Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity

Ran Rubin[a,1], L. F. Abbott[a,b], and Haim Sompolinsky[c,d]

[a]Department of Neuroscience, Columbia University, New York, NY 10027; [b]Department of Physiology and Cellular Biophysics, Columbia University, New York, NY 10027; [c]Edmond and Lily Safra Center for Brain Sciences, Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel; and [d]Center for Brain Science, Harvard University, Cambridge, MA 02138

Neurons and networks in the cerebral cortex must operate reliably despite multiple sources of noise. To evaluate the impact of both input and output noise, we determine the robustness of single-neuron stimulus selective responses, as well as the robustness of attractor states of networks of neurons performing memory tasks. We find that robustness to output noise requires synaptic connections to be in a balanced regime in which excitation and inhibition are strong and largely cancel each other. We evaluate the conditions required for this regime to exist and determine the properties of networks operating within it. A plausible synaptic plasticity rule for learning that balances weight configurations is presented. Our theory predicts an optimal ratio of the number of excitatory and inhibitory synapses for maximizing the encoding capacity of balanced networks for given statistics of afferent activations. Previous work has shown that balanced networks amplify spatiotemporal variability and account for observed asynchronous irregular states. Here we present a distinct type of balanced network that amplifies small changes in the impinging signals and emerges automatically from learning to perform neuronal and network functions robustly.

E/I balance | synaptic learning | associative memory

The response properties of neurons in many brain areas including cerebral cortex are shaped by the balance between coactivated inhibitory and excitatory synaptic inputs (1–5) (for a review see ref. 6). Excitation–inhibition balance may have different forms in different brain areas or species and its emergence likely arises from multiple mechanisms. Theoretical work has shown that, when externally driven, circuits of recurrently connected excitatory and inhibitory neurons with strong synapses settle rapidly into a state in which population activity levels ensure a balance of excitatory and inhibitory currents (7, 8). Experimental evidence in some systems indicates that synaptic plasticity plays a role in maintaining this balance (9–12). Here we address the question of what computational benefits are conferred by the excitation–inhibition balance properties of balanced and unbalanced neuronal circuits. Although it has been shown that networks in the balanced states have advantages in generating a fast and linear response to changing stimuli (7, 8, 13, 14), the advantages and disadvantages of excitation–inhibition balance for general information processing have not been elucidated [except in special architectures (15–17)]. Here we compare the computational properties of neurons operating with and without excitation–inhibition balance and present a constructive computational reason for strong, balanced excitation and inhibition: It is needed for neurons to generate selective responses that are robust to output noise, and it is crucial for the stability of memory states in associative memory networks. The distinct balanced networks we present here naturally and automatically emerge from synaptic learning that endows neurons and networks with robust functionality.

We begin our analysis by considering a single neuron receiving input from a large number of afferents. We characterize its basic task as discriminating patterns of input activation to which

it should respond by firing action potentials from other patterns which should leave it quiescent. Neurons implement this form of response selectivity by applying a threshold to the sum of inputs from their presynaptic afferents. The simplest (parsimonious) model that captures these basic elements is the binary model neuron (18, 19), which has been studied extensively (20–23) and used to model a variety of neuronal circuits (24–28). Our work is based on including and analyzing the implications of four fundamental neuronal features not previously considered together: (*i*) nonnegative input, corresponding to the fact that neuronal activity is characterized by firing rates; (*ii*) a membrane potential threshold for neuronal firing above the resting potential (and hence a silent resting state); (*iii*) sign-constrained and bounded synaptic weights, meaning that individual synapses are either excitatory or inhibitory and the total synaptic strength is limited; and (*iv*) two sources of noise, input and output noise, representing fluctuations arising from variable stimuli and inputs and from processes within the neuron. As will be shown, these features imply that, when the number of input afferents is large, synaptic input must be strong and balanced if the neuron's response selectivity is to be robust. We extend our analysis to recurrently connected networks storing long-term memory and find that similar balanced synaptic patterns are required for the stability of the memory states against noise. In addition, maximizing the performance of neurons and networks in the balanced state yields a prediction for the optimal ratio of excitatory to inhibitory inputs in cortical circuits.

Results

Our model neuron is a binary unit that is either active or quiescent, depending on whether its membrane potential is above

> ### Significance
>
> Neurons and networks in the cerebral cortex must operate reliably despite multiple sources of noise. Using a mathematical analysis and model simulations, we show that noise robustness requires synaptic connections to be in a balanced regime in which excitation and inhibition are strong and largely cancel each other. Our theory predicts an optimal ratio for the number of excitatory and inhibitory synapses that depends on the statistics of afferent activity and is consistent with data. This distinct form of excitation–inhibition balance is essential for robust neuronal selectivity and crucial for stability in associative memory networks, and it emerges automatically from learning in the presence of noise.

or below a firing threshold. The potential, labeled V_{PSP}, is a weighted sum of inputs x_i, $i = 1, 2, ..., N$, that represent afferent firing rates and are thus nonnegative,

$$V_{\text{PSP}}\left(\boldsymbol{x},\, \boldsymbol{w}\right) = V_{\text{rest}} + \sum_{i=1}^{N} w_i x_i, \qquad [1]$$

where V_{rest} is the resting potential of the neuron and $\boldsymbol{x}$ and $\boldsymbol{w}$ are N-component vectors with elements x_i and w_i, respectively. The weight w_i represents the synaptic efficacy of the ith input. If $V_{\text{PSP}} \geq V_{\text{th}}$ the neuron is in an active state; otherwise, it is in a quiescent state. To implement the segregation of excitatory and inhibitory inputs, each weight is constrained so that $w_i \geq 0$ if input i is excitatory and $w_i \leq 0$ if input i is inhibitory.

To function properly in a circuit, a neuron must respond selectively to an appropriate set of inputs. To characterize selectivity, we define a set of P exemplar input vectors $\boldsymbol{x}^\mu$, with $\mu = 1, 2, ..., P$, and randomly assign them to two classes, denoted as "plus" and "minus." The neuron must respond to inputs belonging to the plus class by firing (active state) and to the minus class by remaining quiescent. This means that the neuron is acting as a perceptron (18–22, 25, 27, 29). We assume the P input activations, $\boldsymbol{x}^\mu$, are drawn identically and independently from a distribution with nonnegative means, $\bar{\boldsymbol{x}}$, and covariance matrix, C (when N is large, higher moments of the distribution of $\boldsymbol{x}$ have negligible effect). For simplicity we assume that the stimulus average activities are the same for all input neurons within a population, so that $\bar{x}_i = \bar{x}_{\text{exc(inh)}} \geq 0$, and that C is diagonal with equal variances within a population, $\sigma_i^2 = \sigma_{\text{exc(inh)}}^2$. Note that synaptic weights are in units of membrane potential over input activity levels (firing rates) and hence will be measured in units of $\left(V_{\text{th}} - V_{\text{rest}}\right)/\sigma_{\text{exc}}$.

We call weight vectors that correctly categorize the P exemplar input patterns, $\boldsymbol{x}^\mu$ for $\mu = 1, 2, ..., P$, solutions of the categorization task presented to the neuron. Before describing in detail the properties of the solutions, we outline a broad distinction between two types of possible solutions. One type is characterized by weak synapses, i.e., individual synaptic weights that are inversely proportional to the total number of synaptic inputs, $w_i \sim 1/N$ [note that weights weaker than $\mathcal{O}\left(1/N\right)$ will not enable the neuron to cross the threshold]. For this solution type, the total excitatory and inhibitory parts of the membrane potential are of the same order as the neuron's threshold. An alternative scenario is a solution in which individual synaptic weights are relatively strong, $w_i \sim 1/\sqrt{N}$. In this case, both the total excitatory and inhibitory parts of the potential are, individually, much greater than the threshold, but they make approximately equal contributions, so that excitation and inhibition tend to cancel, and the mean V_{PSP} is close to threshold. We call the first type of solution unbalanced and the second type balanced. Importantly, since both balanced and unbalanced solutions solve the categorization task with the same value of V_{th}, the two solution types are not related to each other by a global scaling of the weights but represent different patterns of $\{w_i\}$. Note that the norm of the weight vector, $|\boldsymbol{w}| = \sqrt{\sum_{i=1}^N w_i^2}$, serves to distinguish the two types of solutions. This norm is of order of $1/\sqrt{N}$ for unbalanced solutions and of order 1 in the balanced case. Weights with norms stronger than $\mathcal{O}\left(1\right)$ lead to membrane potential values that are much larger in magnitude than the neuron's threshold. For biological neurons postsynaptic potentials of such magnitude can result in very high, unreasonable firing rates (although see ref. 30). We therefore impose an upper bound of the weight norm $|\boldsymbol{w}| \leq \Gamma$, where Γ is of order 1. We now argue that the differences between unbalanced and balanced solutions have important consequences for the way the system copes with noise.

