


Activation function dependence of the storage capacity of treelike neural networks

Jacob A. Zavatone-Veth^{1,*} and Cengiz Pehlevan^{2,3,†}

¹*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

²*John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA*

³*Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA*

 (Received 8 July 2020; revised 14 November 2020; accepted 4 February 2021; published 19 February 2021)

The expressive power of artificial neural networks crucially depends on the nonlinearity of their activation functions. Though a wide variety of nonlinear activation functions have been proposed for use in artificial neural networks, a detailed understanding of their role in determining the expressive power of a network has not emerged. Here, we study how activation functions affect the storage capacity of treelike two-layer networks. We relate the boundedness or divergence of the capacity in the infinite-width limit to the smoothness of the activation function, elucidating the relationship between previously studied special cases. Our results show that nonlinearity can both increase capacity and decrease the robustness of classification, and provide simple estimates for the capacity of networks with several commonly used activation functions. Furthermore, they generate a hypothesis for the functional benefit of dendritic spikes in branched neurons.

DOI: [10.1103/PhysRevE.103.L020301](https://doi.org/10.1103/PhysRevE.103.L020301)

The expressive power of artificial neural networks is well known [1–4], but a complete theoretical account of how their remarkable abilities arise is lacking [5–8]. In particular, though a diverse array of nonlinear activation functions have been employed in neural networks [5,6,9–14], our understanding of the relationship between activation function choice and computational capability is incomplete [9–11,15]. Methods from the statistical mechanics of disordered systems have enabled the interrogation of this link in several special cases [11–19], but these previous works have not yielded a general theory.

In this Letter, we characterize how pattern storage capacity depends on activation function in a tractable two-layer network model known as the treelike committee machine (henceforth TCM). In addition to their uses in machine learning, TCMs have been used to model nonlinear computations in dendrite-bearing neurons [20,21]. We find that the storage capacity of a TCM remains finite in the infinite-width limit provided that the activation function is weakly differentiable, and it and its weak derivative are square-integrable with respect to Gaussian measure. For example, the capacity with sign activation functions diverges, while that with rectified linear unit or error function activations is finite. We predict that nonlinearity should increase capacity, but may reduce the robustness of classification. These connections between expressive power and smoothness begin to shed light on the influence of activation functions on the capabilities of neural networks and branched neurons.

The treelike committee machine. The TCM is a two-layer neural network with N inputs divided among K hidden units into disjoint groups of N/K and binary outputs [Fig. 1(a)] [11–14,19]. For a hidden unit activation function g , a set of

hidden unit weight vectors $\{\mathbf{w}_j \in \mathbb{R}^{N/K}\}_{j=1}^K$, a readout weight vector $\mathbf{v} \in \mathbb{R}^K$, and a threshold $\vartheta \in \mathbb{R}$, its output is given as

$$y(\mathbf{x}) = \text{sgn}[s(\mathbf{x})] \quad \text{for} \quad (1)$$

$$s(\mathbf{x}; \{\mathbf{w}_j\}, \mathbf{v}, \vartheta) = \frac{1}{\sqrt{K}} \sum_{j=1}^K v_j g\left(\frac{\mathbf{w}_j \cdot \mathbf{x}_j}{\sqrt{N/K}}\right) - \vartheta, \quad (2)$$

where \mathbf{x}_j denotes the vector of inputs to the j th hidden unit. In this model, the readout weight vector and threshold are fixed, and only the hidden unit weights are learned. The perceptron can thus be viewed as the special case of a TCM with identity activation functions and equal readout weights [16,17].

Statistical mechanics of pattern storage. To characterize this network’s ability to classify a random data set of P examples subject to constraints on the hidden unit weights imposed by a probability measure ρ , we define the Gardner volume [16,17]

$$Z = \int d\rho(\{\mathbf{w}_j\}) \prod_{\mu=1}^P \Theta[y^\mu s(\mathbf{x}^\mu; \{\mathbf{w}_j\}, \mathbf{v}, \vartheta) - \kappa], \quad (3)$$

which measures the fractional volume in weight space such that all examples are classified correctly with margin at least κ . We consider “spherical” committee machines, in which the hidden unit weight vectors lie on the sphere of radius $(N/K)^{1/2}$ [11–14,16–19]. As in most studies of the Gardner volume, we consider a data set in which the components of the inputs and the target outputs are independent and identically distributed as $x_{jk}^\mu = \pm 1$ and $y^\mu = \pm 1$ with equal probability [11–14,16–19].

