Robust associative learning is sufficient to explain structural and dynamical properties of local cortical circuits

Danke Zhang, Chi Zhang, and Armen Stepanyants

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

ABSTRACT

The ability of neural networks to associate successive states of network activity lies at the basis of many cognitive functions. Hence, we hypothesized that many ubiquitous structural and dynamical properties of local cortical networks result from associative learning. To test this hypothesis, we trained recurrent networks of excitatory and inhibitory neurons on memory sequences of varying lengths and compared network properties to those observed experimentally. We show that when the network is robustly loaded with near-maximum amount of associations it can support, it develops properties that are consistent with the observed probabilities of excitatory and inhibitory connections, shapes of connection weight distributions, overrepresentations of specific 3-neuron motifs, distributions of connection numbers in clusters of 3-8 neurons, sustained, irregular, and asynchronous firing activity, and balance of excitation and inhibition. What is more, memories loaded into the network can be retrieved even in the presence of noise comparable to the baseline variations in the postsynaptic potential. Confluence of these results suggests that many structural and dynamical properties of local cortical networks are simply a byproduct of associative learning.

INTRODUCTION

With ever-increasing amounts of data on structure and dynamics of neural circuits, one fundamental question moves into focus: Is there an overarching principle that can account for the multitude of these seemingly unrelated experimental observations? For example, much is known about local connectivity in cortical circuits (see e.g. Supplementary Dataset). Some of the most salient connectivity features can be described based on the classes of excitatory and inhibitory neurons. At the level of pair-wise connectivity, it is known that the probabilities of excitatory connections are generally lower than those for inhibitory. Specifically, the majority of reported probabilities lies in the 0.10-0.19 range (interquartile range based on Supplementary Dataset) if the presynaptic cell is excitatory and in the 0.25-0.56 range for connections originating from inhibitory neurons. It is also known that the distributions of connection weights have stereotypic shapes with the majority of measured coefficients of variation (CV) in the 0.85-1.1 range for excitatory connections and slightly lower values for inhibitory, 0.78-0.96. At the level of connectivity within 3-neuron clusters, several ubiquitously overrepresented connectivity motifs have been discovered [1-4]. Information becomes scarce as one considers larger clusters of neurons, but even here deviations from random connectivity have been reported for clusters of 3-8 neurons [2]. Similarly, many universal features characterize activity of neurons in local cortical networks. For example, individual neurons exhibit highly irregular spiking activity, resembling Poisson processes with close to one CV in inter-spike-intervals [5-9]. Spike trains of nearby neurons are only marginally correlated, 0.04-0.15 [10], and, at the network level, spiking activity can be described as sustained, irregular, and asynchronous.

Two popular models of binary McCulloch and Pitts neuron networks [11] can individually explain some of the above experimental observations. The first model is based on the idea that to have a sustained and irregular activity, excitatory and inhibitory inputs to individual neurons in the network must be balanced [12-17]. This model assumes that excitatory and inhibitory inputs are much larger than the threshold of firing, but their sum lies below the threshold, and firing is driven by fluctuations. The balanced model can produce realistic sustained and irregular spiking activity, however, by taking network connectivity as an input, it generally does not make predictions related to the network structure (but see [18]). The second model, which we will refer to as the associative model, is based on the idea that synaptic connectivity is a product of associative learning [19-23]. This

model can explain many features of local cortical connectivity, but it does not necessarily produce sustained and irregular activity. We show that there is a biologically plausible regime, in which balanced and associative models converge. Therefore, with a single framework one can explain both structural and dynamical properties of cortical circuits and show that these properties emerge as a result of associative learning.

We pursue the idea that associative learning alone is sufficient to explain the above described properties of connectivity. With sensory information continuously impinging on the brain, neural circuits function in a state of perpetual change, recording some of the information in the form of long-term memories. In this process, individual neurons may be operating as independent learning units, constrained by functional and metabolic considerations, such as the requirement to store associative memories, tolerate noise during memory retrieval, and maintain low cost of the underlying connections [24], Figure 1A. In this study, we explore structural and dynamical properties of associative networks in the space of these constraints, and show that there is a unique region of parameters that can explain the above-described experimental observations. In this region, the network is loaded with close to maximum amount of associative memories which can be successfully retrieved even in the presence of significant amount of noise.

RESULTS

Network model of associative learning

We use a McCulloch and Pitts neural network [11] to model a local cortical circuit in which N_{inh} inhibitory neurons and $(N - N_{inh})$ excitatory neurons are all-to-all potentially connected [25,26] (Figure 1B). Associative memories, in the form of temporal sequences of network states, are loaded in the network by modifying the weights of connections between neurons (see Online Methods and *SI* for details). In this process, individual neurons attempt to associate the inputs they receive from the network with predefined outputs (Figure 1C). Several biologically motivated constraints are imposed on the learning process [24]. First, firing thresholds of neurons, h, do not change during learning. Second, the signs of input weights, J, that are determined by the excitatory or inhibitory identities of presynaptic neurons, do not change during learning (Dale's principle) [27]. Third, input

connections of each neuron are homeostatically constrained to have a fixed average absolute weight, w [28-30]. Fourth, each neuron must be able to retrieve the loaded associations even in the presence of noise in its postsynaptic potential. The maximum amount of noise a neuron has to tolerate is referred to as robustness parameter, κ.

Individual neurons in the model learn to associate the presented sequences of network states, and the probability of successfully learning the entire sequence decreases with the sequence length, m, or memory load $\alpha = m/N$ (Figure 1D). Memory load that can be successfully learned with the probability of 0.5 is termed the associative memory storage capacity of the neuron, α_c. This capacity increases with the number of neurons in the network, N, and saturates in the $N \to \infty$ limit at a value that can be determined with the replica theory [20] (see *SI*). Notably, this theoretical solution shows that in the high-weight regime $(Nwf/h \gg 1)$, the neuron's capacity, as well as the shape of its connection weight distribution, depend on the combination of model parameters in the form of $\rho = \frac{\kappa}{w\sqrt{Nf(1-f)}}$, where f is the neuron's firing rate (Figures 1A and S1). The meaning of this combination was elucidated by Brunel et al. [21], who pointed out that ρ can be viewed as a measure of reliability of stored associations to errors in the input. We would like add that ρ can also be viewed as a proxy for the ratio of robustness to fluctuations in postsynaptic potential. Following [22], we will refer to ρ as the rescaled robustness.

Motivated by this theoretical insight we set out to explore the possibility that local networks in the brain function in the high-weight regime. The average absolute connection weight, w, was previously estimated based on experimental data from various cortical systems [24], and the result shows that Nwf/h lies in the range of 4 - 38 (95% confidence interval) with the average of 14. A similar estimate based on the granule to Purkinje cell connectivity in rat cerebellum [21] also results in a relatively high value of this parameter, $Nwf/h \approx 150,000 \times 0.1\,\mathrm{mV} \times 0.0044/10\,\mathrm{mV} = 6.6$. Therefore, high-weight regime may be a general attribute of local circuits, and we will show that this assumption is consistent with a large number of experimental measurements related to network structure and dynamics.

Figure 1E shows that the memory storage capacity of a single neuron is a decreasing function of rescaled robustness. This is expected, as increase in ρ can be thought of as an increase in the strength of the constraint on learning (robustness, κ), or as a decrease in the amount of available resources (absolute connection weight, w). Figure 1E also illustrates that at $N = 800$ single neuron capacity is already sufficiently close to its $N \to \infty$ limit, and therefore, network properties are not expected to change substantially if N is increased beyond 800 (Figure S2).

Properties of neuron-to-neuron connectivity in associative networks

Next, we examined the properties of neuron-to-neuron connectivity in associative networks at different values of memory load and rescaled robustness. One of the most prominent features of connectivity is that substantial fractions of excitatory and inhibitory connections have zero weights, and therefore, connection probabilities are less than one (Figure 2A). The distributions of non-zero connection weights resemble the general shapes of unitary postsynaptic potential (uPSP) distributions, with a notable difference in the frequencies of strong connections. The former have Gaussian or exponential tails [21,24], while the tails of uPSP distributions are often much heavier [1,31]. Several amendments to the associative model have been proposed to correct this discrepancy [21,23]. Here, we would like to point out that heavy tails of experimental distributions can be reproduced within the associative model by considering networks of neurons with inhomogeneous properties, e.g. different values of w or κ.

To compare connection probabilities and widths of non-zero connection weight distributions in associative networks with those reported experimentally, we compiled measurements of connection probabilities and CVs in uPSPs reported for local cortical circuits (Supplementary Dataset). Figures 2B, C show that the average inhibitory connection probability (based on 38 studies, 9,522 connections tested) is significantly higher ($p < 10^{-10}$, two sample t-test) than the average probability for excitatory connections (67 studies, 63,020 connections tested), while CV of inhibitory uPSPs (10 studies, 503 connections recorded) is slightly lower than that for excitatory (36 studies, 3,956 connections recorded). Similar trends are observed in associative networks. Figures 2D, E show that connectivity in associative networks is sparse, with probabilities of

excitatory non-zero connections lower than those for inhibitory connections in the entire considered range of memory load and rescaled robustness. Probabilities of both connection types are decreasing with increasing ρ. This is expected because an increase in ρ can be achieved by lowering w, which is equivalent to reducing the amount of resources needed to make connections. Isocontours in Figures 2D, E demarcate the interquartile ranges of connection probability measurements shown in Figure 2B. There is a region in the α–ρ space of parameters in which both excitatory and inhibitory connection probabilities are in general agreement with the experimental data. Also, consistent with the experimental measurements, CVs of excitatory weights in associative networks are slightly larger than those for inhibitory weights (Figures 2F, G), and there is a wide region in the α–ρ space of parameters in which these values match the experimental data shown in Figure 2C.

Higher order structural properties of associative networks

In addition to specific properties of neuron-to-neuron connectivity, local cortical circuits are known to have non-random patterns of connections in subnetworks of three and more neurons [1-4]. To determine whether associative networks can reproduce some of the known features of higher-order connectivity, we first examined the statistics of connectivity motifs within subnetworks of three excitatory neurons. There are 13 distinct types of connected 3-neuron motifs (Figure 3A). Under/over-expressions of these motif types were quantified with the normalized z-scores, which range from -1 to 1 and are negative/positive for motifs that appear less/more frequently than what is expected by chance (see Methods). Profiles of normalized z-scores in associative networks were compared with data from the Blue Brain project [4]. To gauge the extent of similarity of these profiles we calculated Root-Mean-Square (RMS) difference in the normalized z-scores for various values of load and rescaled robustness (Figure 3B). The results show that there is a region of parameters in which associative networks produce profiles similar to the Blue Brain project data. In particular, motifs 3, 4, 10, and 11 appear to be ubiquitously overexpressed in the region of parameters outlined by the isocontour in Figure 3B, which is in agreement with the Blue Brain project data (see e.g. green line in Figure 3A) and data from other experiments [1,2,4].

Deviations from random connectivity have also been detected in subnetworks of 3 to 8 excitatory neurons by comparing distributions of observed connection numbers with those based on randomly shuffled connectivity [2]. This comparison revealed that the experimental distributions have heavier tails, which is indicative of clustered connectivity (black lines in Figures 3C and S5). This trend was first reproduced by Brunel [22] who considered an associative network of excitatory neurons at capacity. Our model shows that there is a single region of parameters α and ρ in which qualitative agreement is obtained simultaneously for all subnetworks from 3 to 8 neurons (Figures 3D and S5).

Dynamical properties of associative networks

Dynamics in associative networks depends strongly on the values of memory load and rescaled robustness. At small values of ρ, network dynamics quickly terminates at a fixed point in which all neurons are silent (Figure 4A1, red). When ρ is high, associative networks can have long-lasting intrinsic activity, often ending up in a limit cycle of non-zero length. To quantify this behavior, we measured the average number of steps taken by the network to reach a limit cycle or a fixed point from a random initial state (Figure 4A2). The results show that the duration of transient dynamics increases exponentially with memory load and rescaled robustness. Even for moderate values of these parameters, the average length of transient activity can be of the order of network size, N (contour in Figure 4A2).

Individual neurons in associative networks are capable of producing irregular spiking activity, degree of which can be quantified with CV of inter-spike-intervals (ISI) (Figure 4B). According to this measure, neurons exhibit greater irregular activity when memory load and rescaled robustness are high, with CV of ISI values saturating at around 0.9. This is consistent with the range of CV of ISI values reported for different cortical systems, 0.7-1.1 [5-9]. To examine the extent of synchrony in neuron activity we calculated spike train cross-correlation coefficients for pairs of neurons (Figure 4C). The results show that increase in ρ leads to a more asynchronous activity, which can be explained by the reduction in connection probability (Figures 2D, E) and, consequently, reduction in the amount of common input to the neurons. For $\rho > 2.5$, the values of

cross-correlation are consistent with experimental data 0.04-0.15 [10] (interquartile range derived from 26 studies).

Irregular, asynchronous activity can result from balance of excitation and inhibition [12,13]. In the balanced state, the magnitudes of excitatory and inhibitory potentials are typically much greater than the threshold of firing, and, due to a high degree of correlation in these potentials, firing is driven by fluctuations. Consistent with this, the average excitatory and inhibitory inputs in the associative model are much greater than the firing threshold and are tightly anti-correlated (Figures 4D and S7). The degree of anti-correlation decreases with rescaled robustness as the network connectivity becomes sparser. Experimentally, it is difficult to measure anti-correlations of excitatory and inhibitory inputs within a given cell, but such measurements have been performed in nearby cells. The resulting anti-correlations, ~0.4 [32,33], are somewhat below the values observed in associative networks. However, this is expected, as between-cell anti-correlations are likely to be weaker than within-cell anti-correlations.

Cortical circuits are loaded with associative memories close to capacity and can tolerate noise comparable to the baseline variations in postsynaptic input during memory retrieval

Parameter regions described in Figures 2-4, lead to structural and dynamical properties consistent with the experimental observations and have a non-empty intersection. In this biologically plausible region of parameters, associative networks behave qualitatively similar to local cortical circuits. Figure 5A shows the intersection of parameter regions (green dashed line) for the excitatory and inhibitory connection probabilities (red), 3-neuron motifs (green), connections in 3-8 neuron clusters (blue), and duration of transient activity (cyan). The remaining features, i.e. CV of connection weights, CV of ISI, spike cross-correlation coefficient, and excitatory-inhibitory balance, are not shown in Figure 5A both to avoid clutter and because they do not impose additional restrictions on the intersection region. In the biologically plausible region of parameters, individual neurons are loaded with relatively long sequences of network states ($0.2N$ for the green asterisk in Figure 5A), yet it is not clear if the associations learned by individual neurons assemble into memory sequences that can be successfully retrieved at the network level.

To examine this question we first tested memory retrieval in the absence of noise. For this, we initialized the network state at the beginning of the loaded sequence and monitored playout of the memory. The sequence is said to be retrieved successfully if the network states during the retrieval do not deviate substantially from the loaded states (Figure 5B). In practice, there is no need to precisely define the threshold amount of deviation. This is because for large networks, e.g. N = 800, the Hamming distance between the loaded and retrieved sequences either remains within ~ $\sqrt{N}$ or diverges to ~N. Figure 5C shows the probability of successful memory retrieval as function of memory load and rescaled robustness. The transition from successful memory retrieval to inability to retrieve the entire loaded sequence is relatively sharp, making it possible to define network capacity, analogously to single neuron capacity, as the sequence length for which the success rate in memory retrieval equals 0.5 (blue line in Figure 5C). Network capacity deviates from single neuron capacity, but this difference is expected to decrease with network size. Interestingly, the biologically plausible region of parameters is centered on the single neuron capacity curve, implying that individual cortical neurons are loaded with associations close to their capacity. What is more, the biologically plausible region of parameters lies almost entirely below the network capacity curve, indicating that loaded memory sequences can be retrieved with high probability in the absence of noise.

To assess the degree of robustness of the memory retrieval process, we monitored memory playout in the presence of postsynaptic noise. In this experiment, random Gaussian noise of zero mean and standard deviation σ_{noise} was added independently to all neurons at every step of the retrieval process. Network tolerance to noise was defined as σ_{noise} value that results in the memory retrieval probability of 0.5. Figure 5D shows the map of noise tolerance normalized by the baseline variations in postsynaptic input during memory retrieval, $\sigma_{noise}/\sigma_{input}$ (see *SI*). The latter represents standard deviation in postsynaptic input in the absence of noise. We note that the biologically plausible region identified on the basis of structural and dynamical properties of cortical networks (green contour in Figure 5D) has a non-zero overlap with the area in which memory retrieval is robust to noise. In this domain the network can tolerate high noise-to-input ratios (up to 0.5), which serves as an independent validation of the associative model in terms of the hypothesized network function.

DISCUSSION

Our results suggest that local circuits of the mammalian brain operate in a high-weight regime in which individual neurons are loaded with associative memories close to their capacity (Figure 5C) and the network can tolerate relatively large amounts of postsynaptic noise during memory retrieval. In this regime, many structural and dynamical properties of associative networks are in general agreement with experimental measurements from various species and brain regions. It is important to point out that, due to large uncertainties in the reported measurements, we did not attempt to quantitatively fit the associative model to the data. The uncertainties originate from natural variability of network features across individuals, brain areas, and species, and are confounded by experimental biases and measurement errors. Instead, we rely on a large body of qualitative evidence to support our conclusions. These evidence include (1) sparse connectivity, with probability of excitatory connections being lower than that for inhibitory connections, (2) distributions of non-zero connection weights with CVs of excitatory and inhibitory weights being close to 1, (3) overrepresentations of specific 3-neuron motifs, (4) distributions of connection numbers in subnetworks of 3-8 neurons showing clustering behavior, (5) sustained, irregular, and asynchronous firing activity with close to 1 CV of ISI and small positive cross-correlation in neuron activity, and (6) balance of excitatory and inhibitory postsynaptic potentials. Many of these features have been separately reported in various formulations of the associative model [20-24]. Here, we show that with a single set of model parameters it is possible to account for these features collectively. What is more, the identified set of model parameters overlaps with the region in which loaded memories can be successfully recalled even in the presence of postsynaptic noise (Figure 5D), providing an independent functional validation of the theory.

