (for X^{μ}) values.

Target neuron activities, e.g. X_i^{μ}, are independently drawn from neuron-dependent Bernoulli probability distributions: 0 with probability $1 - f_i$ and 1 with probability f_i. Spiking errors in neuron activity states are introduced with Bernoulli trials by making independent and random changes with probabilities $P\left(X_i^{*\mu} = 0 \mid X_i^{\mu} = 1\right) \equiv p_i^-$ for spike failures and $P\left(X_i^{*\mu} = 1 \mid X_i^{\mu} = 0\right) \equiv p_i^+$ for erroneous spikes (Figure 1B). Without loss of generality we assume that these two types of spiking errors are balanced, $f_i p_i^- = \left(1 - f_i\right) p_i^+$, and do not affect the neuron's firing probability, f_i. This relation allows us to describe both types of spiking errors in terms of the neuron's overall spiking error probability, $r_i \equiv f_i p_i^- + \left(1 - f_i\right) p_i^+$.

To describe synaptic noise, we use the quantal model of Del Castillo and Katz [20]. According to this model, connection weights J_{ij}^* on different trials are drawn from a Binomial distribution, in which the variance is proportional to the mean, $\mathrm{var}\left(J_{ij}^*\right) = \frac{h\beta_{syn}}{N}\left|J_{ij}\right|$. The dimensionless coefficient β_{syn} is referred to as the synaptic noise strength, and a factor of h/N is introduced for convenience. We assume that intrinsic noise is Gaussian distributed across trials with the mean

$\langle h^* \rangle \equiv h$ and variance $\mathrm{var}\left(h^*\right) = \frac{h^2 \beta_{int}^2}{N}$. Here, β_{int} is a dimensionless coefficient called intrinsic noise strength, and, as before, a factor of h^2/N is introduced for convenience.

The above described model can be summarized as follows:

$$
\begin{aligned}
&\theta\left(\sum_{j=1}^{N} J_{ij}^* X_j^{*\mu} - h^*\right) = X_i^{*\mu+1}; \quad \mu = 1,\ldots,m, \quad i = 1,\ldots,N \\
&\frac{1}{N}\sum_{j=1}^{N}\left|J_{ij}\right| = w; \quad i = 1,\ldots,N \\
&J_{ij} g_j \geq 0; \quad i,j = 1,\ldots,N \\
&P\left(X_i^{\mu} = 1\right) \equiv f_i; \quad P\left(X_i^{*\mu} = 0 \mid X_i^{\mu} = 1\right) \equiv p_i^-; \quad P\left(X_i^{*\mu} = 1 \mid X_i^{\mu} = 0\right) \equiv p_i^+ \\
&\left\langle J_{ij}^* \right\rangle \equiv J_{ij}; \quad \mathrm{var}\left(J_{ij}^*\right) = \frac{h\beta_{syn}}{N}\left|J_{ij}\right| \\
&\left\langle h^* \right\rangle \equiv h; \quad \mathrm{var}\left(h^*\right) = \frac{h^2 \beta_{int}^2}{N}
\end{aligned}
\tag{1}
$$

A neuron is said to have learned the presented set of associations successfully if, in the presence of input spiking errors, synaptic and intrinsic noise, the fraction of output errors it produces does not exceed its assigned output error probability, r_i. The described associative memory storage model is governed by parameters N, N_{inh}/N, h, w, β_{int}, β_{syn}, $\{f_i\}$, and $\{r_i\}$, and the task is to find connection weights, $\{J_{ij}\}$, that satisfy all the requirements of Eqs. (1).

Solution of the model

Because individual neurons in the model learn independently from each other and have separate sets of constraints, Eqs. (1) can be split into N independent single-neuron learning problems, which can be solved numerically (see Supplemental Material). We verified that memories loaded into individual neurons at certain noise strengths can be successfully recalled at the network level in the presence of the same or slightly lower noise strengths (Figure S1). Figure 1C shows that for a relatively low memory load, the probability of successful learning by a neuron is close to 1. With

increasing load, the learning problems becomes not feasible, and the success probability undergoes a transition from 1 to 0. Memory load corresponding to the success probability of 0.5 is referred to as the neuron's associative memory storage capacity. With increasing network size, N, the transition from successful learning to inability to learn the entire set of associations becomes sharper, and the neuron's capacity monotonically approaches its $N \to \infty$ limit. In this limit, the associative memory storage problem can be solved with the replica method [21,22] (see Supplemental Material), and the capacity is said to be critical.

In the case of homogeneous inputs to the neuron, $f_i = f$ and $r_i = r$, solution of the model depends on a combination of intrinsic and synaptic noise, referred to as the postsynaptic noise strength (see Supplemental Material):

$$\beta_{post} \equiv \sqrt{\beta_{int}^2 + \beta_{syn} f N \frac{w}{h}} \tag{2}$$

Figure 1D illustrates the dependence of single-neuron critical capacity on postsynaptic noise strength and spiking error probability in the homogeneous case. As expected, because intrinsic and synaptic noise make the learning problem more challenging, a neuron's capacity is a decreasing function of postsynaptic noise strength, Eq. (2).

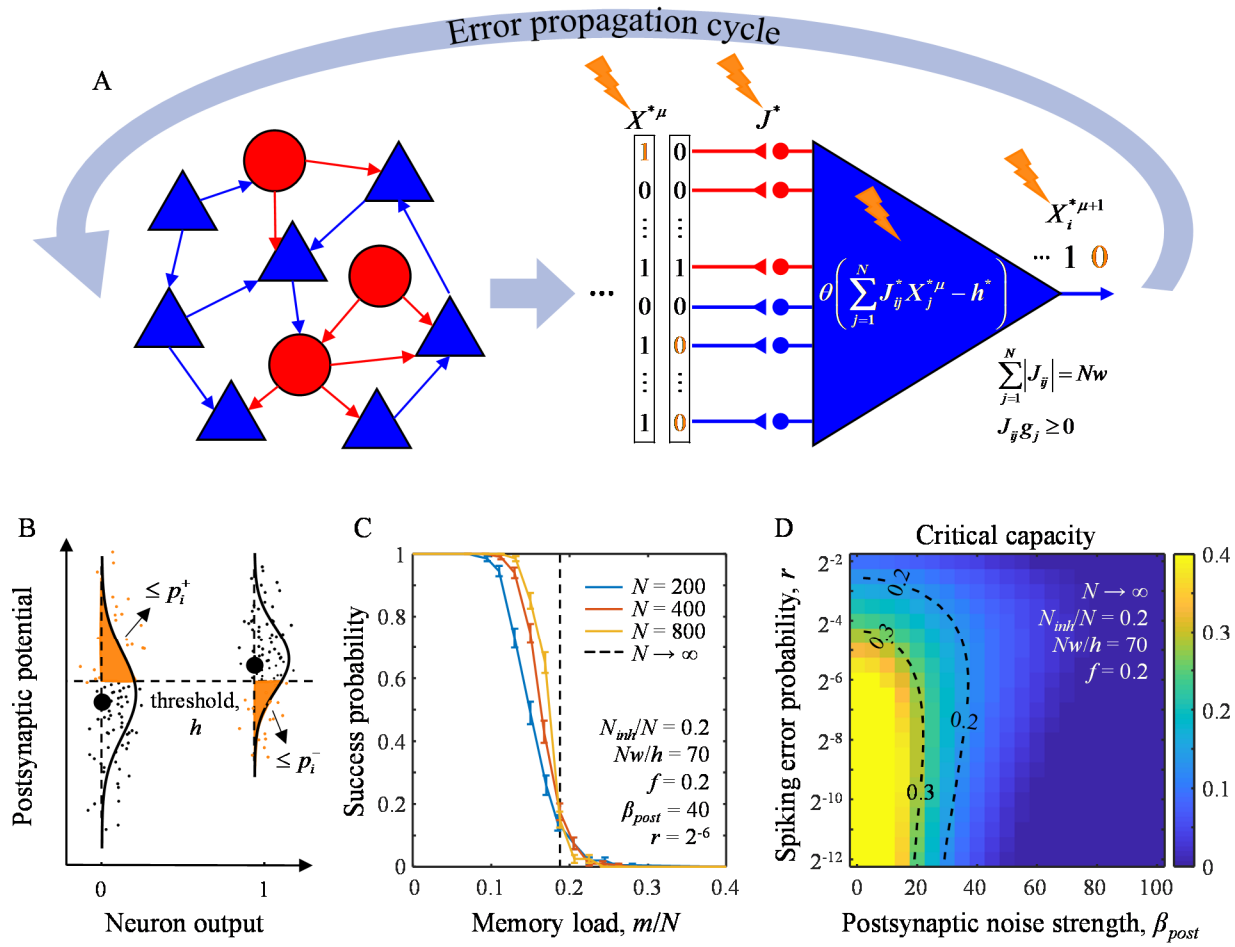

Figure 1: Associative memory storage in a recurrent network of inhibitory and excitatory neurons in the presence of errors and noise. **A.** Cycle of error propagation through the network. Inhibitory neurons (red circles) and excitatory neurons (blue triangles) form an all-to-all potentially (structurally) connected network. Red and blue arrows represent actual (functional) connections. Spiking errors, X^*, synaptic noise, J_{ij}^*, and intrinsic noise, h^*, accompany signal transmission (orange lightning signs). Errors in the neurons' outputs are injected back into the network, becoming spiking errors in the next time step. **B.** Fluctuations in postsynaptic potentials for two associations with target neuron outputs 0 (left) and 1 (right). Large black dots denote postsynaptic potentials in the absence of noise and errors. Small dots represent postsynaptic potentials on different trials in the presence of noise and errors. Orange areas under the postsynaptic potential probability densities (solid lines) represent probabilities of erroneous spiking (left) and spike failures (right). **C.** The probability of successful learning by a neuron is a sharply decreasing function of associative sequence length, m (or memory load m/N). Solid curves represent probabilities of successful learning by a neuron with $N = 200$, 400, and 800 inputs. Values of all other parameters of the neuron are provided in the figure. At 0.5 success probability the neuron is said to be loaded to capacity. The dashed black line represents the theoretical (critical) capacity obtained with the replica method in the $N \to \infty$ limit. **D.** Map of critical capacity as a function of postsynaptic noise strength and spiking error probability.

Neuron-to-neuron connectivity in homogeneous associative networks

One of the most salient features of sign-constrained associative network models, such as the one described in this study, is that finite fractions of inhibitory and excitatory connections assume zero weights at capacity [9], mirroring the trend observed in many local cortical networks. In the following, we compare connection probabilities (P_{con}) and coefficients of variation (CV) of non-zero connection weights in associative networks at capacity to connection probabilities and CV of unitary postsynaptic potentials (uPSP) obtained experimentally. To that end, we used the dataset compiled in [23] based on 87 electrophysiological studies describing neuron-to-neuron connectivity for 420 local cortical projections (lateral distance between neurons < 100 μm). Figure 2A shows that the average inhibitory P_{con} (38 studies, 9,522 connections tested) is significantly larger ($p < 10^{-10}$, two sample t-test) than the average excitatory P_{con} (67 studies, 63,020 connections tested). Associative networks exhibit a similar trend (Figures 2B, C). In particular, in the regions of postsynaptic noise strength and spiking error probability demarcated with dashed isocontours and arrows in Figures 2B, C, the model results are consistent with the middle 50% of experimental measurements.

