eNeuro

Research Article: New Research

Cognition and Behavior

Noise in Neurons and Synapses Enables Reliable Associative Memory Storage in Local Cortical Circuits

Chi Zhang,[1] Danke Zhang,[1,2] and Armen Stepanyants[1]

https://doi.org/10.1523/ENEURO.0302-20.2020

[1]Department of Physics and Center for Interdisciplinary Research on Complex Systems, Northeastern University, Boston, MA 02115 and [2]CAS Key Laboratory of Brain Connectome and Manipulation, Interdisciplinary Center for Brain Information, The Brain Cognition and Brain Disease Institute, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions, Shenzhen, Guangdong China

Abstract

Neural networks in the brain can function reliably despite various sources of errors and noise present at every step of signal transmission. These sources include errors in the presynaptic inputs to the neurons, noise in synaptic transmission, and fluctuations in the neurons' postsynaptic potentials (PSPs). Collectively they lead to errors in the neurons' outputs which are, in turn, injected into the network. Does unreliable network activity hinder fundamental functions of the brain, such as learning and memory retrieval? To explore this question, this article examines the effects of errors and noise on the properties of model networks of inhibitory and excitatory neurons involved in associative sequence learning. The associative learning problem is solved analytically and numerically, and it is also shown how memory sequences can be loaded into the network with a biologically more plausible perceptron-type learning rule. Interestingly, the results reveal that errors and noise during learning increase the probability of memory recall. There is a trade-off between the capacity and reliability of stored memories, and, noise during learning is required for optimal retrieval of stored information. What is more, networks loaded with associative memories to capacity display many structural and dynamical features observed in local cortical circuits in mammals. Based on the similarities between the associative and cortical networks, this article predicts that connections originating from more unreliable neurons or neuron classes in the cortex are more likely to be depressed or eliminated during learning, while connections onto noisier neurons or neuron classes have lower probabilities and higher weights.

Key words: associative learning; memory retrieval; perceptron; replica; spiking errors; synaptic noise

Significance Statement

Signal transmission in the brain is accompanied by many sources of errors and noise, and yet, neural networks can reliably store memories. This article argues that noise should not be viewed as a nuisance, but that it is an essential component of the reliable learning mechanism implemented by the brain. The article describes a network model of associative sequence learning, showing that for optimal retrieval of stored information learning must be conducted in the presence of noise. To validate the model, it is shown that associative memories can be loaded into the network with an online perceptron-type learning rule and that networks loaded to capacity develop many structural and dynamical properties observed in the brain.

Received July 7, 2020; accepted December 16, 2020; First published January 6, 2021.
The authors declare no competing financial interests.

Author contributions: A.S. designed research; C.Z., D.Z., and A.S. performed research; C.Z. analyzed data; C.Z., D.Z., and A.S. wrote the paper.

Introduction

Brain networks can reliably store and retrieve long-term memories despite the facts that various sources of errors and noise accompany every step of signal transmission through the network (Faisal et al., 2008), synaptic connectivity changes over time (Trachtenberg et al., 2002; Holtmaat and Svoboda, 2009; Gala et al., 2017), and extraneous sensory inputs are usually present during memory recall. The brain can reduce the effects of noise and extraneous inputs by attending to the memory retrieval process (Cohen and Maunsell, 2009; Mitchell et al., 2009), but such hindrances cannot be eliminated entirely. Therefore, the reliability required for memory retrieval must be built into the network during learning. This proposal presents an interesting challenge. Traditional supervised learning models, such as the ones that rely on the perceptron rule (Minsky and Papert, 1969; Hertz et al., 1991), modify connectivity only when a neuron's output deviates from its target output. Thus, in such models learning stops as soon as the neuron produces the desired response and, subsequently, there is no possibility for improving the response reliability. The network connection weights in such models may end up near the boundary of the solution region, and a small amount of noise during memory retrieval can lead to errors or completely disrupt the retrieval process. More reliable solutions are located farther away from the solution region boundary, but the perceptron rule is not guaranteed to find them. Thus, it is not clear how the neural networks in the brain manage not only to learn but also to do it reliably.

In the case of associative memory storage, reliability can be incorporated into the perceptron learning rule by means of a generic robustness parameter (Brunel et al., 2004). This traditional description, however, is not biologically motivated and does not account for various types of errors and noise present during learning and memory retrieval (Fig. 1*A*). A more comprehensive account must include errors in the inputs to the neurons, combine them with fluctuations in the neurons' presynaptic connection weights and intrinsic sources of noise, and produce spiking errors in the neurons' outputs. The latter, injected back into the network, give rise to input errors in the next time step. The recurrence of errors presents a clear challenge for the retrieval of associative memory sequences considered in this study. If not corrected at every step of the retrieval process, errors in the network activity can amplify over time and lead to an irreversible deviation of the retrieved trajectory from the loaded sequence, i.e., a partially retrieved memory.

This work was supported by the Air Force Office of Scientific Research Grant FA9550-15–1-0398 and the National Science Foundation Grant IIS-1526642.

Correspondence should be addressed to Armen Stepanyants at a.stepanyants@neu.edu.

https://doi.org/10.1523/ENEURO.0302-20.2020

The premise of this article is that errors and noise are essential components of the reliable learning mechanism implemented in the brain. As different fluctuations accompany the presentation of the same learning example to a neuron on different trials, the neuron in effect never stops learning. Its connection weights move further away from the solution region boundary every time a progressively larger fluctuation is encountered. This process increases the reliability of the loaded memory which can later be retrieved in the presence of noise. Similar ideas have been successfully used in machine learning where an augmentation of training examples with noise (Bishop, 1995) and dropping out neurons and connections (Srivastava et al., 2014) during training have been shown to significantly reduce both overfitting and training time. And, there are many other examples in which noise is put to a constructive use to improve various functions of physical and neural systems (for review, see Gammaitoni et al., 1998; Stein et al., 2005; McDonnell and Abbott, 2009; McDonnell and Ward, 2011). Therefore, the hypothesis that errors and noise are exploited by the brain for reliable memory storage may not be entirely surprising. Still, this hypothesis requires careful quantitative evaluation and validation with experimental data, which is the focus of this study.

Materials and Methods

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

We considered a model of associative sequence learning by a local ($\sim$100 μm in size), all-to-all potentially (structurally) connected (Stepanyants and Chklovskii, 2005; Stepanyants et al., 2008) cortical network, albeit with no synaptic input originating from outside the circuit. The model network consisted of N_{inh} inhibitory and ($N - N_{inh}$) excitatory McCulloch and Pitts neurons (McCulloch and Pitts, 1943; Fig. 1*A*) and was faced with a task of learning a sequence of consecutive network states, $X^1 \rightarrow X^2 \rightarrow \ldots X^{m+1}$, in which X^μ is a binary vector representing target activities of all neurons at a time step μ, and the ratio m/N is referred to as the memory load. Some assumptions and approximations of the model are discussed in (Chapeton et al., 2012). During learning, individual neurons had to independently learn to associate the inputs they received from the network with the corresponding target outputs derived from the associative memory sequence. The neurons learned these input-output associations by adjusting the weights of their input connections, J_{ij} (weight of connection from neuron j to neuron i). In contrast to previous studies, we accounted for the fact that learning in the brain is accompanied by several sources of errors and noise. Within the model, these sources are divided into three categories (Fig. 1*A*, orange lightning signs): (1) input spiking errors, or errors in X^μ, (2) synaptic noise, or noise in J_{ij}, and (3) intrinsic noise, which combines all other sources of noise affecting the neurons' postsynaptic potentials (PSPs). The last category includes background synaptic activity and the stochasticity of ion channels. In the model, this category is equivalent to noise in the neurons' thresholds of firing, h_i

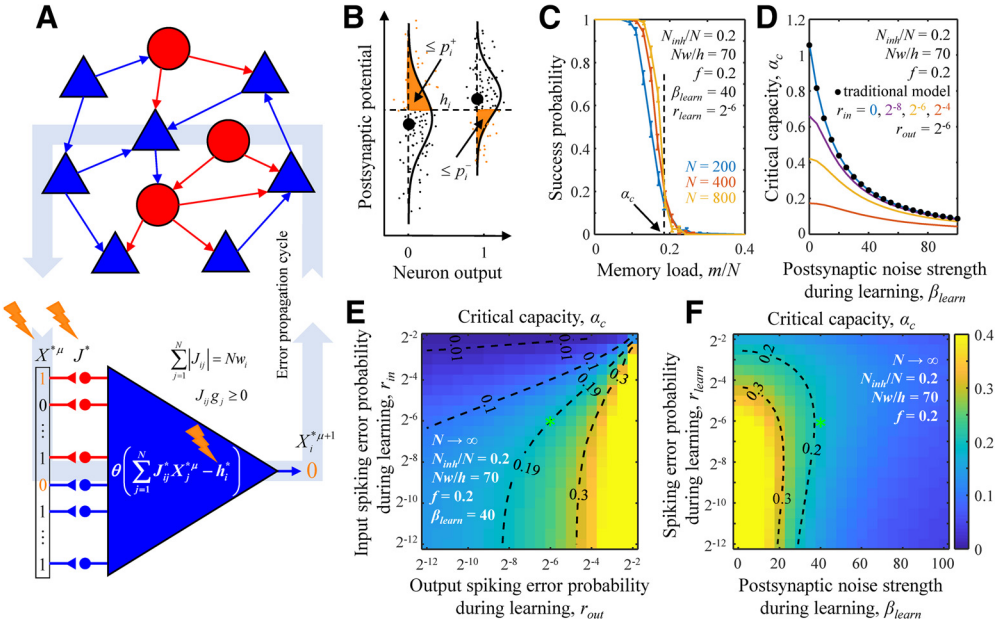

Figure 1. Associative memory storage in a recurrent network of inhibitory and excitatory neurons in the presence of errors and noise. ***A***, Error propagation through the network. Inhibitory neurons (red circles) and excitatory neurons (blue triangles) form an all-to-all potentially (structurally) connected network. Red and blue arrows represent actual (functional) connections. Spiking errors (errors contained in X^*), synaptic noise (J_{ij}^*), and intrinsic noise (h_i^*) accompany signal transmission (orange lightning signs). Errors in the neurons' outputs at a given time step become spiking errors in the next time step. ***B***, Fluctuations in PSPs for two associations with target neuron outputs 0 (left) and 1 (right). Large black dots denote PSPs in the absence of errors and noise. Small dots represent PSPs on different trials in the presence of errors and noise. Orange areas to the left of the PSP probability densities (solid lines) represent the probabilities of erroneous spikes (left) and spike failures (right). ***C***, The probability of successful learning by a neuron is a sharply decreasing function of memory load m/N. Solid curves represent the probabilities of successful learning obtained with nonlinear optimization (see Materials and Methods) for neurons receiving $N = 200$, 400, and 800 homogeneous inputs. The numerical values of β_{learn} and $r_{in} = r_{out} \equiv r_{learn}$ are provided in the figure. The values of all other parameters of the model were adapted from Chapeton et al. (2015). At 0.5 success probability, the neuron is said to be loaded to capacity, α. The dashed black line represents the theoretical (critical) capacity, α_c, obtained with the replica method in the $N \to \infty$ limit. ***D***, α_c as a function of β_{learn} for different input noise strengths (colored lines). In the case of $r_{in} = 0$, solution of Equation 1 (blue line) coincides with the solution of the traditional model (Zhang et al., 2019b), which uses a generic robustness parameter (black dots). ***E***, Map of α_c for a neuron receiving homogeneous input as a function of r_{in} and r_{out}. ***F***, Same as a function of β_{learn} and $r_{in} = r_{out} \equiv r_{learn}$. The maps in ***E***, ***F*** were obtained with the replica method (see Materials and Methods), and the green asterisks correspond to the values of parameters used in ***C***. Dashed isocontours are drawn as a guide to the eye.

(for neuron i). In the following, asterisks are used to denote quantities containing errors or noise (e.g., $X^{*\mu}$), whereas symbols without asterisks represent the mean (for h_i and J_{ij}) or target (for X^μ) values. The three types of errors and noise collectively corrupt the neurons' outputs, $X_i^{*\mu+1} = \theta\left(\sum_{j=1}^{N} J_{ij}^* X_j^{*\mu} - h_i^*\right)$, making them different from the target outputs, $X_i^{\mu+1}$. Here, θ denotes the Heaviside step-function. As the probability of action potential failure in neocortical axons is small (Cox et al., 2000), we assumed that no additional errors affect the neurons' outputs before they become inputs for the next time step.

The target neuron activities (e.g., binary scalar X_i^μ) were independently drawn from neuron-dependent Bernoulli probability distributions: 0 with probability $1 - f_i$ and 1 with probability f_i. Spiking errors in neuron activity states were introduced with the Bernoulli trials by making independent and random 1–0 changes with probabilities $P\left(X_i^{*\mu} = 0 | X_i^\mu = 1\right) \equiv p_i^-$ for spike failures and 0–1 changes with probabilities $P\left(X_i^{*\mu} = 1 | X_i^\mu = 0\right) \equiv p_i^+$ for erroneous spikes. Without loss of generality, we assumed that these two types of spiking errors are balanced, $f_i p_i^- = (1 - f_i) p_i^+$, and do not affect the neuron's firing probability, f_i. This relation allowed us to describe both types of spiking errors in terms of the neuron's overall spiking error probability, $r_i = f_i p_i^- + (1 - f_i) p_i^+$, i.e., $p_i^+ = \dfrac{r_i}{2(1 - f_i)}$ and $p_i^- = \dfrac{r_i}{2f_i}$.

