Impact of Dendritic Nonlinearities on the Computational Capabilities of Neurons

Clarissa Lauditi,[1] Enrico M. Malatesta,[2] Fabrizio Pittorino,[3,*] Carlo Baldassi,[2] Nicolas Brunel,[2,4] and Riccardo Zecchina[2]

[1]*Department of Applied Math, John A. Paulson School of Engineering and Applied Sciences, Harvard University, 02138 Cambridge, Massachusetts, USA*

[2]*Department of Computing Sciences and Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, 20136 Milano, Italy*

[3]*Department of Electronics, Information and Bioengineering, Politecnico di Milano, 20133 Milano, Italy*

[4]*Departments of Neurobiology and Physics, Duke University, Durham, North Carolina, USA*

(Received 9 January 2025; accepted 16 June 2025; published 8 July 2025)

How neurons integrate the myriad synaptic inputs scattered across their dendrites is a fundamental question in neuroscience. Multiple neurophysiological experiments have shown that dendritic nonlinearities can have a strong influence on synaptic input integration. These nonlinearities have motivated mathematical descriptions of single neurons as a two-layer computational units, which have been shown to increase substantially the computational abilities of neurons, compared to linear dendritic integration. However, current analytical studies are restricted to neurons with unconstrained synaptic weights and unplausible dendritic nonlinearities. Here we introduce a two-layer model with sign-constrained synaptic weights and a biologically plausible form of dendritic nonlinearity and investigate its properties using both statistical physics methods and numerical simulations. We find that the dendritic nonlinearity enhances both the number of possible learned input-output associations and the learning velocity. We characterize how capacity and learning speed depend on the implemented nonlinearity and the levels of dendritic and somatic inhibition. We calculate analytically the distribution of synaptic weights in networks close to maximal capacity and find that the dendritic nonlinearity increases the fraction of zero-weight ("silent" or "potential") synapses, compared with the standard perceptron model, when no or weak robustness constraints are present, while the opposite occurs with strong robustness constraints. We test our model on standard real-world benchmark datasets and observe empirically that the nonlinearity provides an enhancement in generalization performance and that it enables to capture more complex input-output relations, compared to the perceptron model.

DOI: 10.1103/d7f1-xc8q

I. INTRODUCTION

Understanding the computational capabilities of single neurons is among the most fundamental open problems in neuroscience. A long-standing question concerns the role of dendrites in shaping neuronal information processing. In the simplest scenario, dendrites are devices that sum synaptic inputs linearly, propagating a dot product of a vector of presynaptic activity with a vector of synaptic weights to the axon initial segment, where a thresholding operation is applied to decide whether the neuron emits an action potential. In this view, neurons are analogous to simple perceptrons, whose learning capabilities have been studied extensively (e.g., Refs. [1–4]).

However, this view ignores the presence of active currents in dendrites, which can potentially lead to nonlinear. integration of synaptic inputs (for reviews, see e.g., Refs. [5–7]). These nonlinearities are due to various types of voltage-gated ionic currents, such as NMDA receptor–mediated synaptic currents [8,9], calcium currents [10,11], or sodium currents [12]. In cortical pyramidal neurons, in particular, it has been shown that inputs to a single dendritic branch sum in a strongly nonlinear fashion, while inputs to distinct dendritic branches sum linearly [13]. These results have led to the idea that neurons could be better described by multilayer devices than by the standard perceptron model [14–20].

Given the treelike morphology of dendrites, and the nonlinear integration of synaptic inputs pertaining to the same dendritic branch, a natural choice is to model single neurons as a particular type of a two-layer neural network called a tree committee machine. The computational properties of this neural architecture have been extensively studied in the statistical physics literature. Early works from the 1990s [21,22] showed that the storage capacity of tree committee machines increases with the size of the hidden layer K. In the case of the sign nonlinearity $g(x) \equiv \mathrm{sgn}(x)$, the maximal number of random input-output associations that can be learned scales as $P_c = \alpha_c(K)N$ where $\alpha_c(K) \propto \sqrt{\ln K}$, and N is the number of inputs [23]. Recently, it was pointed out that these results are valid only for activation functions presenting a discontinuity at the origin [24,25]. In particular, in the case of the rectified linear unit (ReLU) nonlinearity, an activation function commonly

*Contact author: fabrizio.pittorino@polimi.it

used in machine learning, the capacity of the tree committee machine remains finite as the size of the hidden layer K goes to infinity [24]. Moreover, most of the nonlinearities used in machine learning enhance learning by smoothing the corresponding loss landscape and inducing flatter and more robust minima that are attractive for gradient-based algorithms such as stochastic gradient descent (SGD) [24,26–29].

The aforementioned studies on the tree-committee machine from the statistical physics–machine learning community are typically performed without including biological constraints on the excitatory or inhibitory nature of synaptic weights (i.e., constraints on their sign) and without considering the specific dendritic nonlinearities that are observed experimentally. Given that multilayer networks are well known to have more powerful representation and generalization capabilities than single layer ones, as mathematically shown in early works on approximation capabilities of multilayer neural networks [30], a natural question is to what extent these sign constraints and the specific dendritic nonlinearities observed in cortical neurons affect the computational capabilities of single biological neurons.

Here we set out to study the computational capabilities of a single neuron model with dendritic branches implementing experimentally observed nonlinear integration and with sign-constrained positive synapses modeling excitatory connectivity, while inhibitory inputs are incorporated into dendritic and somatic thresholds. Using both analytical methods from statistical physics and numerical simulations we derive the number of possible stored input-output associations (capacity), training speed, noise robustness, and generalization properties and compare these properties to neuronal models with different dendritic nonlinearities. We also compute the distribution of synaptic weights at maximal capacity and show that this distribution contains a large fraction of zero-weight synapses, similarly to experimentally observed synaptic weight distributions in neocortical pyramidal cells. Interestingly, while in standard perceptrons with constrained synapses, sparse synaptic input connectivity can only be obtained when a robustness parameter is introduced [31,32], dendritic nonlinear input integration leads to high sparsity even in the absence of any robustness constraint.

II. SINGLE NEURON MODEL

We consider a single neuron model that transforms N binary synaptic inputs $\xi_i = \{0,1\}^N$ into a binary output $\hat{\sigma} = \{0,1\}$. In the standard perceptron model, the neuronal output is

$$\hat{\sigma} = \Theta\left(\sum_{i=1}^{N} W_i \xi_i - T\right), \tag{1}$$

where Θ is the Heaviside function, $\boldsymbol{W}$ is a vector of synaptic weights, typically optimized by a learning process, and T is a threshold.

Here, motivated by experiments that have revealed significant nonlinearities in the summation of inputs within single dendritic branches, but not across branches [6,8,13], we consider a generalization of the perceptron model with K dendritic branches and nonlinear summation of inputs within each dendritic branch (see Fig. 1). In this model, the neuronal output $\hat{\sigma}$ is

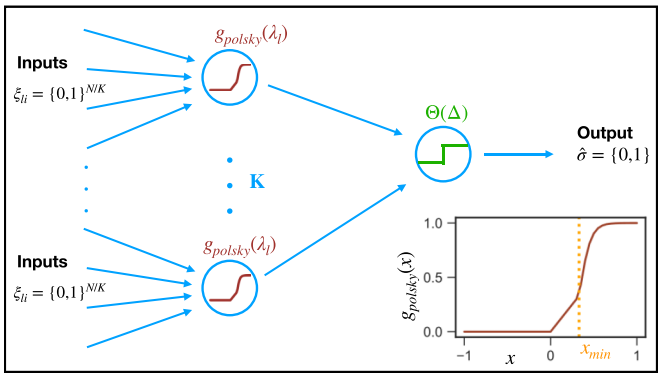


FIG. 1. Single-neuron model with dendritic nonlinearities. The neuron has K dendritic branches. Synaptic inputs to each dendritic branch are summed linearly and then processed through a dendritic nonlinearity depicted in the inset. The outputs of the dendritic branches are then summed linearly and compared to a somatic threshold.

$$\hat{\sigma} = \Theta(\Delta), \tag{2a}$$

$$\Delta = \frac{1}{\sqrt{K}} \sum_{l=1}^{K} g(\lambda_l) - \sqrt{K}\theta_s, \tag{2b}$$

$$\lambda_l = \sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} W_{li}\xi_{li} - \sqrt{\frac{N}{K}}\theta_d, \tag{2c}$$

where Δ is the total input to the soma, proportional to the sum of the outputs of all dendritic branches; g is a nonlinear function describing the dendritic nonlinearity; λ_l is the total input to dendritic branch l, which is a linear sum of inputs to this branch $\xi_{li} \in \{0,1\}^{N/K}$, weighted by synaptic efficacies W_{li}; θ_s is a somatic threshold; and θ_d is a dendritic threshold. Notice that this model is equivalent to a feedforward network with a layer of hidden units endowed with a nonlinear transfer function, and a fixed output summation layer.

A. Constraints on weights, excitation, and inhibition

The efficacy of real synapses is constrained by the identity of the presynaptic neuron. Synaptic weights are non-negative when the presynaptic neuron is excitatory (glutamatergic), while they are nonpositive when the presynaptic neuron is inhibitory (GABAergic). Here we consider for simplicity a scenario in which only the excitatory weights are modeled explicitly and are plastic. Inhibitory synapses are assumed not to be affected by learning and are lumped together in the two thresholds, θ_d and θ_s, describing inhibitory synapses onto dendritic branches and the perisomatic region, respectively. Thus, all synaptic weights W_{li} in Eq. (2c) obey the constraint $W_{li} \geqslant 0$.

B. Dendritic nonlinearity

Experiments in neocortical pyramidal cells have indicated that the dendritic output is roughly linear at low stimulation

intensities and that it then increases in a strongly nonlinear fashion beyond a threshold, before saturating [13].

To capture quantitatively these findings, we consider the following dendritic nonlinear transfer function

$$g_{\text{Polsky}}(x) = \begin{cases} \max(0,x) & x < x_{\min} \\ \frac{2(1-x_{\min})}{1+e^{-\gamma(x-x_{\min})}} - 1 + 2x_{\min} & x \geqslant x_{\min} \end{cases}, \quad (3)$$

where $x_{\min}$ is a dendritic nonlinearity threshold and γ describes the strength of the nonlinearity. We refer to this nonlinearity as the Polsky transfer function. It is plotted in the inset of Fig. 1. In the following we use $x_{\min} = 0.33$ and $\gamma = 15$, which provide a good approximation of the data in Ref. [13] (see Appendix B 2 for a discussion on how these parameters were obtained). Notice that this transfer function interpolates between the ReLU nonlinearity (when $x_{\min} \to \infty$) and the step nonlinearity, obtained for $x_{\min} = 0$, $\gamma \to \infty$. Note also that in the experiments of Ref. [13], only excitatory inputs are considered, and consequently only the positive side of the dendritic nonlinearity is probed. On the negative side, we take for simplicity g to be equal to zero. This scenario can be thought of capturing in a simplified way shunting inhibition.

C. Scaling of inputs and thresholds

Pyramidal cells have on the order of 10 000 synaptic inputs [33,34], scattered along tens to hundreds dendritic branches [35]. In this limit, taking synaptic weights and thresholds to be of order 1 ($W_{li} \sim \theta_d \sim \mathcal{O}(1)$), inputs to dendritic branches scale as N/K due to the sign constraints on the weights, with fluctuations of order $\sqrt{N/K}$ around the mean. To obtain a well-defined limit with finite means and variances, the dendritic threshold should balance the mean inputs, and the difference should be rescaled by $\sqrt{K/N}$. Likewise, at the somatic level, the somatic threshold should cancel the average somatic input, and the difference should be rescaled by $1/\sqrt{K}$. These considerations explain the scalings in Eqs. (2a)–(2c).

D. Learning tasks

We consider first a standard classification task with the objective of learning a dataset $\mathcal{D} = \{\boldsymbol{\xi}^\mu, \sigma^\mu\}_{\mu=1}^P$ composed of $P = \alpha N$ binary random input patterns ξ_{li}^μ that are independent and identically distributed Bernoulli variables with $P(\xi_{li}^\mu = 1) = f_{\text{in}}$ (input coding level) and labels σ^μ that are independent and identically distributed Bernoulli variables with $P(\sigma^\mu = 1) = f_{\text{out}}$ (output coding level). The task of the neuron is to correctly classify all input patterns, i.e., produce the correct output $\hat{\sigma} = \sigma^\mu$ when input $\boldsymbol{\xi}^\mu$ is presented. These input-output associations can be learned by progressively modifying the synaptic weights, to minimize the number of errors, or some surrogate loss function. This classification task (often called "storage problem" in the literature) has been studied extensively for the perceptron architecture [3,36], including also cases with sign-constrained weights [31,32,37]. It has also been studied in tree committee machines with sign nonlinearity of the hidden units [22,23,38], as well as more recently with generic nonlinearity [24,25]. On the numerical side, we study classical benchmark classification tasks in machine learning, such as MNIST [39], Fashion-MNIST [40], and CIFAR-10 [41].

E. Learning algorithm

To evaluate the computational performance of the two-layer nonlinear neuron described in (2), which is endowed with K dendritic branches and the transfer function defined in (3), and compare it with the linear neuron defined in (1) from an algorithmic standpoint, we develop an algorithm capable of learning with sign-constrained synapses. We then proceed to examine its behavior on various paradigmatic learning tasks. This algorithm is a modified version of SGD as detailed in Appendix C. Due to the non-negative nature of excitatory synapses, whenever the learning rule leads them to become negative, they are instantaneously set to zero. Importantly, the definition of the two models, particularly the treelike nature of the dendritic layer of the nonlinear neuron, naturally allows for their comparison at the same number of synaptic parameters, ensuring that computational improvements are exclusively attributable to their architectural and linear or nonlinear properties. The data that support the findings of this article are openly available [42].

III. STORAGE CAPACITY

A. Analytical methods

To investigate the properties of our single neuron model in the storage setting, one can make use of asymptotic methods from statistical physics [4,43]. Given a density of patterns $\alpha = P/N$, the uniform probability measure over all configurations classifying the patterns in $\mathcal{D}$ (or *solutions* to the learning problem) can be expressed, apart from a normalization factor, as

$$\mathbb{X}_{\mathcal{D}}(\boldsymbol{W}) = \prod_{\mu=1}^{P} \Theta[(2\sigma^\mu - 1)\Delta^\mu(\boldsymbol{W};\theta_d,\theta_s) - \kappa], \quad (4)$$

where Δ^μ is the somatic input defined in (2b) and σ^μ is the correct label for input μ. The parameter κ is a margin that imposes a certain degree of robustness on the learned $\boldsymbol{W}$. Exploiting self-averaging properties, the typical Gibbs entropy, i.e., the logarithm of the volume of solutions, can be obtained by taking the average $\langle\cdot\rangle_{\mathcal{D}}$ over the quenched disorder induced by the random realization of patterns and labels,

$$\phi = \lim_{N,K,P\to\infty} \frac{1}{N} \left\langle \ln \int d\mu(\boldsymbol{W})\, \mathbb{X}_{\mathcal{D}}(\boldsymbol{W}) \right\rangle_{\mathcal{D}}. \quad (5)$$

In (5), $\int d\mu(\boldsymbol{W})\,\bullet \equiv \int_0^\infty \prod_{li} dW_{li}\,\bullet$ is the integral over the prior weight measure, with the integration bounds reflecting the constraint over the weights. In order to compute the average $\langle\cdot\rangle_{\mathcal{D}}$ in (5), one can resort to the replica method in the replica symmetric (RS) approximation [44]. By using the saddle-point method [45] one obtains a set of equations for a finite set of order parameters that are sufficient to the describe the large-dimensional limit of the system. We refer to Appendix A for a brief description of the analytical methods and to Appendix D for more detailed derivations.

Pyramidal cells receive roughly 10 000 synaptic inputs, which are distributed across several dozen to several hundred

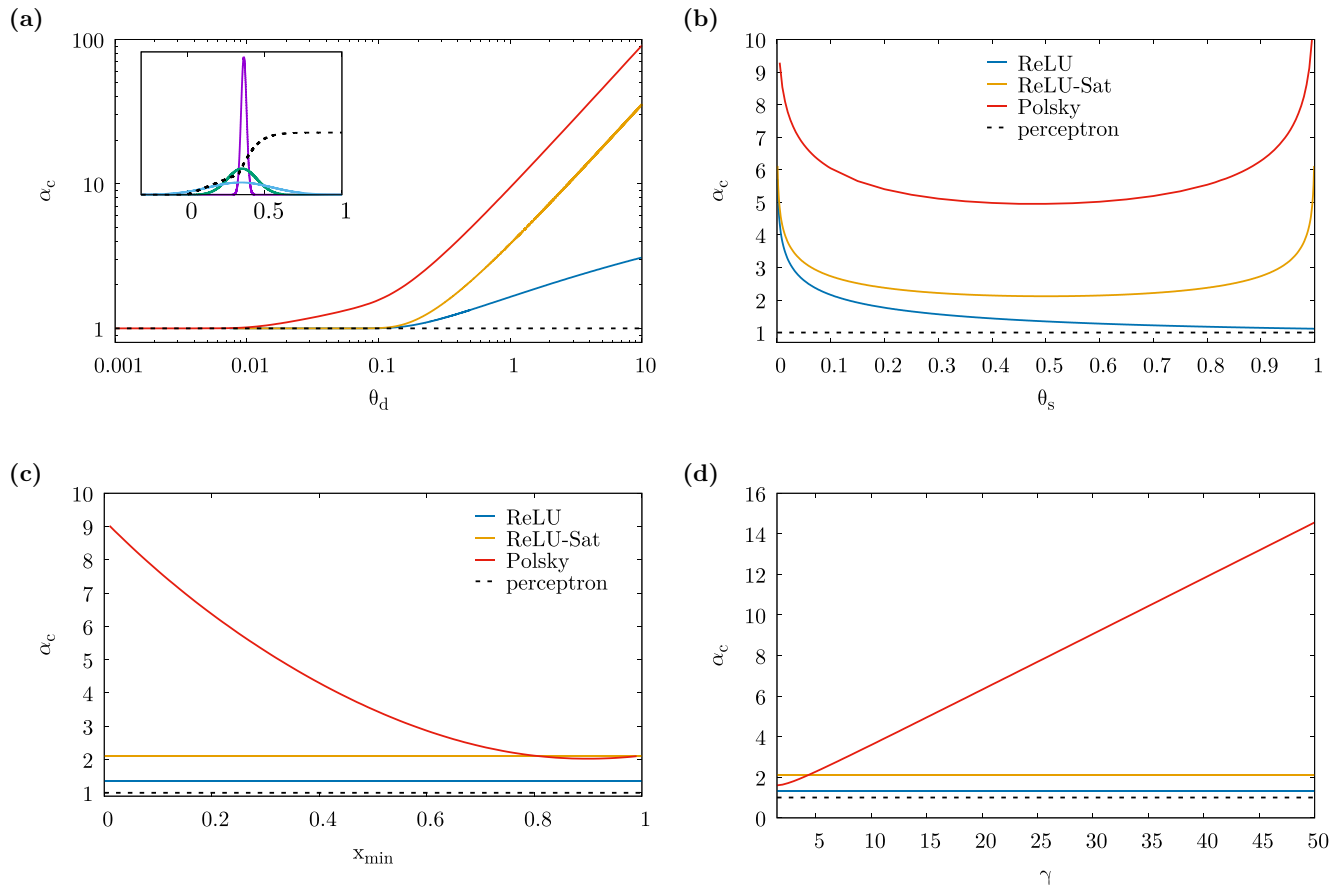


FIG. 2. Critical capacities α_c for ReLU (blue), saturating ReLU (orange), and Polsky (red) nonlinearities a function of various parameters. In all panels $\kappa = 0$, $f_{in} = f_{out} = 0.5$. [(b), (c), and (d)] $\theta_d = 0.5$; [(a), (c), and (d)] $\theta_s = 0.5$; [(a), (b), and (d)] $x_{min} = 0.33$; [(a)–(c)] $\gamma = 15$. Dashed black lines represents the case of the one-layer neuron model, where the critical capacity $\alpha_c^{perc} = 1$. (a) Dependence on the dendritic threshold θ_d. Inset: Distribution of dendritic preactivations for $\theta_d = 0.01$ (violet), 0.05 (green), and 0.1 (cyan); we also plot with the dashed black line the Polsky activation to better show how the nonlinearity is exploited for that value of θ_d. (b) Dependence on the somatic threshold. (c) Dependence on x_{min}. (d) Dependence on γ.

dendritic branches [33–35]. We therefore considered the limit of large number K of dendritic branches in our analytical calculations. At the same time, however, we consider the regime where K is small compared to the total number of synapses N, i.e., $K/N \to 0$. This is not only a realistic assumption, but it also allows us to reduce the computational complexity of the analytical calculations which are valid for a generic nonlinearity (see Appendix A).