Figure 2D shows that the average CV of inhibitory uPSP (10 studies, 503 connections recorded) is slightly lower than that for excitatory (36 studies, 3,956 connections recorded), and this trend is reproduced by associative networks in the entire region of considered postsynaptic noise strength and spiking error probability (Figures 2E, F). As before, there is a region in these maps in which results of the model are consistent with the middle 50% of CV of uPSP measurements.

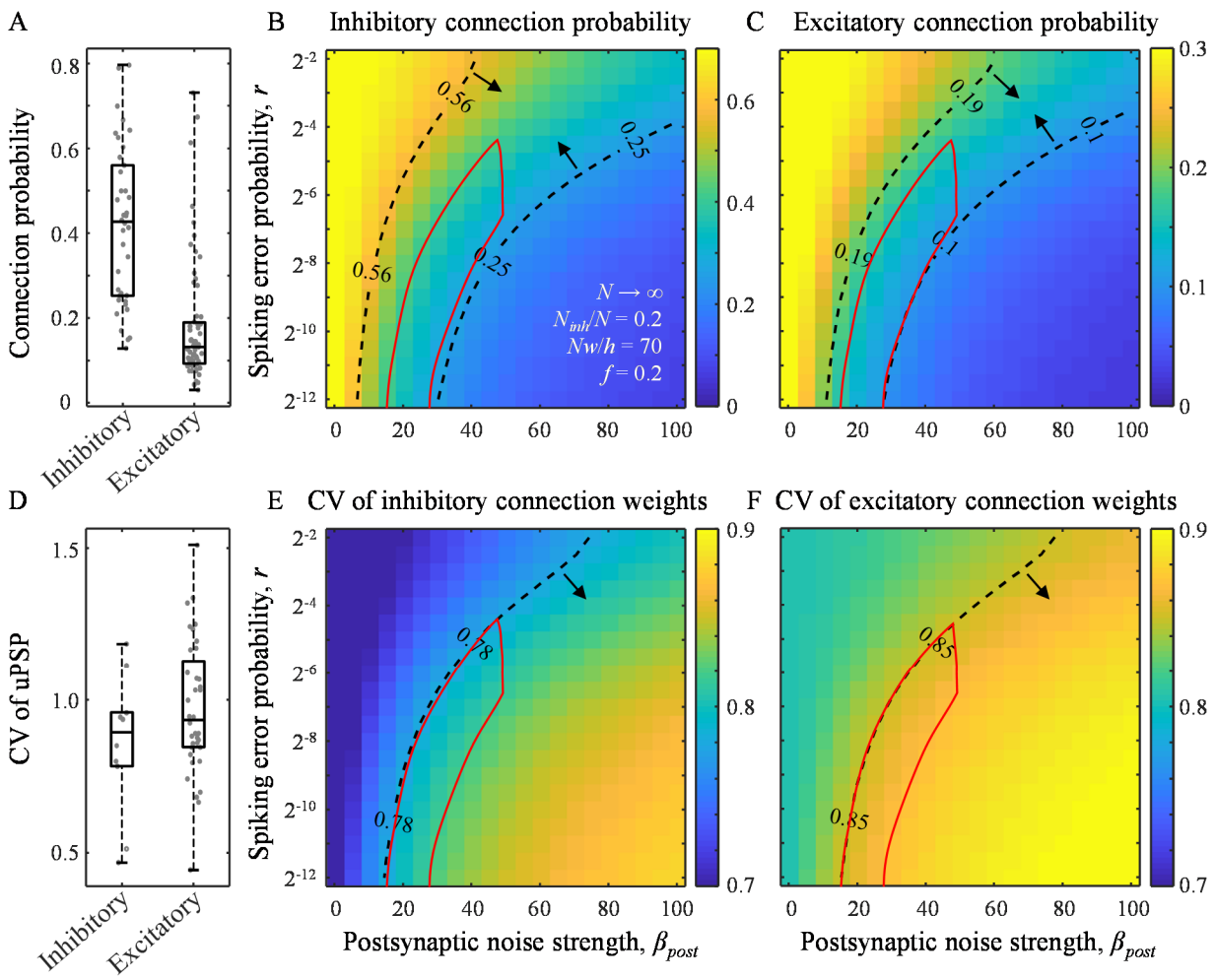

Figure 2: Comparison of structural properties of model and cortical networks. **A.** Inhibitory and excitatory connection probabilities reported in 87 studies describing 420 local cortical projections. Each dot represents the result of a single study/projection. **B, C.** Maps of inhibitory and excitatory connection probabilities as functions of postsynaptic noise strength and spiking error probability. Dashed isocontours and arrows demarcate the interquartile ranges of experimentally observed connection probabilities from (A). The red contour outlines a region of parameters which is consistent with all structural and dynamical measurements in cortical networks considered in this study. **D-F.** Same for CV of inhibitory and excitatory connection weights. (A) and (D) are adapted from [23].

Spontaneous dynamics in homogeneous associative networks

Individual neurons in the model associative network can produce irregular and asynchronous spiking activity, similar to what is observed in cortical networks. To quantify the degree of this

similarity we compared CV of inter-spike-intervals (ISI) and cross-correlation coefficients of randomly initialized spiking network activity in the model to those measurements obtained experimentally. Dashed isocontour in Figure 3A outlines a region of postsynaptic noise strength and spiking error probability in which the model CV of ISI is consistent with the 0.7-1.1 range measured in different cortical systems [24-28]. Similarly, there is a region of postsynaptic noise strength and spiking error probability (Figure 3B) in which calculated spike train cross-correlation coefficients are in agreement with the interquartile range of cortical measurements, 0.04-0.15 [29]. The degree of asynchrony in spontaneous spiking activity in the associative network increases with postsynaptic noise strength, which can be explained by the decrease in connection probability (Figures 2B, C) and, consequently, reduction in the amount of common input to the neurons.

It had been shown that irregular and asynchronous activity can result from balance of inhibitory and excitatory postsynaptic inputs to individual cells [30,31]. In a balanced state, the magnitudes of these inputs are much greater than the threshold of firing, but, due to a high degree of anti-correlation, these inputs nearly cancel, and firing is driven by fluctuations. Figure 3C shows a region of parameters in which neurons in the associative model function in a balanced regime. Because it is difficult to simultaneously measure inhibitory and excitatory postsynaptic inputs to a neuron, anti-correlations of inhibitory and excitatory inputs have only been measured in nearby cells, 0.4 [32,33]. As within-cell anti-correlations are expected to be stronger than between-cell anti-correlations, 0.4 can be viewed as a lower bound for the former (dashed isocontour and arrow in Figure 3C).

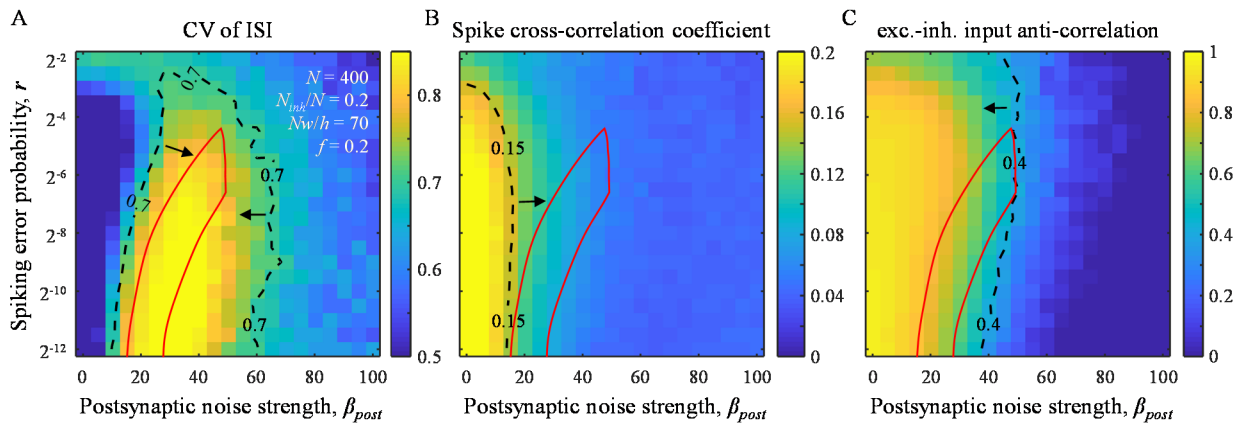

Figure 3: Comparison of dynamical properties of model and cortical networks. **A.** CV of ISI for spontaneous (not learned) activity as a function of postsynaptic noise strength and spiking error probability. Dashed isocontour and arrows demarcates a region of high CV values that is in general agreement with experimental measurements. **B.** Same for cross-correlation coefficient of neuron spike trains. **C.** Same for anti-correlation coefficient of inhibitory and excitatory postsynaptic inputs received by a neuron. The red contour outlines a region of parameters which is consistent with all structural and dynamical measurements considered in Figures 2 and 3.

The seven error-noise regions obtained bases on the properties of neuron-to-neuron connectivity (Figure 2) and network dynamics (Figure 3) have a non-empty intersection (red contour in Figures 2 and 3). In this biologically plausible region, properties of the associative networks are consistent with the considered experimental measurements. This suggests that, during learning and memory retrieval, the postsynaptic noise strength must lie in the 20-50 range, and the spiking error probability must be less than 0.06. The low value of spiking error probability is consistent with experimental observations of reliability of firing patterns evoked by time-varying stimuli *in vivo* [28] and *in vitro* [34].

Properties of heterogeneous associative networks

General formulation and solution of the associative network model, Eqs. (1), makes it possible to investigate the properties of networks composed of heterogeneous populations of inhibitory and excitatory neurons. Specifically, we examined the effects of distributed (neuron class specific) spiking error probabilities and firing probabilities on properties of neuron-to-neuron connectivity at critical capacity. Figures 4B, C show that the probabilities and weights of inhibitory and excitatory connections monotonically decrease with increasing r. Therefore, connections originating from unreliable presynaptic neurons (high r) are significantly depressed. Analysis of connectivity in networks of neurons with distributed firing probabilities (Figures 4D-F) shows that the probabilities of inhibitory connections monotonically increase with f, while the probabilities of excitatory connections exhibit a non-monotonic behavior. One consequence of the results shown in Figures 4B, C and Figures 4E, F is that connection probability must be positively correlated with connection weight, regardless of the source of heterogeneity (r or f). Indeed, this conclusion

is supported by the data from 16 local laminar (same cortical layer) and inter-laminar (different layers) excitatory-to-excitatory projections from mouse barrel cortex [35] (Figure 4G). The same trend was also observed for excitatory-to-inhibitory, inhibitory-to-excitatory, and inhibitory-to-inhibitory projections in mouse visual cortex [36].