Several discrepancies between the results of the associative model and experiment are worth mentioning. First, the model does not produce overexpression of bidirectional connections observed in some experiments [1,34,35]. This can be amended by including point attractors [22] in addition to temporal sequences of network states considered in this study. However, since not all experiments report overexpression of bidirectional connections, this feature may be area specific and/or dependent on the distance between neurons [4]. Second, associative networks did not produce a good agreement with the distribution of inhibitory 3-neuron motifs reported by the Blue Brain project [4]. We believe that this discrepancy can be attributed to a large diversity of inhibitory neuron

population, which is not captured by the presented homogeneous model. Third, our theory does not produce long-tailed distributions of connection weights observed in many experiments [1,31]. Several ways to amend this discrepancy have been previously discussed [21,23]. It is also clear that introducing inhomogeneity in neuron parameters can broaden the tail of the connection weight distribution.

Because local cortical circuits function in the high-weight regime, $Nwf \gg h$, the average excitatory and inhibitory postsynaptic inputs are significantly greater than the threshold of firing (Figure S7). In the identified region of memory load and rescaled robustness, e.g. for the green asterisk in Figure 5, these potentials in magnitude exceed the threshold of firing by factors of 6.3 and 7.8 respectively (Table S1). In this regime, excitatory and inhibitory potentials are strongly anti-correlated (Figure 4D2), which is reminiscent of the balanced state described by many authors [16,32,33,36-39]. We note, however, that there is a difference in how balance of excitatory and inhibitory potentials is realized in the associative vs. balanced networks. The difference originates from scaling of synaptic weight with network size. In associative networks, synaptic weight is inversely proportional to *N*, while in balanced networks inverse proportionality to $\sqrt{N}$ is assumed. In the former model, the average excitatory and inhibitory postsynaptic inputs to a neuron remain unchanged as the network size increases, and balance is the consequence of the high-weight regime, while in the latter model balance emerges with increasing *N* as postsynaptic potentials diverge, which is unsettling. On the other hand, Rubin et al. [18] argue that due to the above scaling difference, synaptic connections in the associative model are weaker, and the network is unstable to large, $O(h)$, noise arising from processes within neurons (e.g. threshold fluctuations). We agree that susceptibility of associative networks to this type of noise is a concern for infinitely large systems. However, there is no biological data on scaling of noise with network size, and having $O(h)$ noise may be unrealistic. More importantly, since local brain networks are finite, robustness to this type of noise can always be achieved by increasing *w*, i.e. in the high-weight regime. For example, an associative network of $N = 800$ neurons, configured at the green asterisk in Figure 5, can tolerate CVs in threshold fluctuations of up 1.1 (see Figure S7). Aside from the issue of robustness to $O(h)$ noise, we show that in the high-weight regime results of the balanced and associative models become independent of the details of scaling, and converge to the same solution (see *SI* text and

Figure S1). Therefore, associative learning in both models will lead to networks with identical structural and dynamical properties.

METHODS

Associative model

We consider an all-to-all potentially connected neural network of N_{inh} inhibitory and $(N - N_{\text{inh}})$ excitatory McCulloch and Pitts neurons [11] involved in an associative learning task, Figure 1B. Biological motivations and assumptions associated with this model have been previously described [24]. Here we only give a concise description of the model, as a more detailed account is provided in *SI*. Neurons in the model may belong to various classes, defined by their excitatory or inhibitory nature, characteristic firing probabilities, homeostatic constraints, and robustness to noise (defined below). The state of the network at time step μ is described by a vector of binary (0 or 1) activities of all neurons, X^{μ}. The network is loaded with a predefined temporal sequence (or a basin) of m+1 network states, $X^{1} \rightarrow X^{2} \rightarrow \ldots X^{m+1}$, and individual neurons, independently from one another, attempt to robustly learn to associate inputs and outputs derived from this sequence, Figure 1C. We assume that the network states to be learned are uncorrelated across neurons and time. Learning in the network is mediated by changing neuron connection weights, $\{J_{ij}\}$ (connection from neuron j to neuron i), in the presence of several biologically inspired constraints. (1) Input connection weights of each neuron are sign-constrained to be non-negative if the presynaptic neuron is excitatory and non-positive if it is inhibitory. (2) Input weights of each neuron are homeostatically constrained to have a predefined l_1-norm. (3) Each neuron must attempt to learn its associations robustly, so that they can be recalled correctly in the presence of a given level of postsynaptic noise. The model can be summarized as follows (see Figure 1C):

$$
\begin{aligned}
&\theta\left(\sum_{j=1}^{N} J_{ij} X_j^{\mu} - h_i + \eta_i\right) = X_i^{\mu+1}; \quad i = 1,\cdots,N; \quad \mu = 1,\cdots,m \\
&J_{ij} g_j \geq 0; \quad j = 1,\cdots,N \\
&\frac{1}{N}\sum_{j=1}^{N} \left|J_{ij}\right| = w_i \\
&\left|\eta_i\right| \leq \kappa_i \\
&\mathrm{Prob}\left(X_i^{\mu}\right) = \begin{cases} 1 - f_i, & X_i^{\mu} = 0 \\ f_i, & X_i^{\mu} = 1 \end{cases}
\end{aligned}
\tag{1}
$$

In these expressions, θ denotes the Heaviside step-function, h_i is firing threshold, and η_i denotes postsynaptic noise which is bounded by the robustness parameter κ_i, i.e. $|\eta_i| < \kappa_i$. To enforce sign-constraints on connection weights we introduce parameter g_j, which equals 1 if the presynaptic neuron j is excitatory and –1 if it is inhibitory. Parameter w_i, referred to as the average absolute connection weight, is introduced to impose the l_1-norm constraint on the neuron's input connection weights. Binary neuron states, X_i^{μ}, are randomly drawn from neuron-class dependent Bernoulli probability distributions: 0 with probability $1 - f_i$ and 1 with probability f_i.

Given a set of associations, the problem outlined in Eqs. (1) may be feasible and have multiple solutions $\{J_{ij}\}$, or non-feasible, in which case no network can satisfy all the constraints of the problem. In the first case, the solutions region is nonempty, and we must employ additional considerations to limit the results to a single, unique solution. We did this by choosing the solution that minimizes the l_2-norm of input connection weights of every neuron. In the non-feasible case, similar to what is done in the formulation of the Support Vector Machine problem [40], for every neuron, we minimized the sum of deviations between the not robustly learned associations and their corresponding margin boundaries (see *SI* for details).

The above model is governed by the following parameters (Table S1): number of neurons in the network, N, fraction of inhibitory neurons, N_{inh}/N, firing probabilities of neurons in the associative sequence $\{f_i\}$, robustness parameters of neurons $\{\kappa_i\}$, their average absolute connection weights, $\{w_i\}$, and the memory load, $\alpha = m/N$. All numerical simulations presented in the main text were performed a homogeneous associative model, in which all excitatory and inhibitory neurons have the same firing threshold, h, average firing rate, f, constraints, w and κ, and memory load, α. We

set the fraction of inhibitory neurons to $N_{inh}/N = 0.2$, the firing rate to $f = 0.2$, and $Nwf/h = 14$, which is similar to what was described previously [24]. We confirmed that the results are not sensitive to small changes in these parameters. The network size was chosen to be $N = 800$. In *SI* we reproduce the results for networks of $N = 200$, $N = 400$, and $N \to \infty$, to illustrate that our conclusions do not depend on the exact value of N (Figures S2-4 and S6).

Numerical and theoretical solutions of the associative model

Because individual neurons in the network learn independently of one another, the problem of sequence learning by a network can be solved individually for each neuron. In addition, since the model outlined in Eqs. (1) is convex, solution for each neuron can be obtained numerically with the methods of convex optimization [41] (see *SI* for details). Briefly, sequences of random, binary, and independent network states were generated with firing probability f. Individual neurons in the network were trained separately on their corresponding input-output associations extracted from these sequences. Probabilities of successful learning for single neurons was calculated by presenting the network with associative sequences of varying lengths (100 times for each length). Memory capacity of a single neuron is defined as the length of sequences the neuron can learn with success probability of 0.5 (Figure 1D).

In the limiting case of large networks, $N \to \infty$, success probability abruptly changes from one to zero with increasing memory load, and neuron's associative capacity is referred to as critical. In this limit, results of the model at critical capacity were obtained with the replica theory [42,43] (see *SI* for details).

Network capacity is defined based on the probability of successfully retrieving loaded associative sequences in the absence of noise (Figure 5C).

Numerical simulations of structural and dynamical properties of associative networks

To examine properties of associative networks for different values of memory load and rescaled robustness, we trained 100 networks for every pair of these parameters, and used the resulting connection weight matrices to characterize network structure and dynamics.

Probability densities of connection weights (Figure 2A) and CV of connection weights (Figures 2F, G) were calculated after excluding small weights, i.e. weights with magnitudes less than $5h/N$. We visually confirmed that this threshold encompasses the central peak in the connection weight distribution, associated with small excitatory and inhibitory weights. We also confirmed that the network properties described in the main text are not sensitive to the exact value of this parameter in the $5h/N$ – $20h/N$ range. To analyze structural properties of associative networks (Figures 2D, E, and 3), weight matrices were converted into adjacency matrices by setting the small weights to zero and the remaining weights to 1. Numbers of 3-neuron motifs (Figures 3A,B) were calculated with the Brain Connectivity Toolbox [44]. Frequencies of 13 connected 3-neuron motifs, n_i, in subnetworks of excitatory neurons were compared with corresponding frequencies in subnetworks with randomly shuffled connections, $n_i^{shuffled}$. We used normalized z-scores, z_i^{norm}, as defined in [4], to characterize the degrees of over- and under-expression of motif types:

$$z_i = \left\langle \frac{n_i - \left\langle n_i^{shuffled} \right\rangle}{\mathrm{std}\left(n_i^{shuffled} \right)} \right\rangle; \quad z_i^{norm} = \frac{z_i}{\sqrt{\sum_{i=1}^{13} z_i^2}} \tag{2}$$

Here, outer angle brackets in the first equation denote averaging over 100 associative networks, angle brackets in the numerator represent averaging over a set of 50 randomly shuffled versions of a given associative network, and std denotes standard deviation over the set of shuffled networks. Normalized z-scores lie in the range of –1 to 1, and a positive z-score indicates that the observed number of motifs of a given type is larger than that expected by chance.

Dynamical properties of associative networks shown in Figure 4 were averaged over 100 networks and 100 random initial states for each network.

Data and code availability

The dataset of connection probabilities and strengths used in this study was compiled from articles published in peer-reviewed journals from year 1990 to 2016. Specifically, we targeted publications reporting connection probabilities between neurons and uPSP amplitudes in local circuits of the mammalian brain. Using Google Scholar queries based on the above criteria, we initially identified 152 articles describing a total of 856 projections. Later we limited our analyses to experiments in which recordings were made in the neocortex, from at least 10 pairs of neurons located in the same layer and separated laterally by less than 100 μm. We also limited the analyses to normal, juvenile or adult animals (no younger than P14 for mouse and rat, and older than that for ferret, cat, monkey, and human). After imposing these limits, the numbers of publications and projections reduced to 87 and 420 respectively. These projections are included in the Supplementary Dataset. Inhibitory and excitatory connection probabilities and CVs based on these projections are shown in Figures 2B, C, in which to reduce bias we averaged the results of individual studies reporting multiple projections of a given type.

MATLAB code for generating theoretical (replica) and numerical (convex optimization) solutions of the associative model is available at https://github.com/neurogeometry/AssociativeLearning.

ACKNOWLEDGEMENTS

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

REFERENCES

1 Song, S., Sjostrom, P. J., Reigl, M., Nelson, S. & Chklovskii, D. B. Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLoS Biol* **3**, e68 (2005).

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

3 Rieubland, S., Roth, A. & Hausser, M. Structured connectivity in cerebellar inhibitory networks. *Neuron* **81**, 913-929, doi:10.1016/j.neuron.2013.12.029 (2014).

4 Gal, E. *et al.* Rich cell-type-specific network topology in neocortical microcircuitry. *Nature neuroscience* **20**, 1004-1013, doi:10.1038/nn.4576 (2017).

5 Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **18**, 3870-3896 (1998).

6 Stevens, C. F. & Zador, A. M. Input synchrony and the irregular firing of cortical neurons. *Nature neuroscience* **1**, 210-217, doi:10.1038/659 (1998).

7 Softky, W. R. & Koch, C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **13**, 334-350 (1993).

8 Holt, G. R., Softky, W. R., Koch, C. & Douglas, R. J. Comparison of discharge variability in vitro and in vivo in cat visual cortex neurons. *Journal of neurophysiology* **75**, 1806-1814, doi:10.1152/jn.1996.75.5.1806 (1996).

9 Buracas, G. T., Zador, A. M., DeWeese, M. R. & Albright, T. D. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron* **20**, 959-969 (1998).

10 Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nature neuroscience* **14**, 811-819, doi:10.1038/nn.2842 (2011).

11 McCulloch, W. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biol* **5**, 115 - 133 (1943).

12 van Vreeswijk, C. & Sompolinsky, H. Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* **274**, 1724-1726 (1996).

13 van Vreeswijk, C. & Sompolinsky, H. Chaotic balanced state in a model of cortical circuits. *Neural Comput* **10**, 1321-1371 (1998).

14 Amit, D. J. & Brunel, N. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cereb Cortex* **7**, 237-252 (1997).

15 Brunel, N. Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J Comput Neurosci* **8**, 183-208 (2000).

16 Deneve, S. & Machens, C. K. Efficient codes and balanced networks. *Nature neuroscience* **19**, 375-382, doi:10.1038/nn.4243 (2016).

17 Renart, A. *et al.* The asynchronous state in cortical circuits. *Science* **327**, 587-590, doi:10.1126/science.1179850 (2010).

18 Rubin, R., Abbott, L. F. & Sompolinsky, H. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *Proc Natl Acad Sci U S A* **114**, E9366-E9375, doi:10.1073/pnas.1705841114 (2017).

19 Hebb, D. O. *The organization of behavior; a neuropsychological theory*. (Wiley, 1949).

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

21 Brunel, N., Hakim, V., Isope, P., Nadal, J. P. & Barbour, B. Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron* **43**, 745-757 (2004).

22 Brunel, N. Is cortical connectivity optimized for storing information? *Nature neuroscience* **19**, 749-755, doi:10.1038/nn.4286 (2016).

23 Chapeton, J., Fares, T., LaSota, D. & Stepanyants, A. Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons. *Proc Natl Acad Sci U S A* **109**, E3614-3622, doi:10.1073/pnas.1211467109 (2012).

24 Chapeton, J., Gala, R. & Stepanyants, A. Effects of homeostatic constraints on associative memory storage and synaptic connectivity of cortical circuits. *Front Comput Neurosci* **9**, 74, doi:10.3389/fncom.2015.00074 (2015).

25 Stepanyants, A. & Chklovskii, D. B. Neurogeometry and potential synaptic connectivity. *Trends Neurosci* **28**, 387-394, doi:S0166-2236(05)00131-1 [pii] 10.1016/j.tins.2005.05.006 (2005).

26 Stepanyants, A. *et al.* Local potential connectivity in cat primary visual cortex. *Cereb Cortex* **18**, 13-28, doi:bhm027 [pii] 10.1093/cercor/bhm027 (2008).

27 Dale, H. Pharmacology and nerve-endings. *Proceedings of the Royal Society of Medicine* **28**, 319-332 (1935).

28 Holtmaat, A., Wilbrecht, L., Knott, G. W., Welker, E. & Svoboda, K. Experience-dependent and cell-type-specific spine growth in the neocortex. *Nature* **441**, 979-983, doi:nature04783 [pii] 10.1038/nature04783 (2006).

29 Kim, S. K. & Nabekura, J. Rapid synaptic remodeling in the adult somatosensory cortex following peripheral nerve injury and its association with neuropathic pain. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **31**, 5477-5482, doi:31/14/5477 [pii] 10.1523/JNEUROSCI.0328-11.2011 (2011).

30 Bourne, J. N. & Harris, K. M. Coordination of size and number of excitatory and inhibitory synapses results in a balanced structural plasticity along mature hippocampal CA1 dendrites during LTP. *Hippocampus* **21**, 354-373, doi:10.1002/hipo.20768 (2011).

31 Lefort, S., Tomm, C., Floyd Sarria, J. C. & Petersen, C. C. The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex. *Neuron* **61**, 301-316, doi:S0896-6273(08)01092-1 [pii] 10.1016/j.neuron.2008.12.020 (2009).

32 Graupner, M. & Reyes, A. D. Synaptic input correlations leading to membrane potential decorrelation of spontaneous activity in cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33**, 15075-15085, doi:10.1523/JNEUROSCI.0347-13.2013 (2013).

33 Okun, M. & Lampl, I. Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nature neuroscience* **11**, 535-537, doi:10.1038/nn.2105 (2008).

34 Markram, H., Lübke, J., Frotscher, M., Roth, A. & Sakmann, B. Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex. *J Physiol* **500 (Pt 2)**, 409-440 (1997).

35 Wang, Y. *et al.* Heterogeneity in the pyramidal network of the medial prefrontal cortex. *Nature neuroscience* **9**, 534-542, doi:nn1670 [pii] 10.1038/nn1670 (2006).

36 Shu, Y., Hasenstaub, A. & McCormick, D. A. Turning on and off recurrent balanced cortical activity. *Nature* **423**, 288-293, doi:10.1038/nature01616 (2003).

37 Haider, B., Duque, A., Hasenstaub, A. R. & McCormick, D. A. Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **26**, 4535-4545, doi:10.1523/JNEUROSCI.5297-05.2006 (2006).

38 Xue, M., Atallah, B. V. & Scanziani, M. Equalizing excitation-inhibition ratios across visual cortical neurons. *Nature* **511**, 596-600, doi:10.1038/nature13321 (2014).

39 Wehr, M. & Zador, A. M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**, 442-446, doi:10.1038/nature02116 (2003).

40 Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning : data mining, inference, and prediction*. 2nd edn, (Springer, 2009).

41 Boyd, S. P. & Vandenberghe, L. *Convex optimization*. (Cambridge University Press, 2004).

42 Edwards, S. F. & Anderson, P. W. Theory of spin glasses. *J. Phys. F: Metal Phys.* **5**, 965-974 (1975).

43 Sherrington, D. & Kirkpatrick, S. Solvable model of a spin glass. *Physical Review Letters* **35**, 1792-1796 (1975).

44 Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *NeuroImage* **52**, 1059-1069, doi:10.1016/j.neuroimage.2009.10.003 (2010).