We will study a sequential infinite-width limit in which we first take $N, P \rightarrow \infty$ with load $\alpha \equiv P/N = \mathcal{O}(1)$ and then take $K \rightarrow \infty$ [22]. The infinite-width limit is of both theoretical and practical interest, as extremely wide networks are now commonly used in applications [7,9,23,24]. In this limit, we

*jzavatoneveth@g.harvard.edu

†cpehlevan@seas.harvard.edu

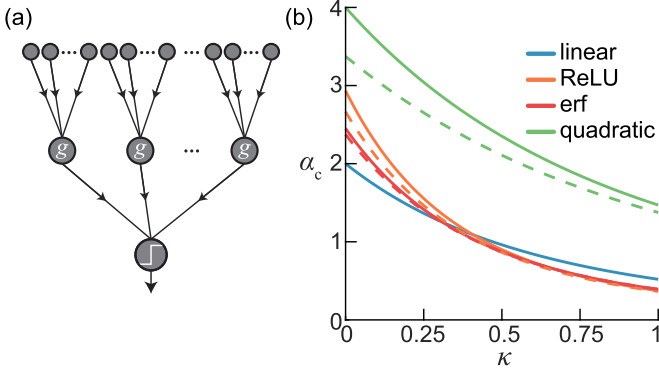


FIG. 1. Pattern storage in treelike committee machines. (a) Network architecture. (b) Capacity α_c as a function of margin κ for several common activation functions. Solid and dashed lines indicate estimates of the capacity under replica-symmetric and one-step replica-symmetry-breaking *Ansätze*, respectively.

expect the free entropy per weight $f = N^{-1} \log Z$ to be self-averaging, and for there to exist a critical load α_c , termed the capacity, below which the classification task is solvable with probability one and above which Z vanishes [14,16–18]. The special case of this model with sign activation functions was intensively studied in the late 20th century, showing that the capacity diverges as $K \rightarrow \infty$ [12,13,19,25,26]. In contrast, Baldassi *et al.* [11] showed in a recent Letter that the capacity with rectified linear unit (ReLU) activations remains bounded in the infinite-width limit. Our primary objective in this work is to identify the class of activation functions for which the capacity remains finite.

We begin our analysis by specifying our choice of general constraints on the activation function, readout weights, and threshold. We will require the $K \rightarrow \infty$ limit to be well defined in the sense that the output preactivation s has finite variance. In this limit, the central limit theorem implies that the hidden unit preactivations converge in distribution to a collection of independent Gaussian random variables [27]. Therefore, the activation function g must lie in the Lebesgue space $\mathcal{L}^2(\gamma)$ of functions that are square-integrable with respect to the Gaussian measure γ on the reals. Furthermore, as $\text{var}(s) \propto \|\mathbf{v}\|_2^2/K$, we must have $\|\mathbf{v}\|_2 = \mathcal{O}(\sqrt{K})$. As $\|\mathbf{v}\|_2$ sets the effective scale of ϑ and κ but does not affect the zero-margin capacity, we fix $\|\mathbf{v}\|_2 = \sqrt{K}$. To ensure that s has mean zero, we set $\vartheta = K^{-1/2}(\mathbb{E}g) \sum_{j=1}^K v_j$, where $\mathbb{E}g = \int d\gamma g$ is the average hidden unit activation. This choice maximizes the capacity for the symmetric data sets of interest [22], and generalizes the conditions on \mathbf{v} and ϑ considered in previous works [11–13,19].

To compute the limiting quenched free entropy, we apply the replica trick, which exploits a limit identity for logarithmic averages and a nonrigorous interchange of limits to write

$$f = \lim_{n \downarrow 0} \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{nN} \log \mathbb{E}_{\mathbf{x}, \mathbf{y}} Z_{N, \alpha N, K}^n, \quad (4)$$

where the validity of analytic continuation of the moments from positive integer n to $n \downarrow 0$ is assumed [16,18,28]. This calculation is standard, and we defer the details to the Supplemental Material [22].

