Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons

Julio Chapeton, Tarec Fares, Darin LaSota, and Armen Stepanyants*

Department of Physics and Center for Interdisciplinary Research on Complex Systems, Northeastern University, Boston, MA 02115

Edited by Terrence J. Sejnowski, Salk Institute for Biological Studies, La Jolla, CA, and approved November 2, 2012 (received for review July 10, 2012)

Many features of synaptic connectivity are ubiquitous among cortical systems. Cortical networks are dominated by excitatory neurons and synapses, are sparsely connected, and function with stereotypically distributed connection weights. We show that these basic structural and functional features of synaptic connectivity arise readily from the requirement of efficient associative memory storage. Our theory makes two fundamental predictions. First, we predict that, despite a large number of neuron classes, functional connections between potentially connected cells must be realized with <50% probability if the presynaptic cell is excitatory and >50% probability if the presynaptic cell is inhibitory. Second, we establish a unique relation between probability of connection and coefficient of variation in connection weights. These predictions are consistent with a dataset of 74 published experiments reporting connection probabilities and distributions of postsynaptic potential amplitudes in various cortical systems. What is more, our theory explains the shapes of the distributions obtained in these experiments.

learning and memory | cortical connectivity | synaptic weight | perceptron | critical capacity

Fundamental functions of the brain, such as learning and memory storage, are mediated by many mechanisms of excitatory (1–4) and inhibitory (5–7) synaptic plasticity. Working together with the genetically encoded developmental mechanisms of circuit formation, synaptic plasticity shapes neural circuits by creating, modifying, and eliminating individual synaptic connections in an experience-dependent manner. It is, therefore, reasonable to hypothesize that many stereotypic features of adult synaptic connectivity, whether established through evolution or the developmental learning process, have arisen to facilitate memory storage.

In this study, we focus on three such features of cortical connectivity. Cortical connectivity is predominantly excitatory; it is mediated by two major classes of neurons—excitatory glutamatergic and inhibitory GABAergic cells. Chemical synapses made by the axons of inhibitory cells in the adult brain are believed to be all inhibitory, whereas those synapses made by the axons of excitatory neurons are believed to be all excitatory (10). The resulting connectivity is largely excitatory, with only about 15–20% of inhibitory neurons and inhibitory synapses (11). The second stereotypic feature of cortical connectivity is sparseness. Networks in the cortex are thought to be organized into relatively small units ranging from hundreds to tens of thousands of neurons in size. Such units may include mini columns (12, 13), structural columns (14, 15), and a variety of functional columns (16, 17). Analysis of neuron morphology (14, 18–21) has shown that cells within such units have the potential of being connected by structural synaptic plasticity (22–24). However, despite this potential, synaptic connectivity within the units is sparse. For example, nearby excitatory neurons in the neocortex are synthetically coupled with less than 50% probability (25–30), and this probability decays with the increase in lateral distance between the neurons beyond the 100-μm range (29). Although connections between inhibitory and excitatory neurons are comparatively more frequent (29, 31), their probabilities also are well below 100%. Lastly, we note that the distributions of excitatory and inhibitory connection weights, measured in terms of the amplitudes of unitary postsynaptic potentials (PSPs), have a stereotypic shape (28–30, 32–34). This shape can be described as a truncated Gaussian, except for a somewhat heavier tail. In this study, we set out to show that the above basic structural and functional features of cortical circuits could have arisen from the hypothesized requirement of efficient memory storage.

Over the years, there has been a great deal of theoretical interest in the problem of associative memory storage. In particular, the perceptron (35, 36) and networks composed of McCulloch and Pitts neurons (37) received much attention (32, 38–46). This is due in part to the existence of a theoretical framework for solving such problems, which was initially developed in the context of statistical physics. Remarkably, most of the models considered thus far do not explicitly constrain the neurons in the network to be either excitatory or inhibitory, and thus, these models are not biologically plausible. Notable exceptions are the studies by Brunel et al. (32) and Barbour et al. (34) that analyzed associative memory storage in a model of a Purkinje cell receiving excitatory parallel fiber inputs. These works show that, at the maximum (critical) storage capacity, the probability of finding a functional parallel fiber to Purkinje cell connection is less than 50% (32, 34). In other words, the majority of parallel fibers that are potentially connected to a given Purkinje cell does not establish functional connections (i.e., connections are either silent or absent). What is more, the shape of the PSP amplitude distribution is consistent with the idea that the parallel fiber Purkinje cell system is functioning at its critical memory storage capacity.