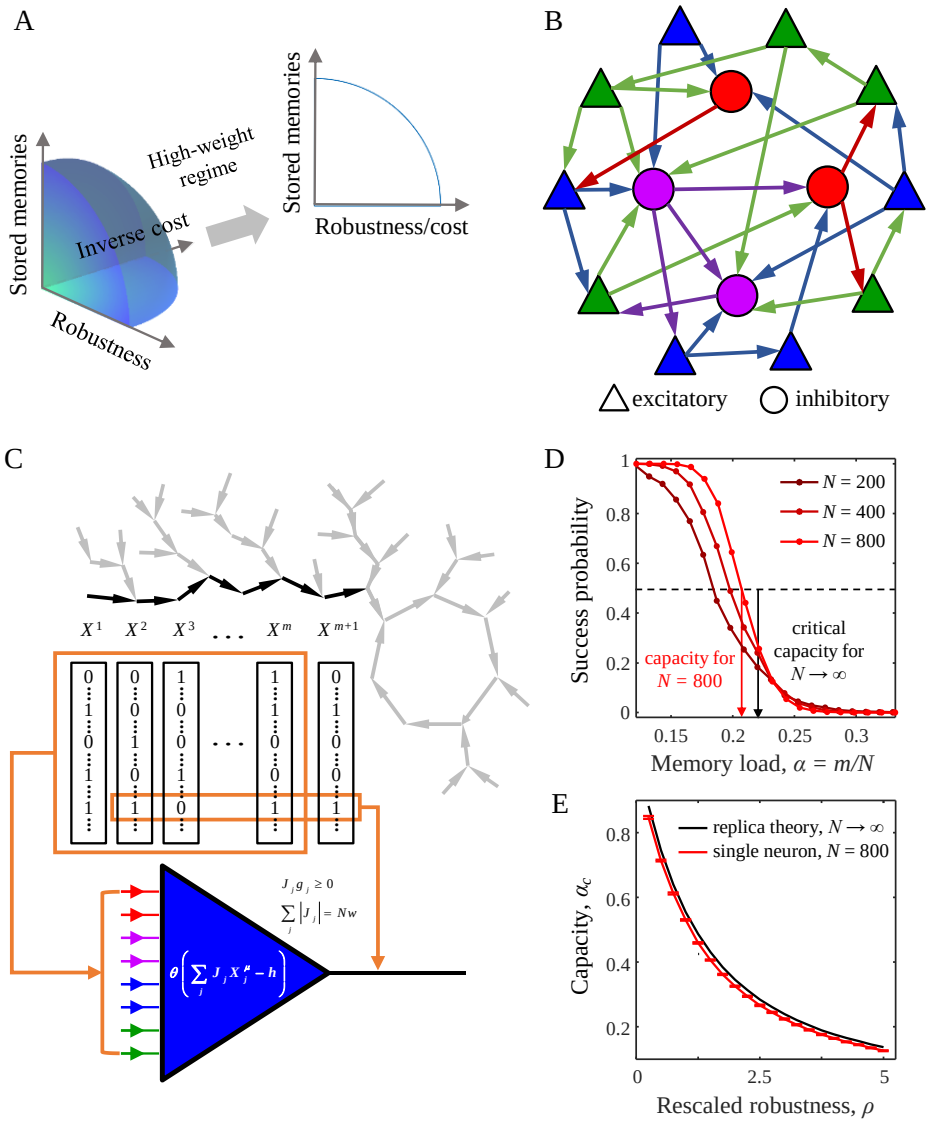

Figure 1: Associative memory storage in recurrent networks of excitatory and inhibitory neurons. **A.** Associative learning in the brain is expected to be constrained by functional and metabolic considerations, such as being able to store large amounts of memories, tolerate noise during memory retrieval, and have a low cost of the underlying connectivity. In the model, these three considerations are represented with memory load, α, robustness parameter, κ, and average absolute connection weight, w. We show that in the biologically plausible regime of high-weight, results of the model depend only on α and κ/w. **B.** Recurrent network of various classes (color) of all-to-all potentially connected excitatory and inhibitory neurons. Note that the arrows indicate actual (or functional) connections. **C.** Associative memory in the model is represented as a sequence (bold arrows), or an entire basin, of network states, $X^1 \rightarrow X^2 \ldots \rightarrow X^\mu$. Vector X^μ represents binary activities of individual neurons at time step μ. Each neuron in the network learns its corresponding set of input-output associations (orange boxes) by modifying the strengths of its input connections, J_j, under the constraints on the signs and l_1-norm of these connections. **D.** A neuron's ability to learn an entire sequence of presented associations decreases with the sequence length, m. Memory storage capacity of the neuron, α_c, (e.g. red arrow for $N = 800$) is defined as the fraction of associations, m/N, that can be learned with success probability of 50%. The transition from perfect learning to inability to learn the entire sequence sharpens with increasing N and approaches the result obtained with the replica theory in the limit of $N \rightarrow \infty$ (black arrow). **E.** Capacity of a single neuron is a decreasing functions of the rescaled robustness, ρ. Error-bars indicate standard deviations calculated based on 100 networks.

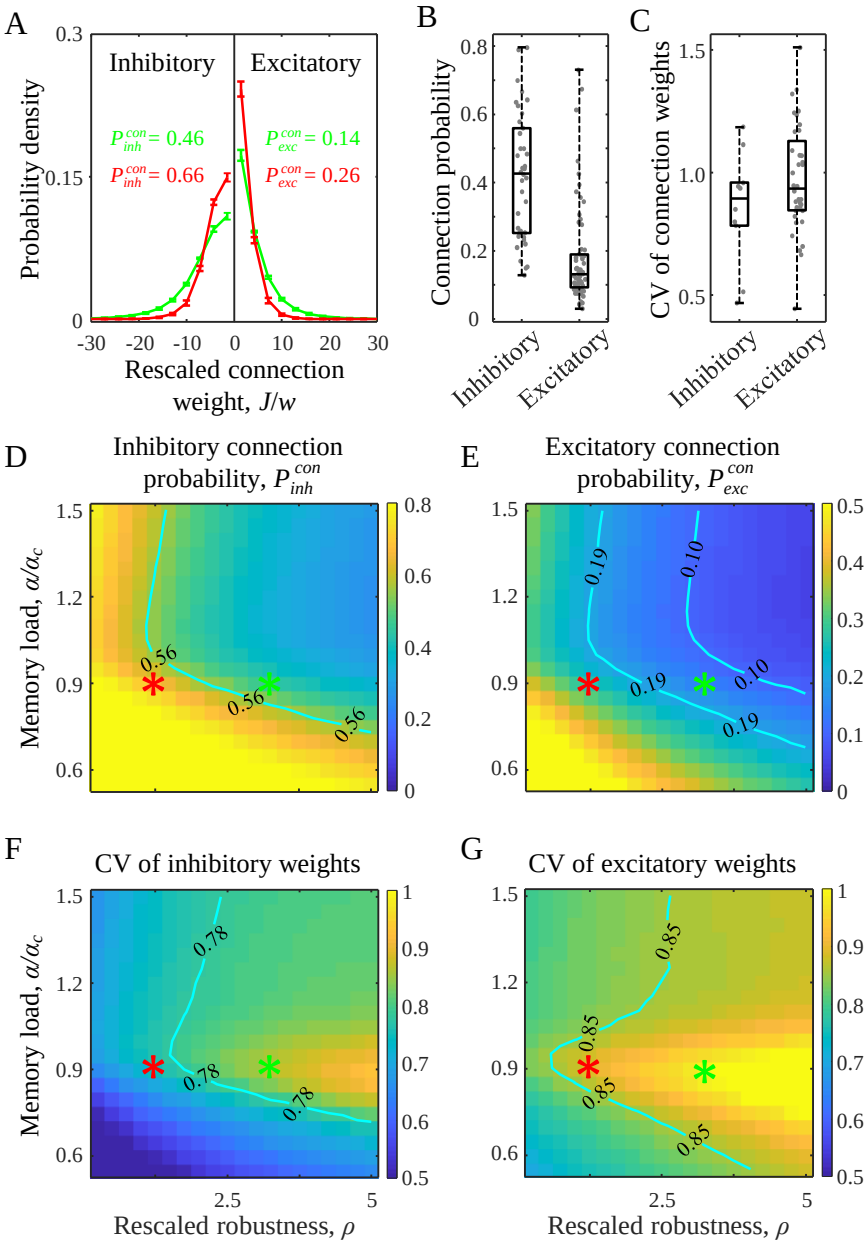

Figure 2: Properties of neuron-to-neuron connectivity in associative networks. **A.** Distributions of weights of inhibitory and excitatory connections for two parameter settings (red and green asterisks in D-G). Note that the distributions contain finite fractions of zero-weight connections. Error-bars indicate standard deviations (based on 100 networks). **B, C.** Connection probabilities and CVs of connection weights for inhibitory and excitatory connections reported in 87 studies describing 420 local cortical projections in mammals (Supplementary Dataset). Each dot represents the result of a single study averaged (with weights equal to the number of connections tested) over the number of reported projections. Maps of probabilities of inhibitory (**D**) and excitatory (**E**) connections as functions of rescaled robustness and relative memory load, i.e. load divided by the theoretical single neuron capacity at $N \to \infty$. Inhibitory connection probability is higher than probability of excitatory connections in the entire region of considered parameters. **F, G.** Maps of coefficients of variation (CV) of non-zero inhibitory and excitatory connection weights as functions of rescaled robustness and relative memory load. Isocontour lines in the maps correspond to the interquartile ranges of experimentally observed connection probabilities and CVs shown in (B) and (C). Numerical results were generated based on networks of $N = 800$ neurons.

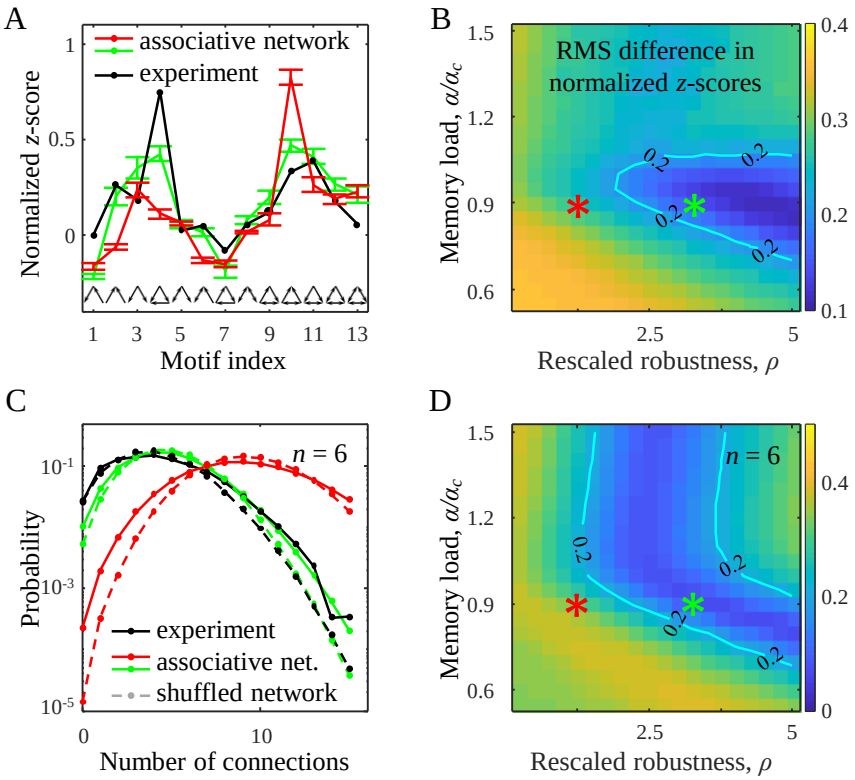

Figure 3: Higher-order structural motifs in associative networks. **A.** Normalized *z*-score of 13 connected 3-neuron motifs in excitatory subnetworks indicate over- and under-expressions of these structures in comparison to the chance level (see Methods for details). Red and green curves show results for the parameter settings specified by the red and green asterisks in (B). Black curve is the result from the Blue Brain project [4]. Error-bars indicate standard deviations (100 networks). **B.** Root-mean-square (RMS) difference between normalized *z*-scores obtained in associative networks and in the Blue Brain project. Isocontour line, demarcating a region of reasonably good solutions, is drawn as a guide to the eye. **C.** Distributions of non-zero connection numbers in clusters of 6 excitatory neurons in associative networks. Solid red and green lines illustrate distributions obtained in associative networks for the parameter settings indicated by the red and green asterisks. Solid black curves indicate the corresponding results for local cortical networks based on electrophysiological measurements [2]. Dashed lines show distributions in randomly shuffled networks (see Methods for details). **D.** Maps of l_2 distances between connection number distributions in associative and cortical networks [2]. Similar results for clusters of 3-8 neurons are shown in Figure S5. Numerical results of the associative model were generated based on networks of N = 800 neurons.

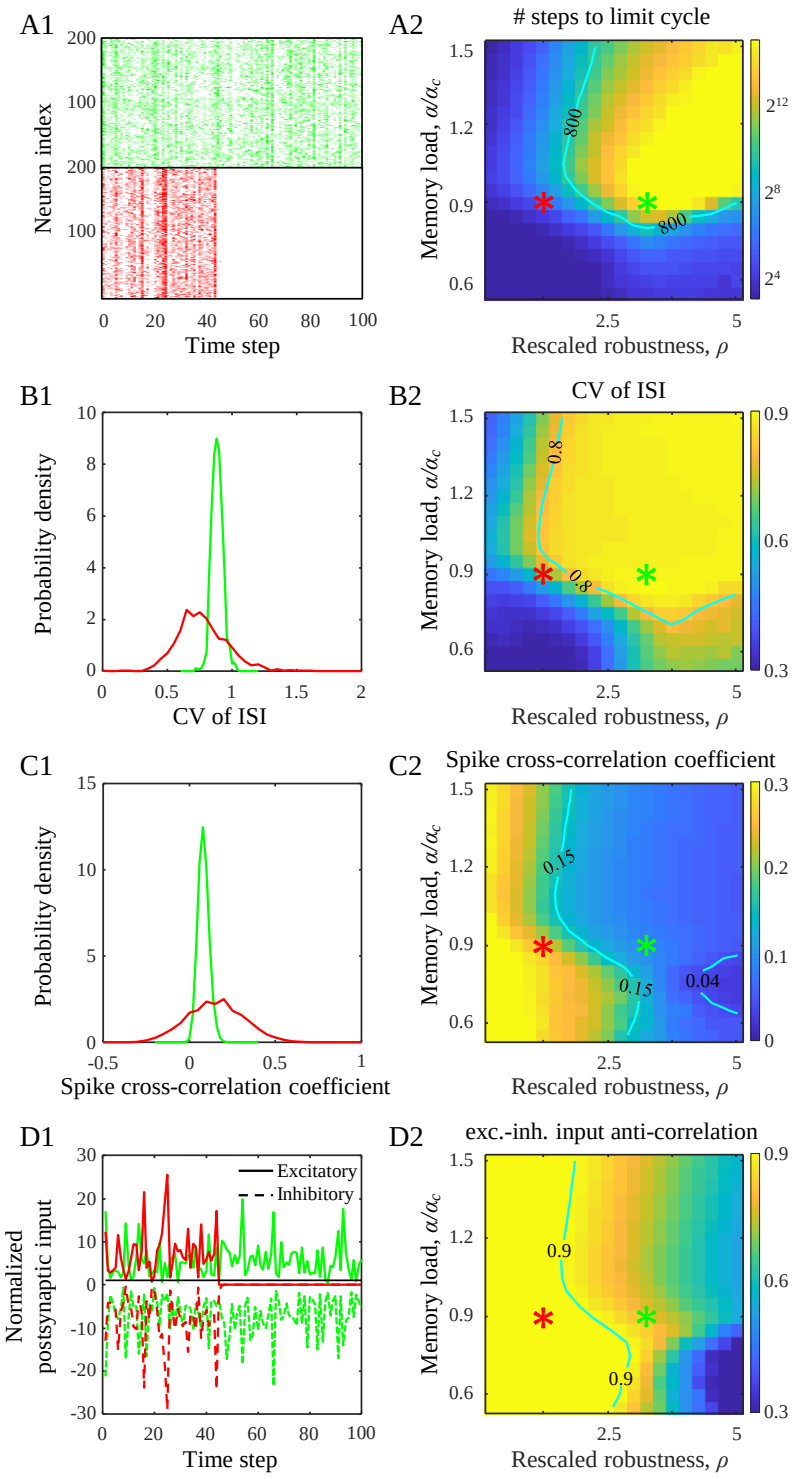

Figure 4: Dynamical properties of associative networks. **A1.** Two examples of spike rasters for associative networks parametrized as indicated with red and green asterisks in (A2). Dynamics at low values of rescaled robustness (red) quickly terminates at a quiescent state. **A2.** Map of the duration of transient dynamics as a function of rescaled robustness and relative memory load (see Methods for details). Note that at high levels of rescaled robustness and memory load, associative networks have long-lasting, transient activity. Isocontour line is drawn as a guide to the eye. **B1.** Distributions of CV in inter-spike-intervals (ISI) for the two parameter settings. Note that the average CV value increases with ρ. **B2.** Map of the average CV of ISI as a function of rescaled robustness and relative memory load. Isocontour line demarcates a region of high CV values that are in general agreement with experimental measurements. **C.** Same for cross-correlation coefficients of neuron spike trains. **D.** Same for anti-correlation coefficient of excitatory and inhibitory postsynaptic inputs received by a neuron. Note that the inputs are normalized by the firing threshold. For the selected parameter configurations, excitatory and inhibitory inputs are tightly balanced (large negative anti-correlation) despite large fluctuations. Maps in (A2, B2, C2, and D2) were generated based on networks of $N = 800$ neurons by averaging the results over 100 networks and 100 trials for each network and parameter setting.