In this limit, the quenched free entropy can be expressed using the method of steepest descent as an extremization over the Edwards-Anderson order parameters $q_j^{ab} = (K/N) \mathbf{w}_j^a \cdot \mathbf{w}_j^b$ [16,18,28], which represent the average overlap between the preactivations of the j th hidden unit in two different replicas a and b . Under a replica- and hidden-unit-symmetric (RS) *Ansatz* $q_j^{ab} = q$, one finds that

$$f_{\text{RS}} = \text{extr}_q \left\{ \alpha \int d\gamma(z) \log H \left(\frac{\kappa + \sqrt{\tilde{q}(q)}z}{\sqrt{\sigma^2 - \tilde{q}(q)}} \right) + \frac{1}{2} \left[\frac{q}{1-q} + \log(1-q) \right] \right\}, \quad (5)$$

where $H(z) = \int_z^\infty d\gamma(x)$ is the Gaussian tail distribution function, $\sigma^2 = \mathbb{E}g^2 - (\mathbb{E}g)^2$ is the variance of the activation, and

$$\tilde{q}(q) = \text{cov} \left[g(x), g(y) : \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} 1 & q \\ q & 1 \end{bmatrix} \right) \right] \quad (6)$$

is an effective order parameter describing the average overlap between the activations of a given hidden unit in two different replicas. This expression for f_{RS} is equivalent to that given in Ref. [11] for ReLU activations, but we adopt a different definition for the effective order parameter that has a clearer statistical interpretation.

To find the replica-symmetric capacity α_{RS} , one must take the limit $q \uparrow 1$ in the saddle-point equation that defines the extremum with respect to q , as the Gardner volume tends to zero in this limit [11–14,16,17]. As $q \uparrow 1$, $\tilde{q} \uparrow \sigma^2$, but the asymptotic properties of \tilde{q} as a function of $\varepsilon \equiv 1 - q$ depend on the choice of activation function. Making the general *Ansatz* that $\sigma^2 - \tilde{q} \sim \varepsilon^\ell$ for some $\ell > 0$, we find that $\alpha_{\text{RS}} \sim \varepsilon^{\ell-1}$ [22]. Therefore, the RS capacity diverges if $\ell < 1$ and vanishes if $\ell > 1$, while the boundary case $\ell = 1$ is special in that the capacity is bounded but nonvanishing. For the special cases of $\text{sgn}(x)$ and $g(x) = \text{ReLU}(x)$, this behavior was noted by Baldassi *et al.* [11]. For sign, one has $\sigma^2 - \tilde{q} \sim \sqrt{\varepsilon}$, and α_{RS} diverges in the infinite-width limit, while for ReLU, $\sigma^2 - \tilde{q} \sim \varepsilon$, and α_{RS} remains finite. However, Ref. [11] and other previous studies [12,13] relied on direct computation of the effective order parameters for all values of q , which is not tractable for most activation functions, and does not yield general insight.

Asymptotics of the effective order parameter. To understand the asymptotic behavior of $\tilde{q}(q)$ as $q \uparrow 1$ for general activation functions g , we apply tools from the theory of Gaussian measures [29]. As g is in $\mathcal{L}^2(\gamma)$ by assumption, it has a Fourier-Hermite series $g(x) = \sum_{k=0}^\infty g_k \text{He}_k(x)$, where $\{\text{He}_k\}$ is the set of orthonormal Hermite polynomials [22]. We note that the $\mathcal{L}^2(\gamma)$ norm of g can then be written as $\|g\|_\gamma^2 = \sum_{k=0}^\infty g_k^2$, and that $g_0 = \mathbb{E}g$. To express $\tilde{q}(q)$ in terms of these coefficients, we recall the Mehler expansion of the standard bivariate Gaussian density $\varphi(x, y; q)$ [30,31], $\varphi(x, y; q) = \varphi(x)\varphi(y) \sum_{k=0}^\infty q^k \text{He}_k(x)\text{He}_k(y)$, where $\varphi(x) = \exp(-x^2/2)/\sqrt{2\pi}$ is the univariate Gaussian density. Then, we can evaluate the expectation in (6), yielding $\tilde{q}(q) + g_0^2 = \sum_{k=0}^\infty g_k^2 q^k$, which, by Abel's theorem, is a bounded, continuous function of $q \in (-1, 1]$ because $\tilde{q}(1) + g_0^2 = \|g\|_\gamma^2$ is finite. Writing $q \equiv 1 - \varepsilon$, we expand $(1 - \varepsilon)^k$ in a