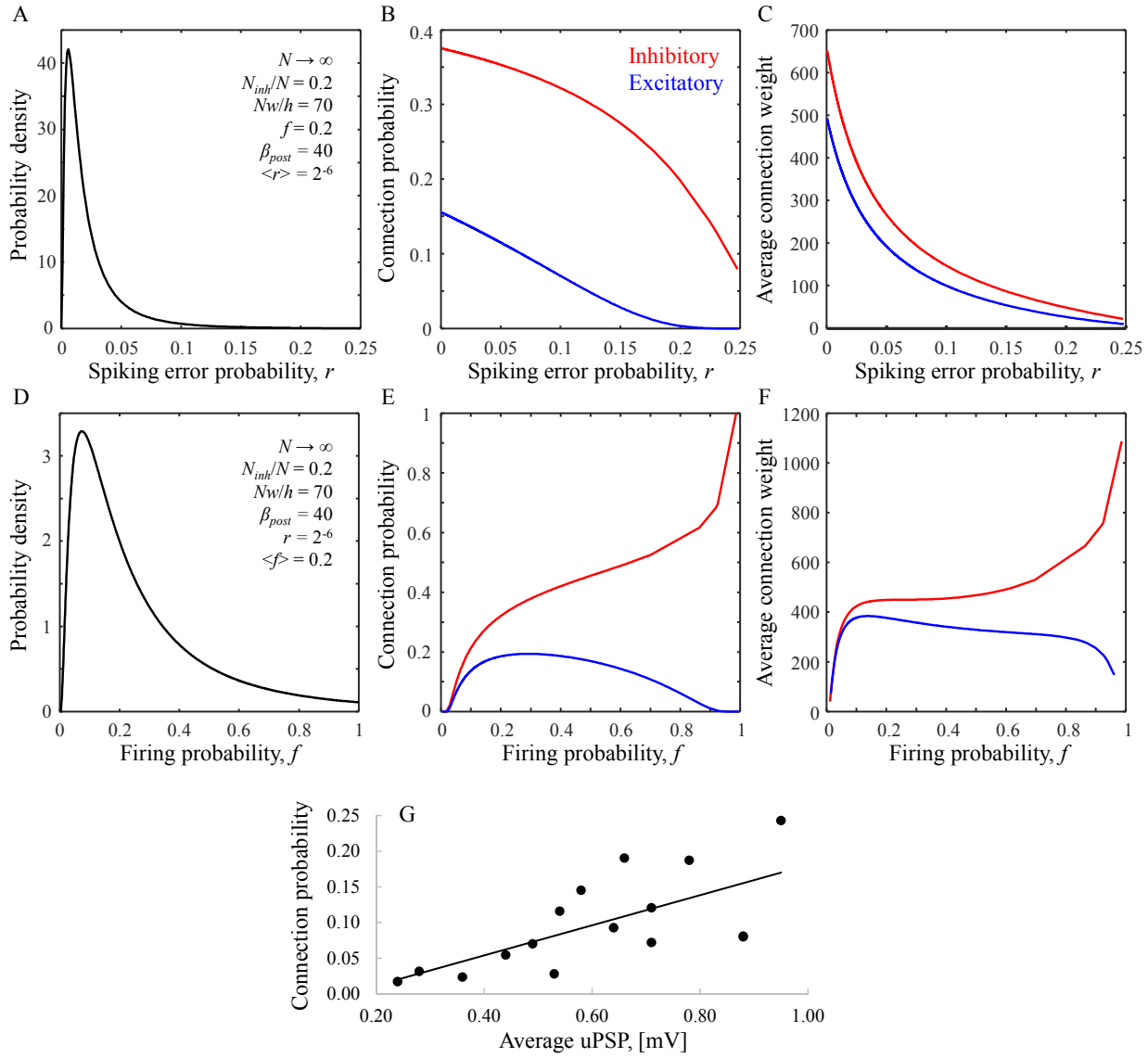

Figure 4: Properties for connections in associative networks of heterogeneous neurons. **A-C.** Connection probability (B) and average non-zero connection weight (C) for inhibitory (red) and excitatory (blue) connections in a network of neurons with distributed spiking errors. Spiking error probabilities of inhibitory and excitatory inputs were randomly drawn from a log-normal distribution with $\langle r \rangle$ belonging to the feasible parameter region (red contour in Figures 2 and 3) (A). Values of all other parameters of the model neurons are shown in (A). Noisy connections have lower

probabilities and weights. **D-F.** Same for a network of neurons with heterogeneous firing probabilities. The postsynaptic noise strength in (A) and (D) and the spiking error probability in (D) belong to the feasible parameter region. Numerical values of all other parameters are explained in [12]. **G.** Connection probability versus average uPSP for 16 local laminar and inter-laminar excitatory projections from mouse barrel cortex [35]. Black line is the best linear fit.

Motivated by the agreement between the results of the associative network model and measurements from various cortical systems, we put forward two falsifiable predictions. First, we predict that in cortical networks, inhibitory and excitatory connections originating from unreliable neurons (or neuron classes) must have lower connection probabilities and average uPSP (Figures 4B, C). We also predict that inhibitory and excitatory neurons (or neuron classes) operating with low firing rates must establish weak and low probability connections (Figures 4E, F).

CONCLUSION

In summary, we built a comprehensive description of the error propagation cycle into the model of associate memory storage. The model was solved theoretically using the replica method in the limit of infinite network size and numerically for large but finites networks. We note that intrinsic noise and postsynaptic spiking errors were first incorporated into the model of associative memory storage by Brunel et al. [10] (see Supplementary Material therein). However, Brunel et al. does not include homeostatic constraints and learning by inhibitory inputs, nor does it examine the complete error propagation cycle, as synaptic noise and presynaptic spiking errors were not considered. We examined the properties of neuron-to-neuron connectivity and dynamics in associative networks at capacity and determined a region of parameters in which the model results are in general agreement with the seven examined features of cortical networks. We show that heterogeneity of neuron population (e.g. different neuron classes) leads to a positive correlation between connection probability and weight, a conclusion supported by measurements from mouse barrel and visual cortices. These results inspired us to put forward two experimentally testable predictions related to cortical connectivity. Finally, we confirmed that memories loaded into individual neurons at a given noise strength can be successfully recalled at the network level in the

presence of the same or slightly lower noise strength (Figure S1). Therefore, errors and noise present during learning can be beneficial, as they increases the robustness of stored memories to fluctuations during memory recall, though, at the expense of memory storage capacity.

ACKNOWLEDGEMENTS

This work is supported by the AFOSR grant FA9550-15-1-0398 and the NSF grant IIS-1526642.

SUPPLEMENTAL MATERIAL

This Supplemental Material describes a model of associative memory storage by a recurrent network of inhibitory and excitatory McCulloch and Pitts neurons. The model builds on previous studies of associative memory (see e.g. [6,7,10-12,18,19]), and provides an all-inclusive account of the error propagation cycle, from pre- and postsynaptic spiking errors, to synaptic noise, to intrinsic noise. In addition, the model incorporates a number of constraints motivated by the experimental data. The model is solved theoretically with the replica method in the limit of infinite network size [21,22], and numerically for large, but finite networks. The model gives rise to a comprehensive list of predictions regarding the structure and dynamics of the neural network at the critical memory storage capacity that are consistent with a large number of experimental studies of connectivity in local cortical circuits.

Perceptron model with biologically inspired constraints in the presence of errors and noise

We consider a single perceptron-like neuron that receives N_{inh} inhibitory and $(N - N_{inh})$ excitatory input connections. The neuron is faced with a task of learning a set of m input-output associations $\left\{X^{\mu}, y^{\mu}\right\}$, where X^{μ} and y^{μ} are a binary vector and a scalar describing the neuron's input and the target output at time step μ. The neuron must learn these input-output associations by adjusting the weights of its connections, J_j, in the presence of two constraints. First, the l_1-norm of

connection weights must remain fixed during learning. Second, the signs of inhibitory and excitatory connection weights must not change during learning.

Learning in the model is accompanied with three types of errors or noise. These include, pre- and postsynaptic spiking errors, or errors in X^{μ} and y^{μ}, synaptic noise, or noise in J, and intrinsic noise, or noise in the neuron's threshold of firing, h. In the following, we use asterisks to denote quantities containing errors or noise (e.g. $X^{*\mu}$), whereas variables without asterisks represent the mean (for h and J_j) or original (for X^{μ} and y^{μ}) values.

Individual elements of X^{μ} (X_j^{μ}), are independently drawn from input dependent Bernoulli probability distributions: 0 with probability $1 - f_j$ and 1 with probability f_j. Errors in X^{μ} are introduced with Bernoulli trials by making independent and random changes with probabilities $P\left(X_j^{*\mu} = 0 \mid X_j^{\mu} = 1\right) \equiv p_j^-$ and $P\left(X_j^{*\mu} = 1 \mid X_j^{\mu} = 0\right) \equiv p_j^+$. Similarly, the target outputs, y^{μ}, are independently drawn from a Bernoulli probability distribution: 0 with probability $1 - f_{out}$ and 1 with probability f_{out}. The probabilities of error in the neuron's output must be bounded, $P\left(y^{*\mu} = 0 \mid y^{\mu} = 1\right) \le p_{out}^-$ and $P\left(y^{*\mu} = 1 \mid y^{\mu} = 0\right) \le p_{out}^+$. In the following, we only consider spiking error probabilities in the ranges, $p_{j,out}^+ < f_{j,out}$ and $p_{j,out}^- < 1 - f_{j,out}$, which is required for the stability of the replica theory calculation.

We assume that the neuron's connection weights fluctuate based on the quantal model of synaptic transmission of Del Castillo and Katz [20]. According to this model, connection weight values J_j^* on different trials are described with a Binomial distribution in which $\left\langle J_j^* \right\rangle \equiv J_j$ and the variance is proportional to the mean, $\mathrm{var}\left(J_j^*\right) = \frac{h\beta_{syn}}{N}\left|J_j\right|$. The dimensionless coefficient β_{syn} is referred to as the synaptic noise strength, and a factor of h/N is introduced for convenience.

All other sources of noise contributing to the neuron's postsynaptic potential (e.g. background synaptic activity and stochasticity of ion channel states) are referred to as intrinsic noise. In the model, these sources are folded into h^*, which is assumed to be Gaussian distributed with $\left\langle h^* \right\rangle \equiv h$

and $\mathrm{var}\left(h^*\right)=\frac{h^2\beta_{int}^2}{N}$. Here, β_{int} is a dimensionless coefficient called intrinsic noise strength, and a factor of h^2/N is introduced for convenience.

The above described model can be summarized as follows:

$$
\begin{aligned}
&\theta\left(\sum_{j=1}^{N} J_j^* X_j^{*\mu} - h^*\right) = y^{*\mu}, \quad \mu = 1,\ldots,m \\
&\frac{1}{N}\sum_{j=1}^{N} J_j g_j = w \\
&J_j g_j \geq 0, \quad j = 1,\ldots,N \\
&P\left(X_j^{\mu}=1\right) \equiv f_j; \quad P\left(X_j^{*\mu}=0 \mid X_j^{\mu}=1\right) \equiv p_j^-; \quad P\left(X_j^{*\mu}=1 \mid X_j^{\mu}=0\right) \equiv p_j^+ \\
&P\left(y^{\mu}=1\right) \equiv f_{out}; \quad P\left(y^{*\mu}=0 \mid y^{\mu}=1\right) \leq p_{out}^-; \quad P\left(y^{*\mu}=1 \mid y^{\mu}=0\right) \leq p_{out}^+ \\
&\left\langle J_j^* \right\rangle \equiv J_j; \quad \mathrm{var}\left(J_j^*\right) = \frac{h\beta_{syn}}{N}\left|J_j\right| \\
&\left\langle h^* \right\rangle \equiv h; \quad \mathrm{var}\left(h^*\right) = \frac{h^2\beta_{int}^2}{N}
\end{aligned}
\tag{S1}
$$

In these expressions, θ is the Heaviside step-function, f_j is the firing probability of input j, f_{out} is the firing probability of the neuron. To enforce sign-constraints on connection weights we introduced parameters g_j, which equal 1 if the input j is excitatory and –1 if it is inhibitory. Parameter w is referred to as the average absolute connection weight. The neuron is faced with the task of finding connection weights, $\left\{J_j\right\}$, that satisfy Eqs. (S1) for a given set of model parameters: $N, m, h, w, \left\{g_j\right\}, \left\{f_j\right\}, f_{out}, \left\{p_j^-\right\}, \left\{p_j^+\right\}, p_{out}^-, p_{out}^+, \beta_{syn}, \beta_{int}$.