As we show in Appendix D, computing the entropy (5) in the large-N and -K limit, in turn, gives access to several physical observables of interest, namely the critical capacity and the distribution of synaptic weights.

B. Critical capacity

The randomness of the labels in the dataset do not make the task learnable for any value of the density α. Indeed, in the large-N limit, there exists a sharp threshold α_c for the probability of finding a solution to the learning problem. For $\alpha < \alpha_c$ this probability is 1, meaning that there exist synaptic weight configurations that are able to classify the inputs correctly; at $\alpha = \alpha_c$ the probability of finding a solution drops abruptly to zero. For $\alpha > \alpha_c$ the complexity of the model is therefore no longer sufficient to classify the activity patterns. α_c can be thought as a measure of expressivity of our single neuron model.

At α_c the typical overlap q between a pair of solutions extracted from the Gibbs measure (4) tends towards the typical squared norm Q of solutions. We have therefore expanded the entropy in terms of the variable $dq \equiv Q - q$. We report in Appendix D the technical details of how α_c can be computed from this scaling for a given value of the external parameters θ_d, θ_s, f_{in}, f_{out}, and κ and for a generic activation function g.

In the case of a linear activation function (i.e., $g(x) \equiv x$), our model is equivalent to the one-layer neuron model whose activity is based on a thresholding operation θ_d applied to the soma. In this case we recover the results on the critical capacity α_c^{perc} [31,32,37,46]. If the margin $\kappa = 0$, then it has been shown in Ref. [31] that the capacity is independent on θ_d; in particular, for $f_{in} = f_{out} = 0.5$, $\alpha_c^{perc} = 1$.

This is not true in the case of the two-layer neuron model, in which changing the dendritic threshold *strongly* alters the expressivity of the model. In the upper left panel of Fig. 2 we show the plot of the critical capacity of our two-layer

neuron model as function of θ_d for different types of dendritic nonlinearities, namely ReLU, a "saturating" ReLU function $\min(\max(0, x), 1)$, and Polsky as in (3). For comparison purposes we also plot the critical capacity of the one-layer neuron model. As the figure shows, the storage capacity of the model is greatly enhanced by the presence of the nonlinearity. As shown analytically in Appendix D, in the limit $\theta_d \to 0$, the two-layer neuron models becomes equivalent to the one-layer perceptron model. When θ_d increases, all models with nonlinear integration increase their capacity but in a strongly non-linearity-dependent way. With a nonsaturating nonlinearity such as ReLU, the increase in capacity is logarithmic in θ_d, and it is much smaller than with saturating nonlinearities where the capacity increases linearly with θ_d. Finally, the model with Polsky nonlinearity outperforms the saturating ReLU function, thanks to the additional nonlinear region for $x > x_{\min}$.

The behavior of α_c as a function of θ_d can be understood through an analysis of the shape of the distribution of dendritic preactivation, which we show in Appendix D to be a Gaussian with a mean and variance that are functions of the norm of the weights Q, and the input coding level f_{in}. We show the shape of the dendritic preactivation for the Polsky activation in the inset of the upper left panel of Fig. 2 for several values of θ_d. If θ_d is small, then the distribution is peaked in a range where the Polsky activation behaves linearly; therefore, the model cannot fully exploit the nonlinearity and behaves as a one-layer model. On the contrary, by increasing θ_d, the dendritic preactivation distribution widens towards the region where the Polsky activation saturates; if one keeps increasing θ_d, then the weight of the active region before saturation becomes negligible. In this limit, we expect the critical capacity to diverge, since the Polsky activation becomes equivalent to the Heaviside theta activation.

The capacity α_c also strongly depends on the somatic threshold θ_s, as shown in the upper right panel of Fig. 2. In the bottom panels of Fig. 2 we show how the capacity of the network with a Polsky nonlinearity depends on the choice of its parameters $x_{\min}$ and γ. The critical capacity increases both with decreasing $x_{\min}$ or increasing γ, as in this case the nonlinearity is closer to the Heaviside theta function. When $x_{\min} = 1$, the Polsky nonlinearity effectively reduces to the "saturating" ReLU function, and so the capacity of the two models coincide in this limit.

Notice that the estimation of the critical capacity that we have done is based on the RS ansatz; in general since the model we are analyzing is nonconvex, the RS ansatz is thought to be only an upper bound to the true result. In order to get more precise results on the critical capacity, one needs to resort to the replica symmetry breaking ansatz (RSB) [47]. 1RSB corrections to the critical capacity estimation have been computed in one and two layer nonconvex neural network models with no constraint on the sign of the weights [22,24,25,48,49]. Recently, the exact capacity of infinitely wide tree committee machines and perceptrons with negative stability has been computed using a full-RSB ansatz [50]. For our two-layer neuron model, computing 1RSB effects on the storage capacity is technically challenging because of the sign constraints. Note also that our calculations are done in the $K \to \infty$ limit. Networks with finite K are expected to have a capacity of at most $16\sqrt{\log(K)}/\pi$, the asymptotic behavior of committee machine with step function nonlinearity and no constraints on weights [23]. For values of K in the range 30–100, this leads to upper bounds in the range 6–7, far below the large K estimates of the capacity shown in Fig. 2 for large θ_d. Thus, the benefits of the specific Polsky nonlinearity are expected to be the strongest in an intermediate region of values of θ_d, x, and γ.

C. Algorithmic capacity and learning speed

In the previous section, we computed an upper bound for the maximal capacity using a RS calculation. We now turn to the question of the capacity of specific learning algorithms, and to the question of the speed of learning. We use the widely used SGD algorithm (see Appendix C for details of the algorithm). It is important to note that, unlike the linear neuron model, the optimization problem in the two-layer nonlinear neuron is highly nonconvex, and there is no guarantee that algorithms can reach the critical capacity, similar to results concerning binary ± 1 weights models [26,51,52].

In Fig. 3, we report the final training error after training with SGD as a function of the control parameter $\alpha = \frac{P}{N}$ (i.e., the density of input patterns) for both the linear and nonlinear neuron models. Figure 4 depicts the training error as a function of the training time for SGD. We select the optimal hyperparameters that maximize the algorithmic capacity and training speed using a grid search procedure (see Appendix C). It is worth noting that, at fixed values of the dendritic and somatic thresholds θ_d and θ_s, the SGD algorithm has two hyperparameters: the learning rate ζ and the cross-entropy parameter γ_{ce}. Appendix C provides an in-depth discussion of algorithmic implementations and hyperparameter selection for both the algorithmic capacity evaluation and the training speed.

Figure 3 shows that the neuron with nonlinear dendritic integration is able to reach algorithmic capacities that are larger than the maximum one achievable by the linear model, i.e., $\alpha_c^{\text{perc}} = 1$. Since the comparison between the two neuron models is performed at the same number of parameters (the number of synapses is $N = 999$ in both cases), the improvement in performance is solely attributable to dendritic nonlinearities. It is worth noting that in the linear model, the SGD algorithm can reach $\alpha_{\text{alg}} = \alpha_c^{\text{perc}} = 1$ due to the convexity of the problem. Concurrently, as reported in Fig. 4, we find that the nonlinear neuron requires fewer steps to learn the training set compared to its linear counterpart.

Despite the fact that the nonlinear neuron algorithmically reaches larger capacities than the maximum capacity theoretically achievable by the linear model, we observe that the algorithmic capacity of our algorithm is generally suboptimal with respect to the analytically calculated critical capacity, $\alpha_c(\theta_d, \theta_s)$ as shown in Fig. 3. The difference between the SGD algorithmic capacity and the analytical estimate is relatively mild at low θ_d, but the gap widens as the dendritic threshold increases. This difference between algorithmic capacity and analytical RS estimate may be due to several factors, including RSB effects on the critical capacity (as the RS estimate represents an upper bound), finite K effects, and algorithmic hardness.

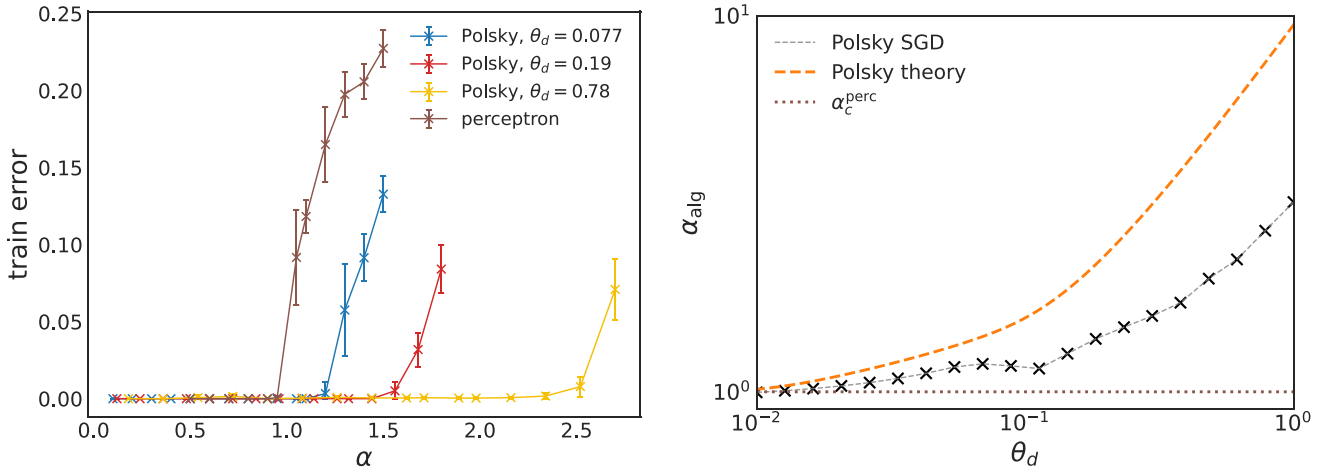


FIG. 3. Left panel: fraction of misclassified patterns on the training set as a function of the total fraction of patterns $\alpha = \frac{P}{N}$ for the nonlinear neuron compared with the linear one, both trained with SGD (Algorithm I). The curves show different representative values of the dendritic threshold θ_d, with $\theta_s = 0.5$. For both neuron models the number of synapses (equivalently, the input size) is $N = 999$, the number of dendritic branches for the nonlinear neuron is $K = 27$, and each curve is averaged over 10 realizations of the initial conditions. Note that the nonlinear neuron achieves capacities greater than the maximal capacity of the linear perceptron model, which is $\alpha_c^{\text{perc}} = 1$. The optimal initial learning rate (see Appendix C for details) for both the linear and nonlinear neurons is $\zeta = 0.01$. Right panel: Algorithmic capacity for SGD (Algorithm I), compared to the analytical RS estimate in function of the dendritic threshold θ_d. Same parameters as the left panel, considering biologically plausible values of $\theta_d \in [0.01,\ 1]$ (see Appendix B 2 for a discussion of the biological values of the parameters). The critical capacity of the perceptron (dashed line) is $\alpha_c^{\text{perc}} = 1$.

IV. DISTRIBUTION OF SYNAPTIC WEIGHTS AND INPUT CONNECTIVITY SPARSITY

We next turned to the calculation of the distribution of synaptic weights in our neuron with nonlinear dendritic integration. In a perceptron with sign-constrained weights, it has been shown that this distribution contains a delta function at zero ("silent" or "potential" synapses) and a truncated Gaussian distribution of non-negative weights at maximal capacity [31,32],

$$P(W) = p_0\,\delta(W) + \frac{1}{\sqrt{2\pi}W_\star} e^{-\frac{(W+BW_\star)^2}{2W_\star^2}}\Theta(W), \qquad (6)$$

where $\Theta(\cdot)$ is the Heaviside function. The fraction of silent weights is a simple function of B, $p_0 = H(-B) \equiv \frac{1}{2}\text{Erfc}(-\frac{B}{\sqrt{2}})$, whereas B, $W_\star$ depend on parameters of the model. In the absence of robustness constraints, the fraction of silent synapses p_0 is exactly 50%, but this fraction increases in the presence of robustness constraints. Furthermore, it has been shown that such a distribution can fit well data from both cerebellar Purkinje cells [31,53], and cortical pyramidal cells [32,54], but only with a large value of κ, consistent with the idea that these networks optimize storage capacity with a strong robustness constraint, or vice versa optimize robustness of stored information. The obtained strong robustness derives from the experimentally observed low connection probabilities, ∼0.2 in granule cell to Purkinje cell connections [53], and ∼0.1 in layer 5 recurrent pyramidal cell connections [54–58].

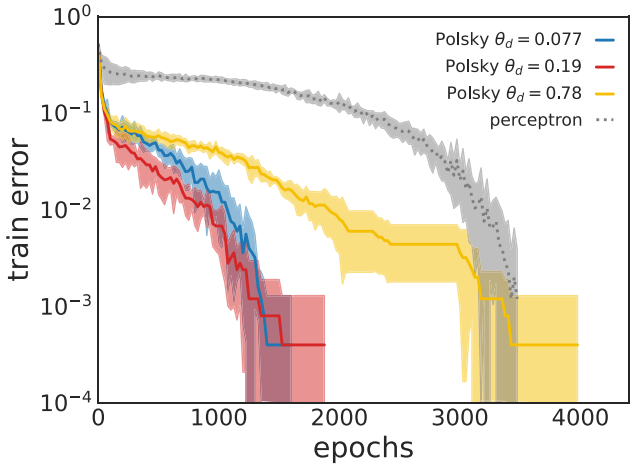


FIG. 4. Training speed. Comparison of the convergence times of linear and nonlinear neuron models using SGD (Algorithm I). The total fraction of patterns is fixed at $\alpha = 0.5$, ensuring that the algorithm can perfectly learn the training set with both the linear and nonlinear neuron models. For both neuron models, the number of synapses (i.e., the input size) is $N = 999$, and the number of dendritic branches for the nonlinear neuron is $K = 27$. The optimal learning rate for both the linear and nonlinear neurons is $\zeta = 0.01$. Each curve is averaged over 10 realizations of the initial conditions.

To investigate the impact of the dendritic nonlinearity on the distribution of synaptic weights, we computed $P(W)$ of models with dendritic nonlinearities. As we show in Appendix D, in the large-K limit, the functional form of the distribution of synaptic weights is exactly the same as the one obtained for the perceptron, for any value of α, and in particular in the critical capacity limit, Eq. (6). However, the values of $p_0, B, W_\star$ depend *strongly* on the type of nonlinearity.