binomial series and formally interchange the order of summation to obtain $\tilde{q}(\varepsilon) + g_0^2 = \sum_{l=0}^{\infty} \frac{(-\varepsilon)^l}{l!} \sum_{k=l}^{\infty} \binom{k}{l} g_k^2$, where $\binom{k}{l} = k(k-1)\cdots(k-l+1)$ is the falling factorial. We recognize the sums over k as the norms of the weak derivatives of g , which have formal Fourier-Hermite series $g^{(l)}(x) = \sum_{k=l}^{\infty} g_k \sqrt{(k)_l} \text{He}_{k-l}(x)$, which follow from the recurrence relation $\text{He}'_k(x) = \sqrt{k} \text{He}_{k-1}(x)$ [29]. Therefore, \tilde{q} admits a formal power series expansion in ε as

$$\tilde{q}(\varepsilon) + g_0^2 = \sum_{l=0}^{\infty} \frac{(-1)^l}{l!} \|g^{(l)}\|_{\gamma}^2 \varepsilon^l. \quad (7)$$

For the RS capacity to remain bounded, we merely require that the first two terms in this series are finite, not for the series to converge at any higher order for nonvanishing ε . Therefore, the RS capacity is finite for once weakly differentiable activations g such that the \mathcal{L}^2 norms of the function and its weak derivative with respect to Gaussian measure, $\|g\|_{\gamma}$ and $\|g'\|_{\gamma}$, are finite. This class of functions is precisely the Sobolev class $\mathcal{H}^1(\gamma)$ [29]. We provide additional background material on $\mathcal{H}^1(\gamma)$ and weak differentiability in the Supplemental Material [22].

Storage capacity. For any activation function in the class $\mathcal{H}^1(\gamma)$, we find that

$$\alpha_{\text{RS}}(\kappa) = \frac{\|g'\|_{\gamma}^2}{\sigma^2} \alpha_{\text{G}}\left(\frac{\kappa}{\sigma}\right), \quad (8)$$

where

$$\alpha_{\text{G}}(\kappa) = \left[\int_{-\kappa}^{\infty} d\gamma(z) (\kappa + z)^2 \right]^{-1} \quad (9)$$

is Gardner's formula for the perceptron capacity [16,22]. In terms of Fourier-Hermite coefficients, we have $\sigma^2 = \sum_{k=1}^{\infty} g_k^2$ and $\|g'\|_{\gamma}^2 = \sum_{k=1}^{\infty} k g_k^2$. Thus, we have $\|g'\|_{\gamma}^2 \geq \sigma^2$, with equality if and only if all nonlinear terms (those corresponding to Hermite polynomials of degree two or greater) vanish. Therefore, introducing nonlinearity always increases the zero-margin RS capacity. However, as $\alpha_{\text{G}}(\kappa)$ is a monotonically decreasing function, the capacity at large margins can be reduced by nonlinearity if $\sigma < 1$. We note that the zero-margin capacity is invariant under rescaling of the activation function and hidden unit weights as $g \mapsto c_1 g$, $\mathbf{v} \mapsto c_2 \mathbf{v}$ for some constants c_1 and c_2 . For finite margin, rescaling can increase or decrease the capacity by changing σ . Thus, in the sense of classification margin, introducing nonlinearity or rescaling can reduce the robustness of classification.

Using this result, we can characterize the RS capacity of wide TCMs for several commonly used activation functions [22]. For a linear activation function, our result reduces to Gardner's perceptron capacity [16], which is expected given the equivalence between such a TCM and the perceptron in the $K \rightarrow \infty$ limit. As the sign function is not weakly differentiable, we recover the result that the capacity diverges [12,13,19]. ReLU is weakly differentiable, and we recover the result of [11] that $\alpha_{\text{RS}} = 2\pi/(\pi-1) \simeq 2.93388$. Considering sigmoidal activations, we find that $\alpha_{\text{RS}} = 2 \arcsin(2/3)/\pi \simeq 2.45140$ for the error function, while $\alpha_{\text{RS}} \simeq 2.35561$ for the hyperbolic tangent and the logistic. As an example of a nonmonotonic activation function, we consider a quadratic, which yields $\alpha_{\text{RS}} = 4$. We plot the RS

capacity as a function of margin for these activation functions in Fig. 1(b), illustrating how nonlinearity can reduce the large-margin capacity while increasing the zero-margin capacity.