Reformulation of the model for the large *N* limit

In the limit of large N, the Central Limit Theorem ensures that the neuron's postsynaptic potential (PSP), $\sum_{j=1}^{N} J_j^* X_j^{*\mu}$, at any time step is Gaussian distributed. Therefore, the deviation of PSP from

the threshold of firing, $I^{*\mu} \equiv \sum_{j=1}^{N} J_j^* X_j^{*\mu} - h^*$, is also Gaussian distributed with mean and standard deviation given by the following expressions:

$$I^{\mu} = \sum_{j=1}^{N} J_j \left[\left(1 - p_j^-\right) X_j^{\mu} + p_j^+ \left(1 - X_j^{\mu}\right) \right] - h$$

$$\sigma^{\mu} = \left(\sum_{j=1}^{N} \left[J_j^2 \left(p_j^- \left(1 - p_j^-\right) X_j^{\mu} + p_j^+ \left(1 - p_j^+\right)\left(1 - X_j^{\mu}\right)\right) + \frac{h\beta_{syn}}{N} J_j g_j \left(\left(1 - p_j^-\right) X_j^{\mu} + p_j^+ \left(1 - X_j^{\mu}\right)\right) + \frac{h^2 \beta_{int}^2}{N^2} \right] \right)^{\frac{1}{2}}$$

(S2)

As a result, the inequality constraints on the probabilities of output spike errors in Eqs. (S1) can be expressed in terms of I^{μ} and σ^{μ}:

$$\begin{aligned} I^{\mu} &\geq \sqrt{2}\,\mathrm{erf}^{-1}\left(1 - 2p_{out}^-\right)\sigma^{\mu}, \quad y^{\mu} = 1 \\ I^{\mu} &\leq -\sqrt{2}\,\mathrm{erf}^{-1}\left(1 - 2p_{out}^+\right)\sigma^{\mu}, \quad y^{\mu} = 0 \end{aligned} \tag{S3}$$

The above two inequalities can be combined into a single expression that must hold, assuming the association μ is successfully learned:

$$\left(2y^{\mu} - 1\right) I^{\mu} \geq \sqrt{2}\left(\mathrm{erf}^{-1}\left(1 - 2p_{out}^-\right) y^{\mu} + \mathrm{erf}^{-1}\left(1 - 2p_{out}^+\right)\left(1 - y^{\mu}\right)\right)\sigma^{\mu} \tag{S4}$$

Additional assumptions required for the replica theory solution

Following the procedure outlined in [11,12], we assume that the model parameters m/N, $\{f_j\}$, f_{out}, $\{p_j^-\}$, $\{p_j^+\}$, p_{out}^-, p_{out}^+, β_{syn}, and β_{int} are intensive, or of order 1 with respect to N. In addition, we assume that connection weights are inversely proportional to the system size, $\left\{J_j = \frac{h}{N}\tilde{J}_j\right\}$, and refer to $\{\tilde{J}_j\}$ as scaled connection weights. This particular scaling is traditionally used in associative memory models [10], and it has been shown that in the biologically plausible high-

weight regime, $Nwf \gg h$, many model results become independent of this assumption [23]. It follows from the second line of Eqs. (S1) that $w = \frac{h}{N}\tilde{w}$, and we refer to $\tilde{w}$ as scaled average absolute connection weight.

The complete model, rewritten in terms of the scaled variables, contains one equality and $m + N$ inequality constraints:

$$
\begin{aligned}
&\left(2y^{\mu}-1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j\left[\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-X_j^{\mu}\right)\right]-1\right)\geq\sqrt{\frac{2}{N}}\left(\mathrm{erf}^{-1}\left(1-2p_{out}^-\right)y^{\mu}+\mathrm{erf}^{-1}\left(1-2p_{out}^+\right)\left(1-y^{\mu}\right)\right)\times\\
&\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_j^2\left(p_j^-\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-p_j^+\right)\left(1-X_j^{\mu}\right)\right)+\beta_{syn}\tilde{J}_j g_j\left(\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-X_j^{\mu}\right)\right)+\beta_{int}^2\right]\right)^{\frac{1}{2}},\quad \mu=1,\ldots,m\\
&\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j g_j=\tilde{w}\\
&\tilde{J}_j g_j\geq 0,\quad j=1,\ldots,N\\
&P\left(X_j^{\mu}=1\right)\equiv f_j;\quad P\left(y^{\mu}=1\right)\equiv f_{out}
\end{aligned}
$$

(S5)

Replica theory solution of the model

We begin by calculating the volume of the connection weight space, $\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)$, in which Eqs. (S5) hold for a given set of associations, $\left\{X^{\mu},y^{\mu}\right\}$:

$$\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)=\int\prod_{j=1}^{N}d\tilde{J}_{j}\prod_{j=1}^{N}\theta\left(\tilde{J}_{j}g_{j}\right)\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_{j}g_{j}-\tilde{w}\right)\prod_{\mu=1}^{m}\theta\left(\left(2y^{\mu}-1\right)\times\right.$$

$$\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_{j}\left[\left(1-p_{j}^{-}\right)X_{j}^{\mu}+p_{j}^{+}\left(1-X_{j}^{\mu}\right)\right]-1\right)-\sqrt{\frac{2}{N}}\left(\mathrm{erf}^{-1}\left(1-2p_{out}^{-}\right)y^{\mu}+\mathrm{erf}^{-1}\left(1-2p_{out}^{+}\right)\left(1-y^{\mu}\right)\right)\times$$

$$\left.\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_{j}^{2}\left(p_{j}^{-}\left(1-p_{j}^{-}\right)X_{j}^{\mu}+p_{j}^{+}\left(1-p_{j}^{+}\right)\left(1-X_{j}^{\mu}\right)\right)+\beta_{syn}\tilde{J}_{j}g_{j}\left(\left(1-p_{j}^{-}\right)X_{j}^{\mu}+p_{j}^{+}\left(1-X_{j}^{\mu}\right)\right)+\beta_{int}^{2}\right]\right)^{\frac{1}{2}}\right)$$

(S6)

The typical volume of this solution space, $\Omega_{typical}$, is defined through the averaging of $\ln\left(\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)\right)$ over the set of associations $\left\{X^{\mu},y^{\mu}\right\}$, and is calculated by introducing n replica systems:

$$\ln\left(\Omega_{typical}\right)=\left\langle\ln\left(\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)\right)\right\rangle_{\left\{X^{\mu},y^{\mu}\right\}}=\lim_{n\to 0}\frac{\left\langle\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)^{n}\right\rangle_{\left\{X^{\mu},y^{\mu}\right\}}-1}{n} \tag{S7}$$

The quantity $\left\langle\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)^{n}\right\rangle_{\left\{X^{\mu},y^{\mu}\right\}}$ can be rewritten as a single multidimensional integral:

$$\left\langle\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)^{n}\right\rangle_{\left\{X^{\mu},y^{\mu}\right\}}=\left\langle\int\prod_{j,a=1}^{N,n}d\tilde{J}_{j}^{a}\prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_{j}^{a}g_{j}\right)\prod_{a=1}^{n}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_{j}^{a}g_{j}-\tilde{w}\right)\prod_{\mu,a=1}^{m,n}\theta\left(\left(2y^{\mu}-1\right)\times\right.\right.$$

$$\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_{j}^{a}\left[\left(1-p_{j}^{-}\right)X_{j}^{\mu}+p_{j}^{+}\left(1-X_{j}^{\mu}\right)\right]-1\right)-\sqrt{\frac{2}{N}}\left(\mathrm{erf}^{-1}\left(1-2p_{out}^{-}\right)y^{\mu}+\mathrm{erf}^{-1}\left(1-2p_{out}^{+}\right)\left(1-y^{\mu}\right)\right)\times$$

$$\left.\left.\left(\frac{1}{N}\sum_{j=1}^{N}\left[\left(\tilde{J}_{j}^{a}\right)^{2}\left(p_{j}^{-}\left(1-p_{j}^{-}\right)X_{j}^{\mu}+p_{j}^{+}\left(1-p_{j}^{+}\right)\left(1-X_{j}^{\mu}\right)\right)+\beta_{syn}\tilde{J}_{j}^{a}g_{j}\left(\left(1-p_{j}^{-}\right)X_{j}^{\mu}+p_{j}^{+}\left(1-X_{j}^{\mu}\right)\right)+\beta_{int}^{2}\right]\right)^{\frac{1}{2}}\right)\right\rangle_{\left\{X^{\mu},y^{\mu}\right\}}$$

(S8)

This integral is calculated by following a previously established procedure [11,12]. Below, we outline the main steps of this calculation.

To calculate the average over the associations, we first decouple X^{μ} and y^{μ} by introducing two new sets of variables:

$$\frac{\lambda^{a,\mu}}{\sqrt{N}} = \frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a\left[\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-X_j^{\mu}\right)\right]-1$$
$$\nu^{a,\mu} = \frac{1}{N}\sum_{j=1}^{N}\left[\left(\tilde{J}_j^a\right)^2\left(p_j^-\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-p_j^+\right)\left(1-X_j^{\mu}\right)\right)+\beta_{syn}\tilde{J}_j^a g_j\left(\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-X_j^{\mu}\right)\right)+\beta_{int}^2\right]$$

(S9)

These variables are next incorporated into Eq. (S8) with the help of Dirac δ-functions:

$$\left\langle\Omega\left(\left\{X^{\mu},y^{\mu}\right\}\right)^n\right\rangle_{\left\{X^{\mu},y^{\mu}\right\}} = \int\prod_{j,a=1}^{N,n}d\tilde{J}_j^a\prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_j^a g_j\right)\prod_{a=1}^{n}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a g_j-\tilde{w}\right)\prod_{\mu,a=1}^{m,n}\frac{d\lambda^{a,\mu}}{\sqrt{N}}\prod_{\mu,a=1}^{m,n}d'\nu^{a,\mu}\times$$
$$\left\langle\prod_{\mu,a=1}^{m,n}\theta\left(\left(2y^{\mu}-1\right)\lambda^{a,\mu}-\sqrt{2}\left(\mathrm{erf}^{-1}\left(1-2p_{out}^-\right)y^{\mu}+\mathrm{erf}^{-1}\left(1-2p_{out}^+\right)\left(1-y^{\mu}\right)\right)\left(\nu^{a,\mu}\right)^{\frac{1}{2}}\right)\right\rangle_{\left\{y^{\mu}\right\}}\times$$
$$\left\langle\prod_{\mu,a=1}^{m,n}\delta\left(1+\frac{\lambda^{a,\mu}}{\sqrt{N}}-\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a\left[\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-X_j^{\mu}\right)\right]\right)\times\right.$$
$$\left.\prod_{\mu,a=1}^{m,n}\delta\left(\nu^{a,\mu}-\frac{1}{N}\sum_{j=1}^{N}\left[\left(\tilde{J}_j^a\right)^2\left(p_j^-\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-p_j^+\right)\left(1-X_j^{\mu}\right)\right)+\beta_{syn}\tilde{J}_j^a g_j\left(\left(1-p_j^-\right)X_j^{\mu}+p_j^+\left(1-X_j^{\mu}\right)\right)+\beta_{int}^2\right]\right)\right\rangle_{\left\{X^{\mu}\right\}}$$

(S10)