We show in Fig. 5 the fraction of silent synapses p_0 as a function of various parameters, for the committee machine with various types of nonlinearities, showing the perceptron as a comparison. We start with the case with no robustness constraints ($\kappa = 0$). Contrary to the perceptron, the fraction

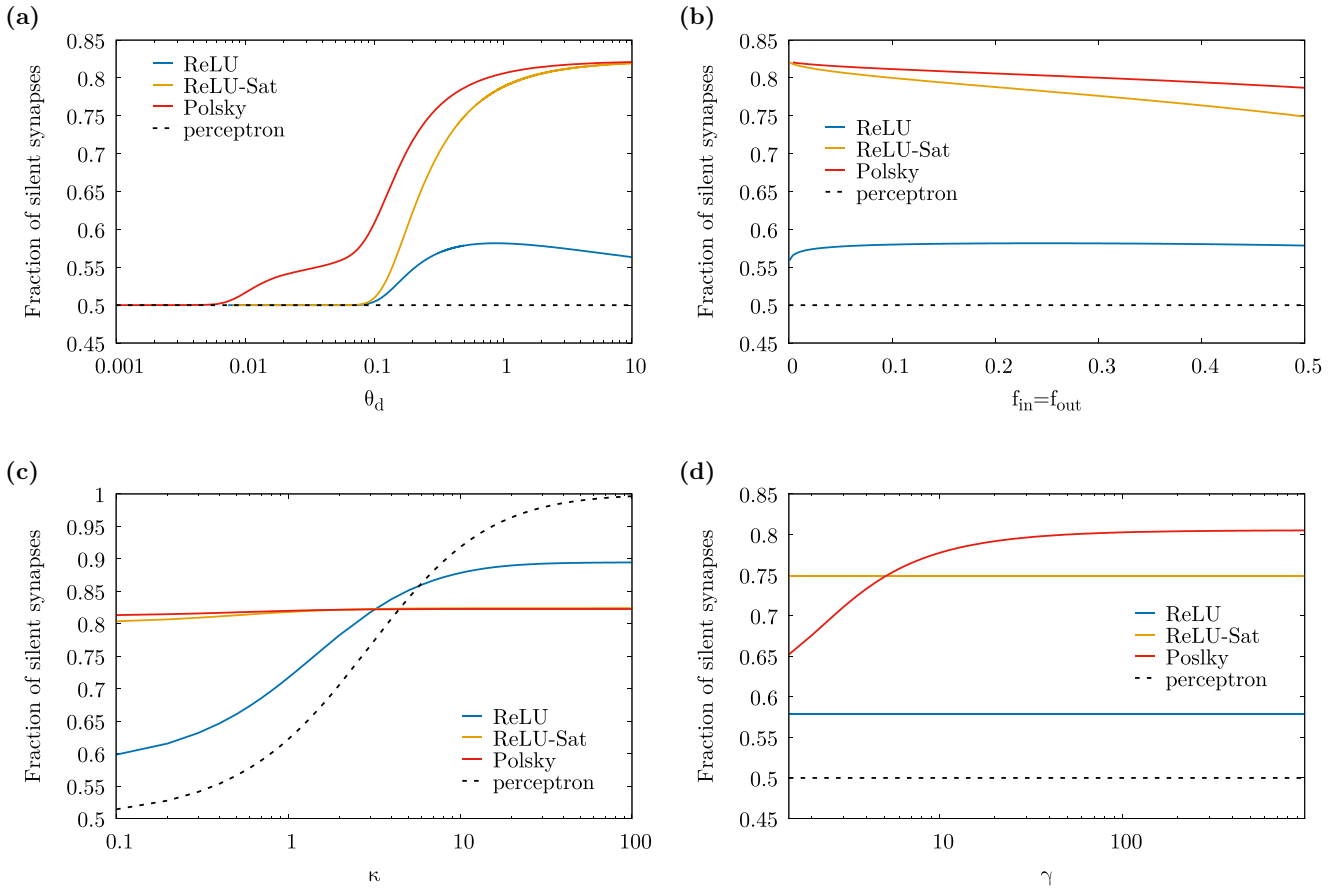


FIG. 5. Fraction of silent synapses at maximal capacity for perceptron (dashed black) and committee machines with ReLU (blue), saturating ReLU (orange), and Polsky ($x_{\min} = 0.33$, red line) nonlinearities as a function of various parameters. In all panels $\theta_s = 0.5$. (a) Dependence on dendritic threshold θ_d for $f_{\text{in}} = f_{\text{out}} = 0.5$, $\kappa = 0$. (b) Dependence on input and output coding levels $f_{\text{in}} = f_{\text{out}}$ for $\kappa = 0$, $\theta_d = 0.5$. (c) Dependence on the robustness parameter κ, for $f_{\text{in}} = f_{\text{out}} = 0.1$, $\theta_d = 0.5$. The biological plausible nonlinearity Polsky has the highest fraction of silent synapses at maximal capacity at small κ, while for large κ the fraction of silent synapses goes to 1 only in the perceptron model. (d) Dependence on γ for $f_{\text{in}} = f_{\text{out}} = \theta_d = 0.5$ and $\kappa = 0$.

of silent synapses is larger than 50% in the nonlinear models. Fig. 5(a) shows that this fraction is close to 0.5 when θ_d is small but then increases as a function of θ_d. In the Polsky and ReLU-Sat cases, the fraction grows until it saturates at around 0.8, for the chosen parameters. This fraction is, however, weakly dependent on coding levels and κ, as shown in Figs. 5(b) and 5(c). In the ReLU, the fraction of silent synapses reaches a maximum close to 0.6 and then decreases in the large-θ_d limit. Thus, the saturating nonlinear models have a drastically different behavior than the perceptron: They have a large fraction of silent synapses at zero robustness, but this fraction depends only weakly on robustness, while the perceptron model has 50% silent synapses at zero robustness, but this fraction increases strongly as a function of κ. Interestingly, the model with ReLU nonlinearity has an intermediate behavior. Finally in Fig. 5(d) we show that increasing the saturation speed of the Polsky nonlinearity, which is controlled by the parameter γ, results in a larger sparsity.

Synaptic weight distribution at algorithmic capacity

Our numerical simulations show that at algorithmic capacity, the distribution becomes well described by a delta function at zero, with a finite fraction of zero weights, and a truncated Gaussian describing positive weights, consistent with analytical calculations. The fraction of zero weight synapses at algorithmic capacity significantly deviates from the 50% one corresponding to the linear neuron, reaching a maximum of 70% for $\theta_d \gtrsim 1$. The synaptic weight distribution at algorithmic capacity, for $\theta_d = 0.78$ and $\theta_s = 0.5$, corresponding to the yellow curve in Fig. 3 at $\alpha = 2.4$, is shown in Fig. 6.

V. NOISE ROBUSTNESS AND GENERALIZATION

A. Robustness to input and synaptic noise

Noise is a ubiquitous feature at all levels of the nervous system, from the molecular to the whole brain level [60]. In particular, single neurons operate in a highly noisy environment, due to background inputs they constantly receive. Thus, robustness to input and synaptic noise are fundamental computational requirements for any realistic single neuron model. We thus turn to an investigation of robustness of our model to noise. In our simulations, we estimate the robustness to input noise by independently flipping the entries of each pattern in the training set with probability ρ maintaining the same label and measuring the train error

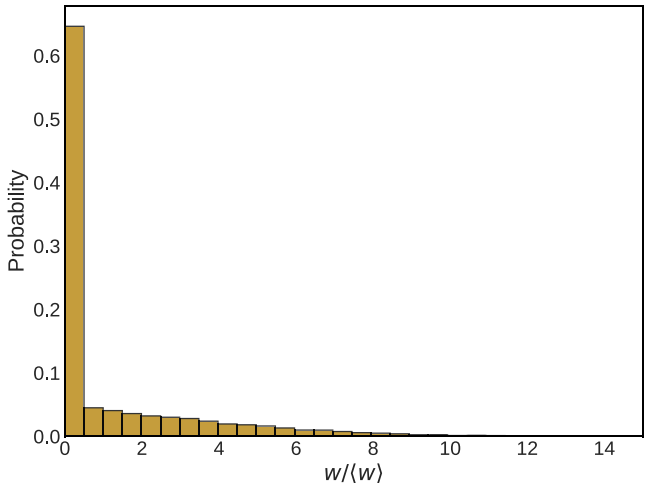


FIG. 6. Synaptic weight distribution at algorithmic capacity for $\theta_d = 0.78$ and $\theta_s = 0.5$ ($\alpha_{\rm alg} = 2.4$ for these values of the parameters).

of optimal synaptic configurations on this corrupted training set. Figure 7 shows the robustness to input noise for the SGD algorithm. We observe that the nonlinear model is more robust to input perturbations compared to the linear model with the same number of synaptic parameters. Similarly, we measure synaptic noise robustness by estimating the number of misclassified patterns when synaptic strengths are perturbed by applying a multiplicative Gaussian noise of amplitude σ to zero-error synaptic configurations $\boldsymbol{W}$. In practice, we measure the quantity $\delta E_{\rm train}(\boldsymbol{W},\sigma) = \mathbb{E}_z\, E_{\rm train}([\boldsymbol{W} + \sigma \boldsymbol{z} \odot \boldsymbol{W}]_+) - E_{\rm train}(\boldsymbol{W})$, where $E_{\rm train}(\boldsymbol{W})$ is the number of errors made by configuration $\boldsymbol{W}$ on the training set, the expectation $\mathbb{E}_z$ is over normally distributed synaptic noise realizations $\boldsymbol{z} \sim \mathcal{N}(0, I_N)$, $\odot$ is the element-wise product, and $[\cdot]_+$ is the ReLU function. The quantity $\delta E_{\rm train}(\boldsymbol{W},\sigma)$ is also known as

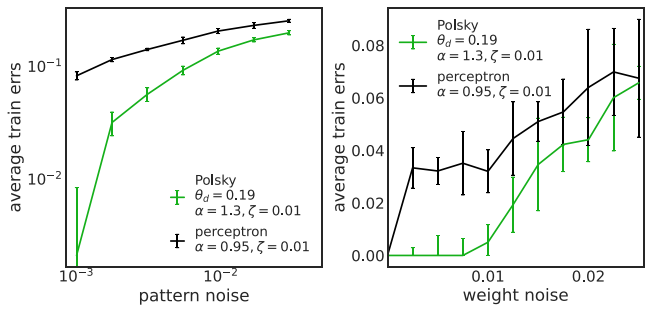


FIG. 7. Robustness to input and synaptic noise in the storage case. Left panel: Robustness to input noise, measured as the increase in the fraction of misclassified patterns as a function of input flipping probability. Right panel: Robustness to synaptic noise, measured as the increase in the fraction of misclassified patterns as a function of the amplitude of Gaussian synaptic noise. Note that synaptic noise robustness is consistent with the *local energy* definition [59]. For both neuron models, the number of synapses (i.e., the input size) is $N = 999$, and the number of dendritic branches for the nonlinear neuron is $K = 27$. Each curve represents the average over 10 realizations of the initial conditions.

the *local energy* in the machine and deep learning literature. It serves as a proxy for the flatness of the energy landscape around a given optimal configuration [59,61]. It is worth noting that a synaptic hard threshold at zero is implemented in this case as well, meaning that synaptic configurations cannot take negative values even under perturbations, i.e., $\boldsymbol{W} > 0$ and $\boldsymbol{W} + \sigma \boldsymbol{z} \odot \boldsymbol{W} > 0$). Figure 7 presents the robustness to synaptic noise of the linear and nonlinear neurons for the SGD algorithm. The results demonstrate that the dendritic nonlinearity enhances synaptic noise robustness, or, equivalently, the local energy landscape around optimal synaptic configurations is flatter in the nonlinear case.

B. Generalization performance on real-world datasets

To study the generalization properties of the neuron model defined in (2), we focus on binary classification learning tasks using the MNIST [39], Fashion-MNIST [40], and CIFAR-10 [41] datasets, which are standard benchmarks in machine learning. The generalization error, a fundamental machine learning observable, can only be estimated in the presence of a test set, which is absent in the storage case. To ensure that the generalization tasks remain reasonably difficult while still allowing for generalization, we divide the MNIST and Fashion-MNIST datasets into an odd-even binary classification task. For MNIST, we separate odd and even digits into two different classes. Similarly, for FashionMNIST, we separate the classes corresponding to even and odd labels into two groups. For the CIFAR-10 dataset, we choose two different classes in order to define a reasonably difficult generalization task, namely Bird and Ship. Appendix C provides details on the dataset binarization procedure and hyperparameter selection. As shown in Fig. 8, the nonlinear neuron demonstrates better performance on all the aforementioned generalization tasks. We have also verified that in more challenging scenarios, such as learning odd and even classes in the CIFAR-10 dataset (separating the classes into two groups based on their even or odd labels), the generalization error is very close to random guessing (around 40%).

VI. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper, we have studied the effect of realistic dendritic nonlinearities on the computational abilities of a single neuron model with sign constrained synaptic weights. We have shown that dendritic nonlinear integration is beneficial for multiple reasons. First, it improves the overall expressivity of a single neuron, measured as the maximum number of input-output associations that it can correctly store. Second, the nonlinearity generates input connectivity sparsity in the model, i.e., it leads to a large (>0.5) fraction of zero weight (silent or potential) synapses, in the absence of any explicit robustness constraint. This is in marked contrast with the standard perceptron model in which 50% of synapses have zero weight in the absence of any robustness constraints. Another marked difference between the model with realistic nonlinearities and the perceptron is that the fraction of zero weight synapses is only weakly dependent of robustness constraints ($\kappa > 0$), while in the perceptron model this fraction increases

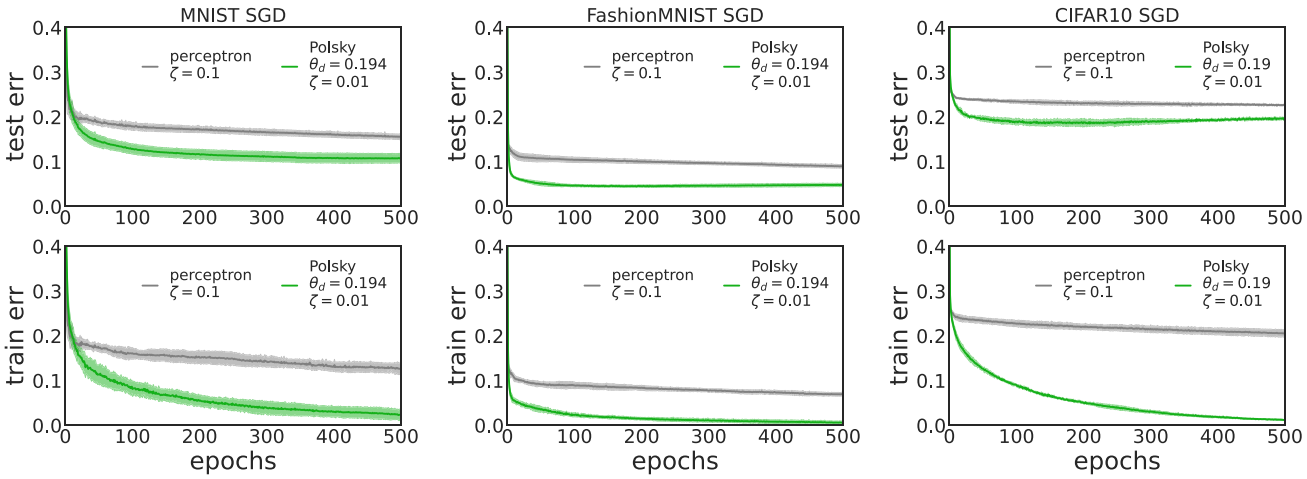


FIG. 8. Generalization capabilities on real-world datasets: MNIST, Fashion-MNIST, CIFAR-10. Comparison of train and generalization errors of the nonlinear neuron and the linear neuron using SGD (Algorithm I). Left column: MNIST (test and train error in the upper and lower panel respectively); middle column: Fashion-MNIST (test and train error in the upper and lower panel respectively); right column: CIFAR-10 (test and train error in the upper and lower panel respectively). For both neuron models, the number of synapses (i.e., the input size) is $N = 1568$ for MNIST and Fashion-MNIST, and $N = 6144$ for CIFAR-10. The number of dendritic branches for the nonlinear neuron is $K = 49$ for MNIST and Fashion-MNIST, and $K = 72$ for CIFAR-10. Each curve represents the average over 10 realizations of the initial conditions.

markedly with κ, going asymptotically to 1 as κ becomes large.

On the algorithmic side, we quantified the benefit of nonlinear dendritic integration from several points of view. First, we have shown that nonlinear dendritic processing enables an algorithm such as SGD to find optimal synaptic configurations at larger density of input patterns with respect to linear integration, and to find them consistently faster. Second, we have found that synaptic configurations found by such algorithms have desirable computational properties in the nonlinear case, such as a stronger robustness to input and synaptic noise, and higher generalization ability, compared to the linear case.

Algorithmically reaching the analytically calculated critical capacity α_c is challenging as shown in Fig. 3. Several factors contribute to the discrepancy between algorithmic and critical capacities. The critical capacity computed using RS serves as an upper bound for the true critical capacity, as RSB is likely to occur at lower values of α, a phenomenon well established for both ReLU and step function nonlinearities [22,24,25,48,50]. Algorithmically, there is no guarantee to reach the optimal capacity, and finite-size effects from simulations with finite K and N could also influence results. Understanding this discrepancy will be the subject of future work.

Our work provides a bridge between two previously disconnected branches of the literature. On one side, committee machines have been studied extensively by statistical physicists as one of the simplest multilayer feedforward architectures [22–25,48,50], but biological constraints are typically ignored in these studies. Here we have generalized previous calculations on committee machines in two ways: (1) by including sign constraints on synaptic weights and (2) by including a more realistic form of dendritic nonlinearity than both the sign [22,23,48,50] and ReLU nonlinearities [24,25] that have been studied previously. Note, however, that calculations in Refs. [24,25] were done for generic activation functions and [50] showed then even a full RSB computation is tractable, so it could be used for the Polsky case mentioned here, though these calculations would need to be extended to sign-constrained synapses. On the other side, computational neuroscientists have long sought to understand the computational roles and/or advantages of dendritic nonlinearities [14,18,19,62], but these studies typically rely on numerical simulations in models that lack analytical tractability. Our work provides instead for the first time analytical estimates of storage capacity and distributions of synaptic weights in a model with a realistic dendritic nonlinearity, and synaptic sign constraints.

Future work will concentrate on dropping some of the modeling assumptions we made here. For example, we have not considered the fact that there may be more than two layers of input processing through dendrites; this would make neurons more similar to deep tree networks. It would be therefore interesting to study the effect of multiple layers of dendritic integration on all the quantities studied in this paper. One should also take into account the effects of potential correlations structures in inputs, e.g., that inputs coming in the same branch are more correlated than inputs to different branches [63]. Similarly, synaptic inputs coming at different regions of the dendritic tree (e.g., basal vs apical) may have different statistical properties and convey different types of inputs.

In the present work, we have focused on excitatory synaptic inputs and absorbed inhibitory inputs in the dendritic and somatic thresholds, assuming that only excitatory inputs participate in learning. This work could be generalized to include plastic inhibitory synapses, as has been done in the case of the standard perceptron [64,65]. In particular, such calculations would be useful to understand the impacts of dendritic and somatic inhibition on the computational abilities of neurons.

We have considered here one of the simplest possible computations performed by a neuron, i.e., classifying sets of inputs into two classes. In such a framework, inputs consist of specific subsets of neurons that are active within a particular

temporal interval that is comparable to NMDA time constants (tens of milliseconds), and the neuron either needs to emit a spike, or a burst of spikes for inputs for which it should be active. A potential future direction would be to generalize this scenario to a scenario in which inputs consist of longer spatiotemporal patterns (e.g., Ref. [66]).

Another important future direction concerns synaptic plasticity algorithms. Here we have investigated a simple plasticity algorithm (SGD) in a standard supervised learning scenario, in which a teaching signal is available to the neuron. In cerebellar Purkinje cells, this teaching signal could be implemented by the climbing fiber input, which has been shown to correlate with error in motor tasks. In cortical pyramidal cells, the presence of error signals is more speculative. In the presence of the error signal, SGD in a single neuron model leads to a local plasticity rule, that could be plausibly implemented in a biological neuron, unlike in multilayer networks (see Appendix C). Finally, it will be interesting to investigate the computational abilities of a neuron with nonlinear dendrites in a purely unsupervised learning setting.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of our colleague and friend Luca Trevisan. F.P. acknowledges support by the PNRR-PE-AI FAIR project funded by the NextGeneration EU program. E.M.M. acknowledges the MUR-Prin 2022 funding Prot. 20229T9EAT, financed by theEuropean Union (Next Generation EU). C.L. acknowledges support by DARPA Award DIAL-FP-038, and The William F. Milton Fund from Harvard University.

DATA AVAILABILITY

The data that support the findings of this article are openly available [42].