As mentioned above, neurons in the central nervous system are subject to multiple sources of noise, and their performance must be robust to its effects. We distinguish two biologically relevant types of noise: input noise resulting from the fluctuations of the stimuli and sensory processes that generate the stimulus-related input $\boldsymbol{x}$ and output noise arising from afferents unrelated to a particular task or from biophysical processes internal to the neuron, including fluctuations in the effective threshold due to spiking history and adaptation (31–33) (for theoretical modeling see ref. 34). Both sources of noise result in trial-by-trial fluctuations of the membrane potential V_{PSP} and, for a robust solution, the probability of changing the state of the output neuron relative to the noise-free condition must be low. The two sources of noise differ in their dependence on the magnitude of the synaptic weights. Because input noise is filtered through the same set of synaptic weights as the signal, its effect on the membrane potential is sensitive to the magnitude of those weights. Specifically, if the trial-to-trial variability of each input x_i^μ is characterized by SD σ_{in}, the fluctuations it generates in the membrane potential have SD $|\boldsymbol{w}|\sigma_{\text{in}}$ (Fig. 1, *Top Left* and *Top Right*). On the other hand, the effect of output noise is independent of the synaptic weights $\boldsymbol{w}$. Output noise characterized by SD σ_{out} induces membrane potential fluctuations with the same SD σ_{out} for both types of solutions (Fig. 1, *Bottom Left* and *Bottom Right*).
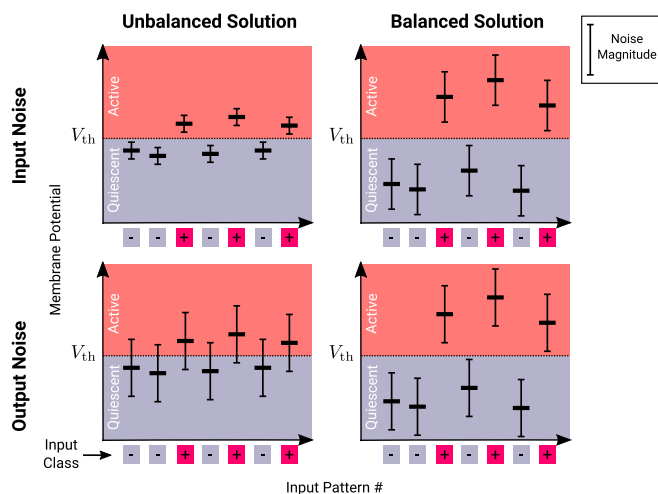

Fig. 1. Only balanced solutions can be robust to both input and output noise. Each panel depicts membrane potentials resulting from different input patterns in a classification task. Weights are unbalanced [$|\boldsymbol{w}| = \mathcal{O}(1/\sqrt{N})$, *Top Left* and *Bottom Left*] or balanced [$|\boldsymbol{w}| = \mathcal{O}(1)$, *Top Right* and *Bottom Right*]. The neuron is in an active state only if the membrane potential is greater than the threshold V_{th}. The input pattern class (plus or minus) is specified by the squares underneath the horizontal axis. Each input pattern determines a membrane potential (mean, horizontal bars) that fluctuates from one presentation to another due to input noise (*Top Left* and *Top Right*) and output noise (*Bottom Left* and *Bottom Right*). Vertical bars depict the magnitude of the noise in each case. The variability of the mean V_{PSP} across input patterns (which is the signal differentiating input pattern classes) is proportional to $|\boldsymbol{w}|$. As a result, the mean V_{PSP}s for unbalanced solutions (*Top Left* and *Bottom Left*) cluster close to the threshold [difference from threshold $\mathcal{O}(1/\sqrt{N})$]. For balanced solutions (*Top Right* and *Bottom Right*), the mean V_{PSP}s have a larger spread [potential difference $\mathcal{O}(1)$]. Input noise (fluctuations of x_i, *Top Left* and *Top Right*) produces membrane potential fluctuations with SD that is proportional to $|\boldsymbol{w}|$, which is of $\mathcal{O}(1/\sqrt{N})$ for unbalanced solutions (*Top Left*) and of $\mathcal{O}(1)$ for balanced solutions (*Top Right*). Output noise (*Bottom Left* and *Bottom Right*) produces membrane potential fluctuations that are independent of $|\boldsymbol{w}|$, so it is of the same magnitude for both solution types. Thus, while both balanced and unbalanced solutions can be robust to input noise, only balanced solutions can also be robust to substantial output noise.

We can now appreciate the basis for the difference in the noise robustness of the two types of solutions. For unbalanced solutions, the difference between the potential induced by typical plus and minus noise-free inputs (the signal) is of the order of $|\mathbf{w}| = \mathcal{O}(1/\sqrt{N})$ (Fig. 1, *Top Left* and *Bottom Left*). Although the fluctuations induced by input noise are of this same order (Fig. 1, *Top Left*), output noise yields fluctuations in the membrane potential of order 1, which is much larger than the magnitude of the weak signal (Fig. 1, *Bottom Left*). In contrast, for balanced solutions, the signal differentiating plus and minus patterns is of order $|\mathbf{w}| = \mathcal{O}(1)$, which is the same order as the fluctuations induced by both types of noise (Fig. 1, *Top Right* and *Bottom Right*). Thus, we are led to the important observation that the balanced solution provides the only hope for producing selectivity that is robust against both types of noise. However, there is no guarantee that robust, balanced solutions exist or that they can be found and maintained in a manner that can be implemented by a biological system. Key questions, therefore, are, Under what conditions does a balanced solution to the selectivity task exist? And what are, in detail, its robustness properties? Below, we derive conditions for the existence of a balanced solution, analyze its properties, and study the implications for single-neuron and network computation. We show that, subject to a small reduction of the total information stored in the network, robust and balanced solutions exist and can emerge naturally when learning occurs in the presence of output noise.

Balanced and Unbalanced Solutions. We begin by presenting the results of an analytic approach (20–22) for determining existence conditions and analyzing properties of weights that generate a specified selectivity, independent of the particular method or learning algorithm used to find the weights (*SI Replica Theory for Sign- and Norm-Constrained Perceptron*). We validate the theoretical results by using numerical methods that can determine the existence of such weights and find them if they exist (*SI Materials and Methods*).

When the number of patterns P is too large, solutions may not exist. The maximal value of P that permits solutions is proportional to the number of synapses, N, so a useful measure is the ratio $\alpha = P/N$, which we call the load. The capacity, denoted as α_c, is the maximal load that permits solutions to the task. The capacity depends on the relative number of plus and minus input patterns. For simplicity we assume throughout that the two classes are equal in size (but see *SI Capacity for Noneven Split of Plus and Minus Patterns*). A classic result for the perceptron with weights that are not sign constrained is that the capacity is $\alpha_c = 2$ (20, 35, 36). For the "constrained perceptron" considered here, we find that α_c depends also on the fraction of excitatory afferents, denoted by f_{exc}. This fraction is an important architectural feature of neuronal circuits and varies in different brain systems. For $f_{exc} = 0$, namely a purely inhibitory circuit, the capacity vanishes, because when all of the input to the neuron is inhibitory, V_{PSP} cannot reach threshold and the neuron is quiescent for all stimuli. When the circuit includes excitatory synapses, the task can be solved by appropriate shaping of the strength of the excitatory and inhibitory synapses, and this ability increases the larger the fraction of excitatory synapses is. Therefore, for $f_{exc} > 0$, α_c increases with f_{exc} up to a maximum of $\alpha_c = 1$ (half the capacity of an unconstrained perceptron) for fractions equal to or greater than a critical fraction $f_{exc} = f_{exc}^\star$. This dependence can be summarized by the capacity curve $\alpha_c (f_{exc})$ (Fig. 2A, solid line) bounding the range of loads which admit solutions for the different excitatory/inhibitory ratios.

Interestingly, $f_{exc}^\star$ depends on the statistics of the inputs (*SI Replica Theory for Sign- and Norm-Constrained Perceptron*). We denote the coefficient of variation (CV) of the excitatory and inhibitory input activities by $\mathrm{CV}_{exc} = \sigma_{exc}/\bar{x}_{exc}$ and $\mathrm{CV}_{inh} = \sigma_{inh}/\bar{x}_{inh}$, respectively. These measure the degree of
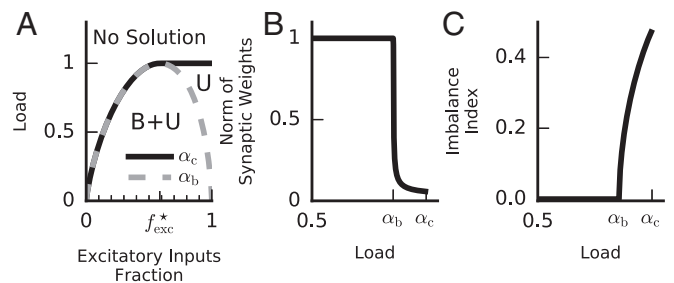

Fig. 2. Balanced and unbalanced solutions. (*A*) Perceptron solutions as a function of load and fraction of excitatory weights. Above the capacity line [$\alpha_c (f_{exc})$, solid line] no solution exists. Balanced solutions exist only below the balanced capacity line [$\alpha_b (f_{exc})$, dashed shaded line]. Between the balanced capacity and maximum capacity lines, only unbalanced solutions exist (U). On the other hand, below the balanced capacity line, unbalanced solutions coexist with balanced ones (B+U). (*B*) The norm of the synaptic weight vector of typical solutions as a function the load [in units of $(V_{th} - V_{rest})/\sigma_{exc}$]. Below α_b the norm is clipped at its upper bound Γ (in this case $\Gamma = 1$). Above α_b the norm collapses and is of order $1/\sqrt{N}$ (shown here for $N = 3,000$). (*C*) The input imbalance index (IB, Eq. 3) of typical solutions as a function of the load. Note the sharp onset of imbalance above α_b. In *B* and *C* $f_{exc} = 0.8$, yielding $\alpha_c = 1$. See *SI Materials and Methods* for other parameters used. For simulation results see Fig. S1.