To describe synaptic noise, we followed the basic model of quantal synaptic transmission (Del Castillo and Katz, 1954) and assumed that the variance of a given connection weight, J_{ij}^*, is proportional to its mean, $\mathrm{var}\left(J_{ij}^*\right) = \dfrac{h_i \beta_{syn,i}}{N} |J_{ij}|$. The dimensionless coefficient $\beta_{syn,\,i}$ is referred to as the synaptic noise strength of neuron i, and the factor of h_i/N was introduced for convenience. We assumed that the intrinsic noise is Gaussian distributed across trials with the mean $\langle h_i^* \rangle = h_i$ and variance

$\mathrm{var}(h_i^*) = \frac{h_i^2 \beta_{int,i}^2}{N}$. Here, $\beta_{int,i}$ is a dimensionless coefficient called the intrinsic noise strength of neuron i, and, as before, a factor of h_i^2/N was introduced for convenience.

Similar to Chapeton et al. (2015), two biologically inspired constraints were imposed on the learning process. First, the l_1-norm of input connection weights of each neuron was fixed during learning, $\frac{1}{N}\sum_{j=1}^{N}|J_{ij}| = w_i$. Here, parameter w_i is referred to as the average absolute connection weight of neuron i. Second, the signs of output connection weights of every neuron (inhibitory or excitatory) were fixed during learning, $J_{ij}g_j \geq 0$. In these N^2 inequalities, parameter $g_j = 1$ if neuron j is excitatory and –1 if it is inhibitory. Biological motivations for these constraints were previously discussed (Chapeton et al., 2015).

Individual neurons (e.g., neuron i) learned independently to associate noisy inputs they received from the network, $X^{*\mu}$, with the corresponding target outputs (not corrupted by noise) derived from the associative memory sequence, $X_i^{\mu+1}$. Neuron i is said to have learned the presented set of associations successfully if, in the presence of input spiking errors, synaptic and intrinsic noise, the fractions of its erroneous and failed spikes do not exceed its assigned spiking error probabilities, p_i^+ and p_i^- (Fig. 1B). The above-described model for neuron i can be summarized as follows:

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

We note that, depending on the loaded associative memory sequence, Equation 1 may have multiple solutions if the learning problem faced by the neuron is feasible or no solution if the problem is not feasible. The neuron's success probability in learning associative sequences of a given length is defined as the average of such binary outcomes (Fig. 1C). It is a decreasing function of the memory load and levels of errors and noise.

At the network level, the described associative memory storage model is governed by the network-related parameters N and $\{g_i\}$, the memory load m/N, and the neuron-related parameters $\{h_i\}$, $\{w_i\}$, $\{f_i\}$, $\{r_i\}$, $\{\beta_{syn,i}\}$, and $\{\beta_{int,i}\}$. The task is to find connection weights, $\{J_{ij}\}$, that satisfy the requirements of Equation 1 for all neurons. In the following, we examine the properties of associative networks composed of inhibitory and excitatory neurons governed by identical ($h_i = h$, $w_i = w$, $f_i = f$, $r_i = r$, $\beta_{int,i} = \beta_{int}$, and $\beta_{syn,i} = \beta_{syn}$) and distributed neuron-related parameters. We refer to these networks as homogeneous and heterogeneous.

Single-neuron model of associative memory storage in the presence of errors and noise

Each neuron in the network (e.g., neuron i) receives N_{inh} inhibitory and $(N - N_{inh})$ excitatory input connections (Fig. 1A) and independently from other neurons attempts to solve the problem outlined by Equation 1. This single-neuron learning problem was solved with the replica method in the limit of infinite network size (Edwards and Anderson, 1975; Sherrington and Kirkpatrick, 1975) and numerically with nonlinear optimization and perceptron-type learning rule for large but finite networks. In contrast to previous studies (Gardner, 1988; Gardner and Derrida, 1988; Brunel et al., 2004; Chapeton et al., 2012, 2015; Brunel, 2016; Rubin et al., 2017; Zhang et al., 2019b), the solution explicitly accounts for several distinct sources of errors and noise present during learning and incorporates two biologically inspired constraints on connectivity.

To simplify the notation in this single-neuron learning problem, in the following, we redefine the variables related to the neuron's output, $X_i^{\mu+1}$ with y^μ, f_i with f_{out}, r_i with r_{out}, and drop index i. The model is then summarized like so:

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

Learning in the model is accompanied by four types of errors and noise. These include presynaptic and output spiking errors, or errors in X^μ and y^μ, synaptic noise, or noise in J, and intrinsic noise, or noise in the neuron's threshold of firing, h. As before, we use asterisks to denote quantities containing errors or noise (e.g., $X^{*\mu}$), whereas variables without asterisks represent the mean (for h and J_j) or target (for X^μ and y^μ) values. The neuron is faced with the task of finding connection weights, $\{J_j\}$, that satisfy Equation 2 for a given set of model parameters: N, m/N, h, w, $\{g_j\}$, $\{f_j\}$, f_{out}, $\{r_j\}$, r_{out}, β_{syn}, β_{int}.

Reformulation of the model in the large N limit

In the limit of large N, the Central Limit Theorem ensures that the neuron's PSP, $\sum_{j=1}^{N} J_j^* X_j^{*\mu}$, is Gaussian distributed at every time step. Therefore, the deviation of PSP from the threshold of firing, $I^{*\mu} = \sum_{j=1}^{N} J_j^* X_j^{*\mu} - h^*$, is also Gaussian distributed with the mean and SD given by the following expressions:

$$
\begin{aligned}
I^{\mu} &= \sum_{j=1}^{N} J_j \left[\left(1 - \frac{r_j}{2f_j}\right) X_j^{\mu} + \frac{r_j(1-X_j^{\mu})}{2(1-f_j)} \right] - h \\
(\sigma^{\mu})^2 &= \sum_{j=1}^{N} J_j^2 \left[\left(1 - \frac{r_j}{2f_j}\right) \frac{r_j X_j^{\mu}}{2f_j} + \left(1 - \frac{r_j}{2(1-f_j)}\right) \frac{r_j(1-X_j^{\mu})}{2(1-f_j)} \right] \\
&+ \frac{h\beta_{syn}}{N} \sum_{j=1}^{N} J_j g_j \left[\left(1 - \frac{r_j}{2f_j}\right) X_j^{\mu} + \frac{r_j(1-X_j^{\mu})}{2(1-f_j)} \right] + \frac{h^2 \beta_{int}^2}{N}.
\end{aligned} \tag{3}
$$

As a result, the inequality constraints on the probabilities of output spiking errors (Eq. 2, line three) can be expressed in terms of I^{μ} and σ^{μ}:

$$
\begin{aligned}
I^{\mu} &\geq \sqrt{2}\,\mathrm{erf}^{-1}\left(1 - \frac{r_{out}}{f_{out}}\right)\sigma^{\mu}, \quad y^{\mu} = 1 \\
I^{\mu} &\leq -\sqrt{2}\,\mathrm{erf}^{-1}\left(1 - \frac{r_{out}}{1-f_{out}}\right)\sigma^{\mu}, \quad y^{\mu} = 0.
\end{aligned} \tag{4}
$$

The above two inequalities can be combined into a single expression that must hold for a successfully learned association μ:

$$
(2y^{\mu}-1) I^{\mu} \geq \sqrt{2}\left(\mathrm{erf}^{-1}\left(1 - \frac{r_{out}}{f_{out}}\right) y^{\mu} + \mathrm{erf}^{-1}\left(1 - \frac{r_{out}}{1-f_{out}}\right)(1-y^{\mu})\right)\sigma^{\mu} \tag{5}
$$

Additional assumptions required for the replica calculation

Following the procedure outlined in Zhang et al. (2019b), we assumed that the model parameters m/N, $\{f_j\}$, f_{out}, $\{r_j\}$, r_{out}, β_{syn}, β_{int} are intensive, or of order 1 in N. Also, we assumed that the connection weights are inversely proportional to the system size, $\left\{J_j = \frac{h}{N}\tilde{J}_j\right\}$, and refer to $\left\{\tilde{J}_j\right\}$ as scaled connection weights. This particular scaling is traditionally used in associative memory models (Brunel et al., 2004), and it has been shown that in the biologically plausible high-weight regime, $Nwf >> h$, many model results become independent of this assumption (Zhang et al., 2019b). It follows from the sixth line of Equation 2 that $w = \frac{h}{N}\tilde{w}$, and we refer to $\tilde{w}$ as scaled average absolute connection weight.

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

$$
\begin{aligned}
&(2y^{\mu}-1)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j\left[\left(1-\frac{r_j}{2f_j}\right)X_j^{\mu} + \frac{r_j(1-X_j^{\mu})}{2(1-f_j)}\right] - 1\right) \\
&\geq \sqrt{\frac{2}{N}}\left(\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right)y^{\mu} + \mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right)(1-y^{\mu})\right) \\
&\times\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_j^2\left(\left(1-\frac{r_j}{2f_j}\right)\frac{r_j X_j^{\mu}}{2f_j} + \left(1-\frac{r_j}{2(1-f_j)}\right)\frac{r_j(1-X_j^{\mu})}{2(1-f_j)}\right)\right.\right. \\
&\left.\left.+\beta_{syn}\tilde{J}_j g_j\left(\left(1-\frac{r_j}{2f_j}\right)X_j^{\mu} + \frac{r_j(1-X_j^{\mu})}{2(1-f_j)}\right) + \beta_{int}^2\right]\right)^{\frac{1}{2}}; \\
&\mu = 1,\ldots,m \\
&\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j g_j = \tilde{w} \\
&\tilde{J}_j g_j \geq 0; \quad j = 1,\ldots,N \\
&P(X_j^{\mu}=1) = f_j; \quad P(y^{\mu}=1) = f_{out}.
\end{aligned} \tag{6}
$$

In the following, we only consider the output spiking error probabilities in the ranges $p_{out}^{+} < f_{out}$ and $p_{out}^{-} < 1 - f_{out}$, which is equivalent to $r_{out} < 2f_{out}(1-f_{out})$. This is required for the stability of the replica solution.

Replica theory solution of the model

We begin by calculating the volume of the connection weight space, $\Omega(\{X^{\mu}, y^{\mu}\})$, in which Equation 6 holds for a given set of associations, $\{X^{\mu}, y^{\mu}\}$:

$$
\begin{aligned}
&\Omega(\{X^{\mu}, y^{\mu}\}) = \int\prod_{j=1}^{N} d\tilde{J}_j \prod_{j=1}^{N}\theta\left(\tilde{J}_j g_j\right)\delta\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j g_j - \tilde{w}\right) \\
&\prod_{\mu=1}^{m}\theta\left((2y^{\mu}-1)\times\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j\left[\left(1-\frac{r_j}{2f_j}\right)X_j^{\mu} + \frac{r_j(1-X_j^{\mu})}{2(1-f_j)}\right] - 1\right)\right. \\
&\left.-\sqrt{\frac{2}{N}}\left(\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right)y^{\mu} + \mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right)(1-y^{\mu})\right)\right. \\
&\left.\times\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_j^2\left(\left(1-\frac{r_j}{2f_j}\right)\frac{r_j X_j^{\mu}}{2f_j} + \left(1-\frac{r_j}{2(1-f_j)}\right)\frac{r_j(1-X_j^{\mu})}{2(1-f_j)}\right)\right.\right.\right. \\
&\left.\left.\left.+\beta_{syn}\tilde{J}_j g_j\left(\left(1-\frac{r_j}{2f_j}\right)X_j^{\mu} + \frac{r_j(1-X_j^{\mu})}{2(1-f_j)}\right) + \beta_{int}^2\right]\right)^{\frac{1}{2}}\right)
\end{aligned} \tag{7}
$$

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

$$
\begin{aligned}
\ln\left(\Omega_{typical}\right) &= \left\langle \ln(\Omega(\{X^{\mu}, y^{\mu}\}))\right\rangle_{\{X^{\mu}, y^{\mu}\}} \\
&= \lim_{n\to 0}\frac{\left\langle \Omega(\{X^{\mu}, y^{\mu}\})^n\right\rangle_{\{X^{\mu}, y^{\mu}\}} - 1}{n}.
\end{aligned} \tag{8}
$$

The quantity $\left\langle \Omega(\{X^{\mu}, y^{\mu}\})^n\right\rangle_{\{X^{\mu}, y^{\mu}\}}$ can be rewritten as a single multidimensional integral and calculated by following a previously established procedure (Zhang et al., 2019b). Below, we only provide the main steps of this calculation, additional details can be found in Zhang et al. (2020):

$$
\begin{aligned}
&\ln(\Omega_{typical}) = N\left((2z+\eta\tilde{w})\sqrt{t} + \tau\kappa t - \frac{\left(D_{out}^{+} + D_{out}^{-}\right)^2}{2(u_{+}+u_{-})^2}(\varepsilon-\delta)\kappa t + \alpha G_E(u_{+}, u_{-}, \varepsilon) + G_S(\eta, t, \tau, z, \delta)\right) \\
&G_E(u_{+}, u_{-}, \varepsilon) = \int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\left(f_{out}\ln\left(\mathrm{erfc}\left(\frac{u_{-}-x}{\sqrt{\varepsilon}}\right)\right) + (1-f_{out})\ln\left(\mathrm{erfc}\left(\frac{u_{+}-x}{\sqrt{\varepsilon}}\right)\right)\right) - \ln 2 \\
&G_S(\eta, t, \tau, z, \delta) = \frac{1}{N}\sum_{j=1}^{N}\int_{-\infty}^{\infty}\frac{e^{-x^2}}{\sqrt{\pi}}dx\ln\left(e^{-\tau t\beta_{int}^2}\frac{\sqrt{\pi}e^{\left(\frac{\eta+2zg_jf_j+\tau\sqrt{t}\beta_{syn}f_j-2x\sqrt{C_j}}{2\sqrt{C_j\delta+B_j\tau}}\right)^2}}{2\sqrt{C_j\delta t + B_j\tau t}}\right. \\
&\left.\mathrm{erfc}\left(\frac{\eta+2zg_jf_j+\tau\sqrt{t}\beta_{syn}f_j-2x\sqrt{C_j}}{2\sqrt{C_j\delta+B_j\tau}}\right)\right) \\
&B_j = r_j\left(1-\frac{r_j}{4f_j(1-f_j)}\right); \quad C_j = f_j(1-f_j)\left(1-\frac{r_j}{2f_j(1-f_j)}\right)^2; \\
&D_{out}^{+} = \sqrt{2}\,\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right); \quad D_{out}^{-} = \sqrt{2}\,\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right).
\end{aligned} \tag{9}
$$