Symbol d' in this expression and thereafter is designated for 0 to ∞ integration, whereas d is used for integration from $-\infty$ to ∞. In the following step, the Heaviside step-functions and the δ-functions are replaced with their Fourier representations, which makes it possible to perform the averaging over the associations:

$$\left\langle \Omega\left(\left\{X^{\mu}, y^{\mu}\right\}\right)^{n}\right\rangle_{\left\{X^{\mu}, y^{\mu}\right\}} = \int \prod_{j,a=1}^{N,n} d\tilde{J}_j^a \prod_{\mu,a=1}^{m,n} \frac{d\lambda^{a,\mu} d\hat{\lambda}^{a,\mu}}{2\pi} \prod_{\mu,a=1}^{m,n} \frac{d'u^{a,\mu} d\hat{u}^{a,\mu}}{2\pi} \prod_{\mu,a=1}^{m,n} \frac{d'\nu^{a,\mu} d\hat{\nu}^{a,\mu}}{2\pi} \prod_{a=1}^{n} \frac{d\hat{k}^a}{2\pi / N} \times$$

$$\prod_{j,a=1}^{N,n} \theta\left(\tilde{J}_j^a g_j\right) \prod_{a=1}^{n} e^{i\hat{k}^a\left(\sum_{j=1}^{N} \tilde{J}_j^a g_j - N\tilde{w}\right)} \prod_{\mu,a=1}^{m,n} e^{-i\hat{u}^{a,\mu} u^{a,\mu}} \prod_{\mu,a=1}^{m,n} \left(f_{out} e^{i\hat{u}^{a,\mu}\left(\lambda^{a,\mu} - C_{out}^{-}\left(\nu^{a,\mu}\right)^{\frac{1}{2}}\right)} + (1 - f_{out}) e^{-i\hat{u}^{a,\mu}\left(\lambda^{a,\mu} + C_{out}^{+}\left(\nu^{a,\mu}\right)^{\frac{1}{2}}\right)} \right) \times$$

$$\prod_{\mu,a=1}^{m,n} e^{i\hat{\lambda}^{a,\mu}\left(\sqrt{N} + \lambda^{a,\mu}\right) + i\hat{\nu}^{a,\mu}\nu^{a,\mu} - i\hat{\nu}^{a,\mu}\beta_{int}^2} \prod_{\mu,j=1}^{m,N} e^{-\frac{iA_j}{\sqrt{N}}\sum_{a=1}^{n}\hat{\lambda}^{a,\mu}\tilde{J}_j^a - \frac{i\beta_{syn} g_j A_j}{N}\sum_{a=1}^{n}\hat{\nu}^{a,\mu}\tilde{J}_j^a - \frac{iB_j}{N}\sum_{a=1}^{n}\hat{\nu}^{a,\mu}\left(\tilde{J}_j^a\right)^2 - \frac{C_j}{2N}\left(\sum_{a=1}^{n}\hat{\lambda}^{a,\mu}\tilde{J}_j^a\right)^2}$$

(S11)

The following notation is used in the above expression:

$$\begin{aligned}
A_j &= \left(1 - p_j^-\right) f_j + p_j^+ \left(1 - f_j\right) \\
B_j &= f_j p_j^- \left(1 - p_j^-\right) + \left(1 - f_j\right) p_j^+ \left(1 - p_j^+\right) \\
C_j &= f_j \left(1 - f_j\right)\left(1 - p_j^+ - p_j^-\right)^2 \\
D_{out}^{\pm} &= \sqrt{2}\,\mathrm{erf}^{-1}\left(1 - 2p_{out}^{\pm}\right)
\end{aligned} \tag{S12}$$

We note that parameters A, B, and C, defined in Eqs. (S12) are nonnegative.

Next, we decouple the products containing indices j and μ by introducing three sets of order parameters and embedding them into Eq. (S11) with the help of Dirac δ-functions as was done in Eqs. (S10, S11):

$$\frac{1}{N}\sum_{j=1}^{N} A_j \tilde{J}_j^a = 1 + \frac{s^a}{\sqrt{N}}, \quad \frac{1}{N}\sum_{j=1}^{N}\left(\beta_{syn} A_j g_j \tilde{J}_j^a + B_j\left(\tilde{J}_j^a\right)^2 + \beta_{int}^2\right) = \kappa^a, \quad \frac{1}{N}\sum_{j=1}^{N} C_j \tilde{J}_j^a \tilde{J}_j^b = q^{a,b}$$

$$\left\langle \Omega\left(\left\{X^\mu, y^\mu\right\}\right)^n \right\rangle_{\left\{X^\mu, y^\mu\right\}} = \int \prod_{a=1}^{n} \frac{d\hat{k}^a}{2\pi/N} \prod_{a=1}^{n} \frac{ds^a d\hat{s}^a}{2\pi/\sqrt{N}} \prod_{a=1}^{n} \frac{d\kappa^a d\hat{\kappa}^a}{2\pi/N} \prod_{a,b=1}^{n} \frac{dq^{a,b} d\hat{q}^{a,b}}{2\pi/N} \times$$

$$\prod_{a=1}^{n} e^{i\hat{s}^a\left(N+\sqrt{N}s^a\right)} \prod_{a=1}^{n} e^{iN\kappa^a\hat{\kappa}^a} \prod_{a,b=1}^{n,n} e^{iN\hat{q}^{a,b}q^{a,b}} \prod_{a=1}^{n} e^{-iN\hat{k}^a\tilde{w}} \times$$

$$\left[\prod_{a=1}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \prod_{a=1}^{n} \frac{d'u^a d\hat{u}^a}{2\pi} \prod_{a=1}^{n} \frac{d'\nu^a d\hat{\nu}^a}{2\pi} \left(f_{out} e^{i\sum_{a=1}^{n}\hat{u}^a\left(\lambda^a - D_{out}^{-}\sqrt{\nu^a}\right)} + (1-f_{out}) e^{-i\sum_{a=1}^{n}\hat{u}^a\left(\lambda^a + D_{out}^{+}\sqrt{\nu^a}\right)} \right) \times \right.$$

$$\left. e^{i\sum_{a=1}^{n}\left(\hat{\lambda}^a\left(\lambda^a - s^a\right) + \hat{\nu}^a\left(\nu^a - \kappa^a\right) - \hat{u}^a u^a\right) - \frac{1}{2}\sum_{a,b=1}^{n,n}\hat{\lambda}^a\hat{\lambda}^b q^{a,b}} \right]^m \times$$

$$\prod_{j,a=1}^{N,n} d\tilde{J}_j^a \prod_{j,a=1}^{N,n} \theta\left(\tilde{J}_j^a g_j\right) \prod_{a=1}^{n} e^{i\hat{k}^a\sum_{j=1}^{N}\tilde{J}_j^a g_j - i\hat{s}^a\sum_{j=1}^{N}A_j\tilde{J}_j^a - i\hat{\kappa}^a\sum_{j=1}^{N}\left(\beta_{syn}A_j g_j \tilde{J}_j^a + B_j\left(\tilde{J}_j^a\right)^2 + \beta_{int}^2\right)} \prod_{a,b=1}^{n,n} e^{-i\hat{q}^{a,b}\sum_{j=1}^{N}C_j\tilde{J}_j^a\tilde{J}_j^b} \tag{S13}$$

After integrating over ν^a, $\hat{\nu}^a$, λ^a, and $\hat{\lambda}^a$ we obtain:

$$\left\langle \Omega\left(\left\{X^\mu, y^\mu\right\}\right)^n \right\rangle_{\left\{X^\mu, y^\mu\right\}} = \int \prod_{a=1}^{n} \frac{d\hat{k}^a}{2\pi/N} \prod_{a=1}^{n} \frac{ds^a d\hat{s}^a}{2\pi/\sqrt{N}} \prod_{a=1}^{n} \frac{d'\kappa^a d\hat{\kappa}^a}{2\pi/N} \prod_{a,b=1}^{n} \frac{dq^{a,b} d\hat{q}^{a,b}}{2\pi/N} \times$$

$$e^{N\left(i\sum_{a=1}^{n}\left(\hat{s}^a - \hat{k}^a\tilde{w}\right) + i\sum_{a=1}^{n}\kappa^a\hat{\kappa}^a + i\sum_{a,b=1}^{n,n}q^{a,b}\hat{q}^{a,b} + n\alpha\tilde{G}_E\left(\left\{s^a\right\},\left\{q^{a,b}\right\}\right) + n\tilde{G}_S\left(\left\{\hat{k}^a\right\},\left\{\hat{s}^a\right\},\left\{\hat{q}^{a,b}\right\}\right)\right)}$$

$$\tilde{G}_E\left(\left\{s^a\right\},\left\{\kappa^a\right\},\left\{q^{a,b}\right\}\right) = \frac{1}{n}\ln\left(\int \prod_{a=1}^{n} \frac{d'u^a d\hat{u}^a}{2\pi} e^{-\frac{1}{2}\sum_{a,b=1}^{n}\hat{u}^a\hat{u}^b q^{a,b}} \left(f_{out} e^{i\sum_{a=1}^{n}\hat{u}^a\left(s^a - u^a - D_{out}^{-}\sqrt{\kappa^a}\right)} + (1-f_{out}) e^{-i\sum_{a=1}^{n}\hat{u}^a\left(s^a + u^a + D_{out}^{+}\sqrt{\kappa^a}\right)} \right)\right)$$

$$\tilde{G}_S\left(\left\{\hat{k}^a\right\},\left\{\hat{s}^a\right\},\left\{\hat{\kappa}^a\right\},\left\{\hat{q}^{a,b}\right\}\right) = \frac{1}{n}\frac{1}{N}\sum_{j=1}^{N}\ln\left(\int \prod_{a=1}^{n} d'\tilde{J}^a e^{i\sum_{a=1}^{n}\left(\hat{k}^a g_j - \hat{s}^a A_j\right)g_j\tilde{J}^a - i\sum_{a=1}^{n}\hat{\kappa}^a\left(\beta_{syn}A_j\tilde{J}^a + B_j\left(\tilde{J}^a\right)^2 + \beta_{int}^2\right) - iC_j\sum_{a,b=1}^{n}\tilde{J}^a\tilde{J}^b\hat{q}^{a,b}}\right)$$

(S14)

The above integral is calculated by using the steepest descent method combined with the assumption of a replica symmetric saddle point, $s^a = s$, $\kappa^a = \kappa \geq 0$, $q^{a,a} = q_0$, $q^{a\neq b} = q$, $\hat{k}^a = \hat{k}$, $\hat{s}^a = \hat{s}$, $\hat{\kappa}^a = \hat{\kappa}$, $\hat{q}^{a,a} = \hat{q}_0$, and $\hat{q}^{a\neq b} = \hat{q}$:

$$\left\langle \Omega\left(\left\{X^{\mu}, y^{\mu}\right\}\right)^{n}\right\rangle_{\left\{X^{\mu}, y^{\mu}\right\}} \sim e^{Nn\left(i\left(\hat{s}-\hat{k}\tilde{w}\right)+i\kappa\hat{\kappa}+iq_0\hat{q}_0-iq\hat{q}+\alpha G_E\left(s,\kappa,q_0,q\right)+G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)\right)}$$

$$G_E\left(s,\kappa,q_0,q\right)=\frac{1}{n}\ln\left(\int\prod_{a=1}^{n}\frac{d'u^a d\hat{u}^a}{2\pi}e^{-\frac{1}{2}q_0\sum_{a=1}^{n}\left(\hat{u}^a\right)^2-\frac{1}{2}q\sum_{a\neq b=1}^{n}\hat{u}^a\hat{u}^b}\left(f_{out}e^{i\sum_{a=1}^{n}\hat{u}^a\left(s-u^a-D_{out}^{-}\sqrt{\kappa}\right)}+(1-f_{out})e^{-i\sum_{a=1}^{n}\hat{u}^a\left(s+u^a+D_{out}^{+}\sqrt{\kappa}\right)}\right)\right)$$

$$G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)=\frac{1}{n}\frac{1}{N}\sum_{j=1}^{N}\ln\left(\int\prod_{a=1}^{n}d'\tilde{J}^a e^{i\left(\hat{k}g_j-\hat{s}A_j\right)g_j\sum_{a=1}^{n}\tilde{J}^a-i\hat{\kappa}\sum_{a=1}^{n}\left(\beta_{syn}A_j\tilde{J}^a+B_j\left(\tilde{J}^a\right)^2+\beta_{int}^2\right)-iC_j\hat{q}_0\sum_{a=1}^{n}\left(\tilde{J}^a\right)^2-iC_j\hat{q}\sum_{a\neq b=1}^{n}\tilde{J}^a\tilde{J}^b}\right)$$