stimulus tuning of the two afferent populations. In terms of these quantities, the critical excitatory fraction is

$$f_{exc}^\star = \frac{\mathrm{CV}_{exc}}{\mathrm{CV}_{exc} + \mathrm{CV}_{inh}}. \quad [2]$$

In other words, the critical ratio between the number of excitatory and inhibitory afferents [$f_{exc}^\star/(1 - f_{exc}^\star)$] equals the ratio of their degree of tuning. To understand the origin of this result, we note that to maximize the encoding capacity, the relative strength of the weights should be inversely proportional to the SD of their afferents, $\bar{w}_{exc(inh)} \propto 1/\sigma_{exc(inh)}$, implying that the mean total synaptic inputs are proportional to $f_{exc}\bar{w}_{exc}\bar{x}_{exc} + f_{inh}\bar{w}_{inh}\bar{x}_{inh} = f_{exc}/\mathrm{CV}_{exc} - f_{inh}/\mathrm{CV}_{inh}$, where $f_{inh} = 1 - f_{exc}$. For excitatory fraction $f_{exc} > f_{exc}^\star$ this mean total synaptic inputs are positive, allowing the voltage to reach the threshold and the neuron to implement the required selectivity task with optimally scaled weights. Thus, the capacity of the neuron is unaffected by changes in f_{exc} in the range $f_{exc}^\star \leq f_{exc} \leq 1$. For excitatory fraction $f_{exc} < f_{exc}^\star$ the neuron cannot remain responsive (reach threshold) with optimally scaled weights, and thus the capacity is reduced.

In cortical circuits, inhibitory neurons tend to fire at higher firing rates and are thought to be more broadly tuned than excitatory neurons (4, 37, 38), implying $f_{exc}^\star > 0.5$ (*SI Effects of E and I Input Statistics*). This is consistent with the abundance of excitatory synapses in cortex. However, input statistics that make $f_{exc}^\star < 0.5$ do not change the qualitative behavior we discuss (*SI Effects of E and I Input Statistics* and Fig. S2A).

For load levels below the capacity, many synaptic weight vectors solve the selectivity task and we now describe the properties of the different solutions. In particular, we investigate the parameter regimes where balanced or unbalanced solutions exist. We find that unbalanced solutions with weight vector norms of order $1/\sqrt{N}$ exist for all load values below α_c. As for the balanced solutions with weight vector norms of order 1, they exist below a critical value α_b which may be smaller than α_c. Specifically, for $f_{exc} \leq f_{exc}^\star$ balanced solutions exist for all load values below capacity; i.e., $\alpha_b = \alpha_c$. For $f_{exc} > f_{exc}^\star$, α_b is smaller than α_c and decreases with f_{exc} until it vanishes at $f_{exc} = 1$ (Fig. 2A, dashed shaded line). The absence of balanced solutions for $f_{exc} = 1$ is clear, as there is no inhibition to balance the excitatory inputs. Furthermore, the synaptic excitatory weights must be weak

(scaling as $1/N$) to ensure that V_{PSP} remains close to threshold (slightly above it for plus patterns and slightly below it for minus ones). For $1 \geq f_{\text{exc}} > f_{\text{exc}}^{\star}$ the predominance of excitatory afferents precludes a balanced solution if the load is high; i.e., $\alpha_{\text{b}} \leq \alpha \leq \alpha_{\text{c}}$. As argued above and shown below, balanced solutions are more robust than unbalanced solutions. Hence, we can identify $f_{\text{exc}}^{\star}$ as the optimal fraction of excitatory input, because it is the fraction of excitatory afferents for which the capacity of balanced solutions is maximal.

For loads below α_{b} both balanced and unbalanced solutions exist, raising the question, What would be the character of a weight vector that is sampled randomly from the space of all possible solutions? Our theory predicts that whenever the balanced solutions exist, the vast majority of the solutions are balanced and furthermore have a weight vector norm that is saturated at the upper bound Γ. This is a consequence of the geometry of high-dimensional spaces in which volumes are dominated by the volume elements with the largest radii (see *SI Replica Theory for Sign- and Norm-Constrained Perceptron* for details). Thus, for $f_{\text{exc}} > f_{\text{exc}}^{\star}$, the typical solution undergoes a transition from balanced to unbalanced weights as α crosses the balanced capacity line $\alpha_{\text{b}}(f_{\text{exc}})$. At this point the norm of the solution collapses from Γ to $|\boldsymbol{w}| \sim 1/\sqrt{N}$ (Fig. 2B).

As explained above, for balanced solutions we expect to find a near cancellation of the total excitatory (E) and inhibitory (I) inputs. Our theory confirms this expectation. To measure the degree of E-I cancellation for any solution, we introduce the imbalance index,

$$\text{IB} = \frac{\sum_i w_i \bar{x}_i}{\sum_{i \in \text{exc}} w_i \bar{x}_i - \sum_{i \in \text{inh}} w_i \bar{x}_i}, \qquad [3]$$

where the overbar symbol denotes an average over all of the input patterns (μ) and, as mentioned above, E weights are nonnegative ($w_i \geq 0$) and I weights are nonpositive ($w_i \leq 0$). Whereas for the unbalanced solution the IB is of order 1, for the balanced solution it is small, of order $1/\sqrt{N}$. Thus, the typical solution below α_{b} has zero imbalance (to leading order in N), but the imbalance increases sharply as α increases beyond α_{b} (Fig. 2C).

Noise Robustness of Balanced and Unbalanced Solutions. To characterize the effect of noise on the different solutions, we introduce two measures, input robustness κ_{in} and output robustness κ_{out}, which characterize the robustness of the noise-free solutions to the addition of two types of noise. To ensure robustness to output noise, the noise-free membrane potential that is the closest to the threshold must be sufficiently far from it. Thus, we define

$$\kappa_{\text{out}} = \min_{\mu} \left| \sum_{i=1}^{N} w_i x_i^{\mu} - 1 \right|, \qquad [4]$$

where the minimum is taken over all of the input patterns in the task and the threshold is 1 [because we measure the weights in units of $(V_{\text{th}} - V_{\text{rest}})/\sigma_{\text{exc}}$]. The second measure, which characterizes robustness to input noise, must take into account the fact that the fluctuations in the membrane potential induced by this form of noise scale with the size of the synaptic weights. Hence, $\kappa_{\text{in}} = \kappa_{\text{out}}/|\boldsymbol{w}|$ [κ_{in} corresponds to the notion of margin in machine learning (39)]. Efficient algorithms for finding the solution with a maximum possible value of κ_{in} have been studied extensively (39, 40). We have developed an efficient algorithm for finding solutions with maximal κ_{out} (*SI Materials and Methods*).

We now ask, What are the possible values of the input and output robustness of unbalanced and balanced solutions? Our theory predicts that the majority of both balanced and unbalanced solutions have vanishingly small values of κ_{in} and κ_{out} and are thus very sensitive to noise. However, for a given load (below

capacity) robust solutions do exist, with a spectrum of robustness values up to maximal values, $\kappa_{\text{in}}^{\text{max}} > 0$ and $\kappa_{\text{out}}^{\text{max}} > 0$. Since the magnitude of $\boldsymbol{w}$ scales both signal and noise in the inputs, $\kappa_{\text{in}}^{\text{max}}$ is not sensitive to $|\boldsymbol{w}|$ and hence is of $\mathcal{O}(1)$ for both unbalanced and balanced solutions. On the other hand, $\kappa_{\text{out}}^{\text{max}} = \kappa_{\text{in}}^{\text{max}} |\boldsymbol{w}|$ is proportional to $|\boldsymbol{w}|$. Thus, we expect $\kappa_{\text{out}}^{\text{max}}$ to be of $\mathcal{O}(1)$ when balanced solutions exist and of $\mathcal{O}(1/\sqrt{N})$ when only unbalanced solutions exist. In addition, we expect that increasing the load will reduce the value of $\kappa_{\text{in}}^{\text{max}}$ and $\kappa_{\text{out}}^{\text{max}}$ as the number of constraints that need to be satisfied by the synaptic weights increases.