However, for nonlinear activation functions, one generically expects the energy landscape to become locally nonconvex, and for replica symmetry breaking (RSB) to occur [11–14,18,28]. The RS estimate of the capacity is therefore only an upper bound, and one must account for RSB effects in order to obtain a more accurate estimate [11–14,18,19,28]. To that end, we have calculated the capacity under a one-step replica-symmetry-breaking (1-RSB) *Ansatz*, extending the results of earlier work [11–13] to arbitrary activation functions. Under the 1-RSB *Ansatz*, the replicas are divided into groups of size m , with intergroup overlap q_0 and intragroup overlap q_1 . Then, the capacity is extracted by taking the limit $q_1 \uparrow 1$, $m \downarrow 0$, with $r \equiv m/(1-q_1)$ finite [11–14,28].

As detailed in the Supplemental Material [22], this calculation yields an expression for $\alpha_{1\text{-RSB}}$ as the solution to a two-dimensional minimization problem over q_0 and r . Importantly, the finite-capacity condition at 1-RSB is the same as that with RS. For functions in $\mathcal{H}^1(\gamma)$, the resulting minimization problem must usually be solved numerically, hence we give results for only a few tractable examples. RSB does not occur for linear activation functions [16–18,32]. For ReLU, we obtain $\alpha_{1\text{-RSB}} \simeq 2.66428$ at $(q_0^*, r^*) \simeq (0.75716, 16.6374)$, which is consistent with the result of Baldassi *et al.* [11] (see Ref. [33]). For erf, we obtain $\alpha_{1\text{-RSB}} \simeq 2.37500$ at $(q_0^*, r^*) \simeq (0.75463, 7.75682)$. Finally, for the quadratic, we have $\alpha_{1\text{-RSB}} \simeq 3.37466$ at $(q_0^*, r^*) \simeq (0.28452, 6.39299)$. In Fig. 1, we plot the 1-RSB capacity for these activation functions at nonzero margins. The gap between the RS and 1-RSB results for the quadratic is larger than that for erf or ReLU, both in the numerical value of the capacity and in the difference between q_0^* and q_1^* . Though the capacities at 1-RSB are reduced relative to the RS result, their ordering for these activation functions is preserved.

For general activation functions in $\mathcal{H}^1(\gamma)$, we can obtain informative upper bounds on $\alpha_{1\text{-RSB}}$ by considering candidate solutions with fixed values of the interblock overlap q_0 . From $q_0 \uparrow 1$, we have $\alpha_{1\text{-RSB}} \leq \alpha_{\text{RS}}$. As shown in the Supplemental Material [22], we can also obtain an upper bound for $\alpha_{1\text{-RSB}}$ at zero margin as a function of α_{RS} by taking $q_0 = 0$ and optimizing over r alone. For $\alpha_{\text{RS}} \leq 5/2$, these two bounds coincide, while the $q_0 = 0$ bound is tighter for $\alpha_{\text{RS}} > 5/2$. In particular, for $\alpha_{\text{RS}} \gg 1$, this yields $\alpha_{1\text{-RSB}} = \mathcal{O}(\log \alpha_{\text{RS}})$. The $q_0 = 0$ bound allows us to define an accessible region in $(\alpha_{\text{RS}}, \alpha_{1\text{-RSB}})$ space, as illustrated in Fig. 2. Our numerical estimates for the 1-RSB capacities of ReLU, erf, and the quadratic all lie within this allowed area, and are relatively close to the $q_0 = 0$ bound [22].

These bounds suggest that RSB strongly affects the capacity for activation functions with large derivative norm and thus large α_{RS} . This is illustrated by the example of Hermite polynomial activation functions. For $g(x) = \text{He}_k(x)$, we have $\alpha_{\text{RS}}(\kappa=0) = 2k$, hence one can obtain an arbitrarily large, but finite, zero-margin RS capacity by taking $k \gg 1$. However, as shown in the Supplemental Material [22], the 1-RSB capacity grows extremely slowly—sublogarithmically—with degree. This result is sensible given the oscillatory nature of