(S15)

The non-redundant, replica symmetric saddle point coordinates $\left(s,\kappa,q_0,q,\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)$ satisfy the following system of nine equations and one inequality:

$$\frac{\partial G_E\left(s,\kappa,q_0,q\right)}{\partial s}=0;\quad \frac{\partial G_E\left(s,\kappa,q_0,q\right)}{\partial \kappa}=-\frac{i\hat{\kappa}}{\alpha};\quad \frac{\partial G_E\left(s,\kappa,q_0,q\right)}{\partial q_0}=-\frac{i\hat{q}_0}{\alpha};\quad \frac{\partial G_E\left(s,\kappa,q_0,q\right)}{\partial q}=\frac{i\hat{q}}{\alpha};$$

$$\frac{\partial G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)}{\partial \hat{k}}=i\tilde{w};\quad \frac{\partial G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)}{\partial \hat{s}}=-i;\quad \frac{\partial G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)}{\partial \hat{\kappa}}=-i\kappa;$$

$$\frac{\partial G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)}{\partial \hat{q}_0}=-iq_0;\quad \frac{\partial G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)}{\partial \hat{q}}=iq;\quad \kappa\geq 0$$

(S16)

To simplify the expressions for G_E and G_S we employ the Hubbard-Stratonovich transformation (see e.g. [12]) and take the $n\to 0$ limit:

$$G_E\left(s,\kappa,q_0,q\right)=\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\left[f_{out}\ln\left(\int\frac{d'ud\hat{u}}{2\pi}e^{-\frac{1}{2}q_0\hat{u}^2+2x\sqrt{-\frac{1}{2}q}\hat{u}+\frac{1}{2}q\hat{u}^2+i\hat{u}\left(s-u-D_{out}^{-}\sqrt{\kappa}\right)}\right)+\right.$$
$$\left.(1-f_{out})\ln\left(\int\frac{d'ud\hat{u}}{2\pi}e^{-\frac{1}{2}q_0\hat{u}^2+2x\sqrt{-\frac{1}{2}q}\hat{u}+\frac{1}{2}q\hat{u}^2-i\hat{u}\left(s+u+D_{out}^{+}\sqrt{\kappa}\right)}\right)\right]$$

$$G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)=\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\ln\left(e^{-i\hat{\kappa}\beta_{int}^2}\int d'\tilde{J}e^{\left(i\left(\hat{k}g_j-\hat{s}A_j\right)g_j+2x\sqrt{-iC_j\hat{q}}-i\hat{\kappa}\beta_{syn}A_j\right)\tilde{J}-i\left(C_j\left(\hat{q}_0-\hat{q}\right)+B_j\hat{\kappa}\right)\tilde{J}^2}\right)$$

(S17)

Integrals in the arguments of the natural logarithm functions can be expressed in terms of complementary error functions:

$$G_E\left(s,\kappa,q_0,q\right)=\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\left(f_{out}\ln\left(\operatorname{erfc}\left(\frac{-s+D_{out}^{-}\sqrt{\kappa}-x\sqrt{2q}}{\sqrt{2\left(q_0-q\right)}}\right)\right)+(1-f_{out})\ln\left(\operatorname{erfc}\left(\frac{s+D_{out}^{+}\sqrt{\kappa}-x\sqrt{2q}}{\sqrt{2\left(q_0-q\right)}}\right)\right)\right)-\ln 2$$

$$G_S\left(\hat{k},\hat{s},\hat{\kappa},\hat{q}_0,\hat{q}\right)=\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\ln\left(e^{-i\hat{\kappa}\beta_{int}^2}\frac{\sqrt{\pi}e^{\frac{\left(i\hat{k}-i\hat{s}g_jA_j-i\hat{\kappa}\beta_{syn}A_j+2x\sqrt{-iC_j\hat{q}}\right)^2}{4i\left(C_j\left(\hat{q}_0-\hat{q}\right)+B_j\hat{\kappa}\right)}}}{2\sqrt{i\left(C_j\left(\hat{q}_0-\hat{q}\right)+B_j\hat{\kappa}\right)}}\operatorname{erfc}\left(-\frac{i\hat{k}-i\hat{s}g_jA_j-i\hat{\kappa}\beta_{syn}A_j+2x\sqrt{-iC_j\hat{q}}}{2\sqrt{i\left(C_j\left(\hat{q}_0-\hat{q}\right)+B_j\hat{\kappa}\right)}}\right)\right)$$

(S18)

The following substitutions transform the replica symmetric saddle point coordinates into the real domain, $u_{\pm}=\frac{D_{out}^{\pm}\sqrt{\kappa}\pm s}{\sqrt{2q}}$, $\varepsilon=\frac{q_0-q}{q}\geq 0$, $t=-i\hat{q}$, $z=\frac{i\hat{s}}{2\sqrt{-i\hat{q}}}$, $\delta=-\frac{\hat{q}_0-\hat{q}}{\hat{q}}$, $\tau=\frac{\hat{\kappa}}{-\hat{q}}$ $\eta=-\frac{i\hat{k}}{\sqrt{-i\hat{q}}}$, where they can be obtained by solving the following problem:

$$G_E\left(u_+,u_-,\varepsilon\right)=\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\left(f_{out}\ln\left(\operatorname{erfc}\left(\frac{u_--x}{\sqrt{\varepsilon}}\right)\right)+(1-f_{out})\ln\left(\operatorname{erfc}\left(\frac{u_+-x}{\sqrt{\varepsilon}}\right)\right)\right)-\ln 2$$

$$G_S\left(\eta,t,\tau,z,\delta\right)=\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\ln\left(e^{-\tau t\beta_{int}^2}\frac{\sqrt{\pi}e^{\left(\frac{\eta+2zg_jA_j+\tau\sqrt{t}\beta_{syn}A_j-2x\sqrt{C_j}}{2\sqrt{C_j\delta+B_j\tau}}\right)^2}}{2\sqrt{C_j\delta t+B_j\tau t}}\operatorname{erfc}\left(\frac{\eta+2zg_jA_j+\tau\sqrt{t}\beta_{syn}A_j-2x\sqrt{C_j}}{2\sqrt{C_j\delta+B_j\tau}}\right)\right)$$

$$\frac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_+}-\frac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_-}=0;\quad \frac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial \varepsilon}=\frac{t}{\alpha}\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2\kappa}{2\left(u_-+u_+\right)^2};\quad \frac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_+}=-\frac{2t\tau\kappa}{\alpha}\frac{1}{\left(u_-+u_+\right)};$$

$$\frac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_+}=\frac{t}{\alpha}\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2\kappa}{\left(u_-+u_+\right)^3}\left(\delta-\varepsilon\right);\quad \frac{\partial G_S\left(\eta,t,\tau,z,\delta\right)}{\partial \eta}=-\tilde{w}\sqrt{t};\quad \frac{\partial G_S\left(\eta,t,\tau,z,\delta\right)}{\partial z}=-2\sqrt{t};$$

$$\frac{\partial G_S\left(\eta,t,\tau,z,\delta\right)}{\partial \tau}=-t\kappa;\quad \frac{\partial G_S\left(\eta,t,\tau,z,\delta\right)}{\partial \delta}=-t\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2\kappa}{2\left(u_-+u_+\right)^2}$$

$$\frac{\partial G_S\left(\eta,t,\tau,z,\delta\right)}{\partial t}=\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2\kappa\left(\varepsilon-\delta\right)}{2\left(u_-+u_+\right)^2}-\frac{\tilde{w}\eta}{2\sqrt{t}}-\frac{z}{\sqrt{t}}-\kappa\tau;\quad \kappa\geq 0;\quad \varepsilon\geq 0;\quad u_++u_-\geq 0$$

(S19)

Replica theory solution at critical capacity

With increasing number of associations m, $\Omega_{typical}$ shrinks and approaches zero at the maximum (critical) capacity of the neuron, $\alpha_c = \frac{m_c}{N}$. In this limit $(q_0 - q)$ goes to zero as well, and the saddle point equations Eqs. (S19) can be expanded asymptotically in terms of $1/\varepsilon$ and $1/\delta$. In the leading order these equations yield:

$$
\begin{aligned}
&(1-f_{out})F(u_+) - f_{out}F(u_-) = 0 \\
&f_{out}D(u_-) + (1-f_{out})D(u_+) = \frac{\varepsilon^2 t\kappa}{\alpha_c}\frac{\left(D_{out}^+ + D_{out}^-\right)^2}{(u_- + u_+)^2} \\
&(1-f_{out})F(u_+) = \frac{2t\tau\varepsilon\kappa}{\alpha_c}\frac{1}{(u_- + u_+)} \\
&(1-f_{out})F(u_+) = -\frac{\varepsilon t\kappa(\delta-\varepsilon)}{\alpha_c}\frac{\left(D_{out}^+ + D_{out}^-\right)^2}{(u_- + u_+)^3} \\
&\frac{1}{N}\sum_{j=1}^{N}\frac{\sqrt{C_j}}{C_j\delta + B_j\tau}F\left(-\frac{\eta + 2zg_jA_j + \tau\sqrt{t}\beta_{syn}A_j}{2\sqrt{C_j}}\right) = 2\tilde{w}\sqrt{t} \\
&\frac{1}{N}\sum_{j=1}^{N}\frac{A_jg_j\sqrt{C_j}}{C_j\delta + B_j\tau}F\left(-\frac{\eta + 2zg_jA_j + \tau\sqrt{t}\beta_{syn}A_j}{2\sqrt{C_j}}\right) = 2\sqrt{t} \\
&\frac{1}{2N}\sum_{j=1}^{N}\frac{B_jC_j}{\left(C_j\delta + B_j\tau\right)^2}D\left(-\frac{\eta + 2zg_jA_j + \tau\sqrt{t}\beta_{syn}A_j}{2\sqrt{C_j}}\right) + t\left(\beta_{int}^2 - \kappa\right) = -\frac{1}{2N}\sum_{j=1}^{N}\frac{\sqrt{C_j}\sqrt{t}\beta_{syn}A_j}{C_j\delta + B_j\tau}F\left(-\frac{\eta + 2zg_jA_j + \tau\sqrt{t}\beta_{syn}A_j}{2\sqrt{C_j}}\right) \\
&\frac{1}{N}\sum_{j=1}^{N}\frac{C_j^2}{\left(C_j\delta + B_j\tau\right)^2}D\left(-\frac{\eta + 2zg_jA_j + \tau\sqrt{t}\beta_{syn}A_j}{2\sqrt{C_j}}\right) = \frac{\left(D_{out}^+ + D_{out}^-\right)^2 t\kappa}{(u_- + u_+)^2} \\
&\frac{1}{4N}\sum_{j=1}^{N}\frac{\tau\beta_{syn}A_j\sqrt{C_j}}{\sqrt{t}\left(C_j\delta + B_j\tau\right)}F\left(-\frac{\eta + 2zg_jA_j + \tau\sqrt{t}\beta_{syn}A_j}{2\sqrt{C_j}}\right) + \tau\left(\beta_{int}^2 - \kappa\right) = -\frac{\left(D_{out}^+ + D_{out}^-\right)^2\kappa(\varepsilon-\delta)}{2(u_- + u_+)^2} + \frac{\tilde{w}\eta + 2z}{2\sqrt{t}} \\
&\kappa \geq 0; \quad \varepsilon \geq 0; \quad u_+ + u_- \geq 0
\end{aligned}
$$