The nine latent variables, u_{+}, u_{-}, κ, ε, η, t, τ, z, and δ are defined by the position of the maximum of $\ln\left(\Omega_{typical}\right)$. They can be obtained by solving the following system of nine equations:

$$\alpha\frac{\partial G_E(u_+,u_-,\varepsilon)}{\partial u_+}+\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2}{\left(u_++u_-\right)^3}(\varepsilon-\delta)\kappa t=0;\quad \alpha\frac{\partial G_E(u_+,u_-,\varepsilon)}{\partial u_-}+\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2}{\left(u_++u_-\right)^3}(\varepsilon-\delta)\kappa t=0$$
$$\tau t-\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2}{2\left(u_++u_-\right)^2}(\varepsilon-\delta)t=0;\quad \alpha\frac{\partial G_E(u_+,u_-,\varepsilon)}{\partial \varepsilon}-\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2}{2\left(u_++u_-\right)^2}\kappa t=0$$
$$\frac{\partial G_S(\eta,t,\tau,z,\delta)}{\partial \eta}+\tilde{w}\sqrt{t}=0;\quad \frac{\partial G_S(\eta,t,\tau,z,\delta)}{\partial t}+\frac{(2z+\eta\tilde{w})}{2\sqrt{t}}+\tau\kappa-\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2}{2\left(u_++u_-\right)^2}(\varepsilon-\delta)\kappa=0$$
$$\frac{\partial G_S(\eta,t,\tau,z,\delta)}{\partial \tau}+\kappa t=0;\quad \frac{\partial G_S(\eta,t,\tau,z,\delta)}{\partial z}+2\sqrt{t}=0;\quad \frac{\partial G_S(\eta,t,\tau,z,\delta)}{\partial \delta}+\frac{\left(D_{out}^{+}+D_{out}^{-}\right)^2}{2\left(u_-+u_+\right)^2}\kappa t=0$$
$$\kappa\geq 0;\quad \varepsilon\geq 0;\quad u_++u_-\geq 0. \tag{10}$$

The three inequality constraints in the last line of Equation 10 ensure that the solution is physical.

Replica theory solution at critical capacity

With an increasing number of associations m, $\Omega_{typical}$ shrinks and approaches zero at the maximum (critical) capacity of the neuron, $\alpha_c=\frac{m_c}{N}$. In this limit, (q_0-q) goes to zero and Equation 10 can be expanded asymptotically in terms of $1/\varepsilon$ and $1/\delta$. After replacing $\tau\sqrt{t}$ with y, δ/τ with x, and eliminating variables, ε, t, κ, τ, and δ, we arrived at the final system of six equations and one inequality. This system contains six latent variables $u_\pm$, x, η, y, and z which determine the critical capacity of the neuron, α_c:

$$\begin{cases}(1-f_{out})F(u_+)-f_{out}F(u_-)=0\\ x=\frac{4(u_-+u_+)}{\left(D_{out}^{+}+D_{out}^{-}\right)^2}\frac{f_{out}E(u_-)+(1-f_{out})E(u_+)}{f_{out}F(u_-)+(1-f_{out})F(u_+)}\\ \frac{1}{N}\sum_{j=1}^{N}\frac{\sqrt{C_j}}{C_jx+B_j}F\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right)=2\tilde{w}y\\ \frac{1}{N}\sum_{j=1}^{N}\frac{f_jg_j\sqrt{C_j}}{C_jx+B_j}F\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right)=2y\\ \frac{1}{N}\sum_{j=1}^{N}\frac{C_j}{\left(C_jx+B_j\right)^2}\left(\frac{B_j}{2}-\frac{(u_-+u_+)^2}{\left(D_{out}^{+}+D_{out}^{-}\right)^2}C_j\right)D\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right)=\beta_{int}^2y^2-y\tilde{w}\eta-2yz\\ \frac{1}{N}\sum_{j=1}^{N}\frac{\beta_{syn}f_j\sqrt{C_j}}{C_jx+B_j}F\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right)=2\tilde{w}\eta+4z-4\beta_{int}^2y\\ u_++u_-\geq 0\end{cases}$$
$$\alpha_c=x^2\frac{f_{out}D(u_-)+(1-f_{out})D(u_+)}{\left(f_{out}E(u_-)+(1-f_{out})E(u_+)\right)^2}\frac{1}{N}\sum_{j=1}^{N}\frac{C_j^2}{\left(C_jx+B_j\right)^2}D\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right). \tag{11}$$

Functions E, F, and D in Equation 11 are defined as follows:

$$\begin{aligned}E(x)&=\frac{1}{2}(1+\mathrm{erf}(x))\\ F(x)&=\frac{1}{\sqrt{\pi}}e^{-x^2}+x(1+\mathrm{erf}(x))\\ D(x)&=xF(x)+E(x)\end{aligned} \tag{12}$$

We note that Equation 11 contains as a limiting case the solution described in Brunel et al. (2004), where a simplified version of the model presented here was solved by minimizing the probability of output spiking errors for a given intrinsic noise strength. Equation 11 expands that result to account for additional features such as the homeostatic constraint, learning by inhibitory inputs, heterogeneity of inputs, synaptic noise, input and output spiking errors.

Distribution of input weights at critical capacity

Connection probabilities, P^{con}, probability densities of non-zero input weights, p^{PSP}, and average weights of these inputs, $\langle\tilde{J}\rangle$, at critical capacity were calculated as previously described (Zhang et al., 2019b). The result depends on the latent variables of Equation 11:

$$\begin{aligned}P_j^{con}&=E\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right)\\ p_j^{PSP}(\tilde{J})&=\frac{\theta\left(g_j\tilde{J}\right)}{\sqrt{2\pi}\sigma_j\tilde{w}E\left(-\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}\right)}e^{-\left(\frac{\tilde{J}}{\sqrt{2}\sigma_j\tilde{w}}+\frac{\eta+2zf_jg_j+y\beta_{syn}f_j}{2\sqrt{C_j}}g_j\right)^2}\\ \langle\tilde{J}_j\rangle&=g_j\sigma_j\tilde{w}\frac{F\left(E^{-1}\left(P_j^{con}\right)\right)}{\sqrt{2}P_j^{con}}\\ \sigma_j&=\frac{\sqrt{C_j}}{\sqrt{2}\tilde{w}y\left(C_jx+B_j\right)}.\end{aligned} \tag{13}$$

A given input, j, has a non-infinitesimal probability of having a connection weight of zero, while its probability density for non-zero connection weights is a truncated Gaussian with a mean $\langle\tilde{J}_j\rangle$ and SD $\sigma_j\tilde{w}$.

Equations 11, 13 were solved in MATLAB to produce the results for heterogeneous networks consisting of inhibitory and excitatory neurons with distributed spiking error probabilities and distributed intrinsic and synaptic noise strengths. The code is available at Zhang et al. (2019a). In both cases, the remaining model parameters were the same for all input connections (e.g., $f_i\equiv f$). In this case, the solutions of Equations 11, 13 depend on β_{int} and β_{syn} only in a combination $\beta=\sqrt{\beta_{int}^2+\tilde{w}\beta_{syn}f}$, referred to as the postsynaptic noise strength.

The solution in the case of two homogeneous classes of inputs

In this case, all inputs have the same firing probability, f_{in}, and the same spiking error probability, r_{in}. Equation 11, 13 simplify significantly after the introduction of two new variables, $v_\pm=\frac{-\eta\pm 2zf_{in}-y\beta_{syn}f_{in}}{2\sqrt{C_{in}}}$:

$$\begin{cases}(1-f_{out})F(u_+)-f_{out}F(u_-)=0\\ \frac{N_{inh}}{N}F(v_+)+\frac{N_{exc}}{N}F(v_-)=\frac{\sqrt{2}}{\sigma}\\ -\frac{N_{inh}}{N}F(v_+)+\frac{N_{exc}}{N}F(v_-)=\frac{\sqrt{2}}{\sigma\tilde{w}f_{in}}\\ \left((u_-+u_+)^2-2\xi\right)\left(\frac{N_{inh}}{N}D(v_+)+\frac{N_{exc}}{N}D(v_-)\right)=\frac{2\beta^2\zeta^2}{\sigma^2}\\ \sigma=\frac{\sqrt{2}\beta^2\zeta^2}{\left((u_-+u_+)\frac{f_{out}E(u_-)+(1-f_{out})E(u_+)}{f_{out}F(u_-)+(1-f_{out})F(u_+)}+\xi\right)\left(\frac{1}{\tilde{w}f_{in}}(v_+-v_-)-(v_++v_-)\right)}\\ u_++u_->0\end{cases}$$
$$\alpha_c=\frac{\sigma^2(u_-+u_+)^2}{2\beta^4\zeta^4}\frac{f_{out}D(u_-)+(1-f_{out})D(u_+)}{\left(f_{out}F(u_-)+(1-f_{out})F(u_+)\right)^2}\left(\frac{1}{\tilde{w}f_{in}}(v_+-v_-)-(v_++v_-)\right)^2\times\left(\frac{N_{inh}}{N}D(v_+)+\frac{N_{exc}}{N}D(v_-)\right)$$
$$P_{inh/exc}^{con}=E(v_\pm)$$
$$p_{inh/exc}^{PSP}(\tilde{J})=\frac{\theta(\mp\tilde{J})}{\sqrt{2\pi}\sigma\tilde{w}E(v_\pm)}e^{-\left(\frac{\tilde{J}}{\sqrt{2}\sigma\tilde{w}}\pm v_\pm\right)^2}$$
$$\langle|\tilde{J}_{inh/exc}|\rangle=\frac{\tilde{w}}{2P_{inh/exc}^{con}}\frac{N}{N_{inh/exc}}\left(1\mp\frac{1}{\tilde{w}f_{in}}\right) \tag{14}$$

The intrinsic and synaptic noises in Equation 14 are entirely contained within the parameter β, while the spiking

error probabilities r_{in} and r_{out} appear only in the parameters ξ and ζ:

$$\begin{aligned}\beta &= \sqrt{\beta_{int}^2 + \tilde{w}\beta_{syn}f_{in}}\\ \xi &= \frac{r_{in}(4f_{in}(1-f_{in})-r_{in})}{2(2f_{in}(1-f_{in})-r_{in})^2}\left(\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right)+\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right)\right)^2\\ \zeta &= \frac{\sqrt{2f_{in}(1-f_{in})}}{\tilde{w}(2f_{in}(1-f_{in})-r_{in})}\left(\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right)+\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right)\right)\end{aligned} \tag{15}$$

We note that in the absence of spiking errors in the input ($r_{in} = 0$), Equation 14 is similar in structure to the solution of a traditional model considered by Zhang et al. (2019b; Fig. 1*D*). That model did not explicitly consider different sources of errors and noise, but instead used a generic robustness parameter κ, or a rescaled robustness parameter $\rho = \frac{\kappa}{w\sqrt{Nf(1-f)}}$, to ensure that memories are recalled reliably in the case when only intrinsic noise is present. Solutions to both models become identical when $r_{in} = 0$ and $\rho = \beta\zeta$. Therefore, Equation 15 explains the nature of parameters κ and ρ, relating them to the output error probability, intrinsic and synaptic noise strengths:

$$\begin{aligned}\kappa &= \frac{h}{\sqrt{2N}}\sqrt{\beta_{int}^2 + \tilde{w}\beta_{syn}f_{in}}\left(\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right)+\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right)\right)\\ \rho &= \frac{\sqrt{\beta_{int}^2 + \tilde{w}\beta_{syn}f_{in}}}{\tilde{w}\sqrt{2f_{in}(1-f_{in})}}\left(\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{1-f_{out}}\right)+\mathrm{erf}^{-1}\left(1-\frac{r_{out}}{f_{out}}\right)\right)\end{aligned} \tag{16}$$

Numerical solution of Equations 14, 15 shows that the critical capacity (Fig. 1) and probabilities of inhibitory and excitatory connections decrease with β_{int}, β_{syn}, and r_{in}, and increase with r_{out}. This is consistent with previous results (Brunel et al., 2004; Zhang et al., 2019b) showing that the critical capacity and connection probabilities are decreasing functions of ρ. The averages and SDs of inhibitory and excitatory connection weight magnitudes exhibit an opposite dependence on errors and noise, which is also consistent with the results of these studies. For homogeneous associative networks, we set $r_{in} = r_{out} \equiv r$ and $f_{in} = f_{out} \equiv f$ in Equations 14, 15, as these parameters must be the same for all neurons in the network. This does not alter the trend of the results related to β, but the dependence on r becomes more complex (Fig. 1*F*). Figures 2–6 show the results for homogeneous networks as functions of β and r.