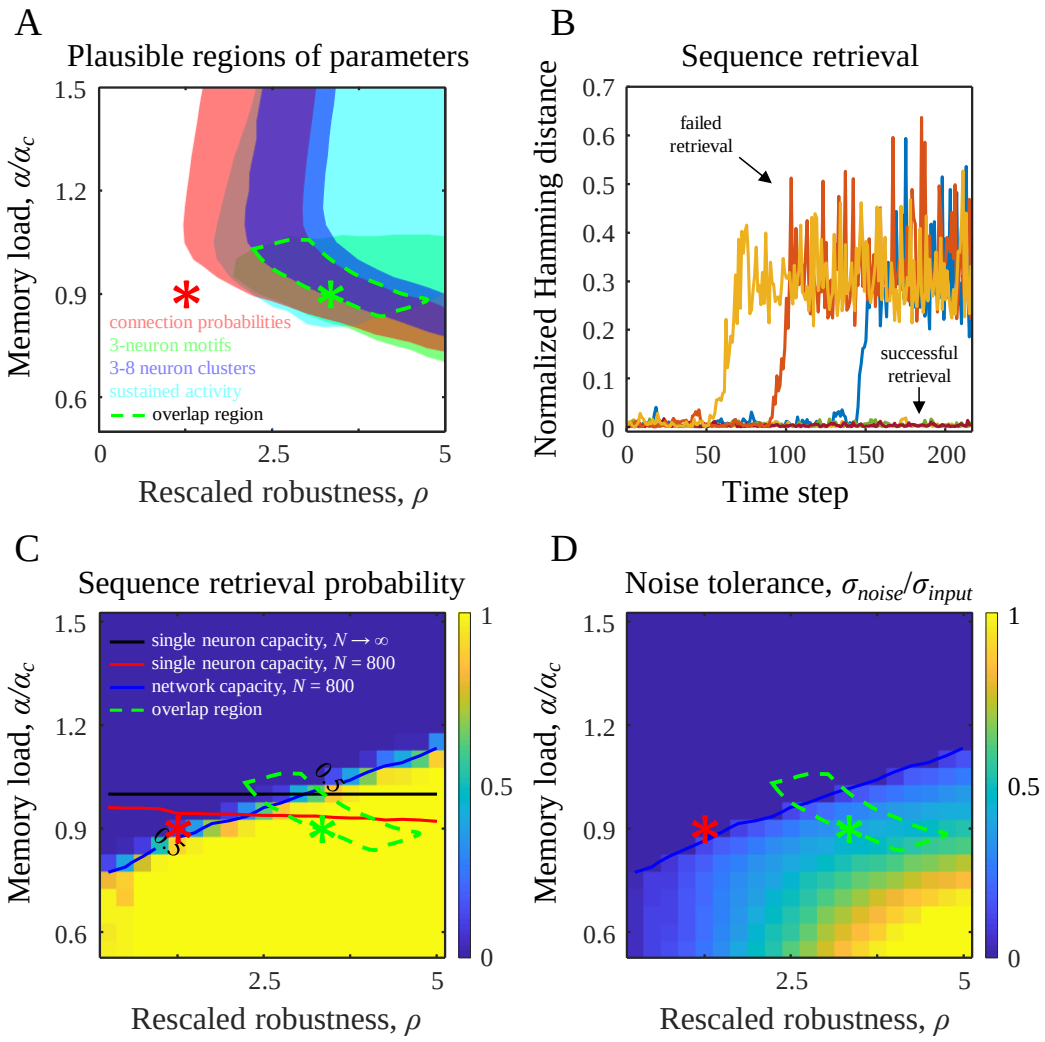

Figure 5: Values of rescaled robustness and memory load identified based on structural and dynamical properties of local cortical networks are consistent with the functional requirement of robust retrieval of stored memories. **A.** Region of parameters (dashed green line) that leads to a general agreement with the experimentally observed excitatory and inhibitory connection probabilities (red), excitatory 3-neuron motifs (green), 3-8 excitatory neuron clusters (blue), and sustained, irregular, asynchronous spiking activity (cyan). **B.** Retrieval of memory sequences in the absence of noise. The network is loaded with a memory sequence that it attempts to learn. Retrieval process is initialized at the start of the sequence, and deviations of subsequent network states from the loaded states are quantified with the Hamming distance normalized by the network size, N. The sequence is said to be successfully retrieved if the deviations are small. **C.** Success probability of sequence retrieval as a function of rescaled robustness and relative memory load. Blue line, corresponding to the success probability of 0.5, defines the network capacity. Single neuron critical capacity for $N \to \infty$ (black line) and capacity for $N = 800$ (red line) are shown for reference. Dashed green contour is the overlap region from (A). **D.** Noise level, σ_{noise}, that can be tolerated by the associative network during memory retrieval. Map shows the relative noise level, $\sigma_{noise}/\sigma_{input}$, corresponding to the retrieval probability of 0.5. Note that in the identified parameter region, the network can tolerate noise that is comparable to the standard deviation in postsynaptic input, σ_{input}.

SUPPLEMENTARY INFORMATION

This Supplementary Information describes the model of associative memory storage by a recurrent network considered in the main text. The model incorporates a number of constraints motivated by the experimental data on connectivity in the cerebral cortex. The model is solved theoretically with the replica method [1,2] in the limit of infinite network size, and numerically with methods of convex optimization [3] for large, but finite networks. The model gives rise to a comprehensive list of predictions regarding the structure and dynamics of neural networks. These predictions are consistent with a large number of experimental studies of connectivity in local cortical circuits. Other models, including only some of the constraints considered in this study, have been previously described [4-9].

General assumptions and approximations

This model is based on a number of assumptions and approximations, some of which have been previously discussed [7,8]:

- We consider a recurrent network of N all-to-all potentially connected [10,11] McCulloch and Pitts neurons [12]. Each neuron may belong to one of several classes defined by the values of the firing threshold, h, firing probability, f, average absolute weight of input connections, w, and robustness parameter, κ.
- The network is presented with associative memory sequences consisting of synchronous network states that are independent across neurons and time.
- Individual neurons in the network attempt to learn, independently from one another, input-output associations derived from these sequences.
- Each neuron learns by modifying its input connection weights, J, in the presence of constraints on the signs and l_1-norm of these connections. Parameters h, f, w, and κ remain fixed during learning.
- In numerical simulations, when the memory load is subcritical and a neuron is presented with a feasible learning problem, we choose the solution with the minimum l_2-norm of input connection weights. For a higher memory load, when the associative learning problem is non-feasible, we choose the solution that minimizes the total error associated with the erroneously learned associations.
- For the replica theory calculations performed in the limit of $N \to \infty$, we consider two specific scenarios of scaling of model parameters with N. We show that in the biologically plausible limit of parameters, both scenarios lead to the same results. We note, that scaling assumptions are not required for numerical simulations performed for finite N.
- Memory retrieval is deemed to be successful if the retrieved sequence does not deviate significantly from the learned associative sequence.

Learning capacity of a single neuron with a fixed threshold, binary (0, 1) input/output, robustness to noise, sign constraints, and multiple equality constraints

We consider a single neuron receiving N potential inputs from the network (enumerated with index j). The neuron attempts to learn a set of m binary (0, 1) input-output associations, $\left\{X^\mu \to y^\mu\right\}$, in the presence of postsynaptic noise, η, and constraints on its input connection weights, J_j:

$$\begin{aligned}
&\theta\left(\sum_{j=1}^{N} J_j X_j^\mu - h + \eta\right) = y^\mu, \quad \mu = 1,\ldots,m \\
&J_j g_j \geq 0, \quad j = 1,\ldots,N \\
&\frac{1}{N}\sum_{j=1}^{N} J_j c_j^\nu = w^\nu, \quad \nu = 1,\ldots,k \\
&\mathrm{Prob}\left(X_j^\mu\right) = \begin{cases} 1-f_j, & X_j^\mu = 0 \\ f_j, & X_j^\mu = 1 \end{cases}; \quad \mathrm{Prob}\left(y^\mu\right) = \begin{cases} 1-f_{out}, & y^\mu = 0 \\ f_{out}, & y^\mu = 1 \end{cases}
\end{aligned} \tag{1}$$

In these expressions, θ denotes the Heaviside step-function and h is the firing threshold. Inputs X_j^μ and outputs y^μ are randomly and independently drawn from distinct Bernoulli distributions, in which the probabilities of having 1 are denoted with f_j and f_{out} respectively. To enforce sign-constraints on connection weights we introduced parameters g_j, which equal 1 for excitatory and –1 for inhibitory inputs. Parameters c_j^ν and w^ν define k linear equality constraints.

We assume that the associations must be learned robustly so that they can be successfully recalled in the presence of postsynaptic noise, η, $|\eta| \leq \kappa$. Parameter $\kappa \geq 0$ is referred to as the robustness parameter. After eliminating η from Eqs. (1), the problem can be rewritten as:

$$\begin{aligned}
&\left(2y^\mu - 1\right)\left(\sum_{j=1}^{N} J_j X_j^\mu - h\right) \geq \kappa, \quad \mu = 1,\ldots,m \\
&\frac{1}{N}\sum_{j=1}^{N} J_j c_j^\nu = w^\nu, \quad \nu = 1,\ldots,k \\
&J_j g_j \geq 0, \quad j = 1,\ldots,N \\
&\mathrm{Prob}\left(X_j^\mu\right) = \begin{cases} 1-f_j, & X_j^\mu = 0 \\ f_j, & X_j^\mu = 1 \end{cases}; \quad \mathrm{Prob}\left(y^\mu\right) = \begin{cases} 1-f_{out}, & y^\mu = 0 \\ f_{out}, & y^\mu = 1 \end{cases}
\end{aligned} \tag{2}$$

Additional assumptions required for the replica theory calculation

We assume that N is large, while m/N, f_j and f_{out} and c_j^ν are $O(1)$ (of order 1 with respect to N). Total input to a neuron can be expressed in terms of the input averaged over the associations, X, plus a deviation from the average:

$$\begin{aligned}\sum_{j=1}^{N} J_j X_j^{\mu} - h &= \left\langle \sum_{j=1}^{N} J_j X_j^{\mu} - h \right\rangle_X + O\left(\sqrt{\underset{X}{\mathrm{var}} \left(\sum_{j=1}^{N} J_j X_j^{\mu} - h \right)} \right) = \\ &= \sum_{j=1}^{N} J_j f_j - h + O\left(\sqrt{\sum_{j=1}^{N} f_j \left(1 - f_j\right) J_j^2} \right)\end{aligned} \tag{3}$$

With this, the associative learning problem of Eqs. (2) can be separated into two categories based on the value of y^{μ}:

$$\begin{cases} \sum_{j=1}^{N} J_j f_j - h + O\left(\sqrt{\sum_{j=1}^{N} J_j^2 f_j \left(1 - f_j\right)} \right) \geq \kappa; \quad y^{\mu} = 1 \\ \sum_{j=1}^{N} J_j f_j - h + O\left(\sqrt{\sum_{j=1}^{N} J_j^2 f_j \left(1 - f_j\right)} \right) \leq -\kappa; \quad y^{\mu} = 0 \end{cases} \tag{4}$$

To guarantee $O(1)$ capacity in the large N limit, it is necessary for the deviation of the average input from the threshold, $\sum_{j=1}^{N} J_j f_j - h$, and the robustness parameter, κ, to be of the same order as (or less than) the width of the input distribution, $\sigma_{input} = \sqrt{\sum_{j=1}^{N} J_j^2 f_j \left(1 - f_j\right)}$. If not, input to the neuron will rarely cross the threshold (if the first conditions is violated) or the robustness margins (if the second condition is violated), and capacity for robust associative memory storage will be close to zero. Therefore,

$$\begin{aligned} \sum_{j=1}^{N} J_j f_j - h &= O\left(\sqrt{\sum_{j=1}^{N} J_j^2 f_j \left(1 - f_j\right)} \right) \\ \kappa &= O\left(\sqrt{\sum_{j=1}^{N} J_j^2 f_j \left(1 - f_j\right)} \right) \end{aligned} \tag{5}$$

For two classes of neurons, one excitatory and one inhibitory, with similar weight magnitudes, Eqs. (5) give rise to various plausible scenarios for scaling of connection weights and robustness parameter with the network size:

$$\begin{aligned} J_j &= \left\{ O\left(\frac{h}{N}\right), O\left(\frac{h}{\sqrt{N}}\right), O(h), \ldots \right\}; \quad \Rightarrow \quad J_j = \left\{ \frac{1}{N}, \frac{1}{\sqrt{N}}, 1, \ldots \right\} \tilde{J}_j h \\ \kappa &= O\left(\sqrt{N} J_j\right); \quad \Rightarrow \quad \kappa = \left\{ \frac{1}{\sqrt{N}}, 1, \sqrt{N}, \ldots \right\} \tilde{\kappa} h \end{aligned} \tag{6}$$

The normalized weights, $\tilde{J}_j$, and the normalized robustness parameter, $\tilde{\kappa}$, in Eqs. (6) do not scale with N.

The first scenario is usually used in associative memory models in conjunction with the replica theory (see e.g. [5]):

$$J_j = \frac{h}{N}\tilde{J}_j; \quad \kappa = \frac{h}{\sqrt{N}}\tilde{\kappa} \tag{7}$$

The second scaling scenario is traditionally used in balanced network models (see e.g. [9]):

$$J_j = \frac{h}{\sqrt{N}}\tilde{J}_j; \quad \kappa = h\tilde{\kappa} \tag{8}$$

The third and the subsequent scenarios, in which J does not scale with N, or J increases with N, can be ruled out because they are biologically unrealistic. What is more, one can see from Eqs. (2) that the firing threshold in these cases can be disregarded, and the results of replica calculation become identical to the second scaling scenario.

In all models, scaling of w^ν is assumed to be the same as that of J, i.e. $w^\nu = \left\{\frac{1}{N}, \frac{1}{\sqrt{N}}\right\}\tilde{w}^\nu h$. Substituting the normalized variables into Eqs. (2) we arrive at two problems, both governed by the same set of intensive parameters, $f_j, f_{out}, \tilde{\kappa}, \tilde{w}^\nu, c_j^\nu, g_j$, as well as parameters m, N, and k:

$$\begin{aligned}
&\left(2y^\mu - 1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j X_j^\mu - \left\{1, \frac{1}{\sqrt{N}}\right\}\right) \geq \frac{\tilde{\kappa}}{\sqrt{N}}, \quad \mu = 1,\ldots,m \\
&\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j c_j^\nu = \tilde{w}^\nu, \quad \nu = 1,\ldots,k \\
&\tilde{J}_j g_j \geq 0, \quad j = 1,\ldots,N \\
&\mathrm{Prob}\left(X_j^\mu\right) = \begin{cases} 1-f_j, & X_j^\mu = 0 \\ f_j, & X_j^\mu = 1 \end{cases}; \quad \mathrm{Prob}\left(y^\mu\right) = \begin{cases} 1-f_{out}, & y^\mu = 0 \\ f_{out}, & y^\mu = 1 \end{cases}
\end{aligned} \tag{9}$$

Note that the two formulations only differ in the threshold term.

Replica theory solution of the model in the large *N* limit

We solve the above two models concurrently by following the steps of the replica theory solution outlined in [7,8]. We begin by calculating the volume of the connection weight space, $\Omega\left(X_j^\mu, y^\mu\right)$, in which Eqs. (9) hold for a given set of associations:

$$\Omega\left(X_j^{\mu}, y^{\mu}\right) = \int \prod_{j=1}^{N} d\tilde{J}_j \prod_{\mu=1}^{m} \theta\left(\left(2y^{\mu}-1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j X_j^{\mu} - \left\{1, \frac{1}{\sqrt{N}}\right\}\right) - \frac{\tilde{\kappa}}{\sqrt{N}}\right) \times \prod_{j=1}^{N}\theta\left(\tilde{J}_j g_j\right)\prod_{\nu=1}^{k}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j c_j^{\nu} - \tilde{w}^{\nu}\right) \tag{10}$$

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

$$\ln\left(\Omega_{typical}\right) = \left\langle \ln\left(\Omega\left(X_j^{\mu}, y^{\mu}\right)\right)\right\rangle_{X_j^{\mu}, y^{\mu}} = \lim_{n\to 0}\frac{\left\langle \Omega\left(X_j^{\mu}, y^{\mu}\right)^n\right\rangle_{X_j^{\mu}, y^{\mu}} - 1}{n} \tag{11}$$

The quantity $\left\langle \Omega\left(X_j^{\mu}, y^{\mu}\right)^n\right\rangle_{X_j^{\mu}, y^{\mu}}$ is then expressed as a single multidimensional integral:

$$\left\langle \Omega\left(X_j^{\mu}, y^{\mu}\right)^n\right\rangle_{X_j^{\mu}, y^{\mu}} = \left\langle \int \prod_{a,j=1}^{n,N} d\tilde{J}_j^a \prod_{\mu,a=1}^{m,n} \theta\left(\left(2y^{\mu}-1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a X_j^{\mu} - \left\{1, \frac{1}{\sqrt{N}}\right\}\right) - \frac{\tilde{\kappa}}{\sqrt{N}}\right) \times \prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_j^a g_j\right)\prod_{a,\nu=1}^{n,k}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a c_j^{\nu} - \tilde{w}^{\nu}\right)\right\rangle_{X_j^{\mu}, y^{\mu}} \tag{12}$$

Input and output associations, X_j^{μ} and y^{μ}, are decoupled through the introduction of a new variable, $\frac{\lambda^{a,\mu}}{\sqrt{N}} = \frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a X_j^{\mu} - \left\{1, \frac{1}{\sqrt{N}}\right\}$:

$$\left\langle \Omega\left(X_j^{\mu}, y^{\mu}\right)^n\right\rangle_{X_j^{\mu}, y^{\mu}} = \int \prod_{a,j=1}^{n,N} d\tilde{J}_j^a \prod_{\mu,a=1}^{m,n} \frac{d\lambda^{a,\mu}}{\sqrt{N}} \left\langle \prod_{\mu,a=1}^{m,n}\theta\left(\left(2y^{\mu}-1\right)\lambda^{a,\mu} - \tilde{\kappa}\right)\right\rangle_{y^{\mu}} \times \left\langle \prod_{\mu,a=1}^{m,n}\delta\left(\left\{1, \frac{1}{\sqrt{N}}\right\} + \frac{\lambda^{a,\mu}}{\sqrt{N}} - \frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a X_j^{\mu}\right)\right\rangle_{X_j^{\mu}} \prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_j^a g_j\right)\prod_{a,\nu=1}^{n,k}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a c_j^{\nu} - \tilde{w}^{\nu}\right) \tag{13}$$