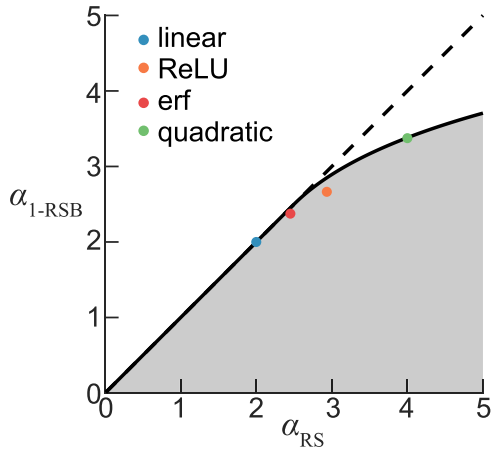


FIG. 2. The accessible region in $(\alpha_{RS}, \alpha_{1-RSB})$ space defined by the $q_0 = 0$ bound. The allowed region is shaded in gray, and the locations of the four example activation functions for which we estimate α_{1-RSB} are indicated by colored dots.

high-degree Hermite polynomials, which one expects to yield a highly nonconvex energy landscape.

Discussion. We have shown that the storage capacity of treelike committee machines with activation functions in $\mathcal{H}^1(\gamma)$ remains bounded in the infinite-width limit. Our results follow from a replica analysis of the Gardner volume, with the capacity given by a simple closed-form expression under a replica-symmetric *Ansatz* and a two-dimensional minimization problem with one-step replica-symmetry breaking. Depending on the activation function, a fully accurate determination of the capacity would likely require higher levels in the Parisi hierarchy of replica-symmetry-breaking *Ansätze* [28]. Furthermore, it can be challenging to rigorously prove that the capacity results obtained using the replica method at any level of the Parisi hierarchy are correct [18,28,32,34,35]. With these caveats in mind, our results begin to elucidate how nonlinear activation functions affect the ability of neural networks to robustly solve classification problems.

Though our analysis focused on a regime in which the input distribution is symmetric, inputs in both biological and artificial neural networks are often only sparsely active [36,37]. Our analysis of the RS capacity can be extended to this regime [22], following Gardner’s [16] work on the perceptron. Provided that the input and target output distributions are not both infinitely sparse, the condition for the capacity to remain finite in the infinite-width limit remains the same. However, if the activation function can be linearized about zero, the zero-margin capacity for a symmetric target distribution decreases to that of the perceptron in the limit of very sparsely active inputs. This holds, for instance, for erf or tanh, but not for ReLU, for which the zero-margin capacity is independent of sparsity. This example illustrates how introducing simple yet realistic forms of data structure can affect pattern storage. Investigating how other forms of data structure affect storage capacity will be an important objective for future work [8,38–40].

In addition to its use as a model system in machine learning, the TCM has been proposed as an abstract model for computation in dendrite-bearing neurons [20,21,41]. In this

application, each hidden unit represents a dendritic unit that integrates some set of synaptic inputs to generate a signal that is transmitted to the soma, which in turn generates a “spike” if the total current exceeds a threshold [20,21]. The most striking form of nonlinearity observed in measurements of dendritic signal processing is the generation of dendritic spikes [42,43]. Though it is difficult to argue that biological nonlinearities can be infinitely sharp, previous works have modeled dendritic spikes using non-weakly-differentiable activation functions [20,21,41]. Our work therefore generates a hypothesis for the functional benefit of dendritic spikes: Nonsmooth dendritic nonlinearities allow the capacity to grow without bound as the number of branches increases and to remain robustly large even when inputs are very sparse. It will be interesting to test this hypothesis using computational models that incorporate greater biophysical realism [21].

The Gardner volume is agnostic to the choice of learning algorithm used to train the weights of the network. This feature makes it a general approach to studying storage capacity, but means that it can provide only limited insight into the practical realizability of the extant solutions [11–14,44]. As a result, it is challenging to directly test theories of the Gardner volume. It is nevertheless possible to experimentally falsify such theories; we have failed to do so [22]. More broadly, this distinction between satisfiability and learnability, combined with its dependence on data and focus on perfect classification, means that the Gardner volume is one of many metrics that should be considered in evaluating activation function choice [9,10,36,44]. In a recent study of least-squares function approximation by wide fully connected networks, Panigrahi *et al.* [9] have shown that the speed and robustness of gradient descent learning is related to activation function smoothness. Their result is suggestively similar to that of this Letter, though it is as yet unclear whether a similar link between smoothness and trainability exists for treelike networks.