The average weights of non-zero inhibitory and excitatory connections are uniquely determined by $\tilde{w}$, $P_{inh/exc}^{con}$, $N_{inh/exc}/N$, and f_{in} (Eq. 14, last line). This result is obtained from the functional form of the input weight distribution, but it also follows from the fact that the input connection weights are homeostatically constrained (Eq. 6, second line) and, at critical capacity, the neuron operates in a balanced regime in which inhibitory and excitatory currents are anti-correlated and largely cancel each other out (Rubin et al., 2017). Experimentally, it has been shown that inhibitory postsynaptic currents are larger in magnitude than excitatory (Atallah and Scanziani, 2009; Salkoff et al., 2015; Feng et al., 2019). Although Equation 14 derived in the $N \to \infty$ limit yield a small positive or zero average postsynaptic input (high-weight regime), associative networks of finite-size loaded with memories to capacity show a trend consistent with the experimental measurements (Zhang et al., 2019b).

Numerical solution of the model with nonlinear optimization

For a finite number of inputs, the solution to the problem outlined in Equation 6 was obtained numerically. To that end, we made the problem feasible by introducing a slack variable $s^\mu \geq 0$ for every association and chose the solution that minimizes the sum of these variables:

$$\begin{aligned}&\underset{\{\tilde{J}_j\}}{\mathrm{argmin}}\left(\sum_{\mu=1}^{m} s^\mu\right)\\ &(2y^\mu-1)\left(\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_j\left[\left(1-\frac{r_j}{2f_j}\right)X_j^\mu+\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right]-1\right)\\ &\quad\geq -s^\mu+\frac{D_{out}^- y^\mu + D_{out}^+(1-y^\mu)}{\sqrt{N}}\\ &\times\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_j^2\left(\left(1-\frac{r_j}{2f_j}\right)\frac{r_jX_j^\mu}{2f_j}+\left(1-\frac{r_j}{2(1-f_j)}\right)\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right)\right.\right.\\ &\qquad\left.\left.+\beta_{syn}\tilde{J}_jg_j\left(\left(1-\frac{r_j}{2f_j}\right)X_j^\mu+\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right)+\beta_{int}^2\right]\right)^{\frac{1}{2}};\\ &\mu=1,\ldots,m\\ &\frac{1}{N}\sum_{j=1}^{N}\tilde{J}_jg_j=\tilde{w}\\ &\tilde{J}_jg_j\geq 0;\quad j=1,\ldots,N\\ &s^\mu\geq 0;\quad \mu=1,\ldots,m.\end{aligned} \tag{17}$$

Equation 17 were solved by using the *fmincon* function of MATLAB and the results are shown in Figures 2, 3, 5, 6. The *fmincon* function utilizes the interior-point technique for finding solutions to constrained nonlinear optimization problems (Byrd et al., 1999, 2000). The code is available at Zhang et al. (2019a).

Numerical solution of the model with a perceptron-type learning rule

In addition to the replica and nonlinear optimization solutions, a biologically more plausible online solution of Equation 17 was devised by approximately stepping in the direction of the negative gradient of the sum of the slack variables. The latter is:

$$\begin{aligned}&-\frac{\partial}{\partial\tilde{J}_j}\sum_{\mu=1}^{m}s^\mu=\frac{1}{N}\sum_{\mu=1}^{m}(2y^\mu-1)\left[\left(1-\frac{r_j}{2f_j}\right)X_j^\mu+\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right]-\frac{1}{N^{3/2}}\sum_{\mu=1}^{m}\left(D_{out}^-y^\mu+D_{out}^+(1-y^\mu)\right)\\ &\times\frac{\tilde{J}_j\left(\left(1-\frac{r_j}{2f_j}\right)\frac{r_jX_j^\mu}{2f_j}+\left(1-\frac{r_j}{2(1-f_j)}\right)\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right)+\frac{1}{2}\beta_{syn}g_j\left(\left(1-\frac{r_j}{2f_j}\right)X_j^\mu+\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^{N}\left[\tilde{J}_j^2\left(\left(1-\frac{r_j}{2f_j}\right)\frac{r_jX_j^\mu}{2f_j}+\left(1-\frac{r_j}{2(1-f_j)}\right)\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right)+\beta_{syn}\tilde{J}_jg_j\left(\left(1-\frac{r_j}{2f_j}\right)X_j^\mu+\frac{r_j(1-X_j^\mu)}{2(1-f_j)}\right)+\beta_{int}^2\right]\right)^{\frac{1}{2}}}\end{aligned} \tag{18}$$

The first approximation to this gradient was made by omitting the second term in the right-hand side of Equation 18. This was done because this term is smaller than the first term (for large enough N) and because there is no clear way of calculating it

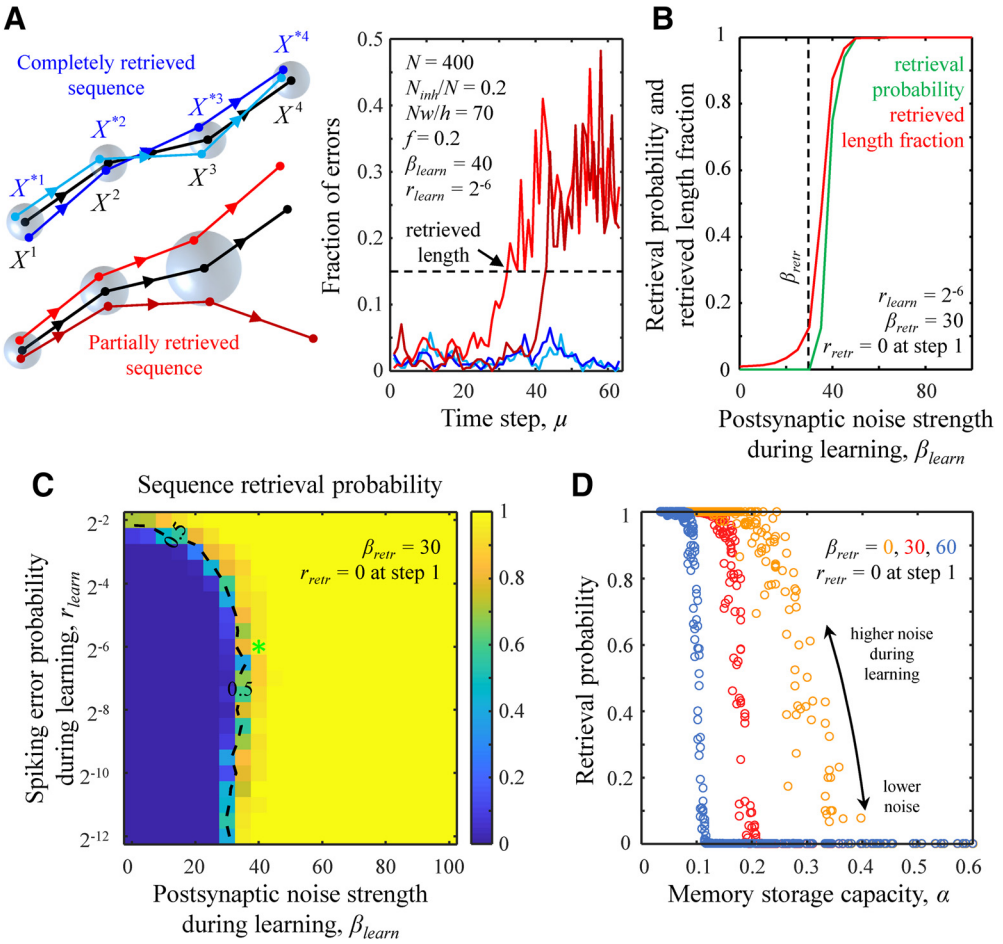

Figure 2. Retrieval of loaded associative memory sequences and the trade-off between capacity and reliability of loaded memories. ***A***, Illustration of memory playout during complete and partial memory retrieval (left). The target memory sequence is shown in black, while the sequences retrieved on different trials are in blue and red. Memory retrieval is incomplete when the retrieved sequence deviates significantly from the target sequence (see text for details). Radii of blue spheres illustrate the root-mean-square Euclidean distances between the retrieved and target states. The fraction of errors as a function of time step during sequence retrieval (right). Successfully retrieved sequences do not deviate from the loaded sequences by more than a threshold amount (dashed line). The parameters of the associative network are provided in the figure. The values of β_{learn} and r_{learn} correspond to the green asterisk from Figure 1. ***B***, The probability of successful memory retrieval (green) and the retrieved fraction of loaded sequence length (red) as a function of β_{learn}. The postsynaptic noise strength $\beta_{retr} = 30$ (dashed line) at every step of memory retrieval and r_{retr} was set to 0 at the first step. ***C***, Map of retrieval probability as a function of β_{learn} and r_{learn}. Dashed isocontour is drawn as a guide to the eye. The location of the green asterisk is the same as in Figure 1*F*. ***D***, The trade-off between memory retrieval probability and α. Individual points correspond to all values of β_{learn} and r_{learn} considered in ***C***. Higher errors and noise during learning result in lower α and higher retrieval probability regardless of the noise strength during memory retrieval (different colors). The results shown in ***A–D*** were obtained with the nonlinear optimization method (see Materials and Methods). For every parameter setting, the results shown in ***B–D*** were averaged over 100 networks and 1000 retrievals of the loaded sequence in each network.

in an online, biologically plausible manner. The second approximation was made by noting that $\left[\left(1-\frac{r_j}{2f_j}\right)X_j^\mu + \frac{r_j\left(1-X_j^\mu\right)}{2\left(1-f_j\right)}\right]$ in the first term in the right-hand side of Equation 18 is the average of $X_j^{*\mu}$ over the spiking errors, and therefore, a stochastic estimate of this gradient direction can be made in an online manner with a perceptron-type learning step $(2y^\mu - 1)X_j^{*\mu}$ (Rosenblatt, 1962). These approximations lead to the learning rule of Equation 22. Related rules, in the absence of errors, noise, or l_1-norm constraint, were previously described (Brunel et al., 2004; Zhang et al., 2019b).

In numerical simulations, we trained neurons on associations presented in the order of their appearance in the associative sequence, one at a time. This constitutes one learning epoch. We set the learning rate $\gamma = 0.1$ and ran the algorithm until a solution was found or the maximum number of 10^6 epochs was reached. The results of this procedure are shown in Figure 6.

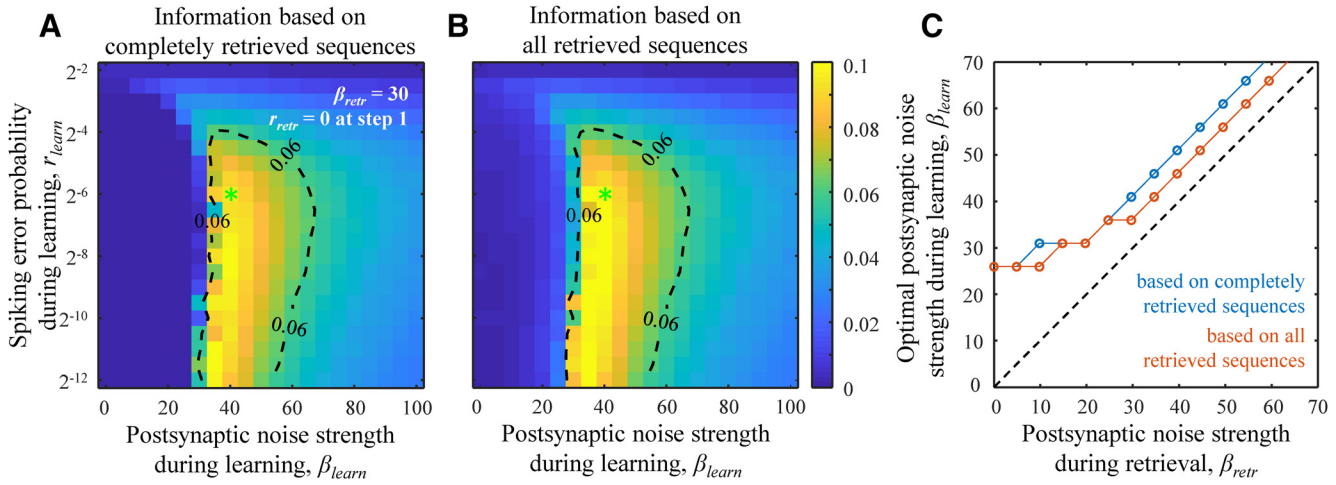

Figure 3. Postsynaptic noise during learning is required for optimal retrieval of stored information. ***A***, ***B***, Maps of expected retrieved information per memory playout calculated based on completely retrieved sequences (***A***) and completely and partially retrieved sequence (***B***) in bits$\times N^2$ as functions of β_{learn} and r_{learn}. β_{retr} = 30 at every step of memory retrieval, and r_{retr} was set to 0 at the first step. Dashed isocontours are drawn as guides to the eye. The locations of the green asterisks are the same as in Figure 1*F*. ***C***, The maximum of retrieved information is achieved when β_{learn} is greater than zero regardless of the value of β_{retr}. The optimal postsynaptic noise strengths were calculated based on the averages of the results from ***A***, blue line, and ***B***, orange line, over the range of r_{learn} values from ***A***, ***B***. All results were obtained with the nonlinear optimization method (see Materials and Methods) and averaged over 100 networks and 1000 retrievals of the loaded sequence in each network for every parameter setting.