In Fig. 3 we present the values of $\kappa_{\text{in}}^{\text{max}}$ and $\kappa_{\text{out}}^{\text{max}}$ vs. the load. As expected, we find that the values of both $\kappa_{\text{in}}^{\text{max}}$ and $\kappa_{\text{out}}^{\text{max}}$ reach zero as the load approaches the capacity, α_{c} (and diverges, as $N \to \infty$, for vanishingly small loads). However, $\kappa_{\text{out}}^{\text{max}}$ is only substantial (of order 1) and proportional to Γ below α_{b} where balanced solutions exist (Fig. 3 A and B). In contrast, $\kappa_{\text{in}}^{\text{max}}$ remains of order 1 up to the full capacity, α_{c} (Fig. 3C). What are the properties of "optimal" solutions that achieve the maximal robustness to either input or output noise? We find that the solutions that achieve the maximal output robustness, $\kappa_{\text{out}}^{\text{max}}$, are balanced for all $\alpha \leq \alpha_{\text{b}}$ and their norm saturates the upper bound, Γ (Fig. S3B). Interestingly, for a wide range of input parameters (*SI Replica Theory for Sign- and Norm-Constrained Perceptron, Effects if E and I Input Statistics*, and Fig. S2B), solutions that achieve the maximal input robustness, $\kappa_{\text{in}}^{\text{max}}$, are unbalanced solutions (Fig. S3C). Nevertheless, we find that below the critical balance load, α_{b}, the κ_{in} values of the balanced maximal κ_{out} solutions are of the same order as, and indeed close to, $\kappa_{\text{in}}^{\text{max}}$ (Fig. 3C, dashed shaded line). In fact, the balanced solution with maximal κ_{out} also poses the maximal value of κ_{in} that is possible for balanced solutions.

We conclude that solutions that are robust to both input and output noise exist for loads less than α_{b} which for $f_{\text{exc}} > f_{\text{exc}}^{\star}$ is smaller than α_{c}. However, as long as f_{exc} is close to $f_{\text{exc}}^{\star}$, the reduction in capacity from α_{c} to α_{b} imposed by the requirement of robustness is small.

Balanced and Unbalanced Solutions for Spiking Neurons. Neurons typically receive their input and communicate their output through action potentials. Thus, a fundamental question is, How will the introduction of spike-based input and spiking output
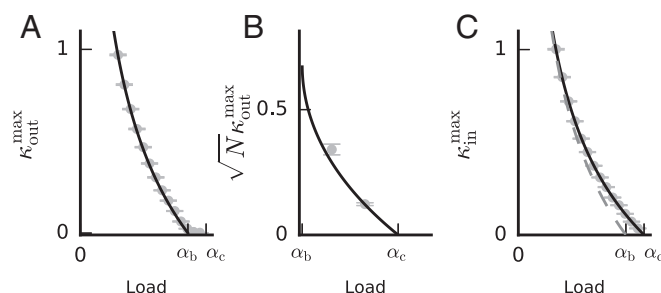

Fig. 3. Maximal values of input and output robustness. (A) Maximal value of κ_{out} vs. load [in units of $\Gamma \sigma_{\text{exc}}/(V_{\text{th}} - V_{\text{rest}})$]. No solutions exist above the maximal κ_{out} line ($\kappa_{\text{out}}^{\text{max}}$, solid line). Below $\kappa_{\text{out}}^{\text{max}}$, for output robustness that is of order 1, only balanced solutions exist. (B) Maximal value of κ_{out} for loads between α_{b} and α_{c} (in units of $\sigma_{\text{exc}}/\bar{x}_{\text{exc}}$). In this range only unbalanced solutions exist and the maximal κ_{out} values (solid line) scale as $1/\sqrt{N}$. (C) Maximal value of κ_{in} vs. load (in units of σ_{exc}). No solutions exist above the maximal κ_{in} line ($\kappa_{\text{in}}^{\text{max}}$, solid line). For the parameters used, solutions that achieve $\kappa_{\text{in}}^{\text{max}}$ are unbalanced. The maximal value of κ_{in} for balanced solutions (dashed shaded line) is not far from the $\kappa_{\text{in}}^{\text{max}}$ and is attained by solutions that maximize κ_{out} for $\alpha < \alpha_{\text{b}}$. In A–C, theory and numerical results are depicted in solid or shaded lines and shaded circles, respectively. Error bars depict SE of the mean. See *SI Materials and Methods* for parameters used. For further simulation results see Fig. S3.

affect our results? Here we show that the main properties of balanced and unbalanced synaptic efficacies, as discussed above, remain when the inputs are spike trains and the model neuron implements spiking and membrane potential reset mechanisms.

We consider a leaky integrate-and-fire (LIF) neuron that is required to perform the same binary classification task we considered using the perceptron. Each input is characterized by a vector of firing rates, $\boldsymbol{x}^{\mu}$. Each afferent generates a Poisson spike train over an interval from time $t = 0$ to $t = T$, with mean rate $r_i \propto x_i^{\mu}$. The LIF neuron integrates these input spikes (*SI Materials and Methods*) and emits an output spike whenever its membrane potential crosses a firing threshold. After each output spike, the membrane potential is reset to the resting potential, and the integration of inputs continues. We define the output state of the LIF neuron, using the total number of output spikes n_{spikes}: The neuron is quiescent if $n_{\text{spikes}} \leq n_{\text{thr}}$ and active if $n_{\text{spikes}} > n_{\text{thr}}$, where n_{thr} is chosen to maximize classification performance. We do not discuss the properties of learning in LIF neurons (41–45), but instead test the properties of the solutions (weights) obtained from the perceptron model when they are used for the LIF neuron. In particular, we compare the performance of the balanced, maximal κ_{out} solution and the unbalanced, maximal κ_{in} solution. When the synaptic weights of the LIF neuron are set according to the two perceptron solutions, the mean output of the LIF neuron correctly classifies the input patterns (according to the desired classification; Fig. S4). Consistent with the results for the perceptron, we find that with no output noise the performance of both solutions is good, even in the pres-

ence of the substantial input noise caused by Poisson fluctuations in the number of input spikes and their timings (Fig. 4 A–C). When the output noise magnitude is increased (*SI Materials and Methods*), however, the performance of the unbalanced maximal κ_{in} solution quickly deteriorates, whereas the performance of the balanced maximal κ_{out} solution remains largely unaffected (Fig. 4 D–F). Thus, the spiking model recapitulates the general results found for the perceptron.

Balanced and Unbalanced Synaptic Weights in Associative Memory Networks. Thus far, we have considered the selectivity of a single neuron, but our results also have important implications for recurrently connected neuronal networks, in particular recurrent networks implementing associative memory functions. Models of associative memory in which stable fixed points of the network dynamics represent memories, and memory retrieval corresponds to the dynamic transformation of an initial state to one of the memory-representing fixed points, have been a major focus of memory research for many years (24, 27, 28, 46–48). For the network to function as an associative memory, memory states must have large basins of attraction so that the network can perform pattern completion, recalling a memory from an initial state that is similar but not identical to it. In addition, memory retrieval must be robust to output noise. As we will show, the variables κ_{in} and κ_{out} for the synaptic weights projecting onto individual neurons in the network are closely related to the sizes of the basins of attraction of the memories and the robustness to output noise, respectively.

We consider a network that consists of Nf_{exc} E and $N\,(1 - f_{\text{exc}})$ I, recurrently connected binary neurons. The network operates in discrete time steps and at each step the state of one randomly chosen neuron, i, is updated according to

$$s_i\,(t + 1) = \Theta\left[\sum_{j \neq i} J_{ij} s_j\,(t) + \eta_{\text{out}}\,(t) - 1\right]. \qquad [5]$$

Here $\Theta(x) = 1$ for $x \geq 0$ and 0 otherwise, J_{ij} is the weight of the synapse from neuron j to neuron i, and $\eta_{\text{out}}(t)$, the output noise, is a Gaussian random variable with SD σ_{out}. P randomly chosen binary activity patterns $\{\boldsymbol{s}^{\mu}\}$, $\mu = 1, 2, ..., P$ (where each $s_i^{\mu} = \{0, 1\}$) representing the stored memories are encoded in the recurrent synaptic matrix J. This is achieved by treating each neuron, say i, as a perceptron with a weight vector $\boldsymbol{w}^i = \{J_{ij}\}_{j \neq i}$ that maps its inputs $\{s_j^{\mu}\}$ from all other neurons to its desired output s_i^{μ} for each memory state (Fig. 5 A and B and *SI Materials and Methods*). This creates an attractor network in which the memory states are fixed points of the dynamics in the noise-free condition ($\sigma_{\text{out}} = 0$) (20).