In this study, we extend the framework in the works by Brunel et al. (32) and Barbour et al. (34), making it applicable to cortical neurons receiving both excitatory and inhibitory inputs. We show that the critical memory storage capacity and the shape of the connection weight distribution in the biologically constrained model considered here exhibit complex dependence on network architecture (i.e., connection probabilities and the fraction of inhibitory neurons). By comparing the theoretical results with the connection probabilities and the distributions of unitary PSP amplitudes measured in numerous cortical systems, we substantiate the hypothesis of efficient memory storage and put forward experimentally verifiable predictions regarding cortical connectivity.

Results

Associative Memory Storage in Local Cortical Circuits. We analyze a recurrent McCulloch and Pitts neural network (37) of \( N_{\text{exc}} \) inhibitory and \( N_{\text{inh}} = N - N_{\text{exc}} \) excitatory neurons (Fig. 1A, Left). This network represents a small cortical unit (e.g., structural or functional column), in which neurons are all to all connected in terms of potential synapses (14, 19). This restriction on the geometric size of the network is necessary to ensure that, if needed,

Author contributions: A.S. designed research; J.C., T.F., D.L., and A.S. performed research; J.C., T.F., D.L., and A.S. analyzed data; and J.C., T.F., D.L., and A.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. E-mail: a.stepanyants@neu.edu.

See Author Summary on page 20794 (volume 109, number 51).

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1211467109/-/DCSupplemental.

E3614-E3622 | PNAS | Published online December 3, 2012

www.pnas.org/cgi/doi/10.1073/pnas.1211467109
any connectivity diagram can be realized during learning. The weights, $J$, of inhibitory and excitatory connections in the model represent the amplitudes of inhibitory and excitatory postsynaptic potentials (IPSP and EPSP). Zero-weight connections correspond to unconnected neurons or neurons connected only with silent synapses. Our theoretical model needs not distinguish between these two possibilities, because both cases are classified as an absence of a connection in electrophysiological experiments. The activity of neurons in the network, where zero represents the resting state and one is the firing state, is described by the probability of firing, $f$. This activity is determined at every time step by comparing the neuron’s total input with its firing threshold, $h$ (Fig. 1A, Right). Both $f$ and $h$ may differ among neuron classes.

The state of the network in our model is described by a set of binary activities of all neurons. The neurons can learn to associate certain subsequent network states by modifying their connection weights. Such learning may be important for processing of sensory information and recall of memories. A number of constraints must be imposed on the learning process to build a biologically plausible model. First, because all synapses made by inhibitory neurons in the adult brain are inhibitory and all synapses made by excitatory neurons are excitatory, the signs of the connection weights in the network should remain fixed throughout learning. Such a network is referred to as sign-constrained (43), whereas in the absence of this constraint, the network is termed unconstrained. Second, our analysis of published electrophysiological data (Experimental Procedures and Tables S1 and S2) shows that the variability in the neurons’ firing thresholds is significantly smaller than the variability in the weights of connections that they receive. This observation motivated us to keep the firing thresholds of all model neurons fixed throughout learning. Third, it is commonly assumed that, similar to many biological networks (47, 48), cortical networks need to be robust (32, 40). Robustness may be required to withstand failures in generation or propagation of presynaptic action potentials, spontaneous neural activity, synaptic failure, and fluctuations in synaptic weight. The requirement of robustness is imposed in the model by the parameter $\kappa \geq 0$ (Experimental Procedures and SI Text). For nonzero $\kappa$, small fluctuations within the network do not necessarily change its dynamics.

In this work, we assume that the network must learn to associate pairs of subsequent states that are random and independent. With this assumption, the problem of learning by the network becomes equivalent to the problem of learning by the $N$ independent inhibitory and excitatory neurons that it contains. Consequently, it is sufficient to analyze learning in a single biologically constrained neuron (Fig. 1A, Right). The ratio between the number of associations ($m$) that such a neuron can learn and the total number of potential inputs ($N$) that it receives is referred to as the neuron’s capacity, $\alpha = m/N$. Fig. 1B shows how the probability of perfect learning depends on the number of presented associations. With increasing $m$, it becomes progressively more difficult to find connection weights that fulfill all associations, and therefore, the probability of successful learning is a decaying function of $m$. The capacity at which the associations can be learned with 50% success probability is known as the critical capacity, $\alpha_c$ (40). Fig. 1B illustrates that $\alpha_c = 2$ for a simple unconstrained perceptron learning.

Fig. 1. Associative memory storage in perceptron-like neural network models. (A) A local neural circuit (e.g., cortical column; Left), consisting of inhibitory (red) and excitatory (blue) neurons, may be modeled as a network of binary threshold units. Each such unit (Right) can learn to associate input patterns with appropriate outputs by adjusting the weights of its connections, $J$. (B) The probability of learning a set of presented associations decreases with the number of associations in the set, $m$. Critical capacity, $\alpha_c$ (dashed arrows), is defined as the number of associations per potential input, $m/N$, that can be learned with 50% probability of success. Critical capacity is two in the unconstrained and one in the sign-constrained perceptron model. The dependence of success probability on number of associations for this model is shown in red. Solid lines show theoretical results (38), and dotted lines are the results of numerical simulations. (C) Geometric interpretation of the presented model for a neuron receiving only two potential inputs. The two sets of inequalities in Eq. 1 parse the space of possible connection weights, $J_1$ and $J_2$, with padded dashed lines for the associations and the figure axes for the sign constraints. The colored areas mark all possible solution regions (if a solution exists). In the limit of a large number of potential inputs, $N$, the solution must be in the vicinity (gray region) of the hyperplane provided by Eq. 2.
random unbiased \( f = 0.5 \) associations (38) and that \( \alpha_c \) reduces to one if a sign constraint is applied (42–44). The red curve in Fig. 1B shows that the success probability in our model exhibits a similar critical dependence on the number of presented associations.

The biologically constrained neuron is faced with the learning task of finding connection weights \( J \) that fulfill the presented set of associations given the above-mentioned constraints (second set of inequalities in Eq. 1). This task is illustrated geometrically in Fig. 1C for a neuron receiving only two potential inputs. The axes in Fig. 1C parse the space of the possible connection weights into positive (excitatory) and negative (inhibitory) regions. The solution has to be confined to one of four quadrants as prescribed by the inhibitory/excitatory nature of the individual inputs (sign constraints). The dashed lines in Fig. 1C represent the individual associations that the neuron must learn. Robust learning restricts the solutions to be on a particular side of the margin surrounding each dashed line (first set of inequalities in Eq. 1). Combined, the two sets of requirements constrain the solutions (if a solution exists) to a single convex region (one of the colored regions in Fig. 1C).

In the limit of large \( N \) and \( m \), the total input received by the biologically constrained neuron is near threshold (SI Text), and the solution region is guaranteed to be on the hyperplane described by Eq. 2 (gray region in Fig. 1C). As the neuron learns a progressively increasing number of associations, the solution volume decreases, and at the critical number of associations, the typical volume of solution regions close as \( \alpha \) increases (Fig. 1C), lowering the solution probability. Any additional learning is not possible without forgetting some of the previously learned associations. Next, we examine the properties of the biologically constrained neuron at its critical capacity.

Properties of the Biologically Constrained Neuron at Critical Capacity. The properties of the biologically constrained neuron at critical capacity depend on the four model parameters: the fraction of potential inhibitory inputs, \( N_{inh}/N \); the firing probabilities of inhibitory and excitatory neurons, \( f_{inh} \) and \( f_{exc} \); and rescaled robustness parameter, \( \kappa = \kappa V/N \). Fig. 2 illustrates some aspects of this dependence. For networks of mostly excitatory neurons (\( N_{inh}/N < 0.5 \)), the critical capacity is a decreasing function of robustness (Fig. 2A). This trend can be explained by the fact that some of the solution regions close as \( \alpha \) increases (Fig. 1C), lowering the solution probability. In the limit of large \( \kappa \), all interior solution regions close, and \( \alpha_c \) converges to its minimum value (black line in Fig. 2A). These theoretical results were confirmed by numerical simulations (dotted lines in Fig. 2A).

A characteristic feature of sign-constrained neural networks at critical capacity is a finite fraction of zero-weight connections (32, 43). The zero-weight connections result from the clear delineation of synapses into excitatory and inhibitory. In the course of learning, the weights of many inhibitory and excitatory connections will approach zero, where many of these weights will remain, because they are unable to cross to the other side. This feature is essential for biological feasibility, because a finite fraction of zero-weight connections (or less than 100% probability of connection between neurons) is a ubiquitous feature of neural circuits in the brain. Fig. 2B illustrates how the connection probability, \( P_{con} \), depends on the value of robustness and the fraction of inhibitory neurons. For \( \kappa = 0 \), both excitatory and inhibitory connection probabilities remain at 50% for all values of \( N_{inh}/N < 0.5 \). In robust networks (\( \kappa > 0 \)), excitatory connection probabilities decrease, whereas the probabilities for inhibitory connections increase with increasing \( N_{inh}/N \) or \( \kappa \). This trend is accompanied with the overall strengthening of inhibitory and excitatory connections (Fig. 2C) and a decline in the neuron’s ability to recall learned associations reliably in the presence of noisy synaptic transmission (Fig. 2D). The reliability of recall is captured by the reliability parameter \( \rho \) (32), which is defined as the ratio of robustness to noise fluctuations in the input (SI Text). For some combinations of \( N_{inh}/N \), \( f_{inh}/f_{exc} \) and \( N_{inh}/N \), the critical capacity is a decreasing function of robustness and the fraction of inhibitory neurons. For \( N_{inh}/N \) or \( \kappa \), the critical capacity reduces to its minimum value (black line in Fig. 2A). These theoretical results were expressed only in terms of experimentally measurable parameters \( N_{inh}/N \), \( f_{inh} \), and \( f_{exc} \) (SI Text, section 6).

Comparison of Theoretical Predictions with Experimental Measurements. In this section, we test the predictive power of the theory by making comparisons with the available synaptic connectivity data. To this end, we examined experimental connection probabilities and shapes of unitary PSP amplitude distributions for inhibitory and excitatory inputs to principle cortical neurons (e.g., pyramidal cell in the cerebral cortex or Purkinje cell in the cerebellum). Although our theoretical results may hold well for nonprinciple postsynaptic cells, this idea is not examined here in detail because of insufficient experimental data (Discussion). To facilitate the comparison, the theoretical results are expressed only in terms of experimentally measurable parameters \( N_{inh}/N \), \( f_{inh} \), and \( f_{exc} \) (SI Text, section 6).

Fig. 3A shows that the volume of biologically plausible solutions, expressed in terms of these parameters, is very limited. It amounts to only 4.9% of the total volume of the parameter space (unit cube in Fig. 3A). Any combination of \( N_{inh}/N \), \( f_{inh} \), and \( f_{exc} \) outside the 4.9% volume corresponds to a subcritical network or a completely unreliable network (\( \rho = 0 \)) with prohibitively large connection weights.

Fig. 3B shows experimental distributions of inhibitory and excitatory connection probabilities gathered from a large number of studies (Experimental Procedures and Table S2). In agreement with the theoretical predictions, probabilities of excitatory connections are entirely within the region of biologically plausible solutions (green square in Fig. 3B); inhibitory connection probabilities, although not entirely within the region, have a significant overlap. It should be noted that connection probabilities obtained in tissue-slice recording experiments may underestimate the in vivo
probabilities (49). This artifact is expected to primarily affect projections with low connection probabilities. Therefore, the low connection probabilities in Fig. 3B may need to be scaled up, which would only improve the agreement between the theory and experiment. Two of the studies from Table S2(29, 50) measured both inhibitory and excitatory connection probabilities. These measurements from rat visual and somatosensory cortices (Fig. 3B, ○ and □) are well within the boundaries of the biologically plausible solution region.

Is there an optimal neural network configuration within the biologically plausible solution volume of Fig. 3A? To examine this question, we looked at the behavior of critical capacity, average absolute connection weight, and reliability of learned associations on three cross-sections of the solution volume (Fig. 3C–E). In an all-excitatory network (32), there is a clear tradeoff between capacity and reliability (Fig. 3CI and EI). With the addition of inhibitory neurons, the tradeoff becomes more complicated, because the average absolute connection weight comes into play. Because connection weight is correlated with the presynaptic bouton volume, the number of synaptic vesicles, the area of the postsynaptic density, and the spine head volume (51–59), stronger connections entail a larger metabolic cost, and they are detrimental to the organism. The three arrows in Fig. 3C2, D2, and E2 illustrate the general directions of increasing capacity and reliability and decreasing connection weight—factors that may be important for efficient memory storage. However, another consideration should be mentioned. At high fractions of inhibitory neurons (greater than 50%), the solution volume is extremely small (0.65% of the total volume of the parameter space), and memory storage may become unstable. For example, small fluctuations in the inhibitory connection probability could push the network out of the solution volume and into the subcritical regime. Balance of the above four functional forces (white arrows in Fig. 3) is expected to play a role in determining the optimal network configuration, whereas the result is likely to be cortical area-dependent.

Fig. 3. Biologically plausible solutions. (A) The region of biologically plausible solutions. In a network operating at critical capacity, excitatory connection probabilities must be less than 50%, whereas probabilities of inhibitory connections must be greater than 50%. (B) This result is in agreement with the available experimental data. Red and blue histograms summarize inhibitory (n = 16) and excitatory (n = 45) connection probabilities from Table S2. The green square delineates the region of biologically plausible solutions; ○ and □ show the results of two experiments, where inhibitory and excitatory connection probabilities were measured in the same systems. (C–E) Critical capacity, average absolute connection weight, and reliability of the model neuron within the biologically plausible solution region from A. Different panels in C–E show the cross-sections of the solution region produced by the planes 1, 2, and 3 from A. White arrows illustrate the general directions of increasing stability, increasing capacity, decreasing connection weight, and increasing reliability.
Next, we compared the shapes of the theoretical and experimental distributions of connection weights. Similar to the study by Brunel et al. (32) performed for all-excitatory connections, our theory predicts that the distributions of inhibitory and excitatory connection weights must consist of finite fractions of zero-weight connections and Gaussians shifted to the left and truncated at zero (Eq. 3). This theoretical finding is validated with a numerical simulation in Fig. 4A. For the first point of comparison, we derived a unique theoretical expression relating (with no free parameters) the coefficient of variation (CV) in connection weight and the connection probability. Fig. 4B shows the experimental data (red) and the theoretical connection weights to the scale of unitary PSP amplitudes recorded in millivolts. Gray bins in all of the histograms in Fig. 5 were deemed unreliable and excluded from fitting (Experimental Procedures). The goodness of all fits, as captured by the adjusted $R^2$ coefficients $0.92 \pm 0.07$ (mean ± SD), was high. As expected, a closer examination of the tails of the distributions revealed the deviations from the theoretical model.

To examine this deviation in more detail, we replotted the two highest count distributions (Fig. 5F and G) along with their fits on a logarithmic scale using logarithmic binning. Fig. 6A and B shows that the theoretical fits appear to be nearly perfect on a semilog scale (adjusted $R^2 = 0.995$ and 0.989). However, on a log-log scale (Fig. 6C), the deviation between the tails of the theoretical Gaussian distributions (Fig. 6, dashed lines) and the experimental distributions (Fig. 6, solid lines) becomes apparent. Remarkably, the two experimental distributions of connection weights obtained from rat visual and mouse barrel cortices (Fig. 6C, solid blue and red lines) have nearly identical shapes on a log-log scale. In fact, the remaining seven distributions from Fig. 5 exhibit very similar nonexponential decays (not confirmed statistically because of the low counts), suggesting that this feature may be inherent to many cortical areas.

Notably, all-excitatory to excitatory connections must receive large ($\sim 1,000$) but finite numbers of potential inputs. The work by Brunel et al. (32) shows that such stereotypic features of adult synaptic connectivity originate from this requirement. To test the hypothesis, we theoretically solved the problem of associative memory storage in a model network of biologically constrained inhibitory and excitatory neurons. Our results quantitatively explain many basic features of cortical connectivity, providing support for the efficient memory storage hypothesis.

Our theoretical analysis builds on the work by Brunel et al. (32), which considers an excitatory feed-forward system consisting of a cerebellar Purkinje cell receiving excitatory parallel fiber inputs. The work by Brunel et al. (32) shows that such a system is analogous to a perceptron functioning at critical memory storage capacity. Motivated by recent observations of inhibitory neuron plasticity (5–9), we developed a more general theory, which can be applied to networks containing not only excitatory but also inhibitory neurons. Our results suggest that both neuron classes are actively involved in efficient associative memory storage.

Specifically, our theory makes two experimentally testable predictions. First, we predict that, despite a large number of neuron classes, all inhibitory to excitatory connections between potentially connected neurons must be realized with $>50\%$ connection probability. In contrast, all-excitatory to excitatory connections must have $<50\%$ connection probability. Hence, connectivity is sparse (i.e., it contains a large fraction of zero-weight connections). A zero-weight connection between potentially connected neu-
rons corresponds to the absence of functional synaptic contacts (unconnected neurons or neurons connected by silent synapses). Second, we predict that the CV in the connection weights is uniquely related to the probability of that connection. These fundamental predictions do not depend on any parameters and must hold for all circuits that are designed for learning. We tested these predictions on a dataset compiled from a large number of published studies, in which connection probabilities and unitary EPSP/IPSP amplitudes had been measured for postsynaptic principle cells in various cortical systems. These measurements, based on a total of 74 experiments, are in good agreement with the theory (Figs. 3B and 4B).

Do the model predictions extend on postsynaptic inhibitory cells? To examine this question, we identified eight studies in which the probabilities of excitatory and inhibitory connections onto postsynaptic inhibitory cells were determined based on recordings from at least 20 connected neuron pairs. These probabilities are 0.60 (n = 73 connected pairs) among inhibitory [fast spiking inhibitory (FS) and somatostatin positive inhibitory cell] cells (64) and 0.47 (n = 110) for excitatory (regular spiking excitatory cell) to

**Table 1.** Nine largest count distributions of unitary PSP amplitudes obtained from published experimental studies

<table>
<thead>
<tr>
<th>No.</th>
<th>Species</th>
<th>Brain area</th>
<th>Projection</th>
<th>Age (d)</th>
<th>( P_{con} ) (no. of connected pairs)</th>
<th>Unitary PSP amplitude mean ± SD (mV; no. of pairs)</th>
<th>CV</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Rat</td>
<td>SS</td>
<td>L5 PC → L5 PC</td>
<td>P14–16</td>
<td>0.10 (138)</td>
<td>1.3 ± 1.1 (138)</td>
<td>0.85</td>
<td>28</td>
</tr>
<tr>
<td>B</td>
<td>Rat</td>
<td>VC and SS</td>
<td>L2/3 PC → L2/3 PC</td>
<td>P14–16</td>
<td>0.06 (83)</td>
<td>0.65 ± 0.64 (83)</td>
<td>0.98</td>
<td>29</td>
</tr>
<tr>
<td>C</td>
<td>Rat</td>
<td>VC and SS</td>
<td>L2/3 FS → L2/3 PC</td>
<td>P14–16</td>
<td>0.55 (109)</td>
<td>3.0 ± 2.5 (109)</td>
<td>0.85</td>
<td>29</td>
</tr>
<tr>
<td>D</td>
<td>Rat</td>
<td>BC</td>
<td>L4 EXC → L4 EXC</td>
<td>P12–15</td>
<td>0.20–0.31 (132)</td>
<td>1.6 ± 1.5 (132)</td>
<td>0.95</td>
<td>26</td>
</tr>
<tr>
<td>E</td>
<td>Rat</td>
<td>VC</td>
<td>L5 PC → L5 PC</td>
<td>P12–21</td>
<td>0.15 (239)</td>
<td>0.73 ± 0.63 (139)</td>
<td>0.86</td>
<td>61</td>
</tr>
<tr>
<td>F</td>
<td>Rat</td>
<td>VC</td>
<td>L5 PC → L5 PC</td>
<td>P14–16</td>
<td>0.12 (931)</td>
<td>0.77 ± 0.84 (931)</td>
<td>1.1</td>
<td>30</td>
</tr>
<tr>
<td>G</td>
<td>Mouse</td>
<td>BC</td>
<td>All layers, EXC → EXC</td>
<td>P18–21</td>
<td>0.10 (909)</td>
<td>0.75 ± 0.94 (909)</td>
<td>1.2</td>
<td>33</td>
</tr>
<tr>
<td>H</td>
<td>Rat</td>
<td>CB</td>
<td>Granule → Purkinje cell</td>
<td>P60–90</td>
<td>0.071 (34)</td>
<td>0.072 ± 0.064 (104)</td>
<td>0.89</td>
<td>32, 62</td>
</tr>
<tr>
<td>I</td>
<td>Guinea pig</td>
<td>HC</td>
<td>CA3 PC → CA1 PC</td>
<td>600–900 g</td>
<td>0.063 (72)</td>
<td>0.13 ± 0.11 (74)</td>
<td>0.83</td>
<td>63</td>
</tr>
</tbody>
</table>

Letters in the first column match Fig. 5. The complete dataset analyzed in this study is described in Table S2. BC, barrel cortex; CB, cerebellum; EXC, excitatory (pyramidal and spiny stellate) cells; HC, hippocampus; PC, pyramidal cell; SS, somatosensory cortex; VC, visual cortex.

Fig. 5. Theoretical distributions of connection weights are consistent with experimental measurements. Histograms in A–I show experimental probability densities of connection weights obtained from the studies listed in Table 1. Note that C describes an inhibitory projection. The green lines are one-parameter fits according to Eq. 3. Because of the experimental uncertainties in the detection of weak connections, gray bars in the histograms were ignored during fitting. In all cases, the area under the fit function (excluding the range of the gray bars) is normalized to match the area of the blue bars. The goodness of fits is captured by the adjusted \( R^2 \) coefficients.
inhibitory (FS and low-threshold spiking inhibitory cell) connections (65) in cortical layer 4 (L4) of rodent barrel cortex; 0.45 (n = 146) in auditory cortex (66) and 0.59 (n = 79) in visual and somatosensory cortices (29) for excitatory to inhibitory (FS) projections in rodent L2/3; 0.43 (n = 26) for pyramidal to Martinotti connections in L5 of rat somatosensory cortex (67); 0.21 (n = 22) for pyramidal to inhibitory connections in L2/3 of rat somatosensory, motor, and visual cortices (68); 0.88 (n = 36) for excitatory to inhibitory (FS) connections in L2/3 of mouse V1 (69); and 0.76 (n = 28) for L4 spiny stellate to L2/3 inhibitory cell connections in rat barrel cortex (70). The last two of these projections deviate significantly from the predictions of the model, indicating that either these projections are not directly involved in efficient associative memory storage or the model assumptions must be revised to extend the theory on these classes of postsynaptic inhibitory neurons, which could be one of the reasons for the presence of inhibitory neurons in cortical circuits. Our results suggest that sparse cortical circuits, built with small fractions of inhibitory neurons and synapses, result from the tradeoff among network capacity, reliability, metabolic cost, and stability (Fig. 3).

**Experimental Procedures**

In this section, we provide a summary of the model, state model assumptions, outline the main theoretical results, and give a brief description of the experimental dataset used to validate these results. A more detailed description can be found in SI Text.

**Formulation of the Model.** In our theoretical model, a neuron receives N potential connections, of which Ninh are inhibitory and Nexc are excitatory (Fig. 1A). The neuron is presented with a set of m binary (0, 1) input–output associations (cij→≡), which are randomly drawn from the probability distributions of inhibitory and excitatory inputs, Xinh,exc. Here, index μ = 1, ..., m enumerates different associations, and j = 1, ..., N enumerates the potential inputs. The associations must be learned by the neuron subject to a number of biologically motivated constraints: (i) learning is mediated through changes in the connection weights, Jj, whereas the firing threshold, h, is held constant; (ii) weights of the inhibitory and excitatory connections cannot change sign throughout learning; and (iii) the associations must be learned robustly, which is enforced by the robustness parameter, ρ.

The model can be written as a set of inequalities—m inequalities to enforce robust implementation of the associations and N inequalities to impose the sign constraints (Eq. 1):

\[
\begin{align*}
(2^\rho - 1) \left( \sum_{j=1}^{N} j \cdot \mathbb{I}_{\text{inh,exc}}(\mu, j) \right) & > x > 0, \quad \mu = 1, \ldots, m \\
J_j g^0 & > 0, \quad j = 1, \ldots, N \\
\mathbb{I}_{\text{inh,exc}} \in X_{\text{inh,exc}}; \quad \mathbb{I}_{\text{exc}} \in X_{\text{exc}} \\
\mathbb{I}_{\text{inh,exc}} &= \{ 0, \quad 1 \mathbb{I}_{\text{inh,exc}} \}.
\end{align*}
\]

Here, f is the probability of firing, and g specifies the sign of input weight j (−1 for inhibitory and +1 for excitatory).

In the large N limit, the total input received by the biologically constrained neuron is near threshold (Eq. S3), and as a result, all solutions of Eq. 1 must be located in the vicinity of a single hyperplane (gray region in Fig. 1C) (Eq. 2):

\[
\sum_{j=1}^{N} \left( \frac{1-g_j}{2} f_{\text{inh}} + \frac{1+g_j}{2} f_{\text{exc}} \right) j = h.
\]

The critical capacity of the model and the distribution of connection weights at critical capacity can be calculated in this limit by using the replica theory from statistical physics (32, 40). In result, the critical capacity αc, connection probabilities pinh,exc, and probability densities of inhibitory and excitatory connection weights, Pinh,exc, are determined in terms of the four model

**Fig. 6.** The distributions of connection weights have heavy, nonexponentially decaying tails. To examine the tails of the distributions, we replotted the two largest count probability densities from Fig. 5 F and G on a semilog scale in A and B and a log-log scale in C. Equal-sized bins on the logarithmic scale were used in these plots. The blue and red solid lines in C correspond to the distributions in A and B, respectively. The dashed lines are the corresponding fits from A and B. (D) The nonexponential decay observed experimentally is reproduced in numerical simulations. To illustrate this point, we replotted the distribution of excitatory connections from Fig. 4A on a log-log scale (solid black line). The dashed line is the theoretical fit from Fig. 4A.

**Table 1.** The distributions of connection weights have heavy, nonexponentially decaying tails. To examine the tails of the distributions, we replotted the two largest count probability densities from Fig. 5 F and G on a semilog scale in A and B and a log-log scale in C. Equal-sized bins on the logarithmic scale were used in these plots. The blue and red solid lines in C correspond to the distributions in A and B, respectively. The dashed lines are the corresponding fits from A and B. (D) The nonexponential decay observed experimentally is reproduced in numerical simulations. To illustrate this point, we replotted the distribution of excitatory connections from Fig. 4A on a log-log scale (solid black line). The dashed line is the theoretical fit from Fig. 4A.
parameters, $N_{inh}$, $f_{inh \, peak}$, and $k = \kappa/N$. In particular, one can show that the probability densities of nonzero inhibitory and excitatory connection weights are Gaussian, shifted to the left by an amount dependent on connection probability, and truncated at zero (Eq. 3):

$$P_{\text{post,exc}}(J) = \frac{1}{\sqrt{2\pi} \sigma_{\text{post,exc}}} e^{-\frac{(J - \text{erfinv}(2\pi J - 1))}{2\sigma_{\text{post,exc}}^2}}$$

Parameter $s$ in this expression, defining the distribution scale, is the only parameter used in the fitting of the experimental distributions in Fig. 5. erfinv denotes the inverse error function.

Model Assumptions and Approximations. Below, we provide some biological justifications for a number of assumptions and approximations made in the theory. McCulloch and Pitts neurons with binary (0, 1) inputs and outputs were just introduced in the 1940s. 

**Parameter Assumptions.** Throughout this study, we discuss only two types of neurons, inhibitory and excitatory. Such binning is possible if the average unitary PSP ($\tau$) is small and reduce to zero at low firing frequencies (ref. 81 and references therein). With this approximation, it is possible to decouple the recurrent connections between neurons in different systems are much more variable than the distributions of the neurons’ firing thresholds. For example, Tables S1 and S2 show that the CVs in connection weights, 0.47–1.7 [0.94 ± 0.03 (mean ± SD), n = 52 systems], are significantly higher than the CVs for firing thresholds, 0.10–0.24 (0.17 ± 0.02, n = 9 systems).

The activities of the model neurons at different time steps are randomly drawn from their respective probability distributions, giving rise to independent network states. This approximation is motivated by the Poisson-like statistics of cortical spike trains recorded in vivo (80). Although successive interspike intervals are often negatively correlated, these correlations are small and reduce to zero at low firing frequencies (ref. B1 and references therein). With this approximation, it is possible to decouple the recurrent network into a set of independent perceptual-like units (Fig. 1A) (24). We also assume that these units are functioning at their critical capacities, which enables us to find a unique analytic solution, independent of the details of the learning rule. Although it is difficult to evaluate the effect of these assumptions directly, they are supported by the agreement of our theoretical results with the experimental measurements.

Throughout this study, we discuss only two types of neurons, inhibitory and excitatory; however, the theory can be easily generalized to include any number of distinct neuron classes. Interestingly, a similar deviation is observed experimentally. Replica measurements obtained in patch-clamp or sharp-electrode recording experiments from a number of laboratories have revealed that the probability distributions of inhibitory connection weights are Gaussian, shifted to the left by an amount dependent on connection probability, and truncated at zero (Eq. 3).

The dataset used in this study was compiled from a number of published articles reporting local connectivity measurements obtained in patch-clamp or sharp-electrode recording experiments. In particular, we targeted studies containing distributions of unitary PSP amplitudes, mean values and SDs of unitary PSP amplitudes, and connection probability distributions. Initially, we selected about 200 articles published in the major neuroscience journals, but later, we limited our analysis to experiments in which recordings were made from at least 20 connected neuron pairs (Table S2). Furthermore, because of very sparse data regarding connections onto postsynaptic inhibitory interneurons (Discussion), we decided to limit our analysis to excitatory and inhibitory connections onto the principle neurons (e.g., pyramidal and spiny stellate cells in the cerebral cortex and Purkinje cells in the cerebellum). These restrictions lead to 42 articles describing 74 projections.

A custom-made MatLab (MathWorks) algorithm was used to extract accurate quantitative information from high-resolution digital images of published distributions of unitary PSP amplitudes. The distributions were fit with Eq. 3 in MatLab using the Nonlinear Least Squares method, and goodness of fits was checked by the adjusted coefficient of determination provided by MatLab’s fit function. Because of fluctuations in baseline recordings, very weak connections between neurons cannot be detected reliably. Such connections are often missed or ignored in experiments, leading to a systematic underestimate of weak connection counts. Therefore, connections with weights below PSP detection thresholds [0.1–0.25 mV in rodent neocortex (26, 28, 82, 83) and 0.017 mV in cerebellum (32)] were deemed unreliable and ignored during fitting (gray bins in Fig. 5 and gray areas in Fig. 6).

Numerical simulations were performed to confirm the results of the theoretical calculations (Figs. 1B, 2A, and 4A) and examine the tail of the connection weight distribution (Fig. 6D). These simulations were performed by using a modified perceptron learning algorithm (SI Text).

**Acknowledgments.** We thank Vincent Hakim, Nicolas Brunel, Boris Barbour, and Dimitri Chklovskii for constructive discussions related to the subject of this study. A portion of this work was done by A.S. while on a sabbatical leave at the Statistical Physics Laboratory of Ecole Normale Supérieure in Paris. This work was supported by National Institutes of Health Grants NS047138 and NS063494.


38. Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition.