(S20)

Special functions E, F, and D, in Eqs. (S20) are defined as follows:

$$
\begin{aligned}
E(x) &= \frac{1}{2}\left(1+\operatorname{erf}(x)\right) \\
F(x) &= \frac{1}{\sqrt{\pi}} e^{-x^2} + x\left(1+\operatorname{erf}(x)\right) \\
D(x) &= xF(x) + E(x)
\end{aligned}
\tag{S21}
$$

After replacing $\tau\sqrt{t}$ with y, and eliminating variables, ε, t, κ, τ, and δ, we arrive at the final system of six equations and one inequality. This system contains six latent variables $u_{\pm}$, x, η, y, and z which determine the critical capacity of the neuron, α_c:

$$
\begin{cases}
(1-f_{out})F(u_+) - f_{out}F(u_-) = 0 \\
x = \dfrac{4(u_- + u_+)}{\left(D_{out}^+ + D_{out}^-\right)^2} \dfrac{f_{out}E(u_-) + (1-f_{out})E(u_+)}{f_{out}F(u_-) + (1-f_{out})F(u_+)} \\
\dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \dfrac{\sqrt{C_j}}{C_j x + B_j} F\left(-\dfrac{\eta + 2zA_j g_j + y\beta_{syn}A_j}{2\sqrt{C_j}}\right) = 2\tilde{w}y \\
\dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \dfrac{A_j g_j \sqrt{C_j}}{C_j x + B_j} F\left(-\dfrac{\eta + 2zA_j g_j + y\beta_{syn}A_j}{2\sqrt{C_j}}\right) = 2y \\
\dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \dfrac{C_j}{\left(C_j x + B_j\right)^2}\left(\dfrac{B_j}{2} - \dfrac{(u_- + u_+)^2}{\left(D_{out}^+ + D_{out}^-\right)^2} C_j\right) D\left(-\dfrac{\eta + 2zA_j g_j + y\beta_{syn}A_j}{2\sqrt{C_j}}\right) = \beta_{int}^2 y^2 - y\tilde{w}\eta - 2yz \\
\dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \dfrac{\beta_{syn}A_j\sqrt{C_j}}{C_j x + B_j} F\left(-\dfrac{\eta + 2zA_j g_j + y\beta_{syn}A_j}{2\sqrt{C_j}}\right) = 2\tilde{w}\eta + 4z - 4\beta_{int}^2 y \\
u_+ + u_- \geq 0
\end{cases}
$$

$$
\alpha_c = x^2 \frac{f_{out}D(u_-) + (1-f_{out})D(u_+)}{\left(f_{out}E(u_-) + (1-f_{out})E(u_+)\right)^2} \frac{1}{N}\sum_{j=1}^{N} \frac{C_j^2}{\left(C_j x + B_j\right)^2} D\left(-\frac{\eta + 2zA_j g_j + y\beta_{syn}A_j}{2\sqrt{C_j}}\right)
$$

(S22)

These equations were solved in MATLAB to produce the results shown in Figure 4 of the main text (network of heterogeneous neurons), and the code is available at https://github.com/neurogeometry/Associative_Learning_with_Noise

We note that Eqs. (S22) are consistent with the solution described in the Supplementary Material of Brunel et al. [10], who solved a similar problem by minimizing the probability of postsynaptic spiking errors for a given intrinsic noise level. However, the model of Brunel et al. does not account for the homeostatic constraints and learning by inhibitory inputs, nor does it consider the complete error propagation cycle by omitting synaptic noise and presynaptic spiking errors.

Distribution of input weights at critical capacity

Connection probabilities, P^{con}, and probability densities for non-zero input weights, p^{PSP}, at critical capacity can be calculated as previously described [10-12]. The result depends on the set of latent variables of Eqs. (S22):

$$
\begin{aligned}
P_i^{con} &= E\left(-\frac{\eta + 2zA_j g_j + y\beta_{syn} A_j}{2\sqrt{C_j}}\right) \\
p_i^{PSP}\left(\tilde{J}\right) &= \frac{\theta\left(g_i \tilde{J}\right)}{\sqrt{2\pi}\sigma_i \tilde{w} E\left(-\frac{\eta + 2zA_j g_j + y\beta_{syn} A_j}{2\sqrt{C_j}}\right)} e^{-\left(\frac{\tilde{J}}{\sqrt{2}\sigma_i \tilde{w}} + \frac{\eta + 2zA_i g_i + y\beta_{syn} A_j}{2\sqrt{C_i}} g_i\right)^2} \\
\sigma_i &= \frac{\sqrt{C_i}}{\sqrt{2}\tilde{w} y\left(C_i x + B_i\right)}
\end{aligned}
\tag{S23}
$$

For a given input, i, there is a finite probability that the connection weight is zero, while the probability density of non-zero connection weights is a truncated Gaussian with a standard deviation $\sigma_i \tilde{w}$.

Solution in the case of balanced spiking errors

Here, we consider a scenario in which the expected numbers of erroneous spikes and spike failures are equal, and, as a result, these spiking errors do not affect the input and output firing probabilities $\{f_j\}$ and f_{out}:

$$\begin{aligned} p_j^+\left(1-f_j\right) &= p_j^- f_j, \quad j=1,\ldots,N \\ p_{out}^+\left(1-f_{out}\right) &= p_{out}^- f_{out} \end{aligned} \tag{S24}$$

In this case, it is more convenient to express the results in terms of spiking error probabilities:

$$\begin{aligned} r_j &\equiv p_j^+\left(1-f_j\right)+p_j^- f_j \\ r_{out} &\equiv p_{out}^+\left(1-f_{out}\right)+p_{out}^- f_{out} \end{aligned} \tag{S25}$$

Eqs. (S22, S23) remain unchanged, but the coefficients defined in Eqs. (S12) transform into:

$$\begin{aligned} A_j &= f_j \\ B_j &= r_j\left(1-\frac{r_j}{4f_j\left(1-f_j\right)}\right) \\ C_j &= f_j\left(1-f_j\right)\left(1-\frac{r_j}{2f_j\left(1-f_j\right)}\right)^2 \\ D_{out}^+ &= \sqrt{2}\,\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{\left(1-f_{out}\right)}\right) \\ D_{out}^- &= \sqrt{2}\,\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right) \end{aligned} \tag{S26}$$

Figure 4 of the main text shows the solutions of Eqs. (S22, S23) for log-normal distributions of spiking error probabilities and firing probabilities. The contribution of noisy inhibitory and excitatory inputs to the neuron's postsynaptic potential is suppressed, as such inputs have lower strengths and connection probabilities (Figures 4B, C).

Solution in the case of two homogeneous classes of inputs and balanced spiking errors

In this case, there are N_{inh} inhibitory inputs and $(N - N_{inh})$ excitatory inputs, and all inputs and the output have the same firing probability, f, and the same spiking error probability, r. Eqs. (S22, S23) simplify significantly after the introduction of new variables, $v_{\pm} = \frac{-\eta \pm 2zf - y\beta_{syn} f}{2\sqrt{C}}$:

$$\begin{cases} (1-f)F(u_+) - fF(u_-) = 0 \\ \frac{N_{inh}}{N} F(v_+) + \frac{N_{exc}}{N} F(v_-) = \frac{\sqrt{2}}{\sigma} \\ -\frac{N_{inh}}{N} F(v_+) + \frac{N_{exc}}{N} F(v_-) = \frac{\sqrt{2}}{\sigma \tilde{w} f} \\ \left((u_- + u_+)^2 - 2\xi\right)\left(\frac{N_{inh}}{N} D(v_+) + \frac{N_{exc}}{N} D(v_-)\right) = \frac{2\beta_{post}^2 \zeta^2}{\sigma^2} \\ \sigma = \frac{\sqrt{2}\beta_{post}^2 \zeta^2}{\left((u_- + u_+)\frac{fE(u_-) + (1-f)E(u_+)}{fF(u_-) + (1-f)F(u_+)} + \xi\right)\left(\frac{1}{\tilde{w} f}(v_+ - v_-) - (v_+ + v_-)\right)} \\ u_+ + u_- > 0 \end{cases}$$

$$\alpha_c = \frac{\sigma^2 (u_- + u_+)^2}{2\beta_{post}^4 \zeta^4} \frac{fD(u_-) + (1-f)D(u_+)}{\left(fF(u_-) + (1-f)F(u_+)\right)^2} \left(\frac{1}{\tilde{w} f}(v_+ - v_-) - (v_+ + v_-)\right)^2 \left(\frac{N_{inh}}{N} D(v_+) + \frac{N_{exc}}{N} D(v_-)\right)$$

$$P_{inh/exc}^{con} = E(v_{\pm})$$

$$p_{inh/exc}^{PSP}(\tilde{J}) = \frac{\theta(\mp \tilde{J})}{\sqrt{2\pi}\sigma \tilde{w} E(v_{\pm})} e^{-\left(\frac{\tilde{J}}{\sqrt{2}\sigma \tilde{w}} \pm v_{\pm}\right)^2}$$

(S27)

Intrinsic and synaptic noise in Eqs. (S27) is entirely contained within the parameter β_{post}, while spiking error probabilities appear only in ξ and ζ:

$$
\begin{aligned}
\beta_{post} &= \sqrt{\beta_{int}^2 + \tilde{w}\beta_{syn} f} \\
\xi &= \frac{r\left(4f\left(1-f\right)-r\right)}{2\left(2f\left(1-f\right)-r\right)^2}\left(\mathrm{erf}^{-1}\left(1-\frac{r}{1-f}\right)+\mathrm{erf}^{-1}\left(1-\frac{r}{f}\right)\right)^2 \\
\zeta &= \frac{\sqrt{2f\left(1-f\right)}}{\tilde{w}\left(2f\left(1-f\right)-r\right)}\left(\mathrm{erf}^{-1}\left(1-\frac{r}{1-f}\right)+\mathrm{erf}^{-1}\left(1-\frac{r}{f}\right)\right)
\end{aligned}
\tag{S28}
$$

We note that due to the imposed constraints on spiking error probabilities, $p_{j,out}^{+} < f_{j,out}$ and $p_{j,out}^{-} < 1 - f_{j,out}$, combined with the definition of r in Eqs. (S25), $r < 2f\left(1-f\right)$, and parameters ξ and ζ are positive. In the main text, parameter β_{post} is referred to as the postsynaptic noise strength. Results shown in Figures 1-3 of the main text are expressed in terms of β_{post} and r.