We do not attempt to perform a complete analysis of the effects of input and output noise in recurrent networks, a difficult challenge. Instead, we link observations from our single-neuron analysis to key features of a recurrent network performing a memory function. The capacity of such a memory network is defined as the maximal load for which the memory patterns can be fixed points of the noise-free dynamics, stable against single-neuron perturbations. This condition is met as long as the single-neuron synaptic weights possess substantial κ_{in} (i.e., $\kappa_{\text{in}} \sim \mathcal{O}\,(1)$) for all neurons. Thus, the single-neuron capacities will determine the overall network capacity. As we showed before, the capacity of a single-neuron perceptron depends on the statistics of its desired output (which in our case is the sparsity of activity across memory states). Since this statistic may be different in E and I populations, the single-neuron capacity of the two populations may vary, and hence the global capacity of the recurrent network is the minimum of the single-neuron capacities of the two neuron types. As long as P is smaller than this critical capacity, a recurrent weight matrix exists for which all P memory states
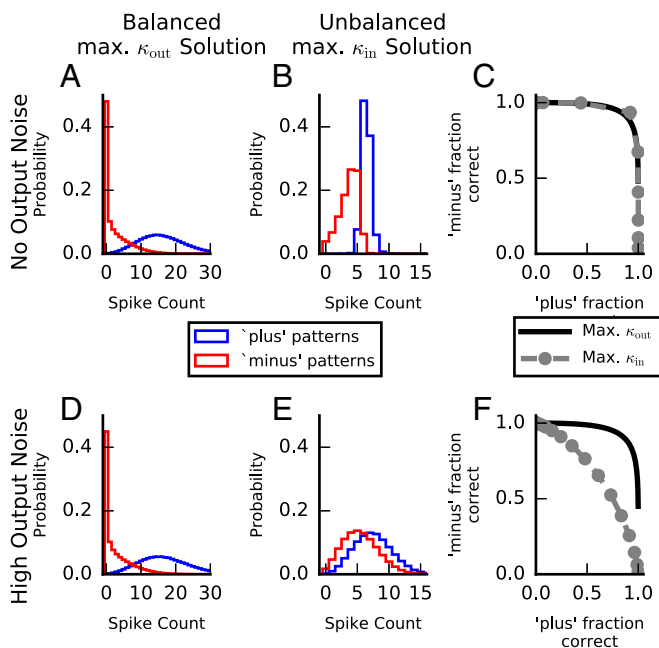

Fig. 4. Selectivity in a spiking model. *A* and *B* (*D* and *E*) depict the output of an LIF neuron with no (high) output noise for the balanced maximal κ_{out} solution (*A* and *D*) and the unbalanced maximal κ_{in} solution (*B* and *E*). *C* and *F* depict the receiver operating characteristic (ROC) curves for the two solutions under the no output noise (*C*) and high output noise (*F*) conditions obtained as the decision threshold (n_{thr}) is modified from 0 to ∞. Consistent with the results of the perceptron, the performances of the two solutions with no output noise are very similar with a slight advantage for the maximal κ_{in} solution. With higher levels of output noise, the performance of the unbalanced maximal κ_{in} solution quickly deteriorates, whereas the performance of the balanced maximal κ_{out} solution is only slightly affected. $|\boldsymbol{w}|$ of the balanced solution was chosen to equalize the mean output spike count across all patterns in both solutions (mean $n_{\text{spike}} \sim 4$). See *SI Materials and Methods* for parameters used.
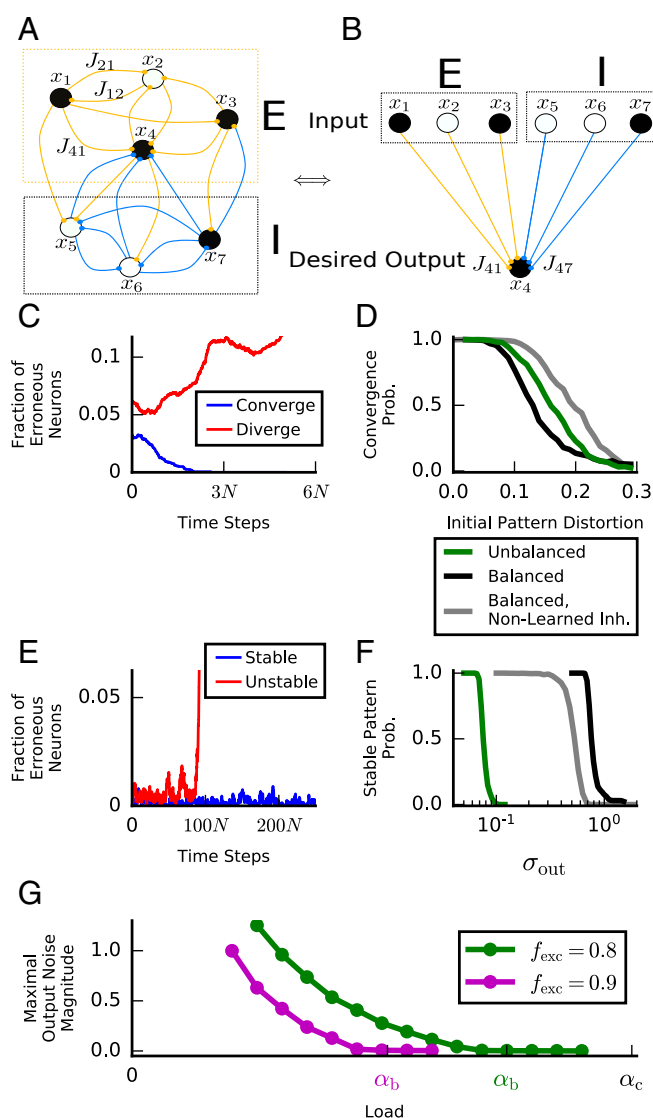
Fig. 5. Recurrent associative memory network constructed using single-neuron feedforward learning. (A) A fully connected recurrent network of E and I neurons in a particular memory state. Active (quiescent) neurons are shown in black (white). E and I synaptic connections (J_{ij}) are shown in yellow and blue, respectively (not all connections are depicted). Lines symbolize axons, and synapses are shown as small circles. (B) To find an appropriate J_{ij}, the postsynaptic weights of each neuron are set using the memory-state activities of the other neurons as input and its own memory state as the desired output. In this example, neuron 4 will implement its desire memory state through modification of the weights J_{4j} for $j = $ 1, 2, 3, 5, 6, 7. *C* and *E* show the fraction of erroneous (different from a given memory pattern) neurons in the network as a function of time. (C) Network dynamics with $\sigma_{out} = 0$. An initial state of the network can either converge to the memory state (blue) or diverge to other network states (red). (D) Probability of converging to a memory state vs. initial pattern distortion (*SI Materials and Methods*) for a network with unbalanced maximal κ_{in} weights (green), a network with balanced maximal κ_{out} weights (black), and a network with balanced maximal κ_{out} weights with unlearned inhibition (gray, main text). (E) Network dynamics with $\sigma_{out} > 0$. The network is initialized at the memory state. The dynamics can be stable (blue; the network remains close to the memory state), or unstable (red; the network diverges to another state). (F) Probability of stable dynamics for at least 500*N* time steps for networks initialized at the memory state in the presence of output noise vs. σ_{out}. Colors are the same as in *D*. (G) Maximal output noise magnitude vs. load for networks with balanced synaptic weights matrix maximizing κ_{out}. Similar to κ_{out}, the maximal output noise magnitude is of order 1 only below α_b. Above it, even though solutions exist they are extremely sensitive to output noise. Results are shown for

are stable fixed points of the noiseless dynamics. However, such solutions are not unique, and the choice of a particular matrix can endow the network with different robustness properties. As stated above, to properly function as an associative memory the fixed points must have large basins of attraction. Corruption of the initial state away from the parent memory pattern introduces variability into the inputs of each neuron for subsequent dynamic iterations and hence is equivalent to injecting input noise in the single-neuron feedforward case. The network propagates this initial input noise in a nontrivial way; however, its magnitude always remains proportional to the magnitude of the norm of the neurons' synaptic weights. We therefore expect that a large basin of attraction is achieved when the matrix J yields a large input noise robustness for each neuron in the (noise-free) fixed points (49, 50). When output noise is introduced to the network dynamics ($\sigma_{out} > 0$), the network may propagate it as input noise to other neurons in subsequent time steps. However, initially its magnitude is proportional only to σ_{out} and is unaffected by the scale of the synaptic weights. Thus, we expect that the requirement that the memory states and retrieval will be robust against output noise is satisfied when J yields a large output noise robustness for each neuron in the (noise-free) fixed points. We therefore consider two types of recurrent connections: one in which each row of J is a weight vector that maximizes κ_{in} and hence, in the chosen parameter regime, is necessarily unbalanced and a second one in which the rows of the connection matrix correspond to balanced solutions that maximize κ_{out}.

We estimate the basins of attraction of the memory patterns numerically by initializing the network in states that are corrupted versions of the memory states (*SI Materials and Methods*) and observing whether the network, with $\sigma_{out} = 0$, converges to the parent memory state (Fig. 5*C*, blue) or diverges away from it (Fig. 5*C*, red). We define the size of the basin of attraction as the maximum distortion in the initial state that ensures convergence to the parent memory with high probability.

Comparing the basins of attraction of the two types of networks, we find that the mean basin of attraction of the unbalanced network is moderately larger than that of the balanced one (Fig. 5*D*), consistent with the slightly lower value of κ_{in} in the balanced case (Fig. 5*D*). On the other hand, the behavior of the two networks is strikingly different in the presence of output noise. To illustrate this, we start each network at a memory state and determine whether it is stable (remains in the vicinity of this state for an extended period), despite the noise in the dynamics (Fig. 5*E*). We estimate the output noise tolerance of the network by measuring the maximal value of σ_{out} for which the memory states are stable (Fig. 5*F*). We find that memory states in the balanced solution with maximal κ_{out} are stable for noise levels that (for the network sizes used in the simulation) are an order of magnitude larger than for the unbalanced network with maximal κ_{in} (Fig. 5*F*).

Finally, we ask how the noise robustness of the memory states in the balanced network depends on the number of memories. As shown in Fig. 5*F*, for a fixed level of load below capacity, memory patterns are stable ($P_{stable} > 0.5$) as long as levels of noise remain below a threshold value, which we denote as $\sigma_{out}^{max}(\alpha)$. When σ_{out} increases beyond $\sigma_{out}^{max}(\alpha)$, stability of the memory states rapidly deteriorates. The critical noise function $\sigma_{out}^{max}(\alpha)$ decreases smoothly from a large value at small α to zero at a level of load, α_b. This load coincides with the maximal load for which both E and I neurons have balanced solutions (Fig. 5*G*). For loads $\alpha_b < \alpha < \alpha_c$, all solutions are unbalanced, and hence

the magnitude of the stochastic dynamical component can be at most of order $1/\sqrt{N}$.