APPENDIX A: ANALYTICAL METHODS

The entropy in (5) can be computed using the replica method; since the average over the log in (5) is difficult to compute one uses the identity $\ln(x) = \lim_{n\to 0} \frac{x^n-1}{n}$. Taking n as an positive integer one ends to deal with an enlarged system of n identical virtual copies of the system, so that the average can easily be performed, at the price to couple the replicas together. However in the large-N limit, it turns out that the properties of the model can be fully characterized by a finite set of quantities called order parameters that are determined self-consistently by solving equations obtained by saddle-point method (see also the recent notes [43] for a more pedagogical introduction). In practice, the averaged entropy can be fully characterized by the order parameters

$$\sum_{i=1}^{N/K} W_{li}^a = \frac{N}{K}\overline{W} + \sqrt{\frac{N}{K}} M_l^a, \tag{A1a}$$

$$q_l^{ab} \equiv \frac{K}{N}\sum_{i=1}^{N} W_{li}^a W_{li}^b, \tag{A1b}$$

$$Q_l^a \equiv \frac{K}{N}\sum_{i=1}^{N} \left(W_{li}^a\right)^2, \tag{A1c}$$

and their conjugated ones $\hat{M}_l^a$, $\hat{q}_l^{ab}$, $\hat{Q}_l^a$, with $a, b \in [n]$ and $l \in [K]$. They represent respectively: the typical average synaptic weight, the most probable overlap between two replicas extracted from the Gibbs measure in (4), and the typical averaged squared norm of a synaptic weights belonging to the dendritic branch $l \in [K]$. Since the fields are nonoverlapping, each branch has access only to a portion of the synaptic input; therefore there is no correlation between hidden units in the same architecture. Notice also that, in (A1a), the average can be expressed with the sum of two contributions: The first represents the average of the scaled synaptic weights $\overline{W} = \frac{\theta_d}{f_{\rm in}}$, while the second represents a $\sqrt{\frac{N}{K}}$ correction needed to fine tune this average with respect to the threshold θ_d. We report in Appendix D the full analytical calculations of the entropy (5) in the case in which the structure of the overlap matrix q_l^{ab} is symmetric under permutation over the replica indices (the so-called RS ansatz), and the order parameters do not depend on replica and dendritic branch indices $l \in [K]$. This means $q_l^{ab} = Q\delta_{ab} + q(1-\delta_{ab})$, $\forall a, b \in [n]$ and $l \in [K]$ and similarly for the other order parameters. The entropy can be therefore obtained by maximizing a function $\phi_{\rm RS}$ with respect to the order parameters

$$\phi = \max_{q,\hat{q},Q,\hat{Q},M,\hat{M}} \phi_{\rm RS}(q, \hat{q}, Q, \hat{Q}, M, \hat{M})$$

$\phi_{\rm RS}$ can be written as

$$\phi_{\rm RS} = \mathcal{G}_S + \alpha\, \mathcal{G}_E. \tag{A2}$$

i.e., as a sum of an entropic contribution $\mathcal{G}_S$ which represents the log of the total volume of configurations $\boldsymbol{W}$, and an energetic part $\mathcal{G}_E$ that corresponds to the log of the fraction of solutions for a given α. The explicit expressions of $\mathcal{G}_S$ and $\mathcal{G}_E$ are reported in Appendix D.

The values of the order parameters q, $\hat{q}$, Q, $\hat{Q}$, M, and $\hat{M}$ can be found by solving a set of coupled saddle-point equations [45] obtained by equating to zero the derivative of $\phi_{\rm RS}$ with respect to each of them.

Limit of large number of dendritic branches

As noted in Refs. [33–35], pyramidal cells receive roughly 10 000 synaptic inputs, which are dispersed across a wide range of dendritic branches, varying from several dozen to several hundred. This motivates considering the limits where N and K tend towards infinity but with K/N approaching zero.

On the technical side, solving the saddle-point equations for generic K is in general a very difficult task, since in order to compute the energetic term one should evaluate $2K$-dimensional integrals. However in the large-K limit the energetic term simplifies considerably. Indeed, because of the central limit theorem, for K large the total inputs to the soma are Gaussian distributed variables with mean and variance that depend on the transfer function implemented by dendrites.

Interestingly, as first observed in Ref. [24], the final expression which we derive in detail in Appendix D, becomes equal to an *effective perceptron*, i.e., the entropy is in form equal to the one presented in Refs. [31,32] for the one-layer neuron model, but where each order parameter is substituted by an integral expression that depends on the activation function g.

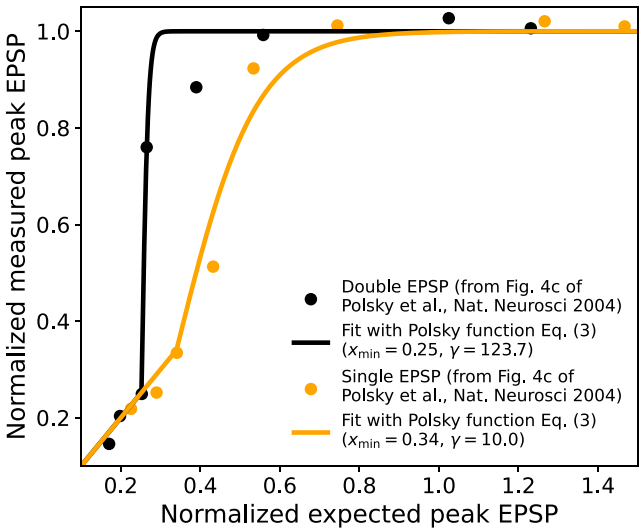


FIG. 9. Fit of experimental data from Fig. 4(c) in the original study by Polsky *et al.* [13].

APPENDIX B: CHOICE OF DENDRITIC NONLINEARITY AND PARAMETER ESTIMATION

1. Fitting the Polsky non-linearity parameters $x_{\min}$ and γ

To fit the Polsky nonlinearity parameters, we use experimental data presented in Fig. 4(c) of Ref. [13]. These data indicates a linear behavior below a threshold $x_{\min}$, followed by a strongly nonlinear behavior, and then a saturation. These observations motivate the choice of the nonlinearity given by Eq. (3). In the model, we use adimensional dendritic inputs and outputs with a saturation at a value of 1, while the data indicates that dendritic potentials saturate at ∼15 mV. We approximate this saturating behavior using a sigmoid function, characterized by a gain parameter γ. The chosen functional form, detailed in Eq. (3), fits the experimental data reasonably well using these two parameters, as shown in Fig. 9. This analysis yields parameter ranges: $x_{\min} \in [0.25, 0.34]$ and $\gamma \in [10, 123.7]$ (in model units). In most of the paper, we use $x_{\min} = 0.33$, $\gamma = 15$. We expect other functional forms for the nonlinearity would lead to qualitatively similar results, provided the main features of the nonlinearity are preserved.

2. Conversion of model parameter to biologically plausible values

In our model, the output of each dendritic branch saturates at 1, while in Polsky *et al.* experiments, the saturation is at ∼15 mV. This means that to convert our parameters to biological values, we need to multiply relevant parameters by 15 mV and to take into account the normalization factors. In the following, we denote parameter values in biophysical units by a subscript "b." In particular, $\bar{W}_b$, the average synaptic weight in millivolts units is given by:

$$\bar{W}_b = \frac{15}{\sqrt{N}}\bar{W}. \tag{B1}$$

ALGORITHM 1. SGD with CE loss and positive weights.

Hyperparameters: learning rate ζ, cross-entropy parameter γ_{ce}
for $t = 1, 2, \ldots$ **do**
  $\xi^\mu, \sigma^\mu \leftarrow$ sample pattern
  $\delta w_{il} \leftarrow \nabla_{w_{il}} \mathscr{L}_{\mathrm{CE}}(w_{il}; \xi^\mu, \gamma_{ce})$
  $w_{il} \leftarrow w_{il} - \zeta \cdot \delta w_{il}$
  if $w_{il} < 0$ **then**
    $w_{il} \leftarrow 0$
  end if
end for

The dendritic threshold θ_{db} in mV units is given by

$$\theta_{db} = \frac{15}{\sqrt{N}}\frac{N}{K}\theta_d = \frac{15\sqrt{N}}{K}\theta_d, \tag{B2}$$

while the somatic threshold is

$$\theta_{sb} = 15K\theta_s. \tag{B3}$$

We assume $N = 10\,000$ (typical cortical neuron synapse count) [34] and $W_b \in [0.3, 1.5]$ mV [55,56], resulting in $W \in [2, 10]$.

We also assume $K \in [20, 100]$ and the dendritic inhibition $\theta_{db} \in [1, 10]$ mV, leading to $\theta_d \in [0.01, 1]$. Additionally, with θ_{sb} as the sum of neuronal threshold (10–20 mV) and somatic inhibition (10–50 mV), we obtain $\theta_s \in [0.01, 0.2]$.

APPENDIX C: LEARNING ALGORITHM

Inspired by machine learning practice, we use a modified version of the SGD algorithm, capable of dealing with strictly positive weights, to train our single neuron models (see Algorithm I). To constrain the synapses to be positive during the learning dynamics, at each gradient step, we reset negative synaptic weights to zero. The SGD algorithm minimizes a differentiable objective function, and a common choice in machine learning is the cross-entropy (CE) loss. For binary outputs, the CE loss is given by $\mathscr{L}_{\mathrm{CE}}(\boldsymbol{W}; \gamma_{ce}, \theta_d, \theta_s) = \sum_{\mu=1}^{P} f_{\gamma_{ce}}(\sigma^\mu \Delta^\mu(\boldsymbol{W}; \theta_d, \theta_s))$ where $f_{\gamma_{ce}}(x) = -\frac{x}{2} + \frac{1}{2\gamma_{ce}}\log(2\cosh(\gamma_{ce}x))$ and the output preactivation is given by $\Delta^\mu(\boldsymbol{W}; \theta_d, \theta_s) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i\xi_i - T\sqrt{N}$ for the linear neuron and (2b) for the nonlinear neuron. The parameter γ_{ce} governs the shape of the CE loss function and, consequently, the training robustness, as discussed in Ref. [27].

Note that unlike in multilayer networks, SGD in a single neuron model leads to a local learning rule, that could be plausibly implemented in a biological neuron, provided an error signal is available. For a single presented pattern μ, a weight w_{il} is changed by an amount proportional to $\delta w_{il} = -\sigma^\mu f'_{\gamma_{ce}}(\sigma^\mu \Delta^\mu) g'(\lambda_l^\mu)\xi_{il}^\mu$. This can be interpreted as a "three factor rule," where $-\sigma^\mu f'_{\gamma_{ce}}$ is a "soft" error signal available to the whole neuron (possibly a "plateau potential" triggered by apical inputs), $g'(\lambda_l^\mu)$ is a local, NMDA-mediated, dendritic signal, and ξ_{il}^μ is the presynaptic activity.

1. Numerical experiments

We provide details and algorithmic considerations used for the numerical experiments reported in the paper for both the linear and nonlinear neuron models.

For the nonlinear neuron model, the hyperparameters are the two thresholds θ_d and θ_s, the learning rate ζ, and the CE robustness parameter γ_{ce} (for the SGD algorithm). For the linear neuron model, there are only two parameters: the learning rate ζ and the robustness parameter γ_{ce} (for SGD). In the linear case, the threshold θ_d governs the synaptic mean value but does not alter the dynamics with a suitable rescaling of the learning rate with θ_d. In the algorithmic capacity simulations, the learning rate is adjusted adaptively: whenever the error stops decreasing, the learning rate is reduced by a factor of two. This process is repeated until the learning rate reaches $\frac{1}{4096 \cdot N}$, which is on the order of 10^{-6}. Otherwise, we perform a simple exponential annealing of the learning rate: at epoch t, the learning rate is $\zeta t = \zeta(1 - d\zeta)^t$ with $d\zeta = 10^{-4}$. Learning rate decay is justified by the fact that in the linear neuron case, adapting [1,67] convergence proof, one can demonstrate that the perceptron algorithm converges below the critical capacity, provided that the variation in weights at each step is smaller than a certain critical value $dw_c > 0$ [31]. During single-neuron training, we present one input pattern at a time (the *minibatch size* is 1). We randomly shuffle the pattern sequence at each dataset presentation (*epoch*). For numerical simulations we use input and output coding levels $f_{\text{in}} = f_{\text{out}} = 0.5$.

Synapses are initialized uniformly at random between zero and twice the theoretical expected value of the mean weight $\bar{w} = \frac{\theta}{f_{\text{in}} N}$. This is the expected value for both the linear and nonlinear neurons with a generic activation function and a generic value of the θ_s threshold in the symmetric $f_{\text{in}} = f_{\text{out}} = 0.5$ case, as analytically shown in Appendix D. If synapses turn negative during training, then a hard boundary condition is enforced, and they are immediately reset to zero. For the nonlinear neuron, the SGD update rule (Algorithm I) with the cross-entropy loss is performed only on the dendritic branches of the first synaptic layer.

For the linear neuron trained with the SGD algorithm, the dynamics is invariant with respect to the rescaling by a positive constant c of the three hyperparameters: ζ, γ_{ce}, and θ_d, i.e., $\zeta \leftarrow \zeta/c$, $\gamma_{ce} \leftarrow c\gamma_{ce}$, $\theta_d \leftarrow \theta_d/c$. As a result, the dynamics effectively depends only on two of these hyperparameters.

Hyperparameter selection

We report the hyperparameter selection used for numerical simulations of the nonlinear neuron (a.k.a. tree committee machine) and the linear one (or perceptron). In the storage case, the input size is $N = 999$, and the number of dendritic branches of the nonlinear neuron is $K = 27$. For numerical simulations on real-world datasets, the input size is twice the number of pixels in each image, i.e., $N = 1568$ for MNIST and FashionMNIST, and $N = 6144$ for CIFAR10. The number of dendritic branches is $K = 7$ for MNIST and FashionMNIST, and $K = 8$ for CIFAR10. The Polsky dendritic nonlinear transfer function has biologically estimated parameters $x_{\min} = 0.33$ and $\gamma = 15$. The nonlinear neuron behavior regarding relevant observables is studied as a function of the dendritic and somatic thresholds θ_d and θ_s.

To optimize neuronal computational performances on relevant observables, we perform a grid search for SGD on the learning rate $\zeta \in \{0.0001, 0.001, 0.01, 0.1, 1.0\}$ and the cross-entropy robustness parameter $\gamma_{ce} \in \{0.001, 0.01, 0.1, 1.0, 10.0, 100.0\}$.

2. Real datasets definition and binarization

Binarization was applied to real-world classification datasets (MNIST, FashionMNIST, and CIFAR10) primarily to align with theoretical calculations presented in this paper, ensuring consistency across analyses. Although not strictly necessary—normalization would have also addressed potential imbalances—binarization helps to maintain uniform input coding levels across classes, avoiding trivial solutions due to correlations between neuronal outputs and overall input intensities, given that both input patterns and synaptic weights are strictly positive. A possible binarization method is to convert pixel intensities into zeros and ones using $\xi_i = \Theta(p_i - T)$, where p_i is the pixel intensity and T is the median intensity. However, in datasets such as MNIST where pixel intensity distribution is already essentially binarized with an imbalance between zeros and ones, this leads to strong heterogeneities in coding levels. To avoid this issue and ensure consistency with the theoretical framework, we double the input dimensionality and use two input neurons per pixel, an "ON" unit with $\xi_i = \Theta(p_i - T)$, and an "OFF" unit with $\xi_i' = \Theta(T - p_i)$, in analogy with biological visual systems where these two types of cells are widespread [68].

APPENDIX D: ANALYTICAL RESULTS

1. Definition of the nonlinear model of the neuron

We recall here the main definitions of the single neuron model studied in the main text of the paper. Given an activity pattern $\boldsymbol{\xi}^\mu \in \{0,1\}^N$, the output of our model of neuron is obtained in two steps. First, the activity pattern is processed by the corresponding dendritic branch; we suppose here that we have $l = 1, \ldots, K$ dendritic branches each having a set of $i = 1, \ldots, N/K$ positive synaptic weights W_{li}. The output activity τ_l^μ of a given branch l that corresponds to the activity pattern $\boldsymbol{\xi}^\mu$ is obtained as

$$\tau_l^\mu = g\left(\sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} W_{li}\xi_{li}^\mu - \sqrt{\frac{N}{K}}\theta_d\right) \equiv g(\lambda^\mu), \tag{D1}$$

where θ_d is a threshold modeling inhibition at the level of the dendritic branch, while $g(\cdot)$ is a generic positive, (possibly) nonlinear function. Second, the output of each branch is combined linearly by using another set of K synaptic weights c_l, $l = 1, \ldots, K$ and the output is obtained as

$$\sigma_{\text{out}}^\mu = \Theta\left[\frac{1}{\sqrt{K}} \sum_{l=1}^{K} c_l\, \tau_l^\mu - \sqrt{K}\theta_s\right], \tag{D2}$$

where $\Theta(x)$ is the Heaviside theta function that is 1 if $x > 0$ and 0 otherwise. The parameter θ_s is a threshold modeling

inhibition coming from inhibitory neurons. In the following we will consider, for simplicity $c_l = 1$ for every $l = 1, \ldots, K$.

2. Training set and partition function

We consider a training set composed of $P = \alpha N$ random independent and identically distributed activity patterns $\boldsymbol{\xi}^\mu \in \{0,1\}^N$ and independent and identically distributed labels $\sigma^\mu \in \{0,1\}$ with $\mu = 1, \ldots, P$. The probability distribution of each component of a pattern is given by

$$P(\xi_{li}^\mu) = f_{\text{in}}\,\delta(\xi_{li}^\mu - 1) + (1 - f_{\text{in}})\,\delta(\xi_{li}^\mu), \tag{D3}$$

where f_{in} is the *input coding level* of the patterns. We consider a probability distribution of labels to be equal in form to (D3) but we allow the possibility to have different coding level in the output f_{out}.

In order to study the volume of synaptic weights that correctly associate to a given pattern of activity $\boldsymbol{\xi}^\mu$ the corresponding label σ^μ we use a standard statistical mechanics approach [3,36]. First, we define the characteristic function

$$X_{\xi,\sigma}(W) = \prod_\mu \Theta\left(\frac{(2\sigma^\mu - 1)}{\sqrt{K}}\left(\sum_{l=1}^K c_l\,\tau_l^\mu - K\theta_s\right) - \kappa\right), \tag{D4}$$

which is 1 when a given weight configuration W_{li} correctly classifies all the patterns (we will call this a *solution*) and 0 otherwise. The volume of the allowed synapses, which in statistical mechanics is known as the *partition function*, is therefore:

$$Z = \int d\mu(W)\, X_{\xi,\sigma}(W), \tag{D5}$$

where $d\mu(W)$ is the measure over the weights. We will consider in the following:

$$\int d\mu(W)\,\bullet \equiv \int_0^\infty \prod_{li} dW_{li}\,\bullet \tag{D6}$$

without giving constraints to the norm of the weights. As we will see the norm will be imposed self-consistently by the learning problem.