Next, the step-functions and the δ-functions are replaced with their Fourier representations, i.e.:

$$\theta\left(\left(2y^{\mu}-1\right)\lambda^{a,\mu}-\tilde{\kappa}\right)=\int\frac{d'u^{a,\mu}d\hat{u}^{a,\mu}}{2\pi}e^{i\hat{u}^{a,\mu}\left(\left(2y^{\mu}-1\right)\lambda^{a,\mu}-\tilde{\kappa}-u^{a,\mu}\right)}$$

$$\delta\left(\left\{1,\frac{1}{\sqrt{N}}\right\}+\frac{\lambda^{a,\mu}}{\sqrt{N}}-\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a X_j^{\mu}\right)=\int\frac{d\hat{\lambda}^{a,\mu}}{2\pi/\sqrt{N}}e^{i\hat{\lambda}^{a,\mu}\left(\left\{\sqrt{N},1\right\}+\lambda^{a,\mu}-\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\tilde{J}_j^a X_j^{\mu}\right)} \tag{14}$$

$$\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^a c_j^{\nu}-\tilde{w}^{\nu}\right)=\int\frac{d\hat{k}^{a,\nu}}{2\pi/N}e^{i\hat{k}^{a,\nu}\left(\sum_{j=1}^{N}\tilde{J}_j^a c_j^{\nu}-N\tilde{w}^{\nu}\right)}$$

Symbol d' in these expressions and thereafter is designated for 0 to ∞ integration, whereas d is used to denote integration from $-\infty$ to ∞.

$$\left\langle\Omega\left(X_j^{\mu},y^{\mu}\right)^n\right\rangle_{X_j^{\mu},y^{\mu}}=\int\prod_{a,j=1}^{n,N}d\tilde{J}_j^a\prod_{\mu,a=1}^{m,n}\frac{d\lambda^{a,\mu}d\hat{\lambda}^{a,\mu}}{2\pi}\prod_{\mu,a=1}^{m,n}\frac{d'u^{a,\mu}d\hat{u}^{a,\mu}}{2\pi}\prod_{a,\nu=1}^{n,k}\frac{d\hat{k}^{a,\nu}}{2\pi/N}\times$$
$$\left\langle\prod_{\mu,a=1}^{m,n}e^{i\hat{u}^{a,\mu}\left(\left(2y^{\mu}-1\right)\lambda^{a,\mu}-\tilde{\kappa}-u^{a,\mu}\right)}\right\rangle_{y^{\mu}}\left\langle\prod_{\mu,a=1}^{m,n}e^{i\hat{\lambda}^{a,\mu}\left(\left\{\sqrt{N},1\right\}+\lambda^{a,\mu}-\frac{1}{\sqrt{N}}\sum_{j=1}^{N}\tilde{J}_j^a X_j^{\mu}\right)}\right\rangle_{X_j^{\mu}}\prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_j^a g_j\right)\prod_{a,\nu=1}^{n,k}e^{i\hat{k}^{a,\nu}\left(\sum_{j=1}^{N}\tilde{J}_j^a c_j^{\nu}-N\tilde{w}^{\nu}\right)} \tag{15}$$

After averaging over the associations we arrive at:

$$\left\langle\Omega\left(X_j^{\mu},y^{\mu}\right)^n\right\rangle_{X_j^{\mu},y^{\mu}}=\int\prod_{a,j=1}^{n,N}d\tilde{J}_j^a\prod_{\mu,a=1}^{m,n}\frac{d\lambda^{a,\mu}d\hat{\lambda}^{a,\mu}}{2\pi}\prod_{\mu,a=1}^{m,n}\frac{d'u^{a,\mu}d\hat{u}^{a,\mu}}{2\pi}\prod_{a,\nu=1}^{n,k}\frac{d\hat{k}^{a,\nu}}{2\pi/N}\times$$
$$\prod_{\mu,a=1}^{m,n}e^{-i\hat{u}^{a,\mu}\left(\tilde{\kappa}+u^{a,\mu}\right)}\prod_{\mu=1}^{m}\left(f_{out}e^{i\sum_{a=1}^{n}\hat{u}^{a,\mu}\lambda^{a,\mu}}+(1-f_{out})e^{-i\sum_{a=1}^{n}\hat{u}^{a,\mu}\lambda^{a,\mu}}\right)\prod_{\mu,a=1}^{m,n}e^{i\hat{\lambda}^{a,\mu}\left(\left\{\sqrt{N},1\right\}+\lambda^{a,\mu}\right)}\times \tag{16}$$
$$\prod_{j,\mu=1}^{N,m}\left(1-f_j+f_j e^{\frac{-i}{\sqrt{N}}\sum_{a=1}^{n}\hat{\lambda}^{a,\mu}\tilde{J}_j^a}\right)\prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_j^a g_j\right)\prod_{a,\nu=1}^{n,k}e^{i\hat{k}^{a,\nu}\left(\sum_{j=1}^{N}\tilde{J}_j^a c_j^{\nu}-N\tilde{w}^{\nu}\right)}$$

Replacing the argument of the first product in line three of Eq. (16) with an exponential expression that approximates it up to the second order in $\frac{-i}{\sqrt{N}}\sum_{a=1}^{n}\hat{\lambda}^{a,\mu}\tilde{J}_j^a$, we obtain:

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

$$\prod_{a,\nu=1}^{n,k} e^{-iN\hat{k}^{a,\nu}\tilde{w}^{\nu}} \prod_{\mu=1}^{m} \left(f_{out} e^{i\sum_{a=1}^{n} \hat{u}^{a,\mu}\lambda^{a,\mu}} + (1-f_{out}) e^{-i\sum_{a=1}^{n} \hat{u}^{a,\mu}\lambda^{a,\mu}} \right) e^{-i\sum_{\mu,a=1}^{m,n} \hat{u}^{a,\mu}\left(\tilde{\kappa}+u^{a,\mu}\right) + i\sum_{\mu,a=1}^{m,n} \hat{\lambda}^{a,\mu}\left(\left\{\sqrt{N},1\right\}+\lambda^{a,\mu}\right)} \times \quad (17)$$

$$\prod_{j,\mu=1}^{N,m} e^{\frac{-i}{\sqrt{N}} f_j \sum_{a=1}^{n} \hat{\lambda}^{a,\mu}\tilde{J}_j^a - \frac{\left(f_j - f_j^2\right)}{2N} \sum_{a,b=1}^{n,n} \hat{\lambda}^{a,\mu}\hat{\lambda}^{b,\mu}\tilde{J}_j^a\tilde{J}_j^b} \prod_{j,a=1}^{N,n} \theta\left(\tilde{J}_j^a g_j\right) \prod_{a,\nu=1}^{n,k} e^{i\hat{k}^{a,\nu}\sum_{j=1}^{N} \tilde{J}_j^a c_j^{\nu}}$$

At this point, we introduce two sets of order parameters which allow us to decouple the products containing indices j and μ,

$$\frac{1}{N}\sum_{j=1}^{N} f_j \tilde{J}_j^a = \left\{1, \frac{1}{\sqrt{N}}\right\} + \frac{s^a}{\sqrt{N}}, \quad \frac{1}{N}\sum_{j=1}^{N} f_j\left(1-f_j\right)\tilde{J}_j^a\tilde{J}_j^b = q^{a,b} \quad (18)$$

Insertion of these order parameters into Eq. (17) leads to the following expression:

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

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

$$\left(\prod_{a=1}^{n} \frac{d\lambda^a d\hat{\lambda}^a}{2\pi} \prod_{a=1}^{n} \frac{d'u^a d\hat{u}^a}{2\pi} \left(f_{out} e^{i\sum_{a=1}^{n} \hat{u}^a\lambda^a} + (1-f_{out}) e^{-i\sum_{a=1}^{n} \hat{u}^a\lambda^a} \right) e^{-i\sum_{a=1}^{n} \hat{\lambda}^a s^a - \frac{1}{2}\sum_{a,b=1}^{n,n} \hat{\lambda}^a\hat{\lambda}^b q^{a,b} - i\sum_{a=1}^{n} \hat{u}^a\left(\tilde{\kappa}+u^a\right) + i\sum_{a=1}^{n} \hat{\lambda}^a\lambda^a} \right)^m \times \quad (19)$$

$$\prod_{j=1}^{N} \left(\prod_{a=1}^{n} d\tilde{J}^a \prod_{a=1}^{n} \theta\left(\tilde{J}^a g_j\right) \prod_{a=1}^{n} e^{i\tilde{J}^a \sum_{\nu=1}^{k} \hat{k}^{a,\nu} c_j^{\nu}} \prod_{a=1}^{n} e^{-i\hat{s}^a f_j \tilde{J}^a} \prod_{a,b=1}^{n} e^{-i\hat{q}^{a,b} f_j\left(1-f_j\right)\tilde{J}^a\tilde{J}^b} \right)$$

After the integration over $d\lambda^a d\hat{\lambda}^a$ we arrive at:

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

$$e^{N\left(i\sum_{a=1}^{n}\left(\{1,0\}\hat{s}^a - \sum_{\nu=1}^{k} \hat{k}^{a,\nu}\tilde{w}^{\nu} \right) + i\sum_{a,b=1}^{n,n} q^{a,b}\hat{q}^{a,b} + n\alpha\tilde{G}_E\left(\{s^a\},\{q^{a,b}\}\right) + n\tilde{G}_S\left(\{\hat{k}^{a,\nu}\},\{\hat{s}^a\},\{\hat{q}^{a,b}\}\right) \right)}$$

$$\tilde{G}_E\left(\{s^a\},\{q^{a,b}\}\right) = \frac{1}{n}\ln\left(\int \prod_{a=1}^{n} \frac{d'u^a d\hat{u}^a}{2\pi} e^{-\frac{1}{2}\sum_{a,b=1}^{n} \hat{u}^a\hat{u}^b q^{a,b}} \left(f_{out} e^{i\sum_{a=1}^{n}\hat{u}^a\left(s^a - u^a - \tilde{\kappa}\right)} + (1-f_{out}) e^{-i\sum_{a=1}^{n}\hat{u}^a\left(s^a + u^a + \tilde{\kappa}\right)} \right) \right) \quad (20)$$

$$\tilde{G}_S\left(\{\hat{k}^{a,\nu}\},\{\hat{s}^a\},\{\hat{q}^{a,b}\}\right) = \frac{1}{n}\frac{1}{N}\sum_{j=1}^{N}\ln\left(\int \prod_{a=1}^{n} d'\tilde{J}^a e^{i\left(\sum_{\nu=1}^{k}\hat{k}^{a,\nu}c_j^{\nu} - \hat{s}^a f_j\right) g_j\tilde{J}^a - if_j\left(1-f_j\right)\sum_{a,b=1}^{n}\tilde{J}^a\tilde{J}^b\hat{q}^{a,b}} \right)$$

The integral in the first line of Eqs. (20) is calculated by using the steepest descent method combined with the assumption of a replica-symmetric saddle point, $s^a = s$, $q^{a,a} = q_0$, $q^{a\neq b} = q$, $\hat{k}^{a,\nu} = \hat{k}^{\nu}$, $\hat{s}^a = \hat{s}$, $\hat{q}^{a,a} = \hat{q}_0$, and $\hat{q}^{a\neq b} = \hat{q}$:

$$\left\langle \Omega\left(X_j^{\mu}, y^{\mu}\right)^n \right\rangle_{X_j^{\mu}, y^{\mu}} \sim e^{Nn\left(i\left(\{1,0\}\hat{s} - \sum_{\nu=1}^{k}\hat{k}^{\nu}\tilde{w}^{\nu} \right) + iq_0\hat{q}_0 - iq\hat{q} + \alpha G_E(s,q_0,q) + G_S\left(\{\hat{k}^{\nu}\},\hat{s},\hat{q}_0,\hat{q}\right) \right)}$$

$$G_E\left(s,q_0,q\right) = \frac{1}{n}\ln\left(\int \prod_{a=1}^{n} \frac{d'u^a d\hat{u}^a}{2\pi} e^{-\frac{1}{2}q_0\sum_{a=1}^{n}\hat{u}^a\hat{u}^a - \frac{1}{2}q\sum_{a\neq b=1}^{n}\hat{u}^a\hat{u}^b} \left(f_{out} e^{i\sum_{a=1}^{n}\hat{u}^a\left(s - u^a - \tilde{\kappa}\right)} + (1-f_{out}) e^{-i\sum_{a=1}^{n}\hat{u}^a\left(s + u^a + \tilde{\kappa}\right)} \right) \right) \quad (21)$$

$$G_S\left(\{\hat{k}^{\nu}\},\hat{s},\hat{q}_0,\hat{q}\right) = \frac{1}{n}\frac{1}{N}\sum_{j=1}^{N}\ln\left(\int \prod_{a=1}^{n} d'\tilde{J}^a e^{i\left(\sum_{\nu=1}^{k}\hat{k}^{\nu}c_j^{\nu} - \hat{s}f_j\right) g_j\tilde{J}^a - if_j\left(1-f_j\right)\hat{q}_0\sum_{a=1}^{n}\tilde{J}^a\tilde{J}^a - if_j\left(1-f_j\right)\hat{q}\sum_{a\neq b=1}^{n}\tilde{J}^a\tilde{J}^b} \right)$$

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

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

$$-i\tilde{w}^{\nu} + \frac{\partial G_S\left(\{\hat{k}^{\nu}\},\hat{s},\hat{q}_0,\hat{q}\right)}{\partial \hat{k}^{\nu}} = 0; \quad i\{1,0\} + \frac{\partial G_S\left(\{\hat{k}^{\nu}\},\hat{s},\hat{q}_0,\hat{q}\right)}{\partial \hat{s}} = 0 \quad (22)$$

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

To simplify the expressions for G_E and G_S, we employ the Hubbard-Stratonovich transformation in the form of,

$$e^{a\sum_{i\neq j}s_i s_j} = \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} dx e^{-x^2+2x\sqrt{a}\sum_i s_i - a\sum_i s_i^2}, \tag{23}$$

and take the $n\to 0$ limit:

$$G_E(s,q_0,q) = \int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\left(f_{out}\ln\left(\int\frac{d'ud\hat{u}}{2\pi}e^{-\frac{1}{2}q_0\hat{u}^2+2x\sqrt{-\frac{1}{2}}q\hat{u}+\frac{1}{2}q\hat{u}^2+i\hat{u}(s-u-\tilde{\kappa})}\right)+\right.$$
$$\left.(1-f_{out})\ln\left(\int\frac{d'ud\hat{u}}{2\pi}e^{-\frac{1}{2}q_0\hat{u}^2+2x\sqrt{-\frac{1}{2}}q\hat{u}+\frac{1}{2}q\hat{u}^2-i\hat{u}(s+u+\tilde{\kappa})}\right)\right) \tag{24}$$
$$G_S\left(\left\{\hat{k}^\nu\right\},\hat{s},\hat{q}_0,\hat{q}\right) = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\ln\left(\int d'\tilde{J}e^{i\left(\sum_{\nu=1}^{k}\hat{k}^\nu c_j^\nu-\hat{s}f_j\right)g_j\tilde{J}+2x\sqrt{-if_j(1-f_j)\hat{q}}\tilde{J}-if_j(1-f_j)(\hat{q}_0-\hat{q})\tilde{J}^2}\right)$$

After calculating the integrals inside the arguments of the natural logarithms we obtain:

$$G_E(s,q_0,q) = \int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\left(f_{out}\ln\left(\operatorname{erfc}\left(\frac{-s+\tilde{\kappa}-x\sqrt{2q}}{\sqrt{2(q_0-q)}}\right)\right)+\right.$$
$$\left.(1-f_{out})\ln\left(\operatorname{erfc}\left(\frac{s+\tilde{\kappa}-x\sqrt{2q}}{\sqrt{2(q_0-q)}}\right)\right)\right)-\ln 2$$
$$G_S\left(\left\{\hat{k}^\nu\right\},\hat{s},\hat{q}_0,\hat{q}\right) = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\ln\left(\frac{\sqrt{\pi}e^{\frac{\left(i\left(\sum_{\nu=1}^{k}\hat{k}^\nu c_j^\nu-\hat{s}f_j\right)g_j+2x\sqrt{-if_j(1-f_j)\hat{q}}\right)^2}{4if_j(1-f_j)(\hat{q}_0-\hat{q})}}}{2\sqrt{if_j(1-f_j)(\hat{q}_0-\hat{q})}}\times\right.$$
$$\left.\operatorname{erfc}\left(-\frac{\left(i\left(\sum_{\nu=1}^{k}\hat{k}^\nu c_j^\nu-\hat{s}f_j\right)g_j+2x\sqrt{-if_j(1-f_j)\hat{q}}\right)}{2\sqrt{if_j(1-f_j)(\hat{q}_0-\hat{q})}}\right)\right) \tag{25}$$

Next, we make substitutions, $u_\pm = \frac{\tilde{\kappa}\pm s}{\sqrt{2q}}$, $\varepsilon = \frac{q_0-q}{q}$, $t=-i\hat{q}$, $z=\frac{i\hat{s}}{2\sqrt{-i\hat{q}}}$, $\delta = -\frac{\hat{q}_0-\hat{q}}{\hat{q}}$, $\eta^\nu = -\frac{i\hat{k}^\nu}{\sqrt{-i\hat{q}}}$, and rewrite G_E, G_S, and the saddle-point equations in terms of the new variables:

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

$$G_S\left(\left\{\eta^\nu\right\},t,z,\delta\right)=\frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\times$$

$$\ln\left(\frac{\sqrt{\pi}e^{\frac{\left(\left(\sum_{\nu=1}^{k}-\eta^\nu c_j^\nu-2zf_j\right)g_j+2x\sqrt{f_j\left(1-f_j\right)}\right)^2}{4f_j\left(1-f_j\right)\delta}}}{2\sqrt{f_j\left(1-f_j\right)\delta t}}\operatorname{erfc}\left(-\frac{\left(\left(\sum_{\nu=1}^{k}-\eta^\nu c_j^\nu-2zf_j\right)g_j+2x\sqrt{f_j\left(1-f_j\right)}\right)}{2\sqrt{f_j\left(1-f_j\right)\delta}}\right)\right)$$

$$\begin{cases}\dfrac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_+}-\dfrac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_-}=0\\ \dfrac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial \varepsilon}=\dfrac{t\left(1-\delta\right)}{\alpha}\dfrac{2\tilde{\kappa}^2}{\left(u_+ +u_-\right)^2}\\ \dfrac{\partial G_E\left(u_+,u_-,\varepsilon\right)}{\partial u_+}=\dfrac{4\tilde{\kappa}^2 t}{\alpha\left(u_+ +u_-\right)^3}\left(\delta-\varepsilon+\delta\varepsilon\right)\\ \dfrac{\partial G_S\left(\left\{\eta^\nu\right\},t,z,\delta\right)}{\partial \eta^\nu}=-\tilde{w}^\nu\sqrt{t},\quad \nu=1,\ldots,k\\ \dfrac{\partial G_S\left(\left\{\eta^\nu\right\},t,z,\delta\right)}{\partial z}=-\left\{1,0\right\}2\sqrt{t}\\ \dfrac{\partial G_S\left(\left\{\eta^\nu\right\},t,z,\delta\right)}{\partial \delta}=-t\left(1+\varepsilon\right)\dfrac{2\tilde{\kappa}^2}{\left(u_+ +u_-\right)^2}\\ \dfrac{\partial G_S\left(\left\{\eta^\nu\right\},t,z,\delta\right)}{\partial t}=-\dfrac{1}{2\sqrt{t}}\sum_{\nu=1}^{k}\tilde{w}^\nu\eta^\nu-\dfrac{z}{\sqrt{t}}\left\{1,0\right\}-\dfrac{2\tilde{\kappa}^2}{\left(u_+ +u_-\right)^2}\left(\delta-\varepsilon+\delta\varepsilon\right)\end{cases}\qquad(26)$$

Replica theory solution at critical capacity

At critical associative memory storage capacity, the saddle point Eqs. (26) can be simplified because as $\Omega_{typical}$ tends to zero, q_0-q goes to zero as well. In this limit, parameters ε and δ are small, and the $k+6$ saddle point equations can be expanded asymptotically to the leading orders in $1/\varepsilon$ and $1/\delta$:

$$
\begin{aligned}
&f_{out}F\left(u_{-}\right)=\left(1-f_{out}\right)F\left(u_{+}\right)\\
&f_{out}D\left(u_{-}\right)+\left(1-f_{out}\right)D\left(u_{+}\right)=\frac{4\tilde{\kappa}^{2}}{\alpha\left(u_{+}+u_{-}\right)^{2}}\varepsilon^{2}t\\
&\left(1-f_{out}\right)F\left(u_{+}\right)=\frac{4\tilde{\kappa}^{2}}{\alpha\left(u_{+}+u_{-}\right)^{3}}\varepsilon t\left(\varepsilon-\delta\right)\\
&\frac{1}{N}\sum_{j=1}^{N}F\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_{j}^{\nu}+2zf_{j}\right)g_{j}}{2\sqrt{f_{j}\left(1-f_{j}\right)}}\right)\frac{c_{j}^{\nu}g_{j}}{\sqrt{f_{j}\left(1-f_{j}\right)}}=2\tilde{w}^{\nu}\delta\sqrt{t},\quad \nu=1,\ldots,k\\
&\frac{1}{N}\sum_{j=1}^{N}F\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_{j}^{\nu}+2zf_{j}\right)g_{j}}{2\sqrt{f_{j}\left(1-f_{j}\right)}}\right)\frac{f_{j}g_{j}}{\sqrt{f_{j}\left(1-f_{j}\right)}}=\left\{1,0\right\}2\delta\sqrt{t}\\
&\frac{1}{N}\sum_{j=1}^{N}D\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_{j}^{\nu}+2zf_{j}\right)g_{j}}{2\sqrt{f_{j}\left(1-f_{j}\right)}}\right)=\frac{4\tilde{\kappa}^{2}}{\left(u_{+}+u_{-}\right)^{2}}\delta^{2}t\\
&\frac{1}{2}\sum_{\nu=1}^{k}\tilde{w}^{\nu}\eta^{\nu}+\left\{1,0\right\}z-\frac{2\tilde{\kappa}^{2}}{\left(u_{+}+u_{-}\right)^{2}}\left(\varepsilon-\delta\right)\sqrt{t}=0
\end{aligned}
\tag{27}
$$

Special functions E, F, and D are introduced in the above expressions for conciseness:

$$
\begin{aligned}
&E\left(x\right)=\frac{1}{2}\left(1+\operatorname{erf}\left(x\right)\right)\\
&F\left(x\right)=\frac{1}{\sqrt{\pi}}e^{-x^{2}}+x\left(1+\operatorname{erf}\left(x\right)\right)\\
&D\left(x\right)=xF\left(x\right)+E\left(x\right)
\end{aligned}
\tag{28}
$$

After eliminating t, δ, and ε from Eqs. (27) we arrive at the final result:

$$\alpha\left(\left\{\tilde{w}^{\nu}\right\},\left\{c_j^{\nu}\right\},\left\{g_j\right\},\left\{f_j\right\},f_{out},\tilde{\kappa}\right)=\left(\frac{\sum_{\nu=1}^{k}\tilde{w}^{\nu}\eta^{\nu}+\{2,0\}z}{\tilde{\kappa}}\right)^2\frac{f_{out}D\left(u_-\right)+(1-f_{out})D\left(u_+\right)}{\left(f_{out}F\left(u_-\right)+(1-f_{out})F\left(u_+\right)\right)^2}$$

$$\begin{cases}
f_{out}F\left(u_-\right)-(1-f_{out})F\left(u_+\right)=0 \\
\frac{1}{N}\sum_{j=1}^{N}\frac{c_j^{\nu}g_j}{\sqrt{f_j\left(1-f_j\right)}}F\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_j^{\nu}+2zf_j\right)g_j}{2\sqrt{f_j\left(1-f_j\right)}}\right)=\frac{\tilde{w}^{\nu}}{\tilde{\kappa}}\left(\frac{\sum_{\nu=1}^{k}\tilde{w}^{\nu}\eta^{\nu}+\{2,0\}z}{\tilde{\kappa}}\right)\times \\
\left(u_++u_-\right)\frac{f_{out}E\left(u_-\right)+(1-f_{out})E\left(u_+\right)}{f_{out}F\left(u_-\right)+(1-f_{out})F\left(u_+\right)},\quad \nu=1,\ldots,k \\
\frac{1}{N}\sum_{j=1}^{N}\frac{f_jg_j}{\sqrt{f_j\left(1-f_j\right)}}F\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_j^{\nu}+2zf_j\right)g_j}{2\sqrt{f_j\left(1-f_j\right)}}\right)=\left\{\frac{1}{\tilde{\kappa}},0\right\}\left(\frac{\sum_{\nu=1}^{k}\tilde{w}^{\nu}\eta^{\nu}+\{2,0\}z}{\tilde{\kappa}}\right)\times \\
\left(u_++u_-\right)\frac{f_{out}E\left(u_-\right)+(1-f_{out})E\left(u_+\right)}{f_{out}F\left(u_-\right)+(1-f_{out})F\left(u_+\right)} \\
\frac{1}{N}\sum_{j=1}^{N}D\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_j^{\nu}+2zf_j\right)g_j}{2\sqrt{f_j\left(1-f_j\right)}}\right)=\left(\frac{\sum_{\nu=1}^{k}\tilde{w}^{\nu}\eta^{\nu}+\{2,0\}z}{\tilde{\kappa}}\right)^2\left(\frac{f_{out}E\left(u_-\right)+(1-f_{out})E\left(u_+\right)}{f_{out}F\left(u_-\right)+(1-f_{out})F\left(u_+\right)}\right)^2 \\
u_++u_-<0;\quad \sum_{\nu=1}^{k}\tilde{w}^{\nu}\eta^{\nu}+\{2,0\}z<0
\end{cases} \tag{29}$$

We note that to have a non-zero capacity, the number of input classes must be greater than the number of homeostatic constraints, k.

Distribution of input weights at critical capacity

The probability density of input weights can be derived from the following general expression:

$$p_i\left(\tilde{J}\right)=\left\langle \frac{1}{\Omega\left(X_j^{\mu},y^{\mu}\right)}\int\prod_{j=1}^{N}d\tilde{J}_j\,\delta\left(\tilde{J}_i-\tilde{J}\right)\prod_{\mu=1}^{m}\theta\left(\left(2y^{\mu}-1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_jX_j^{\mu}-\left\{1,\frac{1}{\sqrt{N}}\right\}\right)-\frac{\tilde{\kappa}}{\sqrt{N}}\right)\times \prod_{j=1}^{N}\theta\left(\tilde{J}_jg_j\right)\prod_{\nu=1}^{k}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_jc_j^{\nu}-\tilde{w}^{\nu}\right)\right\rangle_{X_j^{\mu},y^{\mu}} \tag{30}$$

Eq. (30) can be cast in a form that closely resembles Eq. (12), allowing us to exploit the results of previous section. To that end, we introduce n replicas and take the limit of $n\to 0$ after averaging over the associations:

$$p_i\left(\tilde{J}\right)=\lim_{n\to 0}\left\langle \int\prod_{a,j=1}^{n,N}d\tilde{J}_j^{a}\,\delta\left(\tilde{J}_i^{a=1}-\tilde{J}\right)\prod_{\mu,a=1}^{m,n}\theta\left(\left(2y^{\mu}-1\right)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^{a}X_j^{\mu}-\left\{1,\frac{1}{\sqrt{N}}\right\}\right)-\frac{\tilde{\kappa}}{\sqrt{N}}\right)\times \prod_{j,a=1}^{N,n}\theta\left(\tilde{J}_j^{a}g_j\right)\prod_{a,\nu=1}^{n,k}\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j^{a}c_j^{\nu}-\tilde{w}^{\nu}\right)\right\rangle_{X_j^{\mu},y^{\mu}} \tag{31}$$

Following the steps outlined in Eqs. (12-22) we arrive at:

$$p_i\left(\tilde{J}\right)=A\delta\left(\tilde{J}\right)+\theta\left(g_i\tilde{J}\right)\int_{-\infty}^{\infty}\frac{e^{-x^2}dx}{\sqrt{\pi}}\frac{e^{\sqrt{t}\left(\left(2xg_i\sqrt{f_i(1-f_i)}-\left(\sum_{\nu=1}^{k}\eta^{\nu}c_i^{\nu}+2zf_i\right)\right)\tilde{J}-\sqrt{t}\delta f_i(1-f_i)\tilde{J}^2\right)}}{\int d'\tilde{J}^{a=1}e^{\sqrt{t}\left(\left(2x\sqrt{f_i(1-f_i)}-\left(\sum_{\nu=1}^{k}\eta^{\nu}c_i^{\nu}+2zf_i\right)g_i\right)\tilde{J}^{a=1}-\sqrt{t}\delta f_i(1-f_i)\left(\tilde{J}^{a=1}\right)^2\right)}} \tag{32}$$

Constant A in this expression is defined by the normalization condition.

At critical capacity, the integrals in Eq. (32) can be calculated with the Laplace's method, resulting in the following expression:

$$p_i\left(\tilde{J}\right)=A\delta\left(\tilde{J}\right)+\theta\left(g_i\tilde{J}\right)\left(\delta\sqrt{t}\right)_{+}\sqrt{\frac{f_i(1-f_i)}{\pi}}e^{-\left(\delta\sqrt{t}\sqrt{f_i(1-f_i)}g_i\tilde{J}+\frac{\left(\sum_{\nu=1}^{k}\eta^{\nu}c_i^{\nu}+2zf_i\right)g_i}{2\sqrt{f_i(1-f_i)}}\right)^2} \tag{33}$$

After substituting $\delta\sqrt{t}$ from the replica equations, Eqs. (27), and determining A from the normalization condition, we obtain the distributions of connection weights for different input classes:

$$p_i\left(\tilde{J}\right)=\left(1-E\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^\nu c_i^\nu+2zf_i\right)g_i}{2\sqrt{f_i\left(1-f_i\right)}}\right)\right)\delta\left(\tilde{J}\right)+\theta\left(g_i\tilde{J}\right)\frac{1}{\sqrt{2\pi}\Sigma_i}e^{-\left(\frac{\tilde{J}}{\sqrt{2}\Sigma_i}+\frac{\left(\sum_{\nu=1}^{k}\eta^\nu c_i^\nu+2zf_i\right)}{2\sqrt{f_i(1-f_i)}}\right)^2}$$

$$\Sigma_i=\frac{\sqrt{2}\tilde{\kappa}}{\sqrt{f_i\left(1-f_i\right)}\left(\frac{\sum_{\nu=1}^{k}\tilde{w}^\nu\eta^\nu+\{2,0\}z}{\tilde{\kappa}}\right)}\frac{1}{\left(u_++u_-\right)}\frac{f_{out}F\left(u_-\right)+(1-f_{out})F\left(u_+\right)}{f_{out}E\left(u_-\right)+(1-f_{out})E\left(u_+\right)} \quad (34)$$

These distributions are composed of Gaussian functions truncated at zero and finite fractions of zero-weight connections. Parameters Σ_i describe the distribution widths.

Connection probabilities and distributions of non-zero connection weights follow from Eqs. (34):

$$P_i^{con}=E\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^\nu c_i^\nu+2zf_i\right)g_i}{2\sqrt{f_i\left(1-f_i\right)}}\right)$$

$$p_i^{PSP}\left(\tilde{J}\right)=\frac{\theta\left(g_i\tilde{J}\right)}{\sqrt{2\pi}\Sigma_iE\left(-\frac{\left(\sum_{\nu=1}^{k}\eta^\nu c_i^\nu+2zf_i\right)g_i}{2\sqrt{f_i\left(1-f_i\right)}}\right)}e^{-\left(\frac{\tilde{J}}{\sqrt{2}\Sigma_i}+\frac{\left(\sum_{\nu=1}^{k}\eta^\nu c_i^\nu+2zf_i\right)}{2\sqrt{f_i(1-f_i)}}\right)^2} \quad (35)$$

Solution in the high-weight regime considered in the main text

In the main text, we consider a simplified problem in which there are only two classes of neurons, one inhibitory and one excitatory, each neuron has a single l_1-norm constraint on the weights of its inputs ($\nu=1$, $\tilde{w}^\nu=\tilde{w}$, $\eta^\nu=\eta$, $c_j^1=g_j$), and the firing probabilities are homogeneous ($f_j=f_{out}=f$). In this case, we introduce new variables, $v_\pm=\frac{-\eta\pm 2zf}{2\sqrt{f\left(1-f\right)}}$ and $\sigma=\Sigma_i/\tilde{w}$, and the general solution of Eqs. (29, 35) simplifies significantly:

$$\alpha\left(\tilde{w},\frac{N_{inh}}{N},f,\tilde{\kappa}\right)=\frac{\tilde{\kappa}^2}{\tilde{w}^2 f\left(1-f\right)}\frac{2}{\sigma^2\left(u_+ +u_-\right)^2}\frac{fD\left(u_-\right)+(1-f)D\left(u_+\right)}{\left(fE\left(u_-\right)+(1-f)E\left(u_+\right)\right)^2}$$

$$P^{con}_{exc/inh}\left(\tilde{w},\frac{N_{inh}}{N},f,\tilde{\kappa}\right)=E\left(v_{-/+}\right)$$

$$p^{PSP}_{exc/inh}\left(\tilde{J}\mid\tilde{w},\frac{N_{inh}}{N},f,\tilde{\kappa}\right)=\frac{\theta\left((+/-)\tilde{J}\right)}{\sqrt{2\pi}\sigma\tilde{w}E\left(v_{-/+}\right)}e^{-\left(\frac{\tilde{J}}{\sqrt{2}\sigma\tilde{w}}-(+/-)v_{-/+}\right)^2}$$

$$\begin{cases} fF\left(u_-\right)-(1-f)F\left(u_+\right)=0 \\ \frac{N_{exc}}{N}F\left(v_-\right)+\frac{N_{inh}}{N}F\left(v_+\right)=\frac{\sqrt{2}}{\sigma} \\ \frac{N_{exc}}{N}F\left(v_-\right)-\frac{N_{inh}}{N}F\left(v_+\right)=\left\{\frac{1}{\tilde{w}f},0\right\}\frac{\sqrt{2}}{\sigma} \\ \frac{N_{exc}}{N}D\left(v_-\right)+\frac{N_{inh}}{N}D\left(v_+\right)=\frac{\tilde{\kappa}^2}{\tilde{w}^2 f\left(1-f\right)}\frac{2}{\sigma^2\left(u_+ +u_-\right)^2} \\ \sigma=\frac{\tilde{\kappa}^2}{\tilde{w}^2 f\left(1-f\right)}\frac{\sqrt{2}}{\left(u_+ +u_-\right)\left(\left\{\frac{1}{\tilde{w}f},0\right\}\left(v_+ -v_-\right)-\left(v_+ +v_-\right)\right)}\frac{fF\left(u_-\right)+(1-f)F\left(u_+\right)}{fE\left(u_-\right)+(1-f)E\left(u_+\right)} \\ u_+ +u_- >0;\quad \sigma>0 \end{cases} \tag{36}$$

Note that the difference between the replica solutions of the associative and balanced models explicitly appears only in two places (curly brackets) in Eqs. (36).