Mutual information contained in retrieved associative sequences

The mutual information contained in one successfully retrieved association $(X^{\mu} \to X^{\mu+1})$ can be calculated as a difference of marginal and conditional entropies,

$$I(X^{\mu}; X^{\mu+1}) = H(X^{\mu+1}) - H(X^{\mu+1}|X^{\mu}). \tag{19}$$

For homogeneous networks loaded with associations consisting of random and independent network states, the two entropies reduce to:

$$H(X^{\mu+1}) = -N[f\log_2 f + (1-f)\log_2(1-f)]$$

$$H(X^{\mu+1}|X^{\mu}) = -N\left[f\left(\frac{r}{2f}\log_2\frac{r}{2f} + \left(1-\frac{r}{2f}\right)\log_2\left(1-\frac{r}{2f}\right)\right) + (1-f)\left(\frac{r}{2(1-f)}\log_2\frac{r}{2(1-f)} + \left(1-\frac{r}{2(1-f)}\right)\times\log_2\left(1-\frac{r}{2(1-f)}\right)\right)\right]. \tag{20}$$

As the length of a retrieved sequence may be shorter than the length of the loaded sequence, m, we considered two types of retrieved information. One type is defined as the expected retrieved information per memory playout in which contributions of partially retrieved sequences are set to zero. This information is based on completely retrieved sequences only and is equal to the product of the retrieval probability (Fig. 2*C*) and mI. The other type of retrieved information is calculated based on completely and partially retrieved sequences and is equal to the product of the average retrieved sequence length and I. According to these definitions, the former is always less or equal to the latter.

Dataset of connection probabilities and strengths in local brain circuits in mammals

To compare connection probabilities and widths of non-zero connection weight distributions in associative networks with those reported experimentally, we used the dataset published in (Zhang et al., 2019b). This dataset includes measurements reported in peer-reviewed publications since 1990 in which at least 10 pairs of neurons separated laterally by $<100\,\mu$m were recorded from the same layer of the mammalian neocortex in juvenile or adult animals of either sex. The dataset includes 87 publications describing 420 local projections.

Results

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

We examined a model network consisting of N_{inh} inhibitory and $(N - N_{inh})$ excitatory McCulloch and Pitts neurons (McCulloch and Pitts, 1943; Fig. 1*A*) involved in associative learning. The model is described in detail in Materials and Methods, and in this subsection, we only mention its main features. The network was designed to model a local cortical circuit ($\sim$100 μm in size) of all-to-all potentially (structurally) connected neurons (Stepanyants and Chklovskii, 2005; Stepanyants et al., 2008). The network was presented with a task of learning a sequence of consecutive network states, $X^1 \to X^2 \to \ldots X^{m+1}$, in which X^{μ} is a binary vector representing target activities of all neurons at a time step μ, and the ratio m/N is referred to as the memory load. Network activity in the model was accompanied by several sources of errors and noise (Fig. 1*A*, orange lightning signs), including (1) input spiking errors, or errors in X^{μ}; (2) synaptic noise, or noise in

connection weights, J_{ij} (weight of connection from neuron j to neuron i); and (3) intrinsic noise, which combines all other sources of noise affecting the neurons' PSPs. The last category includes background synaptic activity and the stochasticity of ion channels and in the model is equivalent to noise in the neurons' firing thresholds, h_i. The three types of errors and noise collectively corrupt the neurons' outputs making them different from the target outputs. The strengths of these errors and noise in the model are governed by parameters r_i, $\beta_{syn,\,i}$, and $\beta_{int,\,i}$, respectively.

Individual neurons in the model learned independently to associate noisy inputs they received from the network, $X^{*\mu}$, with the corresponding target outputs (not corrupted by noise) derived from the associative memory sequence, $X_i^{\mu+1}$. The neurons learned such input-output associations by adjusting the weights of their input connections, J_{ij}, in the presence of two biologically inspired constraints (Chapeton et al., 2015). First, the average absolute weight of input connections of each neuron was kept constant, w_i. Second, the output connection weights of neurons (inhibitory or excitatory) did not change signs during learning.

The described associative network model is summarized by Equation 1. It is governed by the network-related parameters N and N_{inh}/N, the memory load m/N, and the neuron-related parameters $\{h_i\}$, $\{w_i\}$, $\{f_i\}$, $\{r_i\}$, $\{\beta_{syn,i}\}$, and $\{\beta_{int,i}\}$. In the following, we examine the properties of associative networks with identical and distributed neuron-related parameters. These networks are referred to as homogeneous and heterogeneous.

Solutions of the model

Equation 1 was solved with the replica method, nonlinear optimization, and a perceptron-type learning rule (see Materials and Methods). Each of these methods has its advantages and drawbacks, and, consequently, all three methods were used in this study. The replica method (Edwards and Anderson, 1975; Sherrington and Kirkpatrick, 1975) provides an analytical solution in the $N \to \infty$ limit. Though neuron networks in the brain are finite, they are thought to be large enough to have many properties that are well described by this limit (Zhang et al., 2019b). More importantly, the analytical solution of the replica method reveals the dependence of the results on combinations of network parameters that can be then explored with other methods. The downside of the replica solution is that it does not provide the full connectivity matrix, J_{ij}, but instead gives the connectivity statistics that is insufficient to calculate all relevant network properties. Nonlinear optimization can be used to solve Equation 1. This method is fast and accurate for small networks, yielding the full connectivity matrix, but is impractical for large networks ($N \sim 1000$). As the replica and nonlinear optimization solutions cannot be readily implemented by neural networks in the brain, we also developed a biologically more plausible perceptron-type learning rule that can be used to approximate the solution of Equation 1. Because simulations based on the perceptron-type learning rule become time-consuming at or near memory storage capacity as the solution region shrinks to a point, results for varying levels of errors and noise were obtained with the replica and nonlinear optimization methods, while the perceptron-type learning rule was used only for a biologically plausible set of parameters to confirm that all three methods lead to similar results.

In the $N \to \infty$ limit, the associative memory storage problem for a neuron loaded to capacity was solved with the replica method. This solution for a neuron in a homogeneous network depends on the following combination of the intrinsic and synaptic noise strengths (see Materials and Methods):

$$\beta = \sqrt{\beta_{int}^2 + \beta_{syn} f N \frac{w}{h}} \tag{21}$$

This quantity is referred to as the postsynaptic noise strength. In the following, we assume that the postsynaptic noise strength, β, and the spiking error probability, r, can differ between the times of learning and memory retrieval and add subscripts "*learn*" and "*retr*" to these parameters to distinguish among the two phases.

Figure 1*C* shows that when the memory load is relatively low, the probability of successful learning by a neuron is close to 1. With increasing load, the learning problem becomes more difficult, and the success probability undergoes a smooth transition from 1 to 0. Memory load corresponding to the success probability of 0.5 is referred to as the neuron's associative memory storage capacity, α. With increasing network size, N, the transition from successful learning to inability to accurately learn the complete memory sequence becomes sharper, and the neuron's capacity monotonically approaches its $N \to \infty$ limit, which is referred to as the critical capacity, α_c. The critical capacity depends on the levels of errors and noise accompanying learning and other parameters of the model. Figure 1*D–F* illustrates the dependence of α_c on the input and output spiking error probabilities and postsynaptic noise strength. As expected, because input spiking errors, intrinsic, and synaptic noise, make the learning problem more challenging, α_c is a decreasing function of r_{in} (Fig. 1*D*,*E*) and β_{learn} (Fig. 1*D*,*F*). On the other hand, the learning problem becomes simpler with increasing r_{out} as more output errors are tolerated, and α_c is an increasing function of r_{out} (Fig. 1*E*). For a neuron in a recurrent homogeneous network, the dependence of α_c on spiking errors is more complex as $r_{in} = r_{out} \equiv r_{learn}$, and both the input and output spiking errors of the neuron are controlled by the same parameter (Fig. 1*F*).

The trade-off between capacity and reliability of loaded memories

Can memories, loaded into individual neurons, be successfully recalled at the network level? To answer this question, we loaded neurons in the network to capacity with associations derived from a single associative sequence by solving Equation 1. The postsynaptic noise and spiking errors during learning were set at the levels β_{learn} and r_{learn} (Fig. 1*F*, green asterisk). During memory

retrieval, the network was initialized at the beginning of the loaded sequence, X^1, and no additional spiking errors, beyond those produced by the network at subsequent steps, were added as the memory played out. At each step of memory playout, synaptic and intrinsic noise were added independently to every connection and every neuron in the network at strengths governed by β_{retr}.

The sequence is said to be retrieved completely if the network states during the retrieval do not deviate substantially from the target states. Otherwise, the sequence is said to be retrieved partially, and the retrieved sequence length is defined by the number of steps taken to the point where the network states begin to deviate substantially from the target states (Fig. 2*A*). In practice, there is no need to precisely define the threshold amount of deviation. This is because for large networks the fraction of errors in a retrieved network state either fluctuates around $r_{learn} \pm \sqrt{r_{learn}(1-r_{learn})/N}$ (mean ± SD) or diverges to $2f(1-f) \pm \sqrt{2f(1-f)(1-2f(1-f))/N}$ (expected fraction of differences between two random network states of firing probability f), which is significantly greater for the chosen values of parameters r_{learn} and f. Figure 2*B* shows the probability of retrieving a complete loaded sequence and the fraction of retrieved sequence length for different values of β_{learn}. It illustrates that memory sequences can be reliably retrieved if they were loaded with the postsynaptic noise strength that is slightly higher than that present during memory retrieval. Likewise, the averaged retrieved sequence length fraction increases with β_{learn} and approaches one as β_{learn} exceeds the noise strength present during retrieval. A similar conclusion can be drawn from Figure 2*C*, which shows the map of the retrieval probability as a function of β_{learn} and r_{learn}. Errors and noise during learning make memory retrieval more reliable. However, the reliability of loaded memories comes at the expense of the memory storage capacity, α. Figure 2*D* shows the trade-off between the retrieval probability and capacity of loaded associative memories in which higher levels of errors and noise during learning enable reliable memory retrieval but reduce α.

Noise during learning is required for optimal retrieval of stored information

Figure 3*A*,*B* shows the maps of expected retrieved information per sequence playout calculated in two different ways. In the first calculation, the contribution of partially retrieved sequences to the expected retrieved information was set to zero, while in the second, partially retrieved sequences contributed in the proportion of the retrieved sequence length (see Materials and Methods). Both maps illustrate that optimal retrieval of stored information is achieved when memories are stored in the presence of noise, $\beta_{learn} > 0$. This conclusion is independent of the postsynaptic noise strength during memory retrieval, which was set to $\beta_{retr} = 30$ in Figure 3*A*,*B*. To illustrate this finding, we averaged the maps over the r_{learn} dimension and determined β_{learn} that correspond to the maxima of the retrieved information. Figure 3*C* illustrates the results of this procedure for different values of β_{retr}, showing that the optimal β_{learn} is greater than zero even when there is no noise during memory retrieval. The optimal β_{learn} increases with β_{retr}, and the two noise strengths become approximately equal in the high noise limit.

Neuron-to-neuron connectivity in associative networks of homogeneous inhibitory and excitatory neurons

One of the most salient features of sign-constrained associative learning models, such as the one described in this study, is that finite fractions of inhibitory and excitatory connections assume zero weights at capacity (Kohler and Widmaier, 1991), mirroring the trend observed in many local cortical networks. We compared the connection probabilities (P_{con}) and the coefficients of variation (CVs) of non-zero connection weights in associative networks at capacity to the connection probabilities and CVs of unitary PSPs (uPSPs) obtained experimentally. To that end, we used the dataset compiled in (Zhang et al., 2019b) based on 87 electrophysiological studies describing neuron-to-neuron connectivity for 420 local cortical projections (lateral distance between neurons $< 100\,\mu$m). Figure 4*A* shows that the average inhibitory P_{con} (38 studies, 9522 connections tested) is significantly larger ($p < 10^{-10}$, two-sample t test) than the average excitatory P_{con} (67 studies, 63,020 connections tested). Associative networks exhibit a similar trend in the entire region of considered β_{learn} and r_{learn} values (Fig. 4*B*,*C*). What is more, in the (β_{learn}, r_{learn}) parameter region demarcated with the dashed isocontours and arrows in Figure 4*B*,*C*, the model results are consistent with the middle 50% of the experimentally measured P_{con} values for inhibitory and excitatory connections.

Figure 4*D* shows that the average CV of inhibitory uPSP (10 studies, 503 connections recorded) is slightly lower than that for excitatory (36 studies, 3956 connections recorded), and this trend is also reproduced by the associative networks in the entire region of considered β_{learn} and r_{learn} values (Fig. 4*E*,*F*). As before, there are (β_{learn}, r_{learn}) parameter regions in these maps in which the results of the model are consistent with the middle 50% of the CV of uPSP measurements for inhibitory and excitatory connections.

Spontaneous dynamics in associative networks of homogeneous inhibitory and excitatory neurons

The model associative networks can exhibit irregular and asynchronous spiking activity like that observed in cortical networks. To analyze such spontaneous (not learned) network dynamics, we used associative networks loaded to capacity, initialized them at random states of firing probability $f = 0.2$, and followed their activity for 1000 time steps. Because the number of available network states, which is exponential in N, is much larger than the number of loaded states, αN, the spontaneous network activity in the numerical simulations never passed through any of the loaded states.