3. Replica method

To compute the average entropy $\langle \ln Z\rangle_{\xi,\sigma}$ of synaptic weights solutions in the large-N limit, we resort the replica method [44] that is based on the following identity:

$$\langle \ln Z\rangle_{\xi,\sigma} = \lim_{n\to 0} \frac{\langle Z^n\rangle_{\xi,\sigma} - 1}{n} = \lim_{n\to 0} \frac{1}{n} \ln \langle Z^n\rangle_{\xi,\sigma}.$$

This trick reconducts the problem of estimating the log of the partition function in (D5) to the computation of the average of n independent copies of the systems with the same realization of the disorder of the activity patterns and labels $\boldsymbol{\xi}^\mu, \sigma^\mu$:

$$\langle Z^n\rangle = \left\langle \int \prod_{a=1}^n d\mu(W^a) \prod_{a,\mu} \Theta\left(\frac{\sigma^\mu}{\sqrt{K}}\left(\sum_{l=1}^K c_l\, g\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}^a \xi_{li}^\mu - \sqrt{\frac{N}{K}}\theta_d\right) - K\theta_s\right) - \kappa\right)\right\rangle_{\xi,\sigma}. \tag{D7}$$

We will denote from now on a and b as the index that run over replicas $a, b = 1, \ldots, n$. Notice also that, because of (D4) we can safely consider having labels $\sigma^\mu = \pm 1$ with the same output coding level as before. The computation follows standard steps [3,4], which we will sketch here. First, we need to perform the average over the activity patterns $\boldsymbol{\xi}^\mu$; we can do that by introducing the auxiliary variables

$$\lambda_l^{\mu a} = \sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi_{li}^\mu - \sqrt{\frac{N}{K}}\theta_d \tag{D8}$$

and the corresponding conjugated variables $\hat{\lambda}_l^{\mu a}$ that arise when we insert the integral representation of the Dirac delta function. The replicated partition function is

$$\begin{aligned}\langle Z^n\rangle = \mathbb{E}_\sigma \int \prod_a d\mu(W^a) \int \prod_{\mu a l} \frac{d\lambda_l^{\mu a} d\hat{\lambda}_l^{\mu a}}{2\pi} e^{i\lambda_l^{\mu a}\hat{\lambda}_l^{\mu a}} \prod_{a,\mu} \Theta\left(\frac{\sigma^\mu}{\sqrt{K}}\left(\sum_{l=1}^K c_l\, g(\lambda_l^{\mu a}) - K\theta_s\right) - \kappa\right) \\ \times e^{i\sqrt{\frac{N}{K}}\theta_d \sum_{\mu a l}\hat{\lambda}_l^{\mu a}} \cdot \prod_{li\mu} \left\langle e^{-i\xi_{li}^\mu \sqrt{\frac{K}{N}}\sum_a W_{li}^a \hat{\lambda}_l^{\mu a}}\right\rangle_{\xi_{li}^\mu}.\end{aligned} \tag{D9}$$

In the large-N limit, averaging over patterns amounts to summing a large number of independent contributions. By the central limit theorem, this sum behaves like a Gaussian variable, justifying the truncation of the cumulant expansion at second order as

$$\begin{aligned}\prod_{li\mu} \left\langle e^{-i\xi_{li}^\mu \sqrt{\frac{K}{N}}\sum_a W_{li}^a \hat{\lambda}_l^{\mu a}}\right\rangle_{\xi_{li}^\mu} &= \prod_{li\mu}\left[1 - f_{\text{in}} + f_{\text{in}} e^{-i\sqrt{\frac{K}{N}}\sum_a W_{li}^a \hat{\lambda}_l^{\mu a}}\right] \simeq \\ &= \prod_{li\mu} e^{-if_{\text{in}}\sqrt{\frac{K}{N}}\sum_a W_{li}^a \hat{\lambda}_l^{\mu a} - \frac{f_{\text{in}}(1-f_{\text{in}})K}{2N}(\sum_a W_{li}^a \hat{\lambda}_l^{\mu a})^2} \\ &= e^{-if_{\text{in}}\sqrt{\frac{K}{N}}\sum_{\mu a l}\hat{\lambda}_l^{\mu a}\sum_i W_{li}^a - \frac{f_{\text{in}}(1-f_{\text{in}})K}{N}\sum_{\mu l}\sum_{a<b}(\sum_i W_{li}^a W_{li}^b)\hat{\lambda}_l^{\mu a}\hat{\lambda}_l^{\mu b}} e^{-\frac{f_{\text{in}}(1-f_{\text{in}})K}{2N}\sum_{\mu a l}\sum_i (W_{li}^a \hat{\lambda}_l^{\mu a})^2}.\end{aligned} \tag{D10}$$

By defining appropriate order parameters, it is possible to conveniently study the problem in the large-N limit. We define:

$$M_l^a = \sqrt{\frac{K}{N}} \sum_i W_{li}^a - \sqrt{\frac{N}{K}} \bar{W}, \tag{D11a}$$

$$q_l^{ab} \equiv \frac{K}{N} \sum_i W_{li}^a W_{li}^b, \tag{D11b}$$

$$Q_l^a \equiv \frac{K}{N} \sum_i \left(W_{li}^a\right)^2. \tag{D11c}$$

(A1a) represents the average synaptic weight; we expressed it in two contributions. The first one represents an averaged *scaled* synaptic weight

$$\overline{W} = \frac{\theta_d}{f_{\text{in}}} \tag{D12}$$

justified by the fact that for each subperceptron of the first layer only $\frac{N}{K} f_{\text{in}}$ synapses contribute. The second term instead is a $\sqrt{\frac{N}{K}}$ correction that is needed in order to fine-tune the average synaptic weight relatively to the threshold θ_d.

The quantity q_l^{ab} is the overlap between the weights of two different replicas a and b belonging to the same dendritic branch l, while Q_l^a represents the averaged squared norm of a synaptic weight belonging to dendritic branch l.

Enforcing the definitions (D11) in (D9), by using Dirac delta functions, we can express the replicated partition function as an integration over the order parameters q_l^{ab}, $\hat{q}_l^{ab}$, Q_l^a, $\hat{Q}_l^a$, M_l^a, and $\hat{M}_l^a$ after performing the rotation of each conjugated variable $\{\hat{q}_l^{ab}, \hat{Q}_l^a, \hat{M}_l^a\} \to i\{\hat{q}_l^{ab}, \hat{Q}_l^a, \hat{M}_l^a\}$:

$$\begin{aligned}\langle Z^n \rangle_{\xi,\sigma} = &\int \prod_{a<b,l} \frac{dq_l^{ab} d\hat{q}_l^{ab}}{2\pi K/N} \int \prod_{a,l} \frac{dQ_l^a d\hat{Q}_l^a}{2\pi K/N} \int \prod_{a,l} \frac{dM_l^a d\hat{M}_l^a}{2\pi\sqrt{K/N}} \, e^{-\frac{N}{K}\sum_{a<b,l} q_l^{ab}\hat{q}_l^{ab} - \frac{N}{K}\sum_{a,l} Q_l^a \hat{Q}_l^a - \frac{N}{K}\overline{W}\sum_{a,l}\hat{M}_l^a} \\ &\times e^{\frac{N}{K} G_S(\hat{q}_l^{ab}, \hat{Q}_l^a, \hat{M}_l^a) + N\alpha G_E(q_l^{ab}, Q_l^a, M_l^a)},\end{aligned} \tag{D13}$$

where we collected the entropic contribution G_S and the energetic one G_E. The first is the usual term that counts how many coupling vectors W^a fulfill the constraints (D11); the second is specific to the learning rule which is used, and depends on the Heaviside function that counts learned patterns:

$$G_S\left(\hat{q}_l^{ab}, \hat{Q}_l^a, \hat{M}_l^a\right) = \ln \int_0^\infty \prod_{a,l} dW_l^a \, e^{\sum_{a<b,l} \hat{q}_l^{ab} W_l^a W_l^b + \sum_{a,l} \hat{Q}_l^a (W_l^a)^2 + \sum_{a,l} \hat{M}_l^a W_l^a}, \tag{D14a}$$

$$\begin{aligned}G_E\left(q_l^{ab}, Q_l^a, M_l^a\right) = &\ln \mathbb{E}_\sigma \int \prod_{a,l} \frac{d\lambda_l^a d\hat{\lambda}_l^a}{2\pi} \Theta\left(\frac{\sigma}{\sqrt{K}}\left(\sum_l c_l \, g\left(\lambda_l^a\right) - K\theta_s\right) - \kappa\right) e^{i\sum_{a,l} \lambda_l^a \hat{\lambda}_l^a - f_{\text{in}}(1-f_{\text{in}})\sum_{a<b,l} q_l^{ab} \hat{\lambda}_l^a \hat{\lambda}_l^b} \\ &\times e^{-\frac{f_{\text{in}}(1-f_{\text{in}})}{2}\sum_{a,l} Q_l^a (\hat{\lambda}_l^a)^2 - i f_{\text{in}} \sum_{a,l} M_l^a \hat{\lambda}_l^a}.\end{aligned} \tag{D14b}$$

Note also that in writing (D13) we have discarded a term of the type $e^{-\sqrt{\frac{N}{K}}\sum_{la} M_l^a \hat{M}_l^a}$ coming from the integral representation of the delta function as it is of order $\sqrt{N}$ and so it is subleading with respect to N. We can now evaluate in the large-N limit (D13) using the saddle-point method [45]. In order restrict the space where to search saddle points we proceed by assuming a particular form of the order parameters, which is the main topic of the next section.

4. Replica symmetric analysis

We use a RS ansatz, i.e., we assume that the order parameters do not depend on the replica indexes and of the index corresponding to the dendritic branch:

$$q_l^{ab} = q, \quad Q_l^a = Q, \quad M_l^a = M, \tag{D15a}$$

$$\hat{q}_l^{ab} = \hat{q}, \quad \hat{Q}_l^a = \hat{Q}, \quad \hat{M}_l^a = \hat{M}. \tag{D15b}$$

In the RS ansatz and in the small-n limit, using *Hubbard-Stratonovich* transformations:

$$e^{\frac{1}{2}bx^2} = \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2} + \sqrt{b}xz},$$

the entropic and energetic terms are the following:

$$\mathcal{G}_S \equiv \lim_{n\to 0} \frac{1}{nK} G_S(\hat{q},\hat{Q},\hat{M}) = \ln\sqrt{\frac{2\pi}{\hat{q}-2\hat{Q}}} + \frac{1}{2}\left(\frac{\hat{M}^2+\hat{q}}{\hat{q}-2\hat{Q}}\right) + \int Dz\, \ln H\left(-\frac{\hat{M}+\sqrt{\hat{q}}z}{\sqrt{\hat{q}-2\hat{Q}}}\right) \tag{D16a}$$

$$\mathcal{G}_E \equiv \lim_{n\to 0} \frac{G_E(q,Q,M)}{n} = \mathop{\mathbb{E}}_{\sigma} \int \prod_l Dt_l\, \ln\left[\int \prod_l D\lambda_l\, \Theta\left(\frac{\sigma}{\sqrt{K}}\left(\sum_l c_l\, g(\sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\lambda_l + a_l) - K\theta_s\right) - \kappa\right)\right], \tag{D16b}$$

where we have introduced the variable

$$a_l \equiv f_{\rm in} M + \sqrt{f_{\rm in}(1-f_{\rm in})q}\, t_l \tag{D17}$$

for convenience. In (D16a) we have also introduced $H(x) \equiv \int_x^\infty Dz = \frac{1}{2}\mathrm{Erfc}(\frac{x}{\sqrt{2}})$ and $Dz \equiv G(z)\,dz$ with $G(z)$ being a standard normal Gaussian $G(z) = \exp(-z^2/2)/\sqrt{2\pi}$.

a. Large-K limit

We focus on the limit $K \to \infty$ (with $\frac{K}{N} \to 0$), for two main reasons:

(i) on the analytical level, it allows to simplify the numerical evaluation of the saddle-point equations corresponding to the RS ansatz;

(ii) it is *biologically* realistic: the number of dendritic branches in neurons is typically large (in some cases even more than a hundred, REF) and the number of synapses in each branch is typically large as well.

To evaluate this limit, we need to do some manipulations on the energetic contribution (D16b) which are based on the central limit theorem. Let us first consider the term in square brackets:

$$\begin{aligned} I &= \int \prod_l D\lambda_l\, \Theta\left(\frac{\sigma}{\sqrt{K}}\left(\sum_l c_l\, g(\sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\lambda_l + a_l\right) - K\theta_s) - \kappa\right) \\ &= \int \frac{dh\, d\hat{h}}{2\pi} e^{-ih\hat{h}}\, \Theta[\sigma(h-\sqrt{K}\theta_s) - \kappa] \int \prod_l D\lambda_l\, e^{\frac{i\hat{h}}{\sqrt{K}}\sum_l c_l\, g(a_l + \sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\,\lambda_l)}. \end{aligned} \tag{D18}$$

In the large-K limit we can therefore expand the exponential up to second order

$$\begin{aligned} I \simeq \int \frac{dh\, d\hat{h}}{2\pi} e^{-ih\hat{h}}\, \Theta(\sigma(h-\sqrt{K}\theta_s) - \kappa) \int \prod_l D\lambda_l \left[1 + \frac{i\hat{h}}{\sqrt{K}}\sum_l c_l\, g(a_l + \sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\,\lambda_l)\right. \\ \left. + -\frac{\hat{h}^2}{2K}\left(\sum_l c_l\, g(a_l + \sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\,\lambda_l)\right)^2\right]. \end{aligned} \tag{D19}$$

and we can integrate with respect to all the λ_l variables term by term. Exponentiating the expression again and integrating over $\hat{h}$ we get

$$I = \int Dh\, \Theta(\sigma(M^{(0)} + \sqrt{\Delta^{(0)}}h - \sqrt{K}\theta_s) - \kappa), \tag{D20}$$

where we have introduced the variables:

$$M^{(0)} = \frac{1}{\sqrt{K}}\sum_l c_l \langle g\rangle_\lambda, \tag{D21a}$$

$$D^{(0)} = \frac{1}{K}\sum_l c_l^2\left[\langle g^2\rangle_\lambda - \langle g\rangle_\lambda^2\right], \tag{D21b}$$

and the notation

$$\langle g\rangle_\lambda = \int D\lambda\, g(f_{\rm in}M + \sqrt{f_{\rm in}(1-f_{\rm in})q}\, t_l + \sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\,\lambda). \tag{D22}$$

Using again the central limit theorem for the K integrals over the variable t_l we get:

$$\mathcal{G}_E = \mathop{\mathbb{E}}_{\sigma} \int Dt\, \ln \int D\lambda\, \Theta[\sigma(M_0 + \sqrt{D_0}\,t + \sqrt{D_1}\,\lambda - \sqrt{K}\theta_s) - \kappa], \tag{D23}$$

where

$$M_0 = m_c \langle \langle g \rangle_\lambda \rangle_t, \tag{D24a}$$

$$D_0 = \sigma_c \left[\langle \langle g \rangle_\lambda^2 \rangle_t - \langle \langle g \rangle_\lambda \rangle_t^2 \right], \tag{D24b}$$

$$D_1 = \sigma_c \left[\langle \langle g^2 \rangle_\lambda \rangle_t - \left\langle \langle g \rangle_\lambda^2 \right\rangle_t \right], \tag{D24c}$$

$$m_c \equiv \frac{1}{\sqrt{K}} \sum_l c_l, \tag{D24d}$$

$$\sigma_c \equiv \frac{1}{K} \sum_l c_l^2, \tag{D24e}$$

and for example

$$\langle \langle g \rangle_\lambda \rangle_t = \int Dt D\lambda \, g(f_{\text{in}} M + \sqrt{f_{\text{in}}(1-f_{\text{in}})q}\, t + \sqrt{f_{\text{in}}(1-f_{\text{in}})(Q-q)}\, \lambda). \tag{D25}$$

Note that in the large-K limit having used the central limit theorem, the resulting expression depend only on the *scaled*[1] mean m_c and the variance σ_c of the second layer (frozen) weights c_l. Specializing to the case $c_l = 1$, we have that the m_c scales as $m_c = \sqrt{K}$ and $\sigma_c = 1$. In order to have a well-defined large-K limit we have to impose (analogously to what we have done on the dendritic threshold θ_d) that the divergence induced by the somatic threshold θ_s cancels with the one coming from M_0. We therefore impose that M scales, in the large-K limit, as

$$M = \overline{M} + \frac{\delta M}{\sqrt{K}}. \tag{D26}$$

M_0 can be simplified by making a rotation over the integration measures λ and t

$$\begin{aligned} M_0 = \sqrt{K} \langle \langle g \rangle_\lambda \rangle_t &= \sqrt{K} \int D\lambda \, Dt \; g(\sqrt{f_{\text{in}}(1-f_{\text{in}})q}\, t + \sqrt{f_{\text{in}}(1-f_{\text{in}})(Q-q)}\lambda + fM) \\ &= \sqrt{K} \int D\lambda \, g(\sqrt{f_{\text{in}}(1-f_{\text{in}})Q}\, \lambda + f_{\text{in}} M). \end{aligned} \tag{D27}$$

We can now insert the scaling in (D26)

$$\begin{aligned} M_0 &= \sqrt{K} \int D\lambda \, g(\sqrt{f_{\text{in}}(1-f_{\text{in}})Q}\, \lambda + f_{\text{in}} M) \\ &\simeq \sqrt{K} \int D\lambda \, g(\sqrt{f_{\text{in}}(1-f_{\text{in}})Q}\lambda + f_{\text{in}} \overline{M}) + f_{\text{in}}\, \delta M \int D\lambda \, g'(\sqrt{f_{\text{in}}(1-f_{\text{in}})Q}\, \lambda + f_{\text{in}} \overline{M}) \\ &= \sqrt{K}\, \theta_s + \Delta. \end{aligned} \tag{D28}$$