The Role of Inhibition in Associative Memory Networks. In our associative memory network model, we assumed that both E and I neurons code desired memory states and that all network connections are modified by learning. Most previous models of associative memory that separate excitation and inhibition assume that memory patterns are restricted to the E population, whereas inhibition provides stabilizing inputs (14, 48, 51–54). To address the emergence of balanced solution in scenarios where the I neurons do not represent long-term memories, we studied an architecture where I to E, I to I, and E to I connections are random sparse matrices with large amplitudes, resulting in I activity patterns driven by the E memory states. In such conditions, the I subnetwork exhibits irregular asynchronous activity with an overall mean activity that is proportional to the mean activity of the driving E population (7, 55, 56). Although the mean I feedback provided to the E neurons can balance the mean excitation, the variability in this feedback injects substantial noise onto the E neurons, which degrades system performance (*SI Recurrent Networks with Nonlearned Inhibition*). This variability stems from the differences in I activity patterns generated by the different E memory states (albeit with the same mean). Additional noise is caused by the temporal irregular activity of the chaotic I dynamics. Next we ask whether the system's performance can be improved through plasticity in the I to E connections for which some experimental evidence exists (23, 57–60). Indeed, we find an appropriate plasticity rule for this pathway (*SI Recurrent Networks with Nonlearned Inhibition*) that suppresses the spatiotemporal fluctuations in the I feedback, yielding a balanced state that behaves similarly to the fully learned networks described above (Fig. 5 *D* and *F*, gray lines). Interestingly, in this case the basins of attraction of the balanced network are comparable to or even larger than the basins of the unbalanced fully learned network (compare gray to green curves in Fig. 5*D*). Despite the fact that no explicit memory patterns are assigned to the the I populations, the I activity plays a computational role that goes beyond providing global I feedback; when the weights of the I to E connections are shuffled, the network's performance significantly degrades (Fig. S5).

Learning Robust Solutions. Thus far, we have presented analytical and numerical investigations of solutions that support selectivity or associative memory and provide substantial robustness to noise. However, we did not address the way in which these robust solutions could be learned by a biological system. In fact, as stated above, the majority of solutions for these tasks have vanishingly small output and input robustness and the above maximum robustness solutions are found numerically by special learning algorithms. Therefore, an important question is whether noise robust weights can emerge naturally from synaptic learning rules that are appropriate for neuronal circuits.

The actual algorithms used for learning in the neural circuits are generally unknown, especially within a supervised learning scenario. Experiments suggest that learning rules may depend on brain area and both pre- and postsynaptic neuron types (for example, refs. 57–59, 61; for reviews see refs. 60, 62–64). From a theoretical perspective, the properties of the solutions found through learning, and in particular their noise robustness, depend on both the type and parameters of the algorithm and the properties of the space of possible solutions. However, our theory suggests that a general, simple way to ensure that learning arrives at a robust solution is to introduce noise during learning. Indeed, this is a common practice in machine learning for increasing generalization abilities [a specific form of data augmentation (65, 66)]. The rationale is that learning algorithms

that achieve low error in the presence of noise necessarily lead to solutions that are robust against noise levels at least as large as those present during learning. In the case we are considering, learning in the presence of substantial input noise should lead to solutions that have substantial κ_{in} and introducing output noise during learning should lead to solutions with substantial κ_{out}. We note that κ_{in} may be large even if κ_{out} remains small (for example, in unbalanced solutions with maximal κ_{in}) but not vice versa [because κ_{out} of order 1 implies $|\mathbf{w}|$ (and as a result κ_{in}) of order 1 as well]. Therefore, learning in the presence of significant output noise should lead to solutions that are robust to both input and output noise, whereas learning in the presence of input noise alone may lead to unbalanced solutions that are sensitive to output noise, depending on the details of the learning algorithm. We therefore predict that performing successful learning in the presence of output noise is a sufficient condition for the emergence of excitation–inhibition balance.

To demonstrate that robust balanced solutions emerge in the presence of output noise, we consider a variant of the perceptron learning algorithm (18) in which we have forced the sign constraints on the weights (29) and, in addition, added a weight decay term implementing a soft constraint on the magnitude of the weights (*SI Materials and Methods*). This supervised learning rule possesses several important properties that are required for biological plausibility: It is online, and weights are modified incrementally after each pattern presentation; it is history independent so that each weight update depends only on the current pattern and error signal; and finally, it is simple and local, and
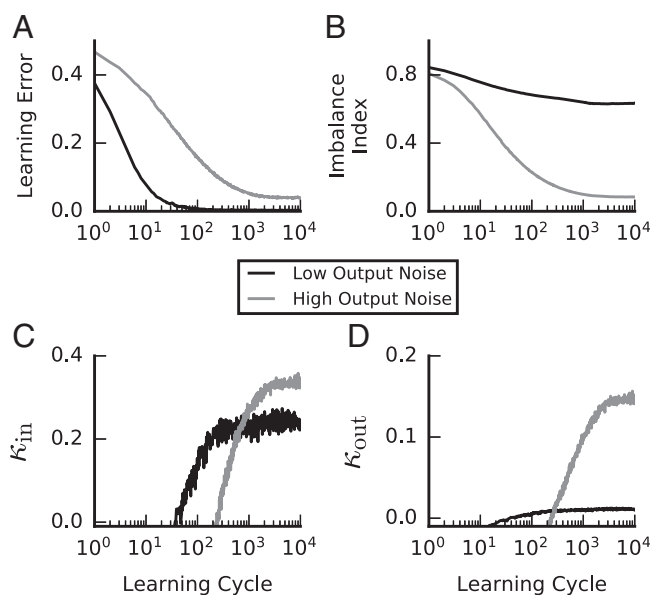

Fig. 6. Emergence of E-I balance from learning in the presence of output noise. All panels show the outcome of perceptron learning for a noisy neuron (*SI Materials and Methods*) under low ($\sigma_{out} = 0.01$, solid lines) and high ($\sigma_{out} = 0.1$, shaded lines) output noise conditions. Except for σ_{out}, all model and learning parameters are identical for the two conditions (including $\sigma_{in} = 0.1$). (*A*) Mean training error vs. learning cycle. On each cycle, all of the input patterns to be learned are presented once. The error decays and plateaus at its minimal value under both low and high output noise conditions. (*B*) Mean IB (Eq. **3**) vs. learning cycle. IB remains of order 1 under low output noise conditions and drops close to zero under high output noise conditions. (*C*) Mean input robustness (κ_{in}) vs. learning cycle. Input robustness is high under both output noise conditions. (*D*) Mean output robustness (κ_{out}) vs. learning cycle. Output robustness is substantial only under the high output noise learning condition. These results demonstrate that robust balanced solutions naturally emerge under learning in the presence of high output noise. See *SI Materials and Methods* for other parameters used.

weight updates are a function of the error signal and quantities that are available locally at the synapse (presynaptic activity and synaptic efficacy). When this learning rule is applied to train a selectivity task in the presence of substantial output noise, the resulting solution has a balanced weight vector with substantial κ_{out} and κ_{in} (Fig. 6, shaded lines). In contrast, if learning occurs with weak output noise, the algorithm's tendency to reduce the magnitude of the weights causes the resulting solution to be unbalanced with small κ_{out}, while its κ_{in} may be large if substantial input noise is present during learning (Fig. 6, solid lines). When this learning rule is applied in the load regime where only unbalanced solutions exist ($\alpha_{\mathrm{b}} < \alpha < \alpha_{\mathrm{c}}$), learning fails to achieve reasonable performance when applied in the presence of large output noise. When noise is scaled down to the value allowed by $\kappa_{\mathrm{out}}^{\max} \propto 1/\sqrt{N}$, learning yields unbalanced solutions with robustness values of the order of the maximum allowed in this region (Fig. S6).

Discussion

The results we have presented come from imposing a set of fundamental biological constraints: fixed-sign synaptic weights, non-negative afferent activities, a positive firing threshold (relative to the resting potential), and both input and output forms of noise. Amit et al. (23) studied the maximal margin solution for the sign-constrained perceptron and showed that it has half the capacity of the unconstrained perceptron. However, this previous work considered afferent activities that were centered around zero and a neuron with zero firing threshold, features that preclude the presence of the behavior exhibited by the more biologically constrained model studied here. Chapeton et al. (27) studied perceptron learning with sign-constrained weights and a preassigned level of robustness, but considered only solutions in the unbalanced regime which, as we have shown, are extremely sensitive to output noise.