In this Letter, we have studied the activation function dependence of the storage capacity of wide TCMs. This network architecture is particularly convenient to study in the infinite-width limit, but it is far removed from the deep networks used in practical applications [5]. As a more realistic model, one could consider a fully connected committee machine (FCM), in which each hidden unit is connected to the full set of inputs. Prior work on such networks with sign activation functions suggests that some qualitative aspects of the behavior of TCMs should still hold true [12,13,45]. However, FCMs possess a symmetry with respect to permutation of the hidden units, which is broken at loads below the RS capacity [12]. This phenomenon and the presence of correlations between hidden units complicate the study of their infinite-width limit. Accurate determination of how FCM storage capacity depends on activation function will therefore require further work, in which the insights developed in this study should prove broadly useful.

J.A.Z.-V. acknowledges support from the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard and the Harvard Quantitative Biology Initiative. C.P. thanks the Harvard Data Science Initiative, Google, and Intel for support.

- [1] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control. Signals, Syst.* **2**, 303 (1989).
- [2] K. Hornik, M. Stinchcombe, H. White *et al.*, Multilayer feed-forward networks are universal approximators, *Neural Netw.* **2**, 359 (1989).
- [3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning requires rethinking generalization, in *5th International Conference on Learning Representations, ICLR* (2017), [arXiv:1611.03530](https://arxiv.org/abs/1611.03530).
- [4] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, Exponential expressivity in deep neural networks through transient chaos, in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16* (Curran Associates, Red Hook, NY, 2016), pp. 3360–3368.
- [5] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature (London)* **521**, 436 (2015).
- [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
- [7] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, *Proc. Natl. Acad. Sci. USA* **116**, 15849 (2019).
- [8] L. Zdeborová, Understanding deep learning is also a job for physicists, *Nat. Phys.* **16**, 602 (2020).
- [9] A. Panigrahi, A. Shetty, and N. Goyal, Effect of activation functions on the training of overparametrized neural nets, in *International Conference on Learning Representations* (2020), [arXiv:1908.05660](https://arxiv.org/abs/1908.05660).
- [10] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions, [arXiv:1710.05941](https://arxiv.org/abs/1710.05941).
- [11] 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).
- [12] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, *Phys. Rev. A* **45**, 4146 (1992).
- [13] 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).
- [14] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, Cambridge, U.K., 2001).
- [15] A. Mozeika, B. Li, and D. Saad, Space of Functions Computed by Deep-Layered Machines, *Phys. Rev. Lett.* **125**, 168301 (2020).
- [16] E. Gardner, The space of interactions in neural network models, *J. Phys. A: Math. Gen.* **21**, 257 (1988).
- [17] E. Gardner and B. Derrida, Optimal storage properties of neural network models, *J. Phys. A: Math. Gen.* **21**, 271 (1988).
- [18] M. Talagrand, *Spin Glasses: A Challenge for Mathematicians: Cavity and Mean Field Models*, Vol. 46 (Springer, Berlin, 2003).
- [19] 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).
- [20] P. Poirazi, T. Brannon, and B. W. Mel, Pyramidal neuron as two-layer neural network, *Neuron* **37**, 989 (2003).
- [21] P. Poirazi and A. Papoutsis, Illuminating dendritic function with computational models, *Nat. Rev. Neurosci.* **21**, 303 (2020).
- [22] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.103.L020301> for the details of the calculations, which includes Refs. [46–51].
- [23] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18* (Curran Associates, Red Hook, NY, 2018), pp. 8571–8580.
- [24] B. Bordelon, A. Canatar, and C. Pehlevan, Spectrum dependent learning curves in kernel regression and wide neural networks, in *Proceedings of the International Conference on Machine Learning*, Vol. 119 (PMLR, 2020), pp. 8135–8145.
- [25] G. Mitchison and R. Durbin, Bounds on the learning capacity of some multi-layer networks, *Biol. Cybern.* **60**, 345 (1989).
- [26] This divergence is slow, with $\alpha_c \sim \sqrt{\log K}$ [19,25]; we provide a detailed discussion of this and other finite-size effects in the Supplemental Material [22].
- [27] D. Pollard, *A User's Guide to Measure Theoretic Probability*, Vol. 8 (Cambridge University Press, Cambridge, U.K., 2002).
- [28] M. Mézard, 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).
- [29] V. I. Bogachev, *Gaussian Measures* (American Mathematical Society, Providence, RI, 1998).
- [30] W. Kibble, An extension of a theorem of Mehler's on Hermite polynomials, *Math. Proc. Cambridge Philos. Soc.* **41**, 12 (1945).
- [31] Y. L. Tong, *The Multivariate Normal Distribution* (Springer, Berlin, 2012).
- [32] M. Shcherbina and B. Tirozzi, Rigorous solution of the Gardner problem, *Commun. Math. Phys.* **234**, 383 (2003).
- [33] In the published version of their Letter, Baldassi *et al.* [11] reported a value of $\alpha_{1\text{-RSB}} \simeq 2.92$. After the appearance of our work in preprint form, they found that this result was incorrect [22]; their revised estimate of $\alpha_{1\text{-RSB}} \simeq 2.6643$ agrees with our results.
- [34] J. Ding and N. Sun, Capacity lower bound for the Ising perceptron, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (ACM, New York, 2019), pp. 816–827.
- [35] B. Aubin, A. Maillard, J. Barbier, F. Krzakala, N. Macris, and L. Zdeborová, The committee machine: Computational to statistical gaps in learning a two-layers neural network, *J. Stat. Mech.: Theory Exp.* (2019) 124023.
- [36] A. Knoblauch, G. Palm, and F. T. Sommer, Memory capacities for synaptic and structural plasticity, *Neural Comput.* **22**, 289 (2010).
- [37] B. Willmore and D. J. Tolhurst, Characterizing the sparseness of neural codes, *Netw., Comput. Neural Syst.* **12**, 255 (2001).
- [38] T. Shinzato and Y. Kabashima, Perceptron capacity revisited: Classification ability for correlated patterns, *J. Phys. A: Math. Theor.* **41**, 324013 (2008).
- [39] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Statistical learning theory of structured data, *Phys. Rev. E* **102**, 032119 (2020).
- [40] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modeling the Influence of Data Structure on Learning in Neural Networks, *Phys. Rev. X* **10**, 041044 (2020).
- [41] D. Breuer, M. Timme, and R.-M. Memmesheimer, Statistical Physics of Neural Systems with Nonadditive Dendritic Coupling, *Phys. Rev. X* **4**, 011053 (2014).