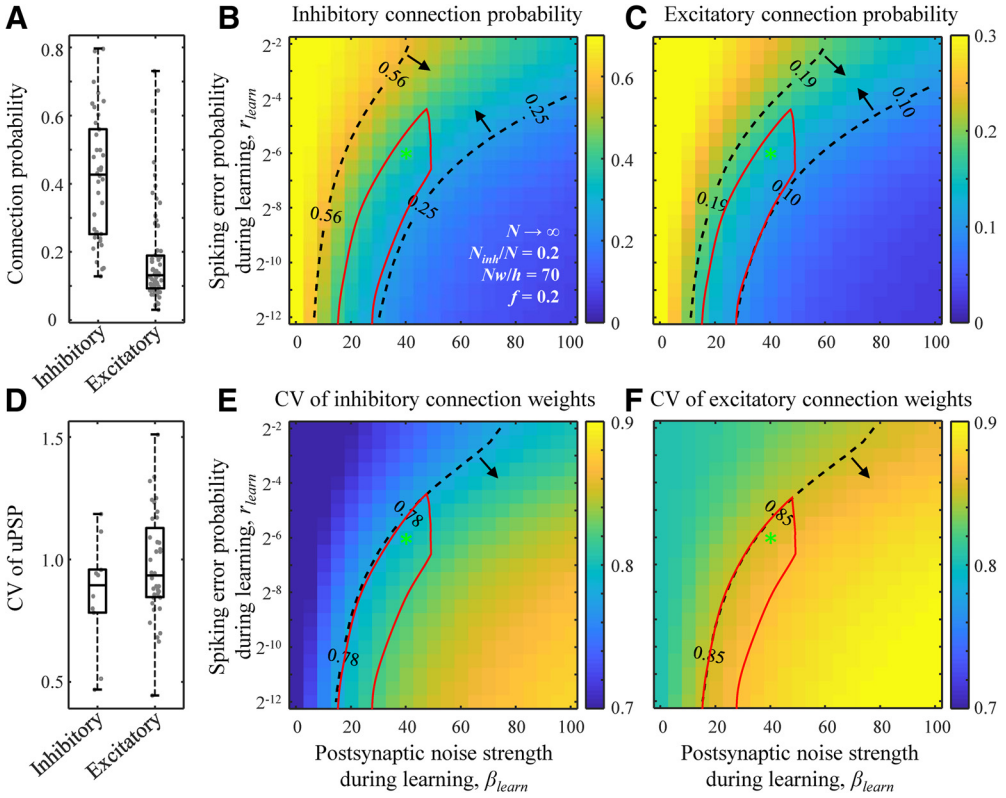

Figure 4. Comparison of structural properties of the model and cortical networks. ***A***, Inhibitory and excitatory connection probabilities reported in 87 studies describing 420 local cortical projections. Each dot represents the result of a single study/projection. ***B***, ***C***, Maps of inhibitory and excitatory connection probabilities as functions of β_{learn} and r_{learn}. The results are based on the replica method (see Materials and Methods). Dashed isocontours and arrows illustrate the interquartile ranges of the experimentally observed connection probabilities from ***A***. The red contour outlines a region of parameters that is consistent with all structural and dynamical measurements in cortical networks considered in this study. The locations of the green asterisks are the same as in Figure 1*F*. ***D–F***, Same for the CV of non-zero inhibitory and excitatory connection weights. ***A***, ***D*** were adapted from Zhang et al. (2019b).

To quantify the degree of similarity in the dynamics of the model and brain networks we compared the CV of interspike-intervals (ISIs) and the cross-correlation coefficient of spiking neuron activity in the model to those measurements obtained experimentally. Figure 5*A*, dashed isocontour, outlines (β_{learn}, r_{learn}) parameter region in which the model CV of ISI is consistent with the 0.7–1.1 range measured in different cortical systems (Softky and Koch, 1993; Holt et al., 1996; Buracas et al., 1998; Shadlen and Newsome, 1998; Stevens and Zador, 1998). Similarly, Figure 5*B* shows that there is a (β_{learn}, r_{learn}) parameter region in which the calculated spike cross-correlation coefficients are in agreement with the interquartile range of the corresponding cortical measurements, 0.04–0.15 (Cohen and Kohn, 2011). The degree of asynchrony in spontaneous spiking activity in associative networks increases with the postsynaptic noise strength, which can be explained by the decrease in connection probability (Fig. 4*B*,*C*) and, consequently, a reduction in the amount of common input to the neurons.

It was shown that irregular and asynchronous activity can result from the balance of inhibitory and excitatory postsynaptic inputs to individual cells (van Vreeswijk and Sompolinsky, 1996, 1998). In a balanced state, the magnitudes of these inputs are much greater than the threshold of firing, but, because of a high degree of anti-correlation, these inputs largely cancel, and firing is driven by fluctuations. Figure 5*C* shows a region of parameters in which neurons in the associative model function in a balanced regime. Because it is difficult to simultaneously measure inhibitory and excitatory postsynaptic inputs to a neuron, the anti-correlation of inhibitory and excitatory inputs has only been measured in nearby cells, averaging to ~0.4 (Okun and Lampl, 2008; Graupner and Reyes, 2013). As within-cell anti-correlations are expected to be stronger than between-cell anti-correlations, 0.4 was used as a lower bound for the former (Fig. 5*C*, dashed isocontour and arrow).

The seven error-noise regions obtained based on the properties of neuron-to-neuron connectivity (Fig. 4) and network dynamics (Fig. 5) have a non-empty intersection (Figs. 4, 5, red contour). In this biologically plausible region of parameters, the considered properties of the associative networks are consistent with the corresponding experimental measurements. This observation suggests that β_{learn} must lie in the 20–50 range and r_{learn} must be <0.06. While we are not aware of direct experimental measurements of these parameters, the low value of r_{learn} is in qualitative agreement with the reliability of firing patterns evoked by time-varying stimuli *in vivo*

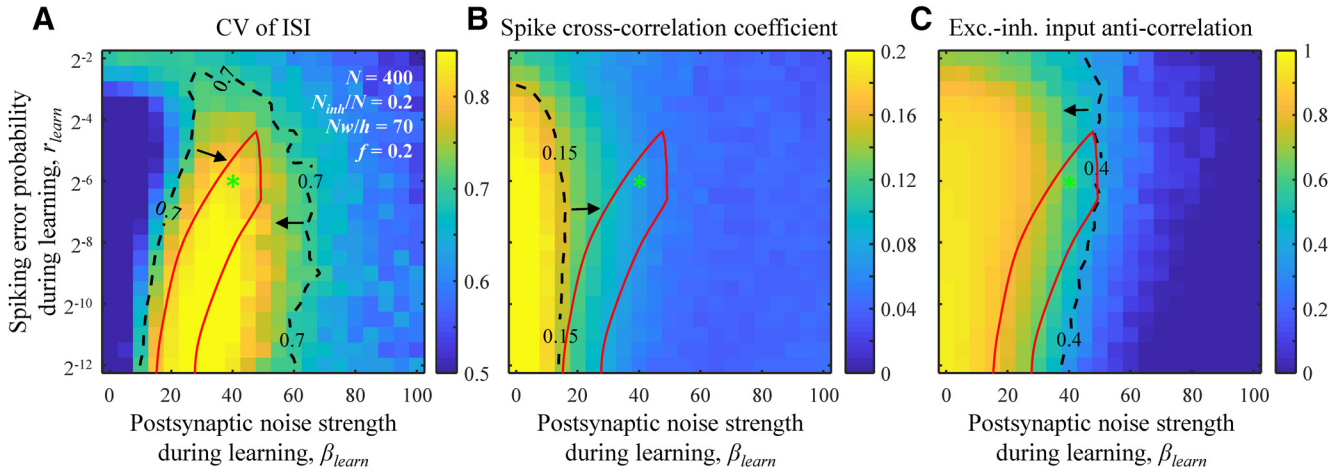

Figure 5. Comparison of dynamical properties of the model and cortical networks. ***A***, The CV of ISI for spontaneous (not learned) activity as a function of β_{learn} and r_{learn}. Dashed isocontour and arrows demarcate a region of CV values that is in general agreement with experimental measurements. ***B***, Same for the cross-correlation coefficient of neuron spike trains. ***C***, Same for the anti-correlation coefficient of inhibitory and excitatory postsynaptic inputs to a neuron. The red contour outlines a region of parameters that is consistent with the considered structural and dynamical measurements. The locations of the green asterisk are the same as in Figure 1*F*. All results were obtained with the nonlinear optimization method (see Materials and Methods) and averaged over 100 networks and 100 runs for each network for every parameter setting.

(Buracas et al., 1998) and *in vitro* (Mainen and Sejnowski, 1995).

Solution of the model with a perceptron-type learning rule

As the replica and nonlinear optimization solutions of Equation 1 cannot be easily implemented by neural networks in the brain, we set out to develop a biologically more plausible online solution to the associative learning problem. The following perceptron-type learning rule was devised to approximate the solution of Equation 1 (see Materials and Methods). At each learning step, e.g., μ, a neuron receives an input containing spiking errors, $X^{*\mu}$, combines it with synaptic and intrinsic noise, and produces an output corrupted by noise, $y^{*\mu} = \theta\left(\sum_{j=1}^{N} J_j^* X_j^{*\mu} - h^*\right)$. If this output differs from the neuron's target output, y^{μ}, which is noise-free, the neuron's input connection weights are updated in four consecutive steps:

$$\begin{aligned} &J_j \mapsto J_j + \frac{\gamma h}{N}(2y^{\mu} - 1)X_j^{*\mu}, \quad j = 1, \ldots, N \\ &J_j \mapsto J_j\,\theta(J_j g_j) \\ &J_j \mapsto J_j + \left(w - \frac{1}{N}\sum_{j=1}^{N}\left|J_j\right|\right) g_j \\ &J_j \mapsto J_j\,\theta(J_j g_j) \end{aligned} \tag{22}$$

The first line in Equation 22 is a stochastic perceptron learning step (Rosenblatt, 1962), in which parameter γ is referred to as the learning rate. The second line enforces the sign constraints, while the last two lines implement the homeostatic l_1-norm constraint and are equivalent to the soft thresholding used in LASSO regression (Tibshirani, 1996). In contrast to the standard perceptron learning rule, Equation 22 uses noisy inputs and enforce sign and homeostatic constraints at every learning step. They can be used to learn temporally correlated input-output network states, including auto-associations.

By including input spiking errors, synaptic and intrinsic noise in the condition that triggers the learning step outlined in Equation 22, the learning rule implicitly depends on the model parameters $\{r_j\}$, $\{\beta_{syn,j}\}$ describing the fluctuations in the neuron's inputs (indexed with j), and the parameter β_{int} which describes the neuron's intrinsic noise. Because Equation 22 is designed to approximately minimize the neuron's output spiking error probability for a given memory load (see Materials and Methods), which at capacity matches the desired output error probability of the neuron, r, the learning rule also depends implicitly on fluctuations in the neuron's output.

Figure 6 compares the theoretical solution obtained with the replica method in the $N \to \infty$ limit with numerical solutions for networks of $N = 200$, 400, and 800 neurons obtained with nonlinear optimization and the perceptron-type learning rule. Figure 6*A* shows that the perceptron-type learning rule sometimes fails to find a solution to a feasible learning problem, i.e., a problem that can be solved with nonlinear optimization. Yet, even in such cases, the perceptron connection weights in a steady state (after 10^6 learning epochs) are well-correlated with the nonlinear optimization weights (Fig. 6*B*). Therefore, though the perceptron-type learning rule is not as efficient as nonlinear optimization, it can find an approximate solution to the learning problem. Consistent with this conclusion, the associative memory storage capacity of a neuron loaded with the perceptron-type learning rule is 15–18% lower than that loaded with nonlinear optimization, and the two methods lead to similar structural and dynamical network properties (Fig. 6*C*, red and blue bars).

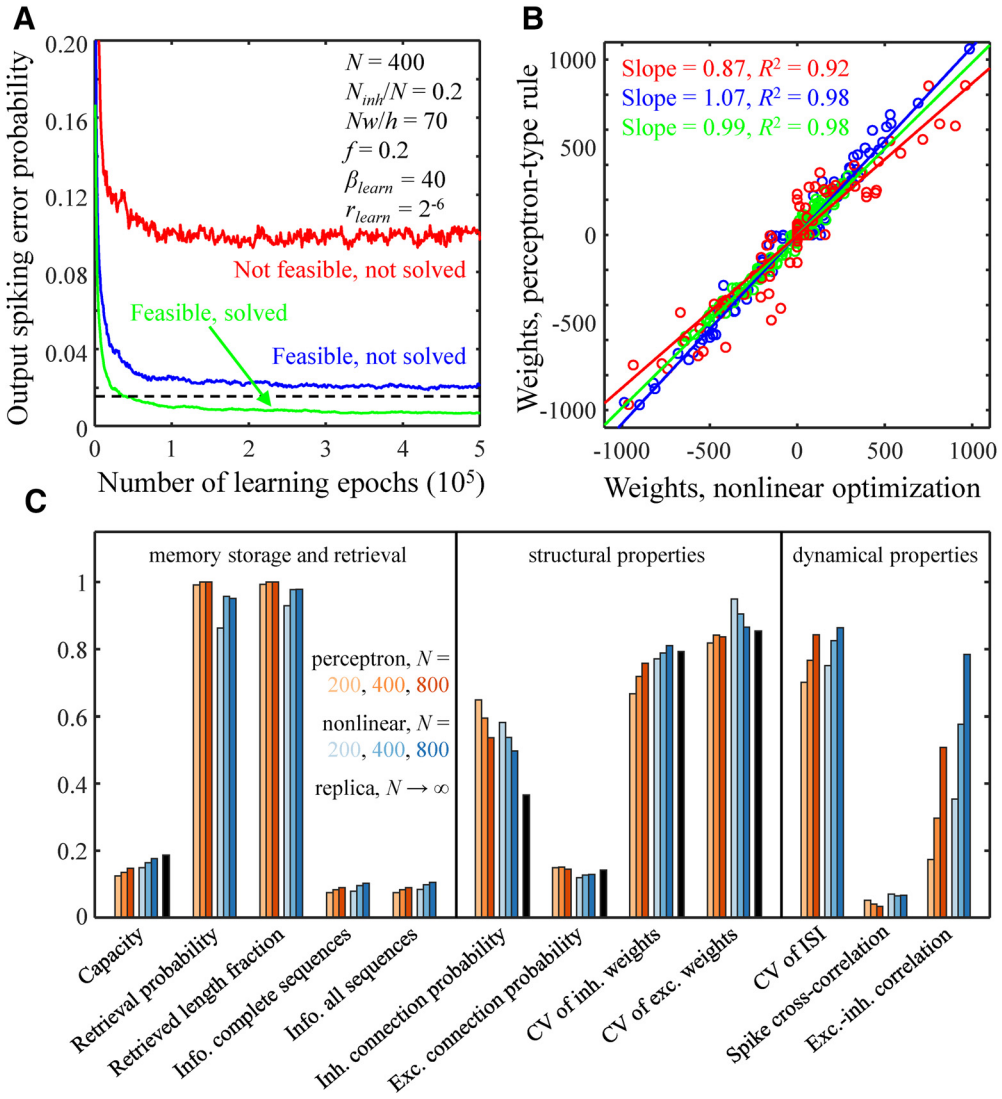

Figure 6. Comparison of solutions obtained with the perceptron-type learning rule, nonlinear optimization, and replica method. ***A***, Output error probability as a function of the number of learning epochs for the perceptron-type learning rule. The black dashed line indicates the target output error probability. Results for three different cases are shown: a not feasible problem (red line), a feasible problem which was not solved with the perceptron-type learning rule (blue line), and a feasible problem which was solved with the perceptron-type learning rule (green line). The parameters of the associative network are provided in the figure. The values of β_{learn} and r_{learn} correspond to the green asterisk from Figure 1*F*. ***B***, Comparisons of connection weights obtained with the perceptron-type learning rule and nonlinear optimization for the three cases shown in ***A***. Straight lines are the best linear fits. ***C***, Comparisons of memory storage capacity, retrieval, structural, and dynamical properties of networks of $N = 200$, 400, and 800 neurons obtained with the perceptron-type learning rule (red colors) and nonlinear optimization (blue colors). The memory storage capacity and structural properties calculated with the replica method in the $N \to \infty$ limit are shown in black.