Learning in neural circuits involves a trade-off between exhausting the system's capacity for implementing complex input–output functions on the one hand and ensuring good generalization properties on the other. A well-known approach in machine learning has been to search for solutions that fit the training examples while maximizing the distance of samples from the decision surface, a strategy known as maximizing the margin (21, 23, 39). The margin being maximized in this case corresponds, in our framework, to κ_{in}. Work in computational neuroscience has implicitly optimized a robustness parameter equivalent to our κ_{out} (25, 27). To our knowledge, the two approaches have not been distinguished before or shown to result in solutions with dramatically different noise sensitivities. In particular, over a wide parameter range, we have shown that maximizing κ_{out} leads to a balanced solution with minimal sensitivity to output noise and robustness to input noise that is almost as good as that of the maximal margin solution, with only a modest trade-off in capacity. On the other hand, maximizing the margin (κ_{in}) often leads to unbalanced solutions with extreme sensitivity to output noise.

The perceptron has long been considered a model of cerebellar learning and computation (67, 68). More recently, Brunel et al. (25) investigated the capacity and robustness of a perceptron model of a cerebellar Purkinje cell, taking all weights to be E. In view of the analysis presented here, balanced solutions are not possible in this case ($f_{\mathrm{exc}} = 1$), and solutions that maximize either input-noise or output-noise robustness both have $\kappa_{\mathrm{out}} \propto 1/\sqrt{N}$. These two types of solutions differ in their weight distributions, with experimentally testable consequences for the predicted circuit structure [SI $\kappa_{\mathrm{out}}^{\max}$ and $\kappa_{\mathrm{in}}^{\max}$ Solutions in Purely E Networks and Fig. S2C; Brunel et al. (25) considered only solutions that maximize κ_{out}]. Output robustness of the unbalanced solutions can be increased by making the input activity patterns sparse.

Denoting by s the mean fraction of active neurons in the input, maximum output robustness scales as $\kappa_{\mathrm{out}} \sim 1/\sqrt{Ns}$ (Fig. 3B and SI Replica Theory for Sign- and Norm-Constrained Perceptron). Thus, the high sparsity in input activation (granule cell activity) of the cerebellum relative to the modest sparsity in the neocortex is consistent with the former being dominated by E modifiable synapses.

Interestingly, our results suggest an optimal ratio of E to I synapses. Capacity in the balanced regime is optimal when $f_{\mathrm{exc}} = f_{\mathrm{exc}}^{\star}$, with $f_{\mathrm{exc}}^{\star}$ determined by the CVs (with respect to stimulus) of the E and I inputs (Eq. 2). Thus, optimality predicts a simple relation between the fraction of E and I inputs and their degree of tuning. Estimating the CVs from existing data is difficult, but it would be interesting to check whether input statistics and connectivity ratios in different brain areas are consistent with this prediction. The commonly observed value in cortex, $f_{\mathrm{exc}} \simeq 0.8$, would be optimal for input statistics with $\mathrm{CV}_{\mathrm{exc}}/\mathrm{CV}_{\mathrm{inh}} \simeq 4$. In general, we expect that $\mathrm{CV}_{\mathrm{exc}}/\mathrm{CV}_{\mathrm{inh}} > 1$, which implies that $f_{\mathrm{exc}}^{\star} > 1/2$.

For most of our work, we assumed that I neurons learn to represent specific sensory and long-term memory information, the same as the E ones, and that all synaptic pathways are learned using similar learning rules. While plasticity in both E and I pathways has been observed (57–59, 61, 63, 64, 69), accumulating experimental evidence indicates a high degree of cell-type and synaptic-type specificity of the plasticity rules. In addition, synaptic plasticity is under tight control of neuromodulatory systems. At present, it is unclear how to interpret our learning rules in terms of concrete experimentally observed synaptic plasticity. Other functional models of neural learning assume learning only within the E population with inhibition acting as a global stabilizing force. In the case of sensory processing, our approach is consistent with the observation of a similar stimulus tuning of E and I postsynaptic currents in many cortical sensory areas. The role of I neurons in memory representations is less known (but see ref. 70). Importantly, we have shown that our main results are valid also in the case in which I neurons do not explicitly participate in the coding of the memories. Interestingly, our work suggests that even if I neurons are only passive observers during learning processes, learning of I synapses onto E cells can amplify the memory stability of the system against fluctuations in the I feedback. Given the diversity of I cell types it is likely that in the real circuits inhibition plays multiple roles, including both conveying information and providing stability.

Several previous models of associative memory have incorporated biological constraints on the sign of the synapses, Dale's law, assuming variants of Hebbian plasticity in the E to E synapses (14, 48, 51–54). The capacity of the these Hebbian models is relatively poor, and their basins of attraction are small, except at extremely sparse activity levels. In contrast, our model applies a more powerful learning rule that, while keeping the sign constraints on the synapses, exhibits significantly superior performance: with high capacity even for moderate sparsity levels, large basins of attraction, and high robustness to output noise.

From a dynamical systems perspective, the associative memory networks we construct exhibit unusual properties. In most associative memory network models large basins of attractions endow the memory state with robustness against stochasticity in the dynamics (i.e., output noise). Here, we found that, for the same set of fixed-point memories, the synaptic weights with the largest possible basins (the unbalanced solutions with maximal κ_{in}) are very sensitive to even mild levels of stochasticity, whereas the balanced synaptic weights with somewhat reduced basins have substantially increased output noise robustness.

At the network level, as at the single-neuron level, imposing basic features of neural circuitry—positive inputs, bounded

synapses of fixed sign, a positive firing threshold, and sources of noise—forces neural circuits into the balanced regime. A recent class of models showing computational benefits of balanced inputs uses extremely strong synapses, which are outside the range we have discussed (16). These models are stabilized by instantaneous transmission of signals between neurons which are not required in the range of synaptic strength we consider.

Previous models of balanced networks have highlighted the ability of networks with strong E and I recurrent synapses to settle into a state in which the total input is dynamically balanced without special tuning of the synaptic strengths. Such a state is characterized by a high degree of intrinsically generated spatiotemporal variability (7). Mean population activities respond fast and in a linear fashion to external inputs. Typically, these networks lack the population-level nonlinearity required to generate multiple attractors. In contrast, we have explored the capacity of the balanced network to support multiple stable fixed points by tuning the synaptic strengths through appropriate learning. We note that fully understanding and characterizing the dynamic properties of these networks and their relation to previously studied models remains an important challenge. Despite the dynamic and functional differences in the two classes of networks, the balancing of excitation and inhibition plays a similar role in both. In the first scenario, synaptic balance amplifies small changes in the spatial or temporal properties of the external drive. Similarly, in the present scenario, balanced synaptic architecture leads to enhanced robustness by amplifying the small variations in the synaptic inputs induced by changes in the stimulus or memory identity. It would be very interesting to combine fast dynamics with robust associative memory capabilities.

In conclusion, we have uncovered a fundamental principle of neuronal learning under basic biological constraints. Our work reveals that excitation–inhibition balance may have a critical computational role in producing robust neuronal functionality that is insensitive to output noise. We showed that this balance is important at the single-neuron level for both spiking and nonspiking neurons and at the level of recurrently connected neural networks. Further, the theory suggests that excitation–inhibition balance may be a collective, self-maintaining, emergent phenomenon of synaptic plasticity. Any successful neuronal learning process in the presence of substantial output noise will lead to strong balanced synaptic efficacies with noise robustness features. The fundamental nature of this result suggests that it should apply across a variety of neuronal circuits that learn in the presence of noise.

Materials and Methods

Detailed methods and simulation parameters are given in *SI Materials and Methods*.

Software. To acknowledge their contribution to scientific work we cite the open source projects that directly and most crucially contributed to the current work: The Python stack of scientific computing [CPython, Numpy, Scipy, Matplotlib (71), Jupyter/Ipython (72), and others], CVXOPT (73) (convex conic optimization), and IPyparallel (parallelization).

Code Availability. Python code for simulations and numerical solution of saddle-point equations is available upon request.