- [42] A. Gidon, T. A. Zolnik, P. Fidzinski, F. Bolduan, A. Papoutsis, 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).
- [43] A. Payeur, J.-C. Béïque, and R. Naud, Classes of dendritic information processing, *Curr. Opin. Neurobiol.* **58**, 78 (2019).
- [44] E. Malach and S. Shalev-Shwartz, Is deeper better only when shallow is good? in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (Curran Associates, Red Hook, NY, 2019), pp. 6429–6438.
- [45] R. Urbanczik, Storage capacity of the fully-connected committee machine, *J. Phys. A: Math. Gen.* **30**, L387 (1997).
- [46] J. E. Kolassa, *Series Approximation Methods in Statistics*, Vol. 88 (Springer, Berlin, 2006).
- [47] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Vol. 55 (U.S. Government Printing Office, Washington, D.C., 1948).
- [48] R. H. Byrd, J. C. Gilbert, and J. Nocedal, A trust region method based on interior point techniques for nonlinear programming, *Math. Program.* **89**, 149 (2000).
- [49] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (CRC Press, Boca Raton, FL, 1991).
- [50] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from [tensorflow.org](https://www.tensorflow.org).
- [51] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).

Supplemental Material for “Activation function dependence of the storage capacity of treelike neural networks”

Jacob A. Zavatone-Veth^{1,*} and Cengiz Pehlevan^{2,3,†}

¹*Department of Physics, Harvard University, Cambridge, Massachusetts 02138, USA*

²*John A. Paulson School of Engineering and Applied Sciences,
Harvard University, Cambridge, Massachusetts 02138, USA*

³*Center for Brain Science, Harvard University, Cambridge, Massachusetts 02138, USA*

CONTENTS

| | |
|---|----|
| A. Gaussian measures, Hermite polynomials, and weak differentiability | 1 |
| B. Finite-size effects in treelike committee machines | 2 |
| C. The Gardner volume of the treelike committee machine | 3 |
| D. Replica-symmetric solution | 5 |
| 1. The replica-symmetric quenched free entropy | 6 |
| 2. The replica-symmetric capacity | 7 |
| E. One-step replica-symmetry-breaking solution | 9 |
| 1. The 1-RSB quenched free entropy | 9 |
| 2. The 1-RSB capacity | 11 |
| F. Computation of the capacity for common activation functions | 13 |
| 1. The rectified linear unit | 14 