Therefore, $\overline{M}$ is fixed by the relation:

$$\theta_s = \int Dy \, g(\sqrt{f_{\text{in}}(1-f_{\text{in}})Q}\, y + f_{\text{in}} \overline{M}), \tag{D29}$$

that involves the output threshold. Hence the energetic term (D23) becomes

$$\mathcal{G}_E = \mathop{\mathbb{E}}_{\sigma} \int Dz \, \ln H\left(\frac{\kappa - \sigma\, \Delta + \sqrt{D_0}\, z}{\sqrt{D_1}} \right) = \int Dz \left[f_{\text{out}} \ln H\left(\frac{\kappa - \Delta + \sqrt{D_0}\, z}{\sqrt{D_1}} \right) + (1 - f_{\text{out}}) \ln H\left(\frac{\kappa + \Delta + \sqrt{D_0}\, z}{\sqrt{D_1}} \right) \right], \tag{D30}$$

where the order parameters, strictly dependent on the choice of the activation function $g(x)$, are simplified as:

$$\Delta \equiv f_{\text{in}} M \int Dx \, g'(\sqrt{f_{\text{in}}(1-f_{\text{in}})Q}\, x + f_{\text{in}} \overline{M}) \tag{D31a}$$

$$D_0 = \Delta_q - \Delta_0 \tag{D31b}$$

$$D_1 = \Delta_Q - \Delta_q, \tag{D31c}$$

[1]Notice the scaling $\frac{1}{\sqrt{K}}$ in the definition of m_c

where we have renamed δM by M with a slight abuse of notation. We have also defined the generic "effective order parameter" or *kernel* functions

$$\Delta_q = \int Dx \left[\int Dy\, g(\sqrt{f_{\rm in}(1-f_{\rm in})q}\, x + \sqrt{f_{\rm in}(1-f_{\rm in})(Q-q)}\, y + f_{\rm in}\overline{M}) \right]^2, \tag{D32}$$

where Δ_Q and Δ_0 are obtained by simply substituting in the previous expression $q \to Q$ and $q \to 0$, respectively. We report them here for clarity

$$\Delta_Q = \int Dx\, g^2(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, x + f_{\rm in}\overline{M}), \tag{D33a}$$

$$\Delta_0 = \left[\int Dy\, g(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, y + f_{\rm in}\overline{M}) \right]^2. \tag{D33b}$$

We have called Δ_q an *effective order parameter* since (D30) (and therefore the whole quenched entropy) is perfectly equivalent to the one found in the perceptron model studied by Brunel in a series of papers [31,32], but where order parameters q and Q are substituted respectively by $\Delta_q - \Delta_0$ and $\Delta_Q - \Delta_0$.

b. Free entropy and saddle-point equations

The average *free entropy* of the dendritic model of a neuron is therefore

$$\phi = \lim_{N\to\infty} \frac{1}{N} \langle \ln Z \rangle_{\xi,\sigma} = \frac{q\hat{q}}{2} - Q\hat{Q} - \overline{W}\hat{M} + \mathcal{G}_S + \alpha\, \mathcal{G}_E. \tag{D34}$$

We now need to compute the saddle-point equations by differentiating (D34) with respect to the order parameters Q, q, M, $\hat{Q}$, $\hat{q}$, and $\hat{M}$. The saddle-point equations involving the entropic term (i.e., taking derivatives with respect to $\hat{M}$, $\hat{Q}$, and $\hat{q}$) are the same as in the case of the perceptron

$$\overline{W} = \int Dz\, \frac{\int_0^\infty dW\, W e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}{\int_0^\infty dW\, e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}, \tag{D35a}$$

$$Q = \int Dz\, \frac{\int_0^\infty dW\, W^2 e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}{\int_0^\infty dW\, e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}, \tag{D35b}$$

$$q = \int Dz\, \frac{\int_0^\infty dW \left(W^2 - \frac{zW}{\sqrt{\hat{q}}}\right) e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}{\int_0^\infty dW\, e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}$$

$$= \int Dz \left[\frac{\int_0^\infty dW\, W e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}}{\int_0^\infty dW\, e^{-(\hat{q}-2\hat{Q})\frac{W^2}{2} + (\hat{M}+\sqrt{\hat{q}}z)W}} \right]^2. \tag{D35c}$$

In order to express the remaining saddle-point equations in a compact way, we define the quantities:

$$a_\sigma(z) = \frac{\sqrt{D_0}z - \sigma\Delta + \kappa}{\sqrt{D_1}} = \sqrt{\frac{D_0}{D_1}}(z - \tau_\sigma) \tag{D36a}$$

$$\tau_\sigma = \frac{\sigma\Delta - \kappa}{\sqrt{D_0}}. \tag{D36b}$$

Deriving (D34) with respect to M, Q, and q lead respectively to

$$0 = \mathbb{E}_\sigma \sigma \int Dz\, \frac{G(a_\sigma(z))}{H(a_\sigma(z))}, \tag{D37a}$$

$$\hat{Q} = \frac{\alpha}{2}\mathbb{E}_\sigma \int Dz\, \frac{G(a_\sigma(z))}{H(a_\sigma(z))} \left[\frac{a_\sigma(z)}{D_1}\frac{dD_1}{dQ} - \frac{z}{\sqrt{D_0 D_1}}\frac{dD_0}{dQ} \right], \tag{D37b}$$

$$\hat{q} = \alpha\mathbb{E}_\sigma \int Dz\, \frac{G(a_\sigma(z))}{H(a_\sigma(z))} \left[-\frac{a_\sigma(z)}{D_1}\frac{dD_1}{dq} + \frac{z}{\sqrt{D_0 D_1}}\frac{dD_0}{dq} \right], \tag{D37c}$$

where we used the saddle-point equation (D37a) when performing the derivative in Q.[2] The six saddle points equation [(D35a), (D35b), (D35c), (D37a), (D37b), and (D37c)], obtained in the large-$K, N \to \infty$ limit with $N \gg K$, have to be numerically solved to obtain the values of the order parameters and represent the final result of our RS analysis.

Notice that imposing $g(x) = x$ we recover the previous saddle-point expressions obtained for the simple linear neuron model [31,32]. Notice also that in the case $f_{\rm out} = \frac{1}{2}$ saddle-point equation (D37a) gives $M = 0$; in this case therefore $\tau_\sigma = 0$.

c. Effective order parameters for some nonlinearities

We report here the analytical expressions of the effective order parameters for several nonlinearities of interest

(i) *Recovering the one-layer neuron model*: If we impose $g(x) = x$, then we recover the one-layer neuron model. We report here the expressions of the corresponding effective order parameters for convenience

$$\Delta = f_{\rm in} M, \tag{D38a}$$

$$\Delta_q = f_{\rm in}(1-f_{\rm in})q + f_{\rm in}^2\overline{M}^2, \tag{D38b}$$

and, in particular,

$$\Delta_Q = f_{\rm in}(1-f_{\rm in})Q + f_{\rm in}^2\overline{M}^2, \tag{D39a}$$

$$\Delta_0 = f_{\rm in}^2\overline{M}^2, \tag{D39b}$$

As a result $\overline{M}$ does not appear anywhere in the energetic term of equation (D30), and $\theta_s = f_{\rm in}\overline{M}$ is also irrelevant.

(i) *Theta nonlinearity*: $g(x) = \Theta(x)$.

[2] This is why no derivative with respect to Q of Δ compares in (D37b).

To evaluate the integrals it is useful to use the following identity:

$$\int Dz\, H^2(a+bz) = H\left(\frac{a}{\sqrt{1+b^2}}\right) - 2T\left(\frac{a}{\sqrt{1+b^2}}, \frac{1}{\sqrt{1+2b^2}}\right), \quad \text{(D40)}$$

where T is the Owen's T function defined as

$$T(h,s) \equiv \frac{1}{2\pi}\int_0^s dx\, \frac{e^{-(1+x^2)\frac{h^2}{2}}}{1+x^2}, \quad \text{(D41)}$$

and has the following important properties:

$$T(h,s) \simeq \frac{G(h)s}{\sqrt{2\pi}} + O(s^2) \quad \text{for } s \to 0, \quad \text{(D42a)}$$

$$T(h,1) = \frac{1}{2}H(h)H(-h), \quad \text{(D42b)}$$

where we remind that $G(x)$ is the Gaussian with mean zero and unit variance. Defining the quantity

$$M_\star = \frac{f_{\rm in}\overline{M}}{\sqrt{f_{\rm in}(1-f_{\rm in})Q}} \quad \text{(D43)}$$

the effective order parameters for the theta nonlinearity are

$$\Delta = M_\star G(-M_\star), \quad \text{(D44a)}$$

$$\Delta_q = H(-M_\star) - 2T\left(M_\star, \sqrt{\frac{Q-q}{Q+q}}\right), \quad \text{(D44b)}$$

and, in particular,

$$\Delta_Q = H(-M_\star), \quad \text{(D45a)}$$

$$\Delta_0 = H(-M_\star)^2. \quad \text{(D45b)}$$

$\overline{M}$ is fixed by the relation

$$\theta_s = H(-M_\star). \quad \text{(D46)}$$

(ii) *ReLU nonlinearity*: $g(x) = x\Theta(x)$: We have

$$\Delta = f_{\rm in}\overline{M}H(-M_\star), \quad \text{(D47a)}$$

$$\Delta_q = f_{\rm in}(1-f_{\rm in})(Q-q)\sqrt{\frac{Q-q}{Q+q}}G^2\left(\frac{f_{\rm in}\overline{M}}{\sqrt{f_{\rm in}(1-f_{\rm in})(Q+q)}}\right) + f_{\rm in}((1-f_{\rm in})q + f_{\rm in}\overline{M}^2)\left[H(-M_\star) - 2T\left(-M_\star, \sqrt{\frac{Q-q}{Q+q}}\right)\right] + 2f_{\rm in}\overline{M}\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, G(M_\star)H\left(-M_\star\sqrt{\frac{Q-q}{Q+q}}\right) + 2f_{\rm in}(1-f_{\rm in})q\sqrt{\frac{Q-q}{Q+q}}G(M_\star)G\left(M_\star\sqrt{\frac{Q-q}{Q+q}}\right), \quad \text{(D47b)}$$

and in particular

$$\Delta_Q = f_{\rm in}((1-f_{\rm in})Q + f_{\rm in}\overline{M}^2)H(-M_\star) + f_{\rm in}\overline{M}\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, G(-M_\star), \quad \text{(D48a)}$$

$$\Delta_0 = [\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, G(M_\star) + f_{\rm in}\overline{M}H(-M_\star)]^2. \quad \text{(D48b)}$$

Note that $\overline{M}$ is fixed by the relation

$$\theta_s = \sqrt{f_{\rm in}(1-f_{\rm in})Q}\, G(M_\star) + f_{\rm in}\overline{M}H(-M_\star). \quad \text{(D49)}$$

d. Small-data regime

In the $\alpha \to 0$ limit the saddle-point equations can be solved exactly. Indeed $\hat{q} = \hat{Q} = 0$ and Eqs. (D35a), (D35b), and (D35c) give respectively

$$\overline{W} = \frac{\theta_d}{f_{\rm in}} = \frac{\int_0^\infty dW\, We^{\hat{M}W}}{\int_0^\infty dW\, e^{\hat{M}W}} = -\frac{1}{\hat{M}}, \quad \text{(D50a)}$$

$$Q = \frac{\int_0^\infty dW\, W^2 e^{\hat{M}W}}{\int_0^\infty dW\, e^{\hat{M}W}} = \frac{2}{\hat{M}^2}, \quad \text{(D50b)}$$

$$q = \left[\frac{\int_0^\infty dW\, We^{\hat{M}W}}{\int_0^\infty dW\, e^{\hat{M}W}}\right]^2 = \frac{1}{\hat{M}^2}. \quad \text{(D50c)}$$

e. Distribution of dendritic preactivations

We derive here the distribution of dendritic preactivations after learning. Since each dendritic branch has access to an independent portion of the input, the distribution of the preactivations if factorized over the K dendritic branches. In the large-K limit the distribution tends to a Gaussian $\mathcal{N}(\mu,\sigma)$ as can be inspected from the postactivation mean (D29) and from the argument of the kernel functions (D32). Denoting by $\lambda_l \equiv \sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi_{li} - \sqrt{\frac{N}{K}}\theta_d$ as in the main text, the mean and the variance of the distribution of the lth dendritic branch are respectively

$$\mu = \mathbb{E}_\xi \lambda_l = f_{\rm in}\overline{M}, \quad \text{(D51a)}$$

$$\sigma^2 = \mathbb{E}_\xi \lambda_l^2 - \mu^2 = f_{\rm in}(1-f_{\rm in})Q, \quad \text{(D51b)}$$

as confirmed by (D29).

In the $\alpha \to 0$ the variance above can be expressed explicitly in terms of θ_d and $f_{\rm in}$ thanks to (D50a) and (D50b). We get that the standard deviation of the dendritic preactivation depends linearly on the dendritic inhibition threshold

$$\sigma = \theta_d\sqrt{\frac{2(1-f_{\rm in})}{f_{\rm in}}}. \quad \text{(D52)}$$

We checked that this linear relation is satisfied even at finite α, or considering a different distribution over the weights at initialization, see below. This shows that if θ_d is small, then one does not completely use the nonlinearity and the model behaves like a one-layer model.

f. Limit of large somatic thresholds

When the somatic threshold θ_s diverges, the only way to satisfy (D29) is that $\overline{M}$ diverges if the nonlinearity g is unbounded. This can be inspected already in (D49) in the case of the ReLU nonlinearity.

Instead, if the nonlinearity is bounded, then the right-hand side of (D29) is a bounded function of $\overline{M}$ as well, therefore above a certain critical value of θ_s it will not possible to find the corresponding value of $\overline{M}$.

Moreover, if the nonlinearity diverges linearly for large arguments, then we expect to recover back the free energy and the expressions for the one-layer neuron model for large θ_s. Indeed expanding equation (D29) one finds $\overline{M} \sim \theta_s / f_{\rm in}$ and the effective order parameters reduce to those ones of the perceptron, see (D38).

5. Critical capacity

In this section we show how in our formalism, it is possible to compute the maximal number of inputs that the neuron is able to classify. We underline that this can be done for a generic form of dendritic nonlinearity.

In the critical capacity limit, the set of possible synaptic weights shrinks towards a single point and q tends to Q:

$$q = Q - dq. \tag{D53}$$

Correspondingly, the other order parameters scale as

$$\hat{q}, \hat{Q} \sim \frac{C}{dq^2}, \tag{D54a}$$

$$\hat{q} - 2\hat{Q} \sim \frac{A}{dq}, \tag{D54b}$$

$$\hat{M} \sim -\frac{B\sqrt{C}}{dq}. \tag{D54c}$$

Using the identities (where a is a positive constant)

$$\frac{\int_0^\infty dx\, x\, e^{-a\frac{x^2}{2}+bx}}{\int_0^\infty dx\, e^{-a\frac{x^2}{2}+bx}} = \frac{b}{a} + \frac{1}{\sqrt{a}} \frac{G\left(-\frac{b}{\sqrt{a}}\right)}{H\left(-\frac{b}{\sqrt{a}}\right)}, \tag{D55a}$$

$$\frac{\int_0^\infty dx\, x^2\, e^{-a\frac{x^2}{2}+bx}}{\int_0^\infty dx\, e^{-a\frac{x^2}{2}+bx}} = \frac{1}{a} + \frac{b^2}{a^2} + \frac{b}{a^{3/2}} \frac{G\left(-\frac{b}{\sqrt{a}}\right)}{H\left(-\frac{b}{\sqrt{a}}\right)}, \tag{D55b}$$

and the expansion

$$\frac{G(x)}{H(x)} \simeq x\theta(x), \quad \text{for } |x| \gg 1, \tag{D56}$$

the saddle-point equations (D35) can be written as

$$\overline{W} = \frac{\sqrt{C}}{A}[G(B) - BH(B)], \tag{D57a}$$

$$Q = \frac{C}{A^2}[(1+B^2)H(B) - BG(B)], \tag{D57b}$$

$$A = H(B). \tag{D57c}$$

More work is required to derive the asymptotic limit of equations (D37). First, we need the expansion of the effective order parameters

$$D_0 = \Delta_q - \Delta_0 = \Delta_Q - \Delta_0 - D_1 \simeq \Gamma_0 - \Gamma_1 dq, \tag{D58a}$$

$$D_1 = \Delta_Q - \Delta_q = \Gamma_1 dq + O(dq^2), \tag{D58b}$$

where we have defined

$$\Gamma_0 = \Delta_Q - \Delta_0 = \int Dx\, g^2(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, x + f_{\rm in}\overline{M}) - \left[\int Dy\, g(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, y + f_{\rm in}\overline{M})\right]^2, \tag{D59a}$$

$$\Gamma_1 = f_{\rm in}(1-f_{\rm in}) \int Dz\, [g'(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, z + f_{\rm in}\overline{M})]^2. \tag{D59b}$$

Now we subtract (D37b) with (D37c) getting

$$\begin{aligned}\hat{q} - 2\hat{Q} &= \alpha \mathbb{E}_\sigma \int Dz\, \frac{G(a_\sigma(z))}{H(a_\sigma(z))} \left[\frac{z}{\sqrt{D_0 D_1}}\left(\frac{dD_0}{dq} + \frac{dD_0}{dQ}\right) - \frac{a_\sigma(z)}{D_1}\left(\frac{dD_1}{dq} + \frac{dD_1}{dQ}\right)\right] \\ &= \alpha \mathbb{E}_\sigma \int Dz\, \frac{G(a_\sigma(z))}{H(a_\sigma(z))} \times \left[\frac{z}{\sqrt{D_0 D_1}} \frac{d\Gamma_0}{dQ} - \frac{a_\sigma(z)}{D_1} \frac{d\Gamma_1}{dQ} dq\right].\end{aligned} \tag{D60}$$