Numerical solution of the model for finite *N*

For a finite number of inputs, solution to the problem outlined in Eqs. (S5) can be obtained numerically. To that end, we make the problem feasible by introducing a slack variable, $s^{\mu} \geq 0$, for every association, and choose the solution that minimizes the sum of these variables:

$$
\begin{aligned}
&\underset{\{\tilde{J}_j\}}{\arg\min}\left(\sum_{\mu=1}^{m} s^{\mu}\right) \\
&\left(2y^{\mu}-1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j\left[\left(1-p_j^{-}\right)X_j^{\mu}+p_j^{+}\left(1-X_j^{\mu}\right)\right]-1\right) \geq -s^{\mu} + \frac{D_{out}^{-}y^{\mu}+D_{out}^{+}\left(1-y^{\mu}\right)}{\sqrt{N}} \times \\
&\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_j^2\left(p_j^{-}\left(1-p_j^{-}\right)X_j^{\mu}+p_j^{+}\left(1-p_j^{+}\right)\left(1-X_j^{\mu}\right)\right)+\beta_{syn}\tilde{J}_j g_j\left(\left(1-p_j^{-}\right)X_j^{\mu}+p_j^{+}\left(1-X_j^{\mu}\right)\right)+\beta_{int}^2\right]\right)^{\frac{1}{2}}, \quad \mu = 1,\ldots,m \\
&\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j g_j = \tilde{w} \\
&\tilde{J}_j g_j \geq 0, \quad j = 1,\ldots,N \\
&s^{\mu} \geq 0, \quad \mu = 1,\ldots,m
\end{aligned}
\tag{S29}
$$

Eqs. (S29) were solved in MATLAB by using the *fmincon.m* function. This code is available at https://github.com/neurogeometry/Associative_Learning_with_Noise

Memory storage and recall in an associative network of model neurons

We consider a McCulloch and Pitts neural network [1] of N_{exc} excitatory and (N - N_{exc}) inhibitory neurons. The network state at time μ is described by a binary vector of neural activities, X^{μ}. The network is loaded with temporal sequences of network states, $X^1 \to X^2 \to \ldots X^{m+1}$, by modifying the weights of connections between neurons $\{J_{ij}\}$ (weight of connection from neuron j to neuron i). As described above, there are sign and homeostatic constraints on connection weights of individual neurons. To learn a temporal sequence, individual neurons (i) must independently associate inputs they receive from the network, X^{μ}, with the corresponding outputs derived from the associative sequence, $X_i^{\mu+1}$. As before, learning in the model is accompanied with spiking errors, or errors in X^{μ}, synaptic noise, or noise in J, and intrinsic noise, or noise in the neuron's threshold of firing, h, and we use asterisks to denote quantities containing errors. We assume that neurons in the network have identical firing thresholds, h, average absolute connection weights, w, synaptic and intrinsic noise strengths, β_{syn} and β_{int}, but may have different firing probabilities, f_i, and spiking error probabilities, r_i. This network model of associative memory storage in the presence of errors and noise can be summarized as follows:

$$
\begin{aligned}
&\theta\left(\sum_{j=1}^{N} J_{ij}^{*} X_j^{*\mu} - h^{*}\right) = X_i^{*\mu+1}; \quad \mu = 1,\ldots,m, \quad i = 1,\ldots,N \\
&\frac{1}{N}\sum_{j=1}^{N} J_{ij} g_j = w; \quad i = 1,\ldots,N \\
&J_{ij} g_j \geq 0; \quad i,j = 1,\ldots,N \\
&P\left(X_i^{\mu} = 1\right) \equiv f_i; \quad P\left(X_i^{*\mu} = 0 \mid X_i^{\mu} = 1\right) = \frac{r_i}{2 f_i}; \quad P\left(X_i^{*\mu} = 1 \mid X_i^{\mu} = 0\right) = \frac{r_i}{2\left(1 - f_i\right)} \\
&\left\langle J_{ij}^{*} \right\rangle \equiv J_{ij}; \quad \mathrm{var}\left(J_{ij}^{*}\right) = \frac{h \beta_{syn}}{N}\left|J_{ij}\right| \\
&\left\langle h^{*} \right\rangle \equiv h; \quad \mathrm{var}\left(h^{*}\right) = \frac{h^2 \beta_{int}^2}{N}
\end{aligned}
\tag{S30}
$$

To load a given memory sequence into the network, Eqs. (S30) were solved independently for every neuron in the network as describe in the previous section. To retrieve a loaded memory sequence, the network is initialized at the beginning of the loaded sequence, X^1, and at each subsequent step of memory playout synaptic and intrinsic noise are added independently to every connection and every neuron in the network. The sequence is said to be retrieved successfully if the network states during the retrieval do not deviate substantially from the loaded states (Figure S1 A). In practice, there is no need to precisely define the threshold amount of deviation. This is because for large networks the Hamming distance between the loaded and retrieved sequences, normalized by N, either remains within $\sim 2r$ or diverges to $\sim 2f(1-f)$, which is much greater in the feasible parameter region considered in the main text (Figure S1 B). Figure S1 C shows the probability of retrieving the entire stored sequence and the average retrieved fraction of the stored sequence length for different values of postsynaptic noise. The transition from successful retrieval to retrieval failure is relatively sharp. It illustrates that the stored sequences can be successfully retrieved at (or slightly below) the post synaptic noise level present during learning.

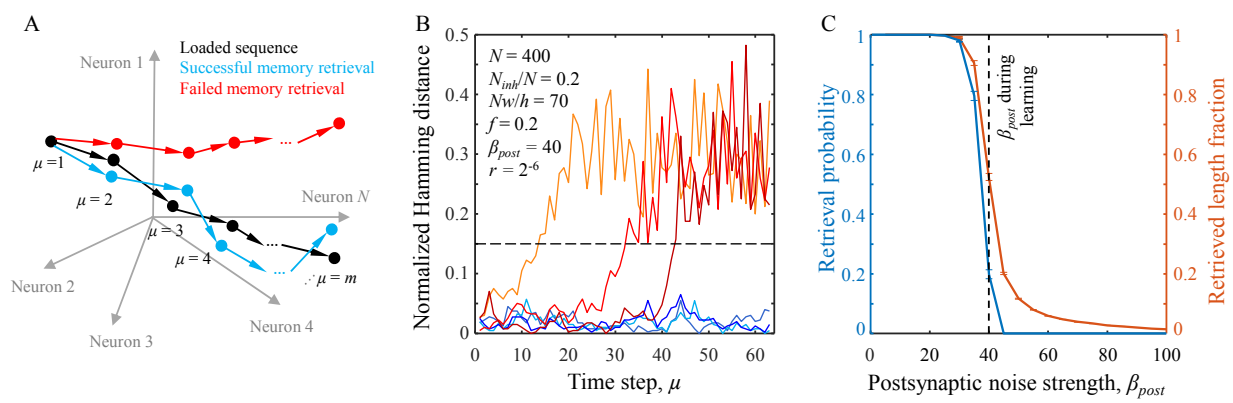

Figure S1: Recall of memory sequences stored in a network of model neurons. **A.** Illustration of memory playout during successful (blue) and failed (red) retrieval. The loaded memory sequence is shown in black. **B.** Normalized Hamming distances between the loaded and retrieved sequences as functions of time step. Successfully retrieved sequences do not deviate from the loaded sequences by more than a threshold amount (dashed line). **C.** Probability of successful memory retrieval (left y-axis) and retrieved length fraction (right y-axis) as functions of postsynaptic noise strength during retrieval. The dashed line indicates the postsynaptic noise strength during learning.

REFERENCES

[1] W. McCulloch and W. Pitts, Bull Math Biol **5**, 115 (1943).

[2] D. O. Hebb, *The organization of behavior; a neuropsychological theory* (Wiley, New York,, 1949).

[3] F. Rosenblatt, Cornell Aeronautical Laboratory **Report 85-460-1** (1957).

[4] M. L. Minsky and S. Papert, *Perceptrons : an introduction to computational geometry* (MIT Press, Cambridge, Mass., 1988), Expanded edn.

[5] T. M. Cover, IEEE Trans. EC **14**, 326 (1965).

[6] E. Gardner, J. Phys. A: Math. Gen. **21**, 257 (1988).

[7] E. Gardner and B. Derrida, J. Phys. A: Math. Gen. **21**, 271 (1988).

[8] J. J. Hopfield, Proc Natl Acad Sci U S A **79**, 2554 (1982).

[9] H. M. Kohler and D. Widmaier, J. Phys. A: Math. Gen. **24**, L495 (1991).

[10] N. Brunel, V. Hakim, P. Isope, J. P. Nadal, and B. Barbour, Neuron **43**, 745 (2004).

[11] J. Chapeton, T. Fares, D. LaSota, and A. Stepanyants, Proc Natl Acad Sci U S A **109**, E3614 (2012).

[12] J. Chapeton, R. Gala, and A. Stepanyants, Front Comput Neurosci **9**, 74 (2015).

[13] C. Allen and C. F. Stevens, Proc Natl Acad Sci U S A **91**, 10380 (1994).

[14] G. B. Ermentrout, R. F. Galan, and N. N. Urban, Trends Neurosci **31**, 428 (2008).

[15] A. A. Faisal, L. P. Selen, and D. M. Wolpert, Nature reviews. Neuroscience **9**, 292 (2008).

[16] Y. Yarom and J. Hounsgaard, Physiological reviews **91**, 917 (2011).

[17] A. Stepanyants and D. B. Chklovskii, Trends Neurosci **28**, 387 (2005).

[18] N. Brunel, Nature neuroscience **19**, 749 (2016).

[19] R. Rubin, L. F. Abbott, and H. Sompolinsky, Proc Natl Acad Sci U S A **114**, E9366 (2017).

[20] J. Del Castillo and B. Katz, The Journal of physiology **124**, 560 (1954).

[21] S. F. Edwards and P. W. Anderson, J. Phys. F: Metal Phys. **5**, 965 (1975).

[22] D. Sherrington and S. Kirkpatrick, Physical Review Letters **35**, 1792 (1975).

[23] D. Zhang, C. Zhang, and A. Stepanyants, bioRxiv, 320432 (2018).

[24] M. N. Shadlen and W. T. Newsome, The Journal of neuroscience : the official journal of the Society for Neuroscience **18**, 3870 (1998).

[25] C. F. Stevens and A. M. Zador, Nature neuroscience **1**, 210 (1998).

[26] G. R. Holt, W. R. Softky, C. Koch, and R. J. Douglas, Journal of neurophysiology **75**, 1806 (1996).

[27] W. R. Softky and C. Koch, The Journal of neuroscience : the official journal of the Society for Neuroscience **13**, 334 (1993).

[28] G. T. Buracas, A. M. Zador, M. R. DeWeese, and T. D. Albright, Neuron **20**, 959 (1998).

[29] M. R. Cohen and A. Kohn, Nature neuroscience **14**, 811 (2011).

[30] C. van Vreeswijk and H. Sompolinsky, Neural computation **10**, 1321 (1998).

[31] C. van Vreeswijk and H. Sompolinsky, Science **274**, 1724 (1996).

[32] M. Graupner and A. D. Reyes, The Journal of neuroscience : the official journal of the Society for Neuroscience **33**, 15075 (2013).

[33] M. Okun and I. Lampl, Nature neuroscience **11**, 535 (2008).

[34] Z. F. Mainen and T. J. Sejnowski, Science **268**, 1503 (1995).

[35] S. Lefort, C. Tomm, J. C. Floyd Sarria, and C. C. Petersen, Neuron **61**, 301 (2009).

[36] X. Jiang, S. Shen, C. R. Cadwell, P. Berens, F. Sinz, A. S. Ecker, S. Patel, and A. S. Tolias, Science **350**, aac9462 (2015).