The scales of non-zero inhibitory and excitatory connection weights according to the replica calculation are primarily determined by w, inhibitory/excitatory connection probabilities, and fractions of these inputs (Eq. 14, last line), and this agrees with the results of nonlinear optimization and perceptron learning.

Properties of heterogeneous associative networks

The associative learning model, Equation 1, makes it possible to investigate the properties of networks composed of heterogeneous populations of inhibitory and excitatory neurons. Specifically, we examined the effects of distributed spiking error probabilities and distributed synaptic and intrinsic noise strengths on properties of connectivity at critical capacity. Figure 7*A–C* shows that in networks of neurons with heterogeneous spiking error probabilities (homogeneous in all other parameters), the probabilities and weights of inhibitory and excitatory connections monotonically decrease with increasing r_{learn}. Therefore, as may have been expected, connections originating from more unreliable neurons (higher r_{learn}) are more likely to be depressed and/or eliminated during learning. Properties of networks of neurons with distributed

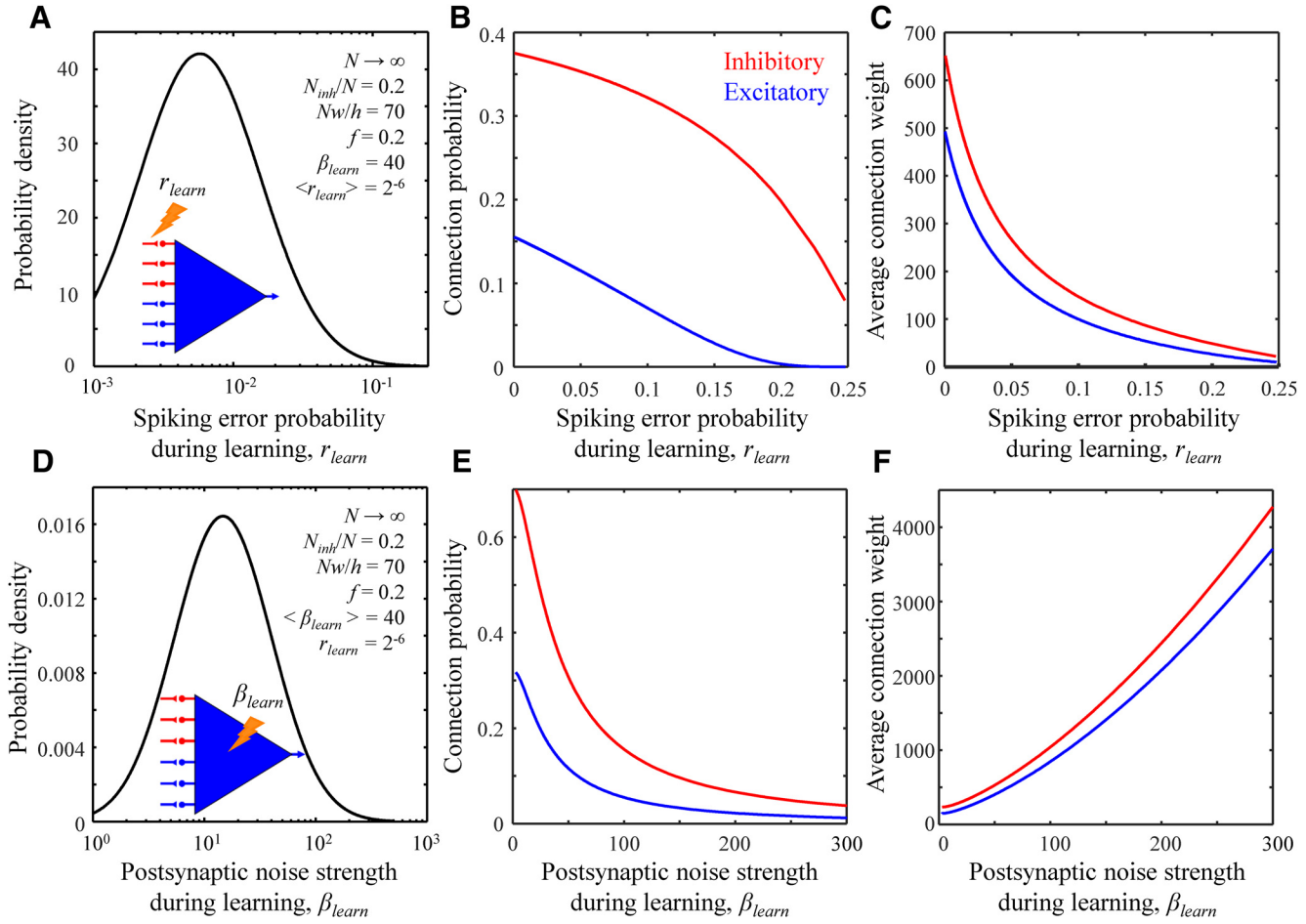

Figure 7. Properties of connections in associative networks of heterogeneous neurons. ***A–C***, Connection probability (***B***) and average non-zero connection weight (***C***) for inhibitory (red) and excitatory (blue) connections in a network of neurons with distributed spiking error probabilities and homogeneous in all other parameters. The spiking error probabilities of inhibitory and excitatory inputs during learning were randomly drawn from the log-normal distribution shown in ***A***. Unreliable inputs have lower probabilities and weights. The parameters of the associative network are shown in ***A***. The values of β_{learn} and $<r_{learn}>$ correspond to the green asterisk from Figure 1*F*. ***D–F***, Same for a network of neurons with heterogeneous postsynaptic noise strengths. The postsynaptic noise strengths of neurons during learning were randomly drawn from the log-normal distribution shown in ***D***. Noisier neurons receive stronger but fewer inhibitory and excitatory inputs.

synaptic and intrinsic noise strengths (homogeneous otherwise) depend on the combination of these parameters in the form of the postsynaptic noise strengths, β_{learn}. Figure 7*D–F* show how connection probabilities and average connection weights depend on β_{learn}. Like in the previous case, connections onto noisier neurons (higher β_{learn}) are less probable. Here, however, the average inhibitory and excitatory connection weights increase with β_{learn} because of the homeostatic l_1-norm constraint (Eq. 1).

Motivated by the agreement between the results of the associative learning model and cortical measurements, we put forward two predictions that can be tested in future experiments. First, we predict that in cortical networks, inhibitory and excitatory connections originating from more unreliable neurons or neuron classes must have lower connection probabilities and average uPSPs (Fig. 7*B*,*C*). Second, we predict that connections onto noisier neurons or neuron classes must have lower connection probabilities but higher average uPSPs (Fig. 7*E*,*F*).

Discussion

We examined a network model of inhibitory and excitatory neurons loaded to capacity with associative memory sequences in the presence of errors and noise. First, we showed that there is a trade-off between the capacity and reliability of stored sequences which is controlled by the levels of errors and noise present during learning. For an optimal trade-off, as judged by the amount of information contained in the retrieved sequences, noise must be present during learning. Second, as synaptic connectivity of neurons changes during learning (Holtmaat and Svoboda, 2009), it is not unreasonable to expect that the requirement of reliable memory retrieval is reflected in the properties of network connectivity and, consequently, the activity of neurons in the brain. Interestingly, local neural networks in the mammalian cortical areas have many common features of connectivity and network activity (Zhang et al., 2019b). We showed that these network properties in the model emerge all at once during reliable

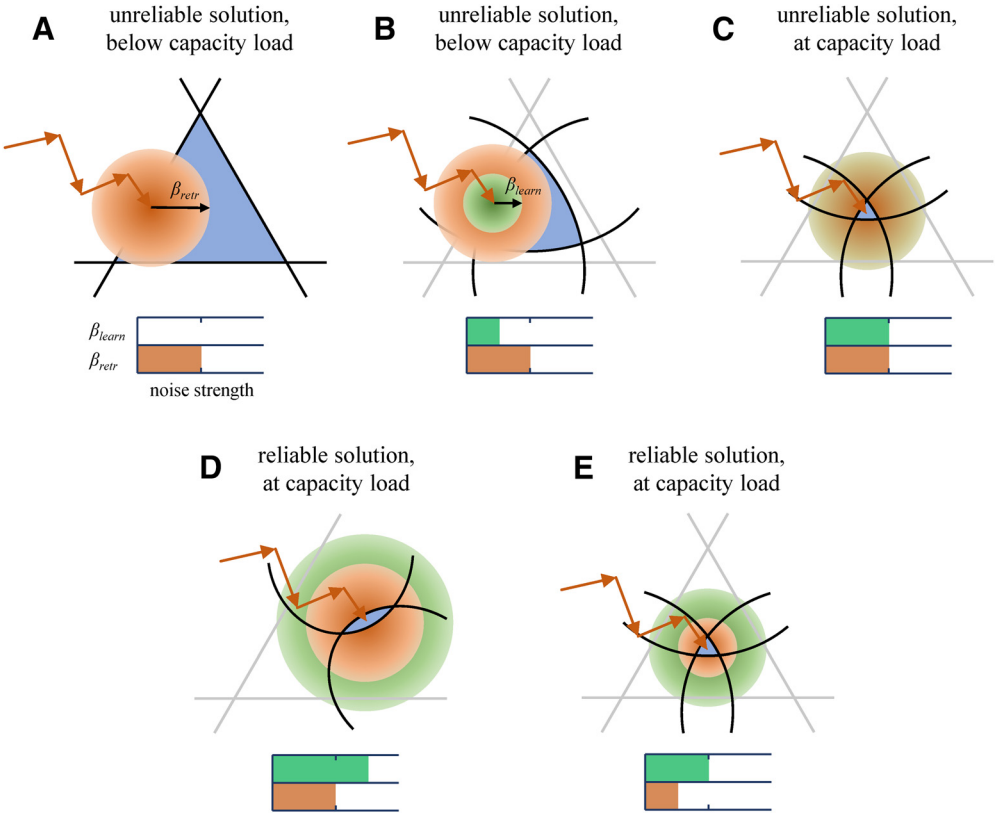

Figure 8. Increasing the noise strength during learning and decreasing it during memory recall lead to more reliable solutions. ***A***, The associative learning problem for a below capacity load in the absence of noise during learning, $\beta_{learn} = 0$. The solution region (blue) is bounded by hyperplanes corresponding to the individual associations (black lines). The learning phase (red arrows) ends as the connection weight vector enters the solution region. The solution shown in ***A*** is unreliable because noise during memory retrieval (red cloud of radius β_{retr}) can move it outside the solution region with high probability. ***B***, Adding noise during learning (green cloud of radius β_{learn}) transforms the association hyperplanes (gray lines) into hypersurfaces (black lines; Eq. 1), reducing the solution region and forcing the connection weight vector further away from the hyperplanes. This increases solution reliability. ***C***, The continued increase of the noise strength improves reliability as the solution region shrinks to zero. At this noise strength, the memory load is at capacity. A further increase in reliability can be achieved by increasing the noise strength during learning (***D***) or decreasing it during retrieval (***E***). In the former case, the memory load must be reduced to match the reduction in capacity.

memory storage. Third, as levels of errors and noise can differ across individual neurons or neuron classes, we examined the properties of model networks composed of heterogeneous neurons and made two salient predictions regarding the connectivity of neurons operating with relatively high levels of errors and noise.