These equations make it clear that the solution of the associative model in the high-weight limit, $\tilde{w}f\gg 1$, converges to the solution of the balanced model. However, since the value of $\tilde{w}f$ estimated from experimental data is large but finite (see the main text), we also examined the agreement between the results of the two models by solving Eqs. (36) numerically for different values of $\tilde{w}$ and $\tilde{\kappa}$. Figure S1 shows that for values of $\tilde{w}$ in the [10 100] range, results of the two models agree within 10%, and the agreement improves with increasing $\tilde{\kappa}$. What is more, in the limit of high-weight the solution depends only on $\tilde{\kappa}/\tilde{w}$ (straight isocontour lines in Figure S1), or alternatively, on a parameter $\rho=\frac{\tilde{\kappa}}{\tilde{w}\sqrt{f\left(1-f\right)}}$. The latter was introduced by Brunel, et al [5,6] and is referred to as the rescaled robustness. Parameter ρ can serve as a proxy for the ratio of robustness and standard deviation in postsynaptic input, κ/σ_{input}. With this, Eqs. (36) simplify further:

$$\alpha\left(\frac{N_{inh}}{N},f,\rho\right)=\frac{2\rho^2}{\sigma^2\left(u_++u_-\right)^2}\frac{fD\left(u_-\right)+(1-f)D\left(u_+\right)}{\left(fE\left(u_-\right)+(1-f)E\left(u_+\right)\right)^2}$$

$$P^{con}_{exc/inh}\left(\frac{N_{inh}}{N},f,\rho\right)=E\left(v_{-/+}\right)$$

$$p^{PSP}_{exc/inh}\left(\tilde{J}\mid\tilde{w},\frac{N_{inh}}{N},f,\rho\right)=\frac{\theta\left((+/-)\tilde{J}\right)}{\sqrt{2\pi}\sigma\tilde{w}E\left(v_{-/+}\right)}e^{-\left(\frac{\tilde{J}}{\sqrt{2}\sigma\tilde{w}}-(+/-)v_{-/+}\right)^2}$$

$$\begin{cases}fF\left(u_-\right)-(1-f)F\left(u_+\right)=0\\ \frac{N_{exc}}{N}F\left(v_-\right)+\frac{N_{inh}}{N}F\left(v_+\right)=\frac{\sqrt{2}}{\sigma}\\ \frac{N_{exc}}{N}F\left(v_-\right)-\frac{N_{inh}}{N}F\left(v_+\right)=0\\ \frac{N_{exc}}{N}D\left(v_-\right)+\frac{N_{inh}}{N}D\left(v_+\right)=\frac{2\rho^2}{\sigma^2\left(u_++u_-\right)^2}\\ \sigma=\frac{-\sqrt{2}\rho^2}{\left(v_++v_-\right)\left(u_++u_-\right)}\frac{fF\left(u_-\right)+(1-f)F\left(u_+\right)}{fE\left(u_-\right)+(1-f)E\left(u_+\right)}\\ u_++u_->0;\quad v_++v_-<0\end{cases}\tag{37}$$

We note that since $\frac{\tilde{\kappa}}{\tilde{w}}=\frac{\kappa}{w\sqrt{N}}$ for both models, rescaled robustness, ρ, and Eqs. (37) are model independent.

The fact that the solutions of the two models converge in the high-weight regime is not surprising. In this regime, $Nwf\gg h$, and, as a result, mean excitatory and inhibitory inputs to the neuron are much greater than the threshold of firing. One can show that in this case h in Eqs. (2) can be disregarded, and the solution becomes independent of scaling of J with N.

Numerical solution for finite *N*

The problem of Eqs. (2) is convex, and hence, it can be solved numerically within the standard constrained optimization framework. Numerical solutions are obtained for finite networks, and the results are independent on the assumptions of scaling of model parameters with N.

Below, we consider two learning scenarios: (*i*) feasible load, in which associations can be learned with specified robustness, and (*ii*) non-feasible load, in which the number of presented associations is so large that Eqs. (2) have no solution.

In the case of feasible load, the region of solutions is nonempty, and one must employ additional considerations to limit the results to a single, "optimal" solution. We do this by choosing the solution that minimizes $\|J\|_2^2$,

$$\begin{aligned}
&\min\left(\sum_{j=1}^{N} J_j^2\right)\\
&\left(2y^\mu - 1\right)\left(\sum_{j=1}^{N} J_j X_j^\mu - h\right) \geq \kappa, \quad \mu = 1,\ldots,m\\
&\sum_{j=1}^{N} J_j c_j^\nu = N w^\nu, \quad \nu = 1,\ldots,k\\
&J_j g_j \geq 0, \quad j = 1,\ldots,N\\
&\text{Prob}\left(X_j^\mu\right) = \begin{cases} 1-f_j, & X_j^\mu = 0 \\ f_j, & X_j^\mu = 1 \end{cases}; \quad \text{Prob}\left(y^\mu\right) = \begin{cases} 1-f_{out}, & y^\mu = 0 \\ f_{out}, & y^\mu = 1 \end{cases}
\end{aligned} \tag{38}$$

In the non-feasible case, we choose the solution that minimizes the sum of deviations, s^μ, of not robustly learned associations from the corresponding margin boundaries. This solution can be obtained by solving the following linear optimization problem:

$$\begin{aligned}
&\min\left(\sum_{\mu=1}^{m} s^\mu\right)\\
&\left(2y^\mu - 1\right)\left(\sum_{j=1}^{N} J_j X_j^\mu - h\right) + s^\mu \geq \kappa, \quad \mu = 1,\ldots,m\\
&s^\mu \geq 0\\
&\sum_{j=1}^{N} J_j c_j^\nu = N w^\nu, \quad \nu = 1,\ldots,k\\
&J_j g_j \geq 0, \quad j = 1,\ldots,N\\
&\text{Prob}\left(X_j^\mu\right) = \begin{cases} 1-f_j, & X_j^\mu = 0 \\ f_j, & X_j^\mu = 1 \end{cases}; \quad \text{Prob}\left(y^\mu\right) = \begin{cases} 1-f_{out}, & y^\mu = 0 \\ f_{out}, & y^\mu = 1 \end{cases}
\end{aligned} \tag{39}$$

The problems outlined in Eqs. (38, 39) were solved in MATLAB in the following sequence of steps. Given the associative memory load, $\alpha = m/N$, we first solved the problem of Eqs. (39), utilizing the *linprog.m* function, to find the distances, s^μ. If some of these distances are greater than zero, the problem is non-feasible. If all $s^\mu = 0$, the problem is feasible, in which case we used connection weights resulting from Eqs. (39) as a starting configuration and solved Eqs. (38) by using the *quadprog.m* function.

Figure S2 shows the results of numerical simulations for neurons loaded to capacity at different values of rescaled robustness. Results for N = 200, 400, and 800 inputs are shown together with the replica theory solution ($N \to \infty$). With increasing N, the numerical solutions gradually approach the results of the replica theory, which serves as an independent validation of numerical and theoretical calculations.

The dependence of structural and dynamical properties of associative networks of $N = 200$, 400, and 800 neurons on memory load and rescaled robustness is shown in Figures S3, S4, and S6. These maps do not depend strongly on N, suggesting that the network properties would not change significantly if the network size was increased further, e.g. to a few thousand neurons.

Numerical values of model parameters and results

In this section, we provide values of various parameters related to network structure and dynamics. This is done for the two network settings corresponding to the red and green asterisks in Figures 2-5 of the main text.

	Parameter name	Red asterisk	Green asterisk
Input parameters	Number of neurons, N	800	800
	Inhibitory neuron fraction, N_{inh}/N	0.20	0.20
	Firing probability, f	0.20	0.20
	Threshold of firing, h	20 mV	20 mV
	Scaled average absolute connection weight, $\tilde{w}$	70	70
	Scaled memory load, α/α_c	0.90	0.90
	Rescaled robustness, ρ	0.50	1.3
Output parameters	Memory load, α	0.38	0.20
	Robustness parameter, κ	25 mV	64 mV
	Excitatory connection probability, P_{exc}^{con}	0.26	0.14
	Inhibitory connection probability, P_{inh}^{con}	0.66	0.46
	CV of excitatory connection weights	0.89	0.99
	CV of inhibitory connection weights	0.75	0.86
	Average number of steps to limit cycle	32	2.3×10^4
	CV of ISI	0.67	0.88
	Spike cross-correlation coefficient	0.24	0.09
	Excitatory input (mean ± s.d.)	133 ± 14 mV	125 ± 38 mV
	Inhibitory input (mean ± s.d.)	−144 ± 35 mV	−155 ± 48 mV
	Total input (mean ± s.d.)	−11 ± 38 mV	−30 ± 60 mV
	Exc.-inh. input correlation coefficient	-0.96	-0.79
	Sequence retrieval probability	0.08	1
	Noise tolerance, $\sigma_{noise}/\sigma_{input}$	0	0.35

Table S1: Input and output model parameters corresponding to the red and green asterisks in Figures 2-5 of the main text.

Figure S7 illustrates partial cancellation of excitatory and inhibitory inputs and the relationships among the total input, firing threshold, and robustness parameter based on the values from Table S1.

REFERENCES

1 Edwards, S. F. & Anderson, P. W. Theory of spin glasses. *J. Phys. F: Metal Phys.* **5**, 965-974 (1975).

2 Sherrington, D. & Kirkpatrick, S. Solvable model of a spin glass. *Physical Review Letters* **35**, 1792-1796 (1975).

3 Boyd, S. P. & Vandenberghe, L. *Convex optimization*. (Cambridge University Press, 2004).

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

5 Brunel, N., Hakim, V., Isope, P., Nadal, J. P. & Barbour, B. Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. *Neuron* **43**, 745-757 (2004).

6 Brunel, N. Is cortical connectivity optimized for storing information? *Nature neuroscience* **19**, 749-755, doi:10.1038/nn.4286 (2016).

7 Chapeton, J., Fares, T., LaSota, D. & Stepanyants, A. Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons. *Proc Natl Acad Sci U S A* **109**, E3614-3622, doi:10.1073/pnas.1211467109 (2012).

8 Chapeton, J., Gala, R. & Stepanyants, A. Effects of homeostatic constraints on associative memory storage and synaptic connectivity of cortical circuits. *Front Comput Neurosci* **9**, 74, doi:10.3389/fncom.2015.00074 (2015).

9 Rubin, R., Abbott, L. F. & Sompolinsky, H. Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. *Proc Natl Acad Sci U S A* **114**, E9366-E9375, doi:10.1073/pnas.1705841114 (2017).

10 Stepanyants, A. & Chklovskii, D. B. Neurogeometry and potential synaptic connectivity. *Trends Neurosci* **28**, 387-394, doi:S0166-2236(05)00131-1 [pii] 10.1016/j.tins.2005.05.006 (2005).

11 Stepanyants, A. *et al.* Local potential connectivity in cat primary visual cortex. *Cereb Cortex* **18**, 13-28, doi:bhm027 [pii] 10.1093/cercor/bhm027 (2008).

12 McCulloch, W. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biol* **5**, 115 - 133 (1943).

13 Gal, E. *et al.* Rich cell-type-specific network topology in neocortical microcircuitry. *Nature neuroscience* **20**, 1004-1013, doi:10.1038/nn.4576 (2017).

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

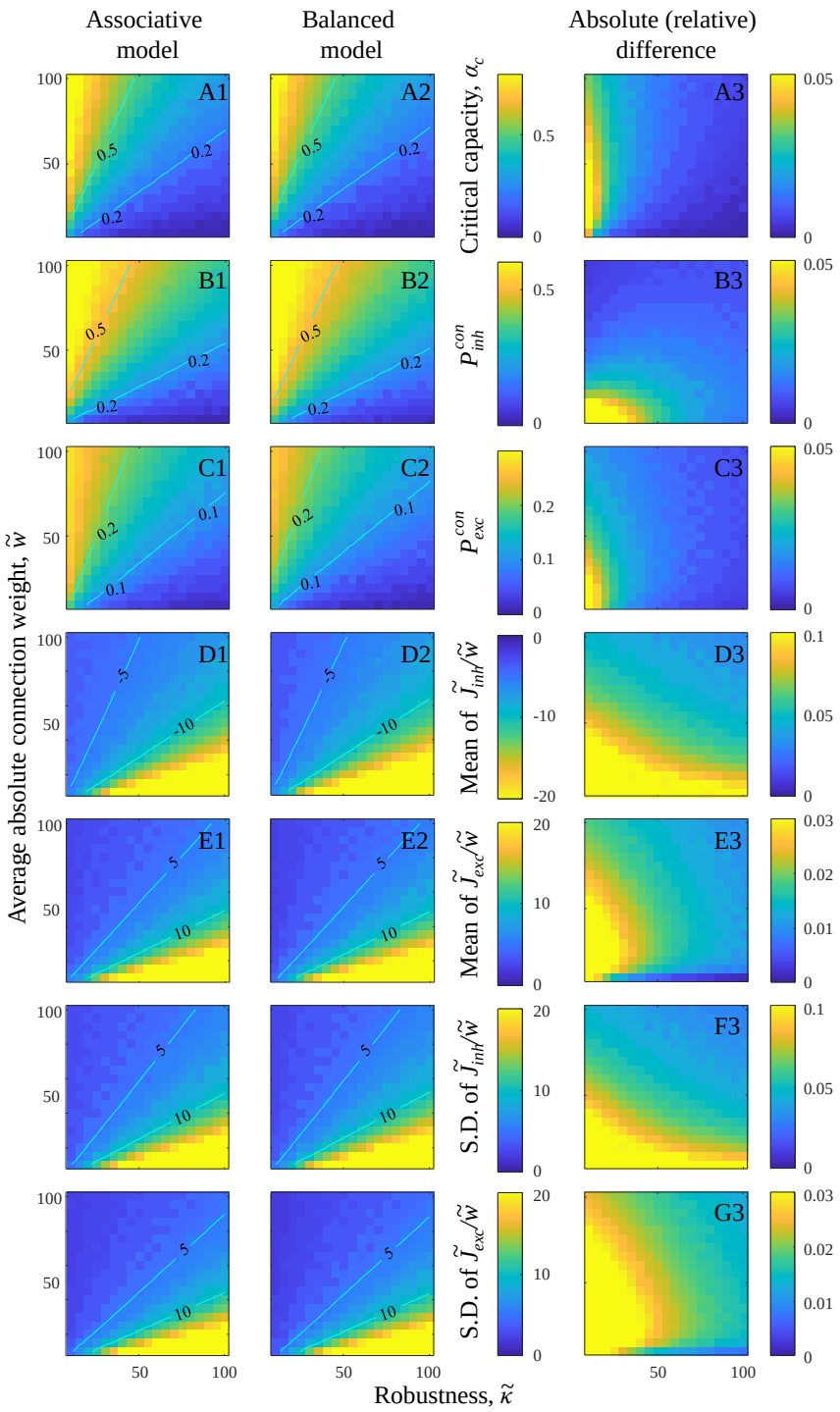

Figure S1: Theoretical solutions for the associative and balanced models converge in the limit of $\tilde{w}f \gg 1$. In this limit, model results depend only on $\tilde{\kappa}/\tilde{w}$, in agreement with Eqs. (37). **A.** Maps of critical capacity as functions of $\tilde{\kappa}$ and $\tilde{w}$ for the associative (A1) and balanced (A2) models. Straight isocontours confirm that the results depend only on $\tilde{\kappa}/\tilde{w}$. Absolute difference of the two maps (A3) shows that critical capacities of the two models converge in the limit of $\tilde{w}f \gg 1$. Same for the probabilities of inhibitory **(B)** and excitatory **(C)** connections. **D.** Maps of mean, non-zero, inhibitory connection weights as a function of $\tilde{\kappa}$ and $\tilde{w}$ for the associative (D1) and balanced (D2) models, as well as the absolute relative difference of these maps (D3). Same for the mean, non-zero, excitatory connection weights **(E)**, and standard deviations of non-zero inhibitory **(F)** and excitatory **(G)** connection weights.

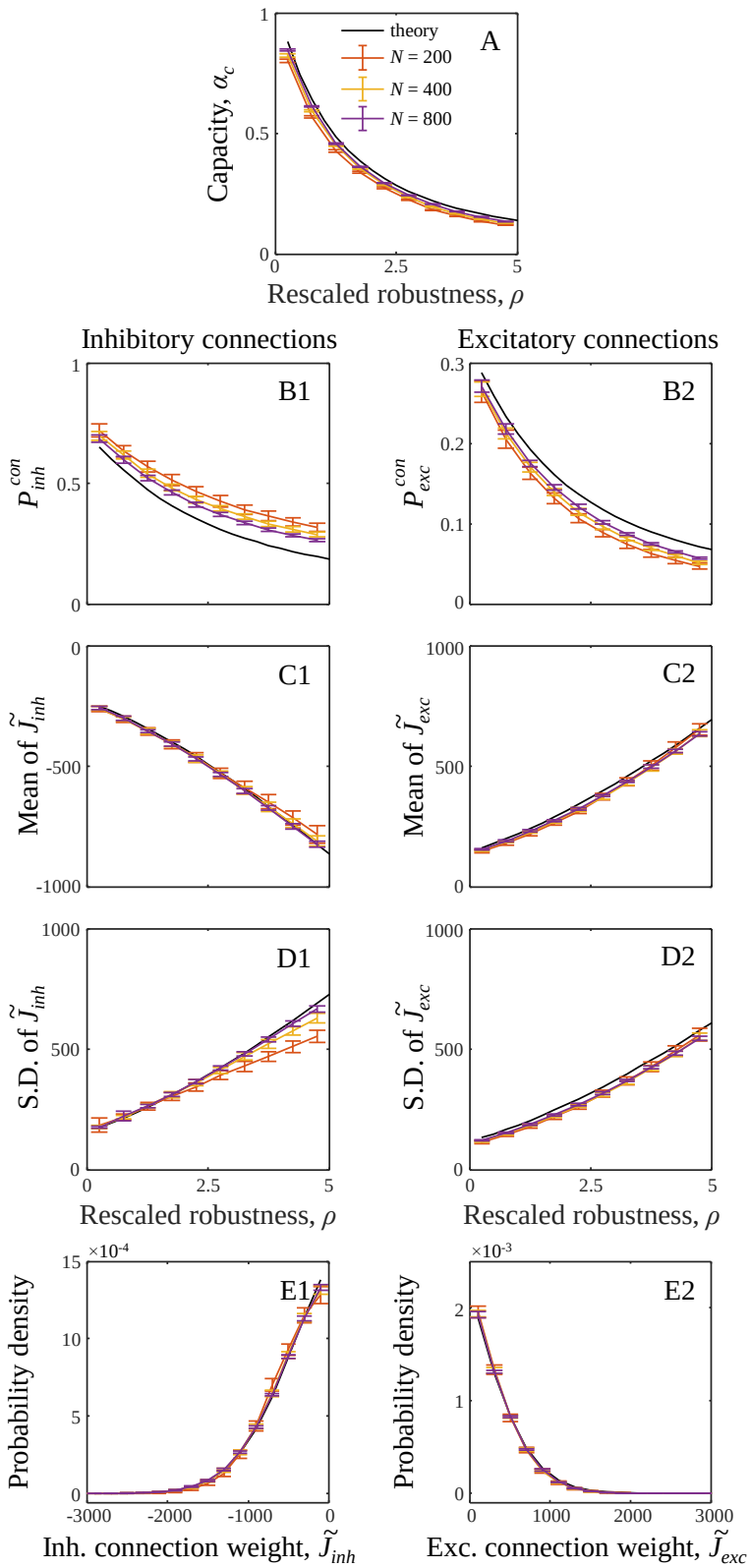

Figure S2: Validation of theoretical results of Eqs. (37) with numerical simulations performed for N = 200, 400, and 800 inputs. **A.** Capacity as a function of rescaled robustness, ρ. With increasing N, numerical results (error-bars) obtained with convex optimization [Eqs. (38, 39)] approach the theoretical solution (black line). Error-bars indicate standard deviations calculated based on $100N$ simulations. Same for the probabilities of non-zero inhibitory **(B1)** and excitatory **(B2)** connections, mean non-zero inhibitory **(C1)** and excitatory **(C2)** connection weights, and standard deviations of non-zero inhibitory **(D1)** and excitatory **(D2)** connection weights. Panels **(E1)** and **(E2)** illustrate the match between theoretical and numerical probability densities of non-zero inhibitory and excitatory connection weights for ρ = 1.