Similarly (D37c) becomes

$$\hat{q} = \alpha \mathbb{E}_\sigma \int Dz\, \frac{G(a_\sigma(z))}{H(a_\sigma(z))} \left[-\frac{a_\sigma(z)}{D_1} \frac{dD_1}{dq}\right] \tag{D61}$$

because the second term in (D37c) is subleading in dq. Using the expansion (D56) and the identities

$$\int Dz\, z^2\, \Theta(z - \tau_\sigma) = H(\tau_\sigma) + \tau_\sigma G(\tau_\sigma), \tag{D62a}$$

$$\int Dz\, z\, \Theta(z - \tau_\sigma) = G(\tau_\sigma), \tag{D62b}$$

we obtain the following saddle-point equations:

$$0 = \mathbb{E}_\sigma \sigma [G(\tau_\sigma) - \tau_\sigma H(\tau_\sigma)], \tag{D63a}$$

$$A = \alpha_c \left[\frac{1}{\Gamma_1} \frac{d\Gamma_0}{dQ} \mathbb{E}_\sigma H(\tau_\sigma) - \frac{\Gamma_0}{\Gamma_1^2} \frac{d\Gamma_1}{dQ} \mathbb{E}_\sigma \times \left[\left(1+\tau_\sigma^2\right) H(\tau_\sigma) - \tau_\sigma G(\tau_\sigma)\right]\right], \tag{D63b}$$

$$C = \frac{\alpha_c \Gamma_0}{\Gamma_1} \mathbb{E}_\sigma \left[\left(1+\tau_\sigma^2\right) H(\tau_\sigma) - \tau_\sigma G(\tau_\sigma)\right], \tag{D63c}$$

which involve the critical capacity as an unknown parameter to find. Notice that in the previous equation we have redefined $\tau_\sigma = (\sigma\Delta - \kappa)/\sqrt{\Gamma_0}$. The full set of saddle-point equations for the order parameters A, B, C, M, Q, and $\overline{M}$ and for α_c are

$$\overline{W} = \frac{\sqrt{C}}{A}[G(B) - BH(B)], \tag{D64a}$$

$$Q = \frac{C}{A^2}[(1+B^2)H(B) - BG(B)], \tag{D64b}$$

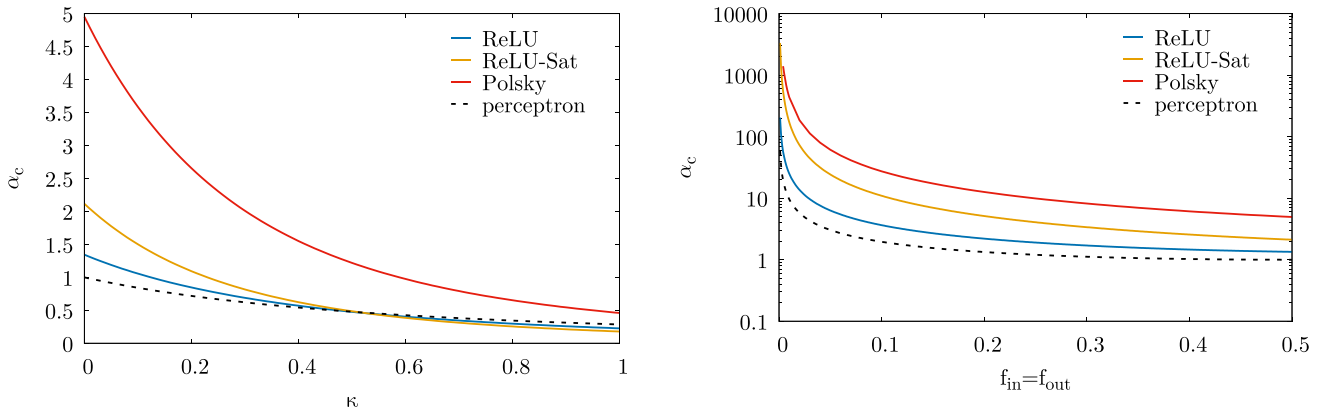


FIG. 10. Critical capacity α_c as a function of the margin κ (left panel) and input/output coding level (right) for different activation functions. As expected, increasing the margin has the effect of decreasing the critical capacity, as it restricts the system to a smaller subset of configurations capable of meeting the stricter classification criteria.

$$A = H(B), \tag{D64c}$$

$$0 = \mathbb{E}_\sigma \sigma[G(\tau_\sigma) - \tau_\sigma H(\tau_\sigma)], \tag{D64d}$$

$$A = \frac{\alpha_c}{\Gamma_1}\frac{d\Gamma_0}{dQ}\mathbb{E}_\sigma H(\tau_\sigma) - \frac{1}{\Gamma_1}\frac{d\Gamma_1}{dQ}C, \tag{D64e}$$

$$C = \frac{\alpha_c \Gamma_0}{\Gamma_1}\mathbb{E}_\sigma\left[(1+\tau_\sigma^2)H(\tau_\sigma) - \tau_\sigma G(\tau_\sigma)\right], \tag{D64f}$$

$$\theta_s = \int Dy\, g(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, y + f_{\rm in}\overline{M}). \tag{D64g}$$

As we have anticipated before, in the case $f_{\rm out} = 0.5$ the equations can be further simplified, since $M = 0$; if also $\kappa = 0$ therefore $\tau_\sigma = 0$; the saddle-point equations (D64) then reduce to

$$\overline{W} = \sqrt{\frac{\alpha_c \Gamma_0}{2\Gamma_1}}\frac{1}{H(B)}[G(B) - BH(B)], \tag{D65a}$$

$$Q = \frac{\alpha_c \Gamma_0}{2\Gamma_1}\frac{1}{H^2(B)}[(1+B^2)H(B) - BG(B)], \tag{D65b}$$

$$\alpha_c = \frac{2\Gamma_1 H(-B)}{\frac{d\Gamma_0}{dQ} - \frac{\Gamma_0}{\Gamma_1}\frac{d\Gamma_1}{dQ}}, \tag{D65c}$$

$$\theta_s = \int Dy\, g(\sqrt{f_{\rm in}(1-f_{\rm in})Q}\, y + f_{\rm in}\overline{M}). \tag{D65d}$$

The critical capacity α_c as a function of the margin κ and input/output coding level for different activation functions is reported in Fig. 10.

a. Limit of large dendritic threshold

As mentioned in the main text, the critical capacity of the model depends strongly on the shape of the nonlinearity. In particular, in the limit of large dendritic threshold the critical capacity can diverge differently depending if the nonlinearity saturates or not for a sufficiently large stimulus. We show in Fig. 11 how the critical capacity diverges logarithmically in θ_d for the ReLU activation function whereas the divergence is linear in the ReLU-Sat activation function. Performing a fit we get when $\theta_d \to \infty$

$$\alpha_c^{\text{ReLU-Sat}} \simeq 3.518\,\theta_d, \tag{D66a}$$

$$\alpha_c^{\text{ReLU}} \simeq 0.9602\,\ln\theta_d. \tag{D66b}$$

b. Limit of small dendritic threshold

As shown in the main text numerically and above for $\alpha < \alpha_c$, in the low dendritic threshold regime for fixed θ_s the model with Polsky and ReLU activation function behaves like a one-layer model. We give another quantitative argument here for $\alpha = \alpha_c$. Since $\theta_d \to 0$ the right-hand side of the first of (D64) should go to zero. Since the function $G(B) - BH(B)$ does not go to zero for finite values of B, by necessity $C \to 0$. By the last of (D64), this requires $\Gamma_0 \to 0$, i.e., $Q \to 0$. In this limit we get

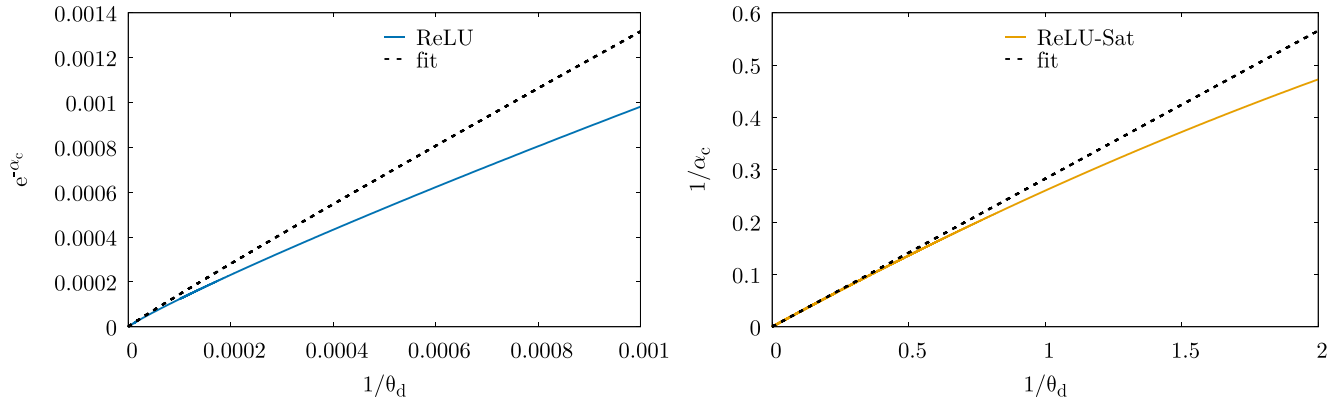
$$\Gamma_0 = \Delta_Q - \Delta_0 \simeq \Gamma_1 Q, \tag{D67}$$

which holds in the perceptron case. The saddle-point equations therefore become equivalent to the one found in the perceptron.

6. Distribution of synaptic weights

The distribution of synaptic weights is as follows:

$$P(W) = \left\langle\left\langle \frac{1}{\Omega}\int_0^\infty \prod_{li} dW_{li} \prod_{\mu=1}^{\alpha N} \Theta\left[\frac{\sigma^\mu}{\sqrt{K}}\left(\sum_{l=1}^{K} c_l\, g\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi_{li}^\mu - \sqrt{\frac{N}{K}}\theta d\right) - K\theta s\right) - \kappa\right] \delta(W - W_{11})\right\rangle\right\rangle_{\{\xi^\mu,\sigma^\mu\}}, \tag{D68}$$



(a) ReLU. (b) ReLU-Sat.

FIG. 11. Fit to the critical capacity for large dendritic thresholds for the ReLU (a) and the ReLU-Sat (b) nonlinearities. The external parameters are the same used in the corresponding figures of the main text, i.e., $f_{\rm in} = f_{\rm out} = \theta_s = 0.5$ and $\kappa = 0$. Notice that in the case of the ReLU function (left panel) we are plotting $e^{-\alpha_c}$ versus θ_d at variance to the ReLU-Sat. In each plot the dashed black line represents a linear fit $a + bx$ of the analytical data for large θ_d.

with respect to the first weight of the first subperceptron W_{11} for simplicity. Resorting to the replica method by introducing n replicas we have:

$$P(W) = \lim_{n\to 0} \mathop{\mathbb{E}}_{\sigma^\mu} \int_0^\infty \prod_{i,l,a} dW_{li}^a \int \prod_{\mu,a,l} \frac{d\lambda_{l\mu}^a \, d\hat{\lambda}_{l\mu}^a}{2\pi} \, e^{i\lambda_{l\mu}^a \hat{\lambda}_{l\mu}^a} \prod_{\mu=1}^{\alpha N} \Theta\left[\frac{\sigma^\mu}{\sqrt{K}} \left(\sum_l c_l \, g\left(\lambda_{l\mu}^a\right) - K\theta s \right) - \kappa \right]$$
$$\times \, \delta(W - W_{11}) e^{i\sqrt{\frac{N}{K}}\theta d \sum_{\mu,a,l} \hat{\lambda}_{l\mu}^a} \prod_{l,i,\mu} \left\langle e^{-i\xi_{li}^\mu \sqrt{\frac{K}{N}} \sum_a W_{li}^a \hat{\lambda}_{l\mu}^a} \right\rangle_{\xi_{li}^\mu}. \tag{D69}$$

Repeating the same steps as in Appendix D 3 we have

$$P(W) = \lim_{n\to 0} \int \prod_{a<b,l} \frac{dq_l^{ab} \, d\hat{q}_l^{ab}}{2\pi K/N} \int \prod_{a,l} \frac{dQ_l^a \, d\hat{Q}_l^a}{2\pi K/N} \int \prod_{a,l} \frac{dM_l^a \, d\hat{M}_l^a}{2\pi\sqrt{K/N}} \, e^{-\frac{N}{K}\sum_{a<b,l} q_l^{ab}\hat{q}_l^{ab} - \frac{N}{K}\sum_{a,l} Q_l^a \hat{Q}_l^a - \frac{N}{K}\overline{W}\sum_{a,l}\hat{M}_l^a}$$
$$\times \, e^{N\alpha \, G_E(q_l^{ab}, Q_l^a, M_l^a) + (\frac{N}{K}-1) G_S(\hat{q}_l^{ab}, \hat{Q}_l^a, \hat{M}_l^a)} \int_0^\infty \prod_{l,a} dW_l^a \, \delta(W - W_1) \, e^{\sum_{a<b,l} \hat{q}_l^{ab} W_l^a W_l^b + \sum_{a,l} \hat{Q}_l^a (W_l^a)^2 + \sum_{a,l} \hat{M}_l^a W_l^a} \tag{D70}$$

where the entropic and energetic terms are the same as in Eqs. (D14a), (D14b), and the order parameters are those defined in (D11). In the limit $n \to 0$ therefore

$$P(W) = \lim_{n\to 0} \int_0^\infty \prod_{l,a} dW_l^a \, \delta(W - W_1) \, e^{\sum_{a<b,l} \hat{q}_l^{ab} W_l^a W_l^b + \sum_{a,l} \hat{Q}_l^a (W_l^a)^2 + \sum_{a,l} \hat{M}_l^a W_l^a} \tag{D71}$$

provided the order parameters satisfy the same saddle-point equations as before. Under the RS ansatz expression (D71) becomes

$$P(W) = \Theta(W) \int Dz \, \frac{e^{-\frac{1}{2}(\hat{q}-2\hat{Q})W^2 + (\sqrt{\hat{q}}z + \hat{M})W}}{\int_0^\infty dW \, e^{-\frac{1}{2}(\hat{q}-2\hat{Q})W^2 + (\sqrt{\hat{q}}z + \hat{M})W}}$$
$$= \Theta(W) \sqrt{\hat{q} - 2\hat{Q}} \, e^{-\frac{1}{2}(\hat{q}-2\hat{Q})W^2 + \hat{M}W} \int Dz \, e^{\sqrt{\hat{q}}Wz} \, \frac{G\left(-\frac{\sqrt{\hat{q}}z + \hat{M}}{\sqrt{\hat{q}-2\hat{Q}}}\right)}{H\left(-\frac{\sqrt{\hat{q}}z + \hat{M}}{\sqrt{\hat{q}-2\hat{Q}}}\right)}. \tag{D72}$$

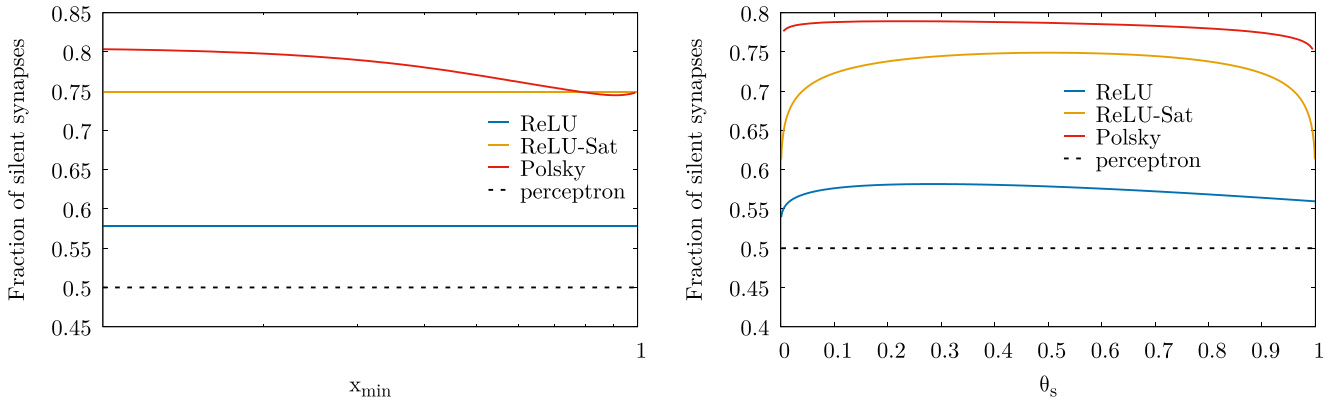
Notice that the dependence of $P(W)$ on K and on the activation function is not explicit, but is concealed inside the order parameters that clearly depend on them through the saddle point they have to satisfy. Notice also that for $\alpha = 0$ the synaptic weight satisfies an exponential distribution

$$P(W) = \Theta(W)\hat{M} e^{-\hat{M}W} = \Theta(W) \frac{f_{\rm in}}{\theta_d} e^{-\frac{f_{\rm in}}{\theta_d} W}. \tag{D73}$$

This is to be expected, since at $\alpha = 0$ the only constrain that is required apart for the fact that the synapses are non-negative, is that their average is $\overline{W} = \frac{\theta_d}{f_{\rm in}}$.

Distribution of synaptic weights in the maximal storage limit

In the critical capacity limit $\alpha \to \alpha_c$ the expression of the distribution of synaptic weight greatly simplifies. Using the



(a) $\kappa = 0$, $f_{\rm in} = f_{\rm out} = 0.5$ and $\theta_s = \theta_d = 0.5$, $\gamma = 15$. (b) $\kappa = 0$, $f_{\rm in} = f_{\rm out} = 0.5$ and $\theta_d = 0.5$, $x_{\rm min} = 0.33$, $\gamma = 15$.

FIG. 12. Fraction of silent synapses as a function the parameter $x_{\rm min}$ of the Polsky nonlinearity (left) and the the somatic threshold. In the plot we compare the Polsky to the to ReLU, the "saturating" ReLU. The dashed black line represents the case of the one-layer neuron model, where the critical capacity $\alpha_c^{\rm perc} = 1$. In the captions of the panels we show the value of the fixed external parameters. Notice that in both plots no robustness κ is imposed.

scalings in (D54) we find

$$P(W) = \Theta(W)\, e^{-\frac{A}{2dq}W^2 - \frac{B\sqrt{C}}{dq}W} \sqrt{\frac{A}{dq}} \int Dz\, e^{\frac{\sqrt{C}}{dq}Wz} \left[G\left(\sqrt{\frac{C}{A}} \frac{z-B}{\sqrt{\hat{q}-2\hat{Q}}} \right) \Theta(z-B) - \sqrt{\frac{C}{Adq}} (z-B)\Theta(B-z) \right]. \tag{D74}$$

Using the identity

$$\int Dz\, e^{az}\, (z+b)\, \Theta(-b-z) = (a+b)\, e^{\frac{a^2}{2}} H(a+b) - e^{-ab}\, G(b), \tag{D75}$$

we obtain

$$P(W) = H(-B)\, \delta(W) + \frac{1}{\sqrt{2\pi} W_\star} e^{-\frac{(W+BW_\star)^2}{2W_\star^2}} \Theta(W), \tag{D76}$$

where $W_\star \equiv \frac{\sqrt{C}}{A}$. As showed in Ref. [31] in the one layer neuron model the synaptic weight distribution changes from being exponential to being a Gaussian plus a spike consisting to a fraction $H(-B)$ of "silent" weights at the critical capacity. Indeed as constraints due to the training set are added, more and more synapses tend to assume low weight. This is the case also in our two layer neuron model. It is interesting to note that both the distribution of synaptic weight at finite α (D72) and at critical capacity are in form exactly the same as the one derived in Ref. [31] for the one layer neuron model; the dependence on the nonlinearity induced by the dendrites is actually implicit in the order parameters.

In Fig. 2 we show how the fraction of silent synapses $p_0 = H(-B)$ depends on the somatic threshold for the ReLU, ReLU-Sat, and Polsky nonlinearities. We also show in the Polsky case, how p_0 depends on the parameters defining the shape of the function itself, $x_{\rm min}$ and γ. The fraction of silent synapses as a function of the parameter $x_{\rm min}$ of the Polsky nonlinearity and the the somatic threshold θ_s is reported in Fig. 12.