This study incorporates a comprehensive description of errors and noise into the model of associate sequence learning by recurrent networks of neurons with biologically inspired constraints. It shows that errors and noise during learning can be beneficial, as they can increase the reliability of loaded memories to fluctuations during memory retrieval. Because errors and noise are both free and unavoidable harnessing their power, rather than trying to suppress it, may be an efficient way of improving the reliability of memories in the brain. This mechanism is illustrated in Figure 8. When the associative memories are loaded at a below capacity level, the solution region of Equation 1 is comparatively large. A solution, e.g., a vector of connection weights of a neuron obtained with a perceptron-type learning rule, may be located near the solution region boundary. Such a solution is deemed unreliable because a small amount of noise during memory retrieval can move it outside the solution region, resulting in spiking errors that can disrupt the associative sequence retrieval process (Fig. 8*A*). By adding noise during learning, the solution can be forced to move away from the boundary, thus making it more reliable (Fig. 8*B*). However, increasing the noise strength reduces the neuron's capacity, and at a certain strength, the capacity and memory load are guaranteed to match (Fig. 8*C*). A further increase in noise strength can improve the reliability even more, but at the expense of the memory load as the latter must remain at or below the capacity (Fig. 8*D*). An alternative way of improving reliability is by suppressing noise during memory retrieval (Fig. 8*E*). Incidentally, it has been shown that visual attention that improves behavioral performance reduces the variability in spike counts of individual neurons in Macaque V4 (Cohen and Maunsell, 2009; Mitchell et al., 2009). Though significant, the amount of reduction is relatively small, suggesting that this mechanism has physical limitations. Using noise during learning can enhance the reliability of stored memories beyond what can be accomplished by attending to the memory retrieval process.

The study of associative memory storage by artificial neural networks has a long history dating back to the seminal works of McCulloch and Pitts, Hebb, Rosenblatt, Steinbuch, Cover, Minsky, and Papert (McCulloch and Pitts, 1943; Hebb, 1949; Rosenblatt, 1957; Steinbuch, 1961; Cover, 1965; Minsky and Papert, 1969). Associative models of binary neurons can be generally categorized into learning models, in which memories are loaded into the network over time using activity-dependent learning rules, and memory storage models, which often bypass the learning phase and focus on memory storage capacity and properties of learned networks. Models of the first type often rely on Hebbian-type learning rules in which connection weights are modified based on activities of presynaptic and postsynaptic neurons (Willshaw et al., 1969; Hopfield, 1982; Tsodyks and Feigel'man, 1988; Amit, 1989; Palm, 2013). Although the general idea of Hebbian learning has been corroborated experimentally and characterized as long-term potentiation/long-term depression, recent studies demonstrated that changes in synaptic efficacy can have a complicated dependence on spike timing, spike frequency, and PSP (Sjöström et al., 2001).

Memory storage models make no assumptions as to the details of the learning rules, provided that they are powerful enough to load memories into the network, and analyze network properties as functions of the memory load and network parameters. An advantage of such models is that they often yield closed-form analytically solutions. One of the first models of this type was solved by Cover (Cover, 1965) who used a geometrical argument to show that a simple perceptron with N inputs can learn $2N$ unbiased associations. Later, a general framework for the analysis of memory storage capacity was established by Gardner and Derrida (Gardner, 1988; Gardner and Derrida, 1988) who used the replica theory to solve the problem of robust learning of arbitrarily biased associations. Subsequent studies incorporated sources of noise into the associative learning model and examined the effects of learning on neural network properties. In these studies, the basic associative learning model was extended to include biologically inspired elements, such as sign-constrained postsynaptic connections (inhibitory and excitatory; Kohler and Widmaier, 1991; Brunel et al., 2004; Chapeton et al., 2012), homeostatically constrained presynaptic connections (Chapeton et al., 2015), and robustness to noise which is traditionally enforced through a generic robustness parameter κ (Gardner, 1988; Gardner and Derrida, 1988). In particular, Brunel et al. (2004) and Brunel (2016) showed that sparse excitatory connectivity and certain two-neuron and three-neuron motifs develop in networks robustly loaded with associations to capacity and that similar results can be obtained in a model which, in place of κ, includes Gaussian intrinsic noise and output spiking errors (see their supplementary material). Rubin et al. (2017) considered presynaptic and intrinsic noise and showed that the balance of inhibitory and excitatory currents emerges at capacity. Zhang et al. (2019b) showed that many structural and dynamical properties of local cortical networks emerge in associative networks robustly loaded to capacity.

This article significantly differs from the above-mentioned studies both in terms of the model and results. First, the model introduced in this article provides a more systematic account of errors and noise by combining input and output spiking errors, synaptic and intrinsic noise. Second, the model allows for the possibility of having different levels of errors and noise during learning and memory retrieval. Third, the model makes it possible to analyze networks of neurons with heterogeneous properties. In terms of model results, we first show how errors and noise during learning facilitate reliable memory retrieval and next produce a comprehensive list of results related to network structure and dynamics that are then compared with the data from local cortical networks to validate the model and make predictions. What is more, our results explain the nature of the robustness parameter, κ, used in traditional models (Eq. 16) and show explicitly how it is related to errors and noise present during learning.

The model described in this study assumes that individual neurons learn independently from one another and are loaded with memories to capacity. There is no direct support for these assumptions, but they have been shown to lead to structural and dynamical network properties that are consistent with experimental data (Brunel et al., 2004; Clopath et al., 2010; Chapeton et al., 2012; Brunel, 2016; Zhang et al., 2019b). This study corroborates these assumptions by matching a variety of experimental results with a single set of model parameters. The derived perceptron-type rule mediates learning by modifying connection weights based on local activities of presynaptic and postsynaptic neurons in the presence of errors and noise, which is biologically feasible. However, a supervision signal must be fed to every neuron during learning. This is a major drawback of the presented approach and the supervised learning models in general, as the origins of this signal in the brain remain unknown. The problem can be minimized by feeding the supervision signal to a fraction of neurons in the network while letting the remaining neurons learn in an unsupervised manner (Krotov and Hopfield, 2019). Unsupervised learning can be mediated by local spike timing, frequency, and voltage-dependent rules that are biologically more plausible and can explain many experiments describing functional properties of individual neurons (Clopath et al., 2010). However, unsupervised learning rules are not known to produce the host of structural and dynamical properties of local cortical circuits examined in this study. It would be interesting to find out if a recurrent network composed of unsupervised and supervised neurons can satisfy all the requirements of a biologically realistic learning network.

References

Amit DJ (1989) Modeling brain function: the world of attractor neural networks. Cambridge; New York: Cambridge University Press.

Atallah BV, Scanziani M (2009) Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. Neuron 62:566–577.

Bishop CM (1995) Neural networks for pattern recognition. Oxford; New York: Clarendon Press; Oxford University Press.

Brunel N (2016) Is cortical connectivity optimized for storing information? Nat Neurosci 19:749–755.
Brunel N, Hakim V, Isope P, Nadal JP, Barbour B (2004) Optimal information storage and the distribution of synaptic weights: perceptron versus Purkinje cell. Neuron 43:745–757.
Buracas GT, Zador AM, DeWeese MR, Albright TD (1998) Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. Neuron 20:959–969.
Byrd RH, Hribar ME, Nocedal J (1999) An interior point algorithm for large-scale nonlinear programming. SIAM J Optim 9:877–900.
Byrd RH, Gilbert JC, Nocedal J (2000) A trust region method based on interior point techniques for nonlinear programming. Math Program 89:149–185.
Chapeton J, Fares T, LaSota D, Stepanyants A (2012) Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons. Proc Natl Acad Sci USA 109:E3614–E3622.
Chapeton J, Gala R, Stepanyants A (2015) Effects of homeostatic constraints on associative memory storage and synaptic connectivity of cortical circuits. Front Comput Neurosci 9:74.
Clopath C, Büsing L, Vasilaki E, Gerstner W (2010) Connectivity reflects coding: a model of voltage-based STDP with homeostasis. Nat Neurosci 13:344–352.
Cohen MR, Maunsell JH (2009) Attention improves performance primarily by reducing interneuronal correlations. Nat Neurosci 12:1594–1600.
Cohen MR, Kohn A (2011) Measuring and interpreting neuronal correlations. Nat Neurosci 14:811–819.
Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans Electron Comput 14:326–334.
Cox CL, Denk W, Tank DW, Svoboda K (2000) Action potentials reliably invade axonal arbors of rat neocortical neurons. Proc Natl Acad Sci USA 97:9724–9728.
Del Castillo J, Katz B (1954) Quantal components of the end-plate potential. J Physiol 124:560–573.
Edwards SF, Anderson PW (1975) Theory of spin glasses. J Phys F Met Phys 5:965–974.
Faisal AA, Selen LP, Wolpert DM (2008) Noise in the nervous system. Nat Rev Neurosci 9:292–303.
Feng F, Headley DB, Amir A, Kanta V, Chen Z, Paré D, Nair SS (2019) Gamma oscillations in the basolateral amygdala: biophysical mechanisms and computational consequences. eNeuro 6: ENEURO.0388-18.2018.
Gala R, Lebrecht D, Sahlender DA, Jorstad A, Knott G, Holtmaat A, Stepanyants A (2017) Computer assisted detection of axonal bouton structural plasticity in in vivo time-lapse images. Elife 6.
Gammaitoni L, Hänggi P, Jung P, Marchesoni F (1998) Stochastic resonance. Rev Mod Phys 70:223–287.
Gardner E (1988) The space of interactions in neural network models. J Phys A Math Gen 21:257–270.
Gardner E, Derrida B (1988) Optimal storage properties of neural network models. J Phys A Math Gen 21:271–284.
Graupner M, Reyes AD (2013) Synaptic input correlations leading to membrane potential decorrelation of spontaneous activity in cortex. J Neurosci 33:15075–15085.
Hebb DO (1949) The organization of behavior; a neuropsychological theory. New York: Wiley.
Hertz J, Krogh A, Palmer RG (1991) Introduction to the theory of neural computation. Redwood City: Addison-Wesley Publ Co.
Holt GR, Softky WR, Koch C, Douglas RJ (1996) Comparison of discharge variability in vitro and in vivo in cat visual cortex neurons. J Neurophysiol 75:1806–1814.
Holtmaat A, Svoboda K (2009) Experience-dependent structural synaptic plasticity in the mammalian brain. Nat Rev Neurosci 10:647–658.
Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci USA 79:2554–2558.
Kohler HM, Widmaier D (1991) Sign-constrained linear learning and diluting in neural networks. J Phys A Math Gen 24:L495–L502.
Krotov D, Hopfield JJ (2019) Unsupervised learning by competing hidden units. Proc Natl Acad Sci USA 116:7723–7731.
Mainen ZF, Sejnowski TJ (1995) Reliability of spike timing in neocortical neurons. Science 268:1503–1506.
McCulloch W, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115– 133.
McDonnell MD, Abbott D (2009) What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology. PLoS Comput Biol 5:e1000348.
McDonnell MD, Ward LM (2011) The benefits of noise in neural systems: bridging theory and experiment. Nat Rev Neurosci 12:415–426.
Minsky ML, Papert S (1969) Perceptrons: an introduction to computational geometry. Cambridge: The MIT Press.
Mitchell JF, Sundberg KA, Reynolds JH (2009) Spatial attention decorrelates intrinsic activity fluctuations in macaque area V4. Neuron 63:879–888.
Okun M, Lampl I (2008) Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. Nat Neurosci 11:535–537.
Palm G (2013) Neural associative memories and sparse coding. Neural Netw 37:165–171.
Rosenblatt F (1957) The Perceptron–a perceiving and recognizing automaton (Project PARA). Report No. 85–460-1. New York: Cornell Aeronautical Laboratory.
Rosenblatt F (1962) Principles of neurodynamics; perceptrons and the theory of brain mechanisms. Washington: Spartan Books.
Rubin R, Abbott LF, Sompolinsky H (2017) Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity. Proc Natl Acad Sci USA 114:E9366–E9375.
Salkoff DB, Zagha E, Yuzgec O, McCormick DA (2015) Synaptic mechanisms of tight spike synchrony at gamma frequency in cerebral cortex. J Neurosci 35:10236–10251.
Shadlen MN, Newsome WT (1998) The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. J Neurosci 18:3870–3896.
Sherrington D, Kirkpatrick S (1975) Solvable model of a spin glass. Phys Rev Lett 35:1792–1796.
Sjöström PJ, Turrigiano GG, Nelson SB (2001) Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. Neuron 32:1149–1164.
Softky WR, Koch C (1993) The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. J Neurosci 13:334–350.
Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15:1929–1958.
Stein RB, Gossen ER, Jones KE (2005) Neuronal variability: noise or part of the signal? Nat Rev Neurosci 6:389–397.
Steinbuch K (1961) Automat und Mensch; über menschliche und maschinelle Intelligenz. Berlin: Springer.
Stepanyants A, Chklovskii DB (2005) Neurogeometry and potential synaptic connectivity. Trends Neurosci 28:387–394.
Stepanyants A, Hirsch JA, Martinez LM, Kisvárday ZF, Ferecskó AS, Chklovskii DB (2008) Local potential connectivity in cat primary visual cortex. Cereb Cortex 18:13–28.
Stevens CF, Zador AM (1998) Input synchrony and the irregular firing of cortical neurons. Nat Neurosci 1:210–217.
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 58:267–288.
Trachtenberg JT, Chen BE, Knott GW, Feng G, Sanes JR, Welker E, Svoboda K (2002) Long-term in vivo imaging of experience-dependent synaptic plasticity in adult cortex. Nature 420:788–794.
Tsodyks MV, Feigel'man MV (1988) The enhanced storage capacity in neural networks with low activity level. Europhys Lett 6:101–105.
van Vreeswijk C, Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. Science 274:1724–1726.

van Vreeswijk C, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. Neural Comput 10:1321–1371.
Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Non-holographic associative memory. Nature 222:960–962.
Zhang C, Zhang D, Stepanyants A (2019a) Associative learning in the presence of errors and noise. Available from https://github.com/neurogeometry/Associative_Learning_with_Noise.
Zhang D, Zhang C, Stepanyants A (2019b) Robust associative learning is sufficient to explain the structural and dynamical properties of local cortical circuits. J Neurosci 39: 6888–6904.
Zhang C, Zhang D, Stepanyants A (2020) Noise in neurons and synapses enables reliable associative memory storage in local cortical circuits. bioRxiv 583922.