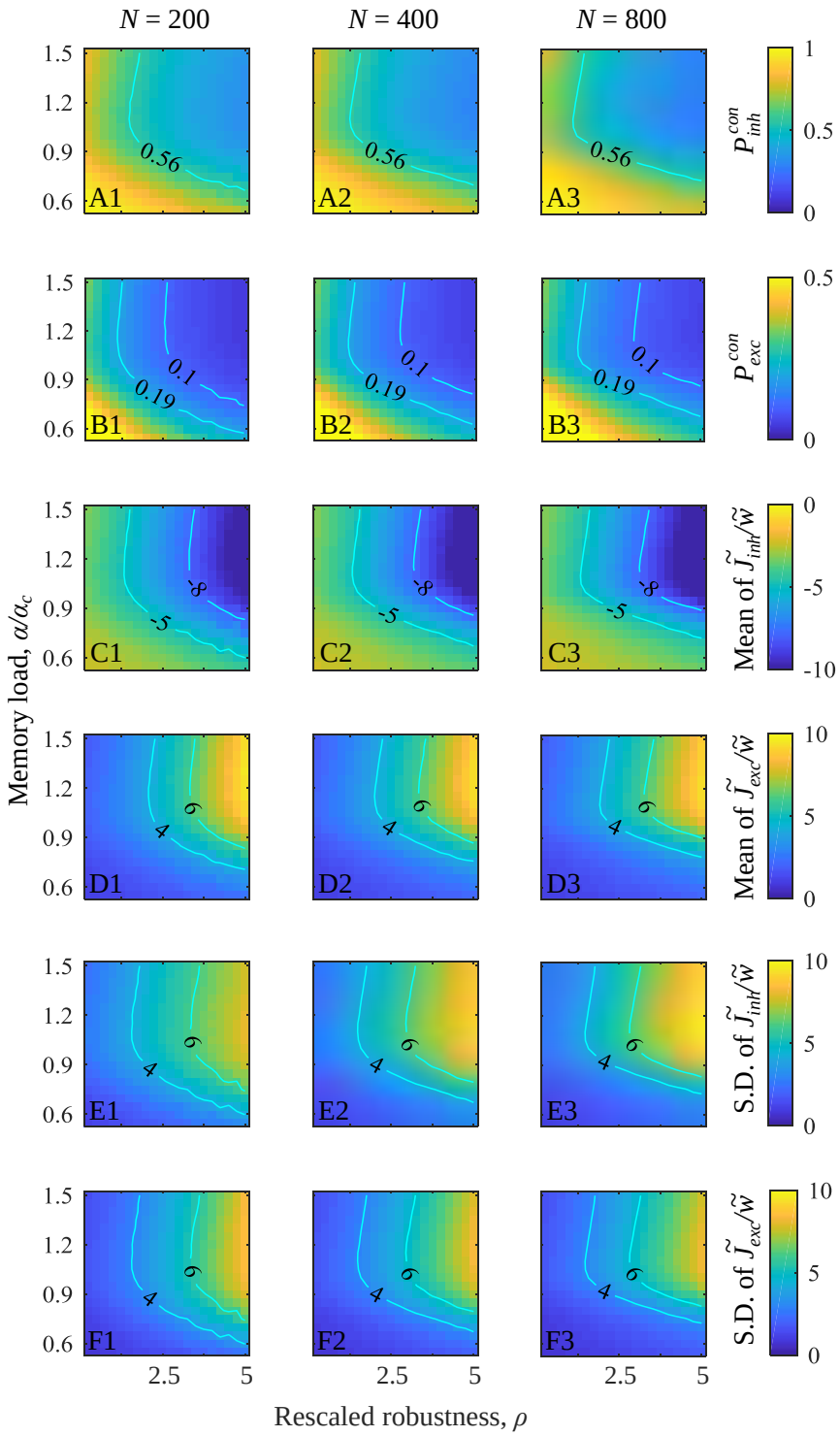

Figure S3: Properties of neuron-to-neuron connectivity in associative networks. **A.** Inhibitory connection probabilities as functions of rescaled robustness and relative memory load in networks of N = 200 (A1), N = 400 (A2), and N = 800 (A3) neurons. Same for excitatory connection probabilities (**B**), means of non-zero inhibitory (**C**) and excitatory (**D**) weights ($\tilde{J}_{inh} / \tilde{w}$ and $\tilde{J}_{exc} / \tilde{w}$), and standard deviations of inhibitory (**E**) and excitatory (**F**) weights. Isocontour lines match those used in the main text.

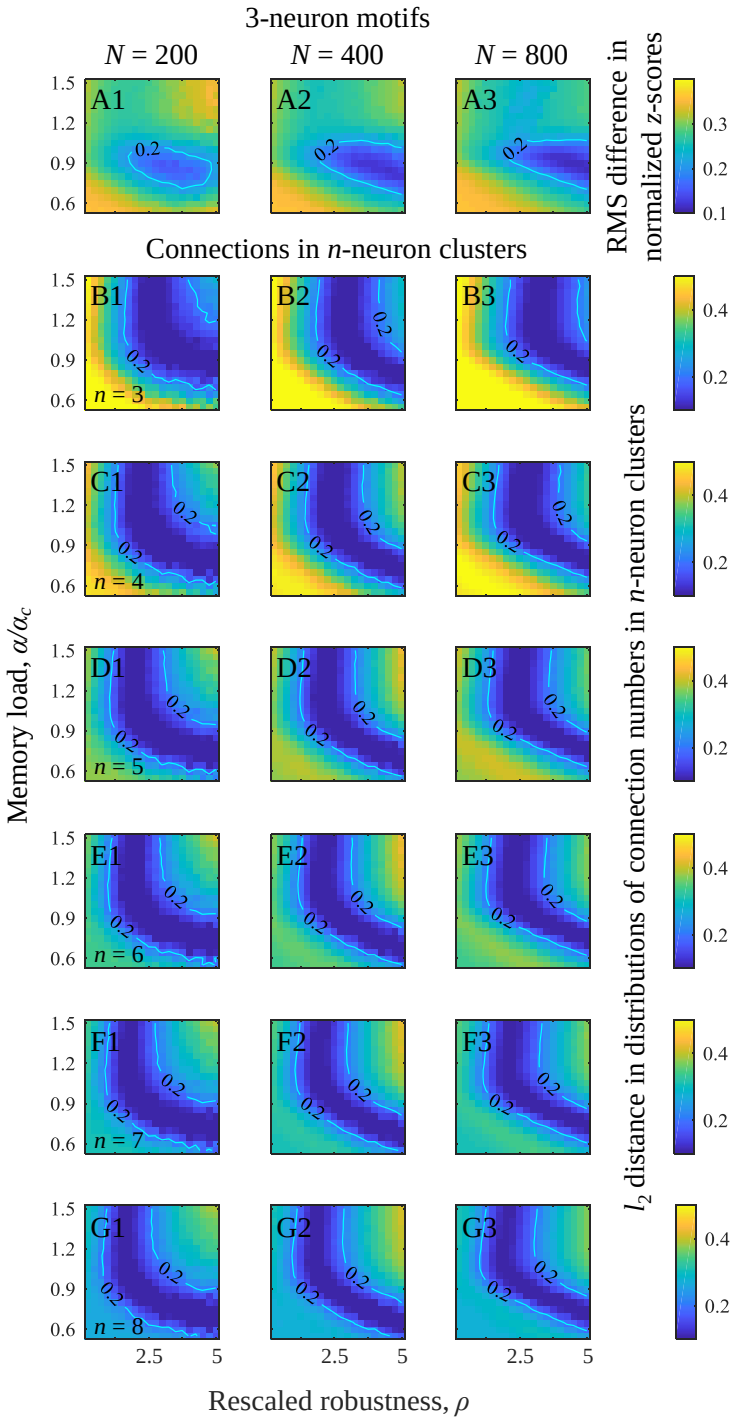

Figure S4: Dependence of higher-order structural properties of associative networks on network size. **A.** Maps of root-mean-square (RMS) differences between normalized *z*-scores of excitatory 3-neuron motifs observed in associative networks of N = 200, 400, and 800 neurons and those reported by the Blue Brain project [13]. **B.-G.** Maps of l_2 distances between distributions of connection numbers in 3-8 excitatory neuron clusters in associative networks of N = 200, 400, and 800 neurons and those observed experimentally [14]. Isocontour lines match those used in the main text.

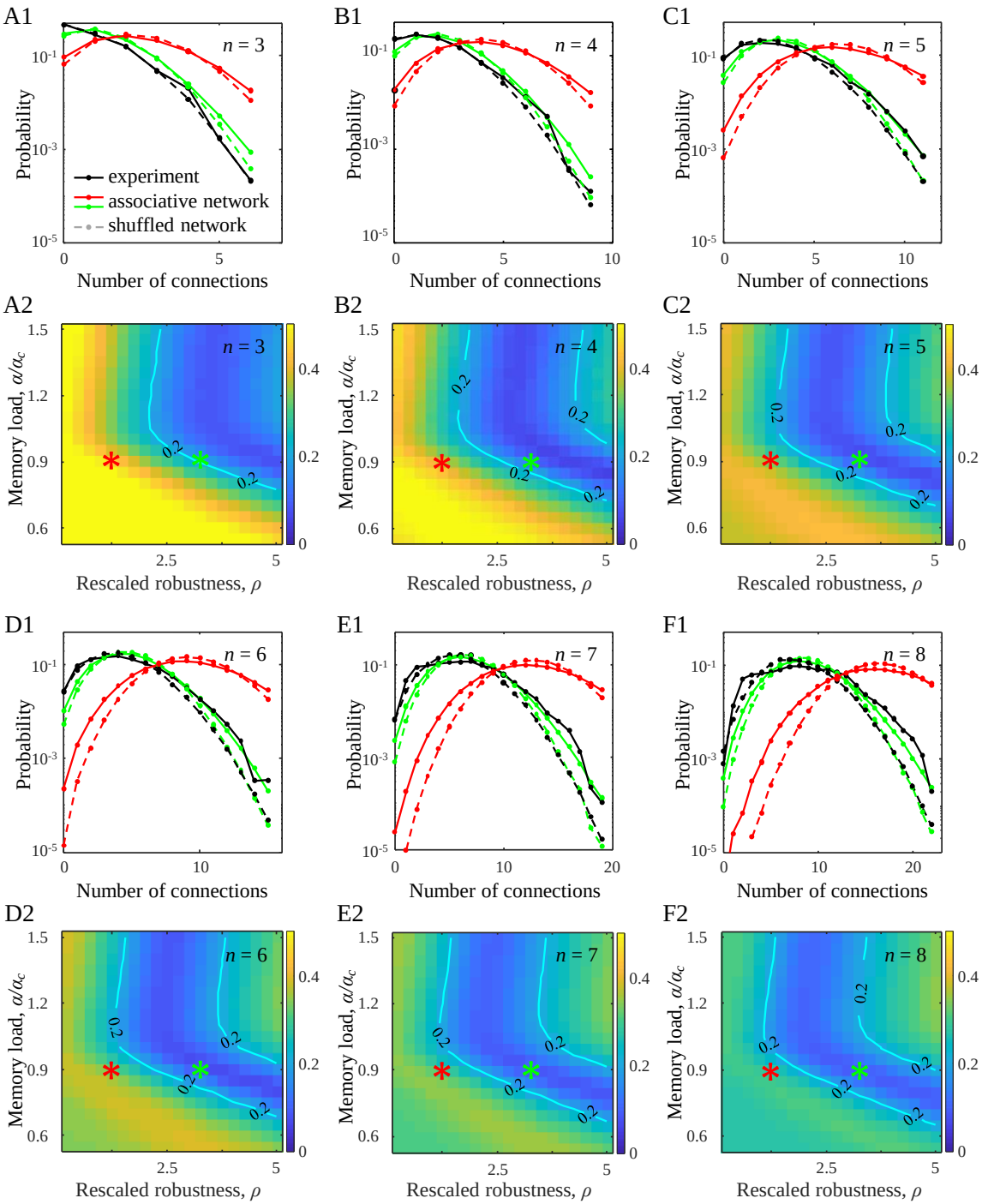

Figure S5: Distributions of non-zero connection numbers in clusters of 3-8 excitatory neurons in associative networks. **A1-F1.** Solid red and green lines illustrate distributions obtained in associative networks for the parameter settings indicated by the red and green asterisks. Solid black curves indicate the corresponding results for local cortical networks based on electrophysiological measurements [2]. Dashed lines show distributions in randomly shuffled networks (see Methods for details). **A2-F2.** Maps of l_2 distances between connection number distributions in associative and cortical networks [2]. Numerical results of the associative model were generated based on networks of $N = 800$ neurons.

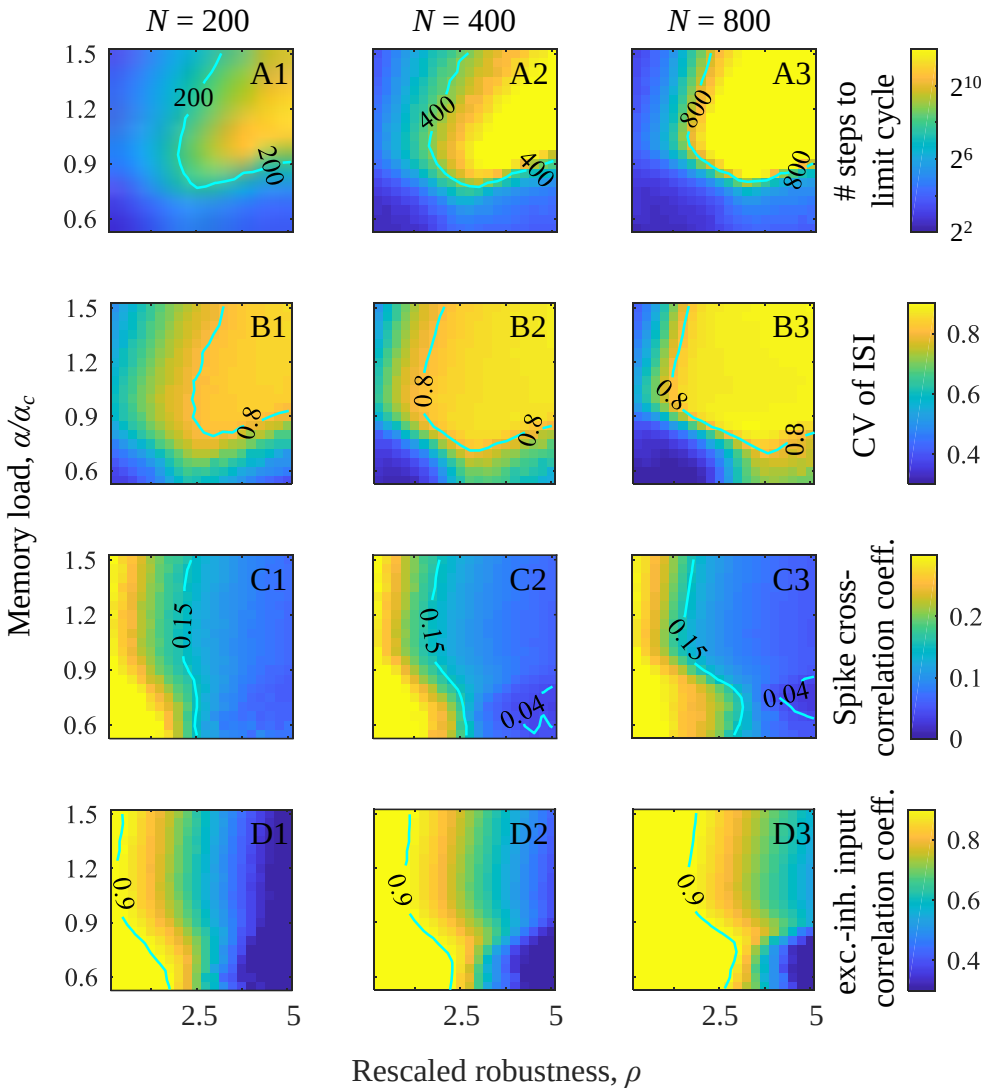

Figure S6: Dependence of dynamical properties of associative networks on network size. **A.** Maps for the average durations of transient network dynamics as functions of rescaled robustness and relative memory load for networks of N = 200, 400, and 800 neurons. Same for CV values of inter-spike-intervals (ISI) (**B**), spike cross-correlation coefficients (**C**), and correlation coefficients of excitatory and inhibitory inputs (**D**). Isocontour lines match those used in the main text.

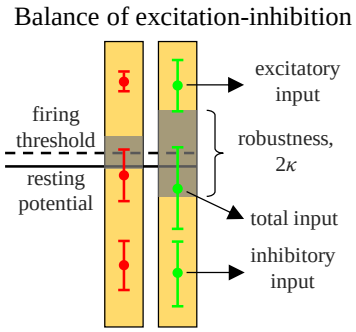

Figure S7: Excitatory and inhibitory inputs in relation to firing threshold and robustness. Left and right halves of the figure are based on the data from Table S1, and correspond to the red and green asterisks from Figures 2-5 of the main text. The average excitatory and inhibitory inputs are much larger than the threshold of firing. However, the total input lies within one standard deviation from the firing threshold due to a partial cancellation of its excitatory and inhibitory components.