1. Anderson JS, Carandini M, Ferster D (2000) Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *J Neurophysiol* 84:909–926.
2. Wehr M, Zador AM (2003) Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* 426:442–446.
3. Okun M, Lampl I (2008) Instantaneous correlation of excitation and inhibition during ongoing and sensory-evoked activities. *Nat Neurosci* 11:535–537.
4. Poo C, Isaacson JS (2009) Odor representations in olfactory cortex: "Sparse" coding, global inhibition, and oscillations. *Neuron* 62:850–861.
5. Atallah BV, Scanziani M (2009) Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition. *Neuron* 62:566–577.
6. Isaacson J, Scanziani M (2011) How inhibition shapes cortical activity. *Neuron* 72:231–243.
7. van Vreeswijk C, Sompolinsky H (1996) Chaos in neuronal networks with balanced excitatory and inhibitory activity. *Science* 274:1724–1726.
8. Vreeswijk Cv, Sompolinsky H (1998) Chaotic balanced state in a model of cortical circuits. *Neural Comput* 10:1321–1371.
9. Froemke RC, Merzenich MM, Schreiner CE (2007) A synaptic memory trace for cortical receptive field plasticity. *Nature* 450:425–429.
10. Dorrn AL, Yuan K, Barker AJ, Schreiner CE, Froemke RC (2010) Developmental sensory experience balances cortical excitation and inhibition. *Nature* 465:932–936.
11. Sun YJ, et al. (2010) Fine-tuning of pre-balanced excitation and inhibition during auditory cortical development. *Nature* 465:927–931.
12. Li Yt, Ma Wp, Pan Cj, Zhang LI, Tao HW (2012) Broadening of cortical inhibition mediates developmental sharpening of orientation selectivity. *J Neurosci* 32:3981–3991.
13. Tsodyks MV, Sejnowski T (1995) Rapid state switching in balanced cortical network models. *Netw Comput Neural Syst* 6:111–124.
14. van Vreeswijk C, Sompolinsky H (2005) Course 9-irregular activity in large networks of neurons in Les Houches. *Methods and Models in Neurophysics*, eds Carson C, Boris G, David H, Claude M, Jean D (Elsevier, Amsterdam), Vol 80, pp 341–406.
15. Lim S, Goldman MS (2013) Balanced cortical microcircuitry for maintaining information in working memory. *Nat Neurosci* 16:1306–1314.
16. Boerlin M, Machens CK, Denève S (2013) Predictive coding of dynamical variables in balanced spiking networks. *PLoS Comput Biol* 9:e1003258.
17. Lajoie G, Lin KK, Thivierge JP, Shea-Brown E (2016) Encoding in balanced networks: Revisiting spike patterns and chaos in stimulus-driven systems. *PLoS Comput Biol* 12:e1005258.
18. Rosenblatt F (1962) *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Washington, DC).
19. Minsky ML, Papert SA (1988) *Perceptrons: Expanded Edition* (MIT Press Cambridge, MA).
20. Gardner E (1987) Maximum storage capacity in neural networks. *Europhys Lett* 4:481–485.
21. Gardner E (1988) The space of interactions in neural network models. *J Phys A Math Gen* 21:257–270.
22. Gardner E, Derrida B (1988) Optimal storage properties of neural network models. *J Phys A Math Gen* 21:271–284.
23. Amit DJ, Campbell C, Wong KYM (1989) The interaction space of neural networks with sign-constrained synapses. *J Phys A Math Gen* 22:4687–4693.
24. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558.
25. 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.
26. Clopath C, Nadal JP, Brunel N (2012) Storage of correlated patterns in standard and bistable Purkinje cell models. *PLoS Comput Biol* 8:e1002448.
27. 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.
28. Brunel N (2016) Is cortical connectivity optimized for storing information? *Nat Neurosci* 19:749–755.
29. Amit DJ, Wong KYM, Campbell C (1989) Perceptron learning with sign-constrained weights. *J Phys A Math Gen* 22:2039–2045.
30. Denève S, Machens CK (2016) Efficient codes and balanced networks. *Nat Neurosci* 19:375–382.
31. Brown DA, Adams PR (1980) Muscarinic suppression of a novel voltage-sensitive K+ current in a vertebrate neurone. *Nature* 283:673–676.
32. Madison DV, Nicoll RA (1984) Control of the repetitive discharge of rat CA 1 pyramidal neurones in vitro. *J Physiol* 354:319–331.
33. Fleidervish IA, Friedman A, Gutnick MJ (1996) Slow inactivation of Na+ current and slow cumulative spike adaptation in mouse and guinea-pig neocortical neurones in slices. *J Physiol* 493:83–97.
34. Benda J, Herz AVM (2003) A universal model for spike-frequency adaptation. *Neural Comput* 15:2523–2564.
35. Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans Electron Comput* EC-14:326–334.
36. Venkatesh SS (1986) Epsilon capacity of neural networks. *AIP Conference Proceedings*, ed Denker JS (AIP Publishing, Melville, NY), Vol 151, pp 440–445.
37. Liu Bh, et al. (2009) Visual receptive field structure of cortical inhibitory neurons revealed by two-photon imaging guided recording. *J Neurosci* 29:10520–10532.

38. Kerlin AM, Andermann ML, Berezovskii VK, Reid RC (2010) Broadly tuned response properties of diverse inhibitory neuron subtypes in mouse visual cortex. *Neuron* 67:858–871.

39. Vapnik V (2000) *The Nature of Statistical Learning Theory* (Springer, New York).

40. Bottou L, Lin CJ (2007) Support vector machine solvers. *Large Scale Kernel Machines*, eds Bottou L, Chapelle O, DeCoste D, Weston J (MIT Press, Cambridge, MA), pp 301–320.

41. Gütig R, Sompolinsky H (2006) The tempotron: A neuron that learns spike timing-based decisions. *Nat Neurosci* 9:420–428.

42. Memmesheimer RM, Rubin R, Ölveczky B, Sompolinsky H (2014) Learning precisely timed spikes. *Neuron* 82:925–938.

43. Gütig R (2016) Spiking neurons can discover predictive features by aggregate-label learning. *Science* 351:aab4113.

44. Rubin R, Gütig R, Sompolinsky H (2013) Neural coding and decoding with spike times. *Spike Timing: Mechanisms and Function*, eds DiLorenzo PM, Victor JD (CRC Press, Boca Raton, FL), pp 35–64.

45. Gütig R (2014) To spike, or when to spike? *Curr Opin Neurobiol* 25:134–139.

46. Amit DJ, Gutfreund H, Sompolinsky H (1985) Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys Rev Lett* 55:1530–1533.

47. Tsodyks MV, Feigel'man MV (1988) The enhanced storage capacity in neural networks with low activity level. *Europhys Lett* 6:101–105.

48. Roudi Y, Latham PE (2007) A balanced memory network. *PLOS Comput Biol* 3:e141.

49. Krauth W, Nadal JP, Mezard M (1988) Basins of attraction in a perceptron-like neural network. *Complex Syst* 2:387–408.

50. Krauth W, Nadal JP, Mezard M (1988) The roles of stability and symmetry in the dynamics of neural networks. *J Phys A Math Gen* 21:2995–3011.

51. Amit DJ, Treves A (1989) Associative memory neural network with low temporal spiking rates. *Proc Natl Acad Sci USA* 86:7871–7875.

52. Golomb D, Rubin N, Sompolinsky H (1990) Willshaw model: Associative memory with sparse coding and low firing rates. *Phys Rev A* 41:1843–1854.

53. Hasselmo ME (1993) Acetylcholine and learning in a cortical associative memory. *Neural Comput* 5:32–44.

54. Barkai E, Bergman RE, Horwitz G, Hasselmo ME (1994) Modulation of associative memory function in a biophysical simulation of rat piriform cortex. *J Neurophysiol* 72:659–677.

55. Kadmon J, Sompolinsky H (2015) Transition to chaos in random neuronal networks. *Phys Rev X* 5:041030.

56. Harish O, Hansel D (2015) Asynchronous rate chaos in spiking neuronal circuits. *PLoS Comput Biol* 11:e1004266.

57. Nugent FS, Kauer JA (2008) LTP of GABAergic synapses in the ventral tegmental area and beyond. *J Physiol* 586:1487–1493.

58. Chevaleyre V, Castillo PE (2003) Heterosynaptic LTD of hippocampal GABAergic synapses: A novel role of endocannabinoids in regulating excitability. *Neuron* 38:461–472.

59. D'amour J, Froemke R (2015) Inhibitory and excitatory spike-timing-dependent plasticity in the auditory cortex. *Neuron* 86:514–528.

60. McBain CJ, Kauer JA (2009) Presynaptic plasticity: Targeted control of inhibitory networks. *Curr Opin Neurobiol* 19:254–262.

61. Lu Jt, Li Cy, Zhao JP, Poo Mm, Zhang Xh (2007) Spike-timing-dependent plasticity of neocortical excitatory synapses on inhibitory interneurons depends on target cell type. *J Neurosci* 27:9711–9720.

62. Kullmann DM, Lamsa KP (2007) Long-term synaptic plasticity in hippocampal interneurons. *Nat Rev Neurosci* 8:687–699.

63. Lamsa KP, Kullmann DM, Woodin MA (2010) Spike-timing dependent plasticity in inhibitory circuits. *Front Synaptic Neurosci* 2:8.

64. Larsen RS, Sjöström PJ (2015) Synapse-type-specific plasticity in local circuits. *Curr Opin Neurobiol* 35:127–135.

65. 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.

66. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.

67. Marr D (1969) A theory of cerebellar cortex. *J Physiol* 202:437–470.

68. Albus JS (1971) A theory of cerebellar function. *Math Biosci* 10:25–61.

69. Hennequin G, Everton AJ, Vogels TP (2017) Inhibitory plasticity: Balance, control and codependence. *Annu Rev Neurosci* 40:557–579.

70. Wilent WB, Nitz DA (2007) Discrete place fields of hippocampal formation interneurons. *J Neurophysiol* 97:4152–4161.

71. Hunter JD (2007) Matplotlib: A 2d graphics environment. *Comput Sci Eng* 9:90–95.

72. Pérez F, Granger BE (2007) IPython: A system for interactive scientific computing. *Comput Sci Eng* 9:21–29.

73. Andersen M, Dahl J, Liu Z, Vandenberghe L (2011) Interior-point methods for large-scale cone programming. *Optimization for Machine Learning*, eds Sra S, Nowozin S, Wright SJ (MIT Press, Cambridge, MA), pp 55–83.

74. Litwin-Kumar A, Harris KD, Axel R, Sompolinsky H, Abbott LF (2017) Optimal degrees of synaptic connectivity. *Neuron* 93:1153–1164.e7.

75. Engel A, Broeck C (2001) *Statistical Mechanics of Learning* (Cambridge Univ Press, Cambridge, UK).