APPENDIX E: CHOICE AND SCALING OF THE NUMERICAL SIMULATIONS HYPER-PARAMETERS

Consider the transfer function implemented by the neuron with nonlinear dendritic branches, which represents the output preactivation prior to the thresholding operation performed by the Θ function:

$$\Delta_{\rm out}^{\mu} = \frac{1}{\sqrt{K}} \sum_{l=1}^{K} c_l\, g\left(\sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} W_{li} \xi_{li}^{\mu} - \sqrt{\frac{N}{K}} \theta_d \right) - \sqrt{K}\theta_s, \tag{E1}$$

we impose that $W \in [0, \frac{2\theta_d}{f_{\rm in}}]$ and consequently:

$$W \sim O\left(\frac{\theta_d}{f_{\rm in}} \right).$$

If we assume that $\theta_d \sim O(1)$ and $\theta_s \sim O(1)$, then we have $W \sim O(\frac{1}{f_{\rm in}})$.

We also have $\sum_{i=1}^{N/K} W_{li} \xi_{li}^{\mu} \sim O(\frac{N}{K})$. Consequently, from (E1), we observe that the dendritic preactivations scale as:

$$\left(\sqrt{\frac{K}{N}} \sum_{i=1}^{N/K} W_{li} \xi_{li}^{\mu} - \sqrt{\frac{N}{K}} \theta_d \right) \sim O\left(\sqrt{\frac{N}{K}} \right). \tag{E2}$$

Consequently, the preactivation variance remains finite in the limit $N \to \infty$, which is the primary motivation for choosing these scalings, as the preactivation variance is a crucial factor in ensuring the consistency of the learning setting when vary-

ing the input dimensionality N and the number of dendritic branches K.

Applying the same type of consideration to the preactivation of the single output node in (E1) and recalling that $c_l = 1\ \forall\, l$ and $g(\cdot) \sim O(1)$, we find that it scales as:

$$\Delta^{\mu}_{\text{out}} \sim O(\theta_s\sqrt{K}). \tag{E3}$$

1. SGD with cross-entropy loss

Recalling the expression for the cross-entropy loss used to investigate the performance of SGD on the neuron:

$$\mathscr{L}_{ce}\left(\Delta^{\mu}_{\text{out}}\right) = \frac{1}{2\gamma_{ce}} \log\left(1 + \exp\left(-2\gamma_{ce}\Delta^{\mu}_{\text{out}}\right)\right), \tag{E4}$$

we can estimate the gradients of the cross-entropy loss, which are used in the SGD update, as follows:

$$\frac{\partial \mathscr{L}_{ce}(\Delta)}{\partial \Delta} = -\frac{1}{1 + \exp\left(2\gamma_{ce}\Delta\right)} \tag{E5}$$

$$\sim O(1), \tag{E6}$$

where the last step holds [due to (E3)] if:

$$\gamma_{ce} \sim O\left(\frac{1}{\theta_s\sqrt{K}}\right). \tag{E7}$$

Proceeding with the derivative with respect to to the weights, we obtain:

$$\frac{\partial \mathscr{L}_{ce}(\Delta(W))}{\partial W} = \frac{\partial \Delta(W)}{\partial W}\frac{\partial \mathscr{L}_{ce}(\Delta)}{\partial \Delta}, \tag{E8}$$

$$= \frac{\partial}{\partial W}\left(\frac{1}{\sqrt{K}}\sum_{l=1}^{K} c_l\, g\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi^{\mu}_{li} - \sqrt{\frac{N}{K}}\theta_d\right) - \sqrt{K}\theta_s\right)\frac{\partial \mathscr{L}_{ce}(\Delta)}{\partial \Delta}, \tag{E9}$$

$$= \left[\frac{1}{\sqrt{K}}\sum_{l=1}^{K}\left(c_l\, g'\left(\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K} W_{li}\xi^{\mu}_{li} - \sqrt{\frac{N}{K}}\theta_d\right)\sqrt{\frac{K}{N}}\sum_{i=1}^{N/K}\xi^{\mu}_{il}\right)\right]\frac{\partial \mathscr{L}_{ce}(\Delta)}{\partial \Delta}, \tag{E10}$$

$$\sim O\left(\frac{1}{\sqrt{K}}\sqrt{\frac{K}{N}} f_{\text{in}} N\right) O(1), \tag{E11}$$

$$\sim O(f_{\text{in}}\sqrt{N}), x \tag{E12}$$

where in the penultimate step, we have used the fact that the derivative of the function g is bounded within the interval [0,1] so that $g'(\cdot) \sim O(1)$; that $\sum_{i=1}^{N/K}\xi^{\mu}_{il} \sim O(f_{\text{in}}\frac{N}{K})$; and (E6).

At this point, recalling that the SGD update rule is as follows:

$$w_{il} \leftarrow w_{il} - \zeta\nabla_{w_{il}}\mathscr{L}(w_{il}) \tag{E13}$$

and imposing an update of the same order as the weights, i.e., $w \sim O(\theta_d) \Rightarrow \zeta\nabla_w\mathscr{L}(w) \sim O(\theta_d)$, we find the desired scaling for the learning rate ζ:

$$\zeta \sim O\left(\frac{\theta_d}{f_{\text{in}}\sqrt{N}}\right). \tag{E14}$$

2. Comparison between the dendritic and the linear neuron

The transfer function of the linear neuron (i.e., the perceptron model), is given by:

$$\Delta^{\mu}_{\text{perc, out}} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} W_i\xi^{\mu}_i - \sqrt{N}\theta_s. \tag{E15}$$

From this expression, and observing that the weights scale with θ_s, we obtain the following scalings for the learning rate ζ and cross-entropy parameter γ_{ce}:

$$\zeta \sim O\left(\frac{\theta_s}{f_{\text{in}}\sqrt{N}}\right), \quad \gamma_{ce} \sim O\left(\frac{1}{\theta_s\sqrt{N}}\right). \tag{E16}$$

[1] M. Minsky and S. A. Papert, *Perceptrons—Expanded Edition* (MIT Press, Cambridge, MA, 1988).

[2] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition., IEEE Trans **EC-14**, 326 (1965).

[3] E. Gardner, The space of interactions in neural network models, J. Phys. A: Math. Gen. **21**, 257 (1988).

[4] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, UK, 2001).

[5] M. London and M. Häusser, Dendritic computation, Annu. Rev. Neurosci. **28**, 503 (2005).

[6] M. Larkum, A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex, Trends Neurosci. **36**, 141 (2013).

[7] G. Major, M. E. Larkum, and J. Schiller, Active properties of neocortical pyramidal neuron dendrites, Annu. Rev. Neurosci. **36**, 1 (2013).

[8] J. Schiller, G. Major, H. J. Koester, and Y. Schiller, NMDA spikes in basal dendrites of cortical pyramidal neurons, Nature (Lond.) **404**, 285 (2000).

[9] T. Branco and M. Häusser, Synaptic integration gradients in single cortical pyramidal cell dendrites, Neuron **69**, 885 (2011).

[10] M. E. Larkum, J. J. Zhu, and B. Sakmann, A new cellular mechanism for coupling inputs arriving at different cortical layers, Nature (Lond.) **398**, 338 (1999).

[11] M. E. Larkum, W. Senn, and H.-R. Luscher, Top-down dendritic input increases the gain of layer 5 pyramidal neurons, Cereb. Cortex **14**, 1059 (2004).

[12] T. Nevian, M. E. Larkum, A. Polsky, and J. Schiller, Properties of basal dendrites of layer 5 pyramidal neurons: A direct patch-clamp recording study, Nat Neurosci. **10**, 206 (2007).

[13] A. Polsky, B. W. Mel, and J. Schiller, Computational subunits in thin dendrites of pyramidal cells, Nat. Neurosci. **7**, 621 (2004).

[14] P. Poirazi, T. Brannon, and B. W. Mel, Pyramidal neuron as two-layer neural network, Neuron **37**, 989 (2003).

[15] N. Brunel, V. Hakim, and M. J. Richardson, Single neuron dynamics and computation, Curr. Opin. Neurobiol. **25**, 149 (2014).

[16] B. B. Ujfalussy, J. K. Makara, M. Lengyel, and T. Branco, Global and multiplexed dendritic domputations under in-vivo-like conditions, Neuron **100**, 579 (2018).

[17] M. E. Larkum, T. Nevian, M. Sandler, A. Polsky, and J. Schiller, Synaptic integration in tuft dendrites of layer 5 pyramidal neurons: A new unifying principle, Science **325**, 756 (2009).

[18] D. Beniaguev, I. Segev, and M. London, Single cortical neurons as deep artificial neural networks, Neuron **109**, 2727 (2021).

[19] M. Pagkalos, R. Makarov, and P. Poirazi, Leveraging dendritic properties to advance machine learning and neuro-inspired computing, Curr. Opin. Neurobiol. **85**, 102853 (2024).

[20] S. Sardi, R. Vardi, A. Sheinin, A. Goldental, and I. Kanter, New types of experiments reveal that a neuron functions as multiple independent threshold units, Sci. Rep. **7**, 18036 (2017).

[21] G. J. Mitchison and R. M. Durbin, Bounds on the learning capacity of some multi-layer networks, Biol. Cybern. **60**, 345 (1989).

[22] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, Phys. Rev. A **45**, 4146 (1992).

[23] R. Monasson and R. Zecchina, Weight space structure and internal representations: A direct approach to learning and generalization in multilayer neural networks, Phys. Rev. Lett. **75**, 2432 (1995).

[24] C. Baldassi, E. M. Malatesta, and R. Zecchina, Properties of the geometry of solutions and capacity of multilayer neural networks with rectified linear unit activations, Phys. Rev. Lett. **123**, 170602 (2019).

[25] J. A. Zavatone-Veth and C. Pehlevan, Activation function dependence of the storage capacity of treelike neural networks, Phys. Rev. E **103**, L020301 (2021).

[26] C. Baldassi, C. Borgs, J. T. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes, Proc. Natl. Acad. Sci. USA **113**, E7655 (2016).

[27] C. Baldassi, F. Pittorino, and R. Zecchina, Shaping the learning landscape in neural networks around wide flat minima, Proc. Natl. Acad. Sci. USA **117**, 161 (2020).

[28] C. Lucibello, F. Pittorino, G. Perugini, and R. Zecchina, Deep learning via message passing algorithms based on belief propagation, Mach. Learn.: Sci. Technol. **3**, 035005 (2022).

[29] B. L. Annesi, C. Lauditi, C. Lucibello, E. M. Malatesta, G. Perugini, F. Pittorino, and L. Saglietti, Star-shaped space of solutions of the spherical negative perceptron, Phys. Rev. Lett. **131**, 227301 (2023).

[30] K. Hornik, Approximation capabilities of multilayer feedforward networks, Neural Netw. **4**, 251 (1991).

[31] N. Brunel, V. Hakim, P. Isope, J.-P. Nadal, and B. Barbour, Optimal information storage and the distribution of synaptic weights: Perceptron versus purkinje cell, Neuron **43**, 745 (2004).

[32] N. Brunel, Is cortical connectivity optimized for storing information? Nat. Neurosci. **19**, 749 (2016).

[33] V. Braitenberg and A. Schütz, *Anatomy of the Cortex* (Springer-Verlag, Berlin, 1991).

[34] D. M. Iascone, Y. Li, U. l, M. Doron, H. Chen, V. Andreu, F. Goudy, H. Blockus, L. F. Abbott, I. Segev, H. Peng, and F. Polleux, Whole-neuron synaptic mapping reveals spatially precise excitatory/inhibitory balance limiting dendritic and somatic spiking, Neuron **106**, 566 (2020).

[35] G. N. Elston, R. Benavides-Piccione, A. Elston, P. R. Manger, and J. Defelipe, Pyramidal cells in prefrontal cortex of primates: Marked differences in neuronal structure among species, Front. Neuroanat. **5**, 2 (2011).

[36] E. Gardner and B. Derrida, Optimal storage properties of neural network models, J. Phys. A: Math. Gen. **21**, 271 (1988).

[37] D. Amit, K. Y. Wong, and C. Campbell, Perceptron learning with sign-constrained weights, J. Phys. A: Math. Gen. **22**, 2039 (1989).

[38] M. Stojnic, Capacity of the treelike sign perceptrons neural networks with one hidden layer–RDT based upper bounds, arXiv:2312.08244.

[39] Y. LeCun, L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. A. Muller, E. Sackinger, P. Simard, and V. Vapnik, Comparison of learning algorithms for handwritten digit recognition, in *International Conference on Artificial Neural Networks*, edited by F. Fogelman and P. Gallinari (EC2 & Cie, Paris, 1995), pp. 53–60.

[40] H. Xiao, K. Rasul, and R. Vollgraf, Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms, arXiv:1708.07747.

[41] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images (Canadian Institute for Advanced Research, Toronto, 2009).

[42] Code available at https://gitlab.com/bocconi-artlab/biologicalneuron.

[43] E. M. Malatesta, High-dimensional manifold of solutions in neural networks: Insights from statistical physics, arXiv:2309.09240.

[44] M. Mezard, G. Parisi, and M. Virasoro, *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific, Singapore, 1987).

[45] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory* (Springer Science & Business Media, New York, 2013).

[46] C. Clopath, J.-P. Nadal, and N. Brunel, Storage of correlated patterns in standard and bistable Purkinje cell models, PLoS Comput. Biol. **8**, e1002448 (2012).

[47] G. Parisi, Infinite number of order parameters for spin-glasses, Phys. Rev. Lett. **43**, 1754 (1979).

[48] A. Engel, H. M. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Storage capacity and learning algorithms for two-layer neural networks, Phys. Rev. A **45**, 7590 (1992).

[49] C. Baldassi, E. M. Malatesta, G. Perugini, and R. Zecchina, Typical and atypical solutions in nonconvex neural networks with discrete and continuous weights, Phys. Rev. E **108**, 024310 (2023).

[50] B. L. Annesi, E. M. Malatesta, and F. Zamponi, Exact full-RSB SAT/UNSAT transition in infinitely wide two-layer neural networks, SciPost Phys. **18**, 118 (2025).

[51] C. Baldassi, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses, Phys. Rev. Lett. **115**, 128101 (2015).

[52] C. Baldassi, C. Lauditi, E. M. Malatesta, G. Perugini, and R. Zecchina, Unveiling the structure of wide flat minima in neural networks, Phys. Rev. Lett. **127**, 278301 (2021).

[53] P. Isope and B. Barbour, Properties of unitary granule cell: Purkinje cell synapses in adult rat cerebellar slices, J. Neurosci. **22**, 9668 (2002).

[54] S. Song, P. J. Sjöström, M. Reigl, S. Nelson, and D. B. Chklovskii, Highly nonrandom features of synaptic connectivity in local cortical circuits, PLoS Biol. **3**, e68 (2005).

[55] H. Markram, J. Lubke, M. Frotscher, A. Roth, and B. Sakmann, Physiology and anatomy of synaptic connections between thick tufted pyramidal neurones in the developing rat neocortex, J. Physiol. **500**, 409 (1997).

[56] P. J. Sjöström, G. G. Turrigiano, and S. Nelson, Rate, timing, and cooperativity jointly determine cortical synaptic plasticity, Neuron **32**, 1149 (2001).

[57] S. Lefort, C. Tomm, J. C. Floyd Sarria, and C. C. Petersen, The excitatory neuronal network of the C2 barrel column in mouse primary somatosensory cortex, Neuron **61**, 301 (2009).

[58] L. Campagnola, S. C. Seeman, T. Chartrand, L. Kim, A. Hoggarth, C. Gamlin, S. Ito, J. Trinh, P. Davoudian, C. Radaelli, M.-H. Kim, T. Hage, T. Braun, L. Alfiler, J. Andrade, P. Bohn, R. Dalley, A. Henry, S. Kebede, A. Mukora, D. Sandman, G. Williams, R. Larsen, C. Teeter, T. L. Daigle, K. Berry, N. Dotson, R. Enstrom, M. Gorham, M. Hupp, S. D. Lee, K. Ngo, P. R. Nicovich, L. Potekhina, S. Ransford, A. Gary, J. Goldy, D. McMillen, T. Pham, M. Tieu, L. Siverts, M. Walker, C. Farrell, M. Schroedter, C. Slaughterbeck, C. Cobb, R. Ellenbogen, R. P. Gwinn, C. D. Keene, A. L. Ko, J. G. Ojemann, D. L. Silbergeld, D. Carey, T. Casper, K. Crichton, M. Clark, N. Dee, L. Ellingwood, J. Gloe, M. Kroll, J. Sulc, H. Tung, K. Wadhwani, K. Brouner, T. Egdorf, M. Maxwell, M. McGraw, C. A. Pom, A. Ruiz, J. Bomben, D. Feng, N. Hejazinia, S. Shi, A. Szafer, W. Wakeman, J. Phillips, A. Bernard, L. Esposito, F. D. D'Orazi, S. Sunkin, K. Smith, B. Tasic, A. Arkhipov, S. Sorensen, E. Lein, C. Koch, G. Murphy, H. Zeng, and T. Jarsky, Local connectivity and synaptic dynamics in mouse and human neocortex, Science **375**, eabj5861 (2022).

[59] F. Pittorino, C. Lucibello, C. Feinauer, G. Perugini, C. Baldassi, E. Demyanenko, and R. Zecchina, Entropic gradient descent algorithms and wide flat minima, in *International Conference on Learning Representations* (ICLR, Vienna, Austria, 2021).

[60] A. A. Faisal, L. P. Selen, and D. M. Wolpert, Noise in the nervous system, Nat. Rev. Neurosci. **9**, 292 (2008).

[61] F. Pittorino, A. Ferraro, G. Perugini, C. Feinauer, C. Baldassi, and R. Zecchina, Deep networks on toroids: Removing symmetries reveals the structure of flat regions in the landscape geometry, in *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 162, edited by K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato (PMLR, Baltimore, MD, 2022), pp. 17759–17781.

[62] A. Gidon, T. A. Zolnik, P. Fidzinski, F. Bolduan, A. Papoutsi, P. Poirazi, M. Holtkamp, I. Vida, and M. E. Larkum, Dendritic action potentials and computation in human layer 2/3 cortical neurons, Science **367**, 83 (2020).

[63] J. H. Kirchner and J. Gjorgjieva, Emergence of synaptic organization and computation in dendrites, Neuroforum **28**, 21 (2022).

[64] J. Chapeton, T. Fares, D. LaSota, and A. Stepanyants, Efficient associative memory storage in cortical circuits of inhibitory and excitatory neurons, Proc. Natl. Acad. Sci. USA **109**, E3614 (2012).

[65] R. Rubin, L. F. Abbott, and H. Sompolinsky, Balanced excitation and inhibition are required for high-capacity, noise-robust neuronal selectivity, Proc. Natl. Acad. Sci. USA **114**, E9366 (2017).

[66] R. Gütig and H. Sompolinsky, The tempotron: A neuron that learns spike timing-based decisions, Nat. Neurosci. **9**, 420 (2006).

[67] F. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms* (Spartan Books, Washington, DC, 1962).

[68] F. S. Werblin and J. E. Dowling, Organization of the retina of the mudpuppy necturus maculosus. ii. intracellular recordings, J. Neurophysiol. **32**, 339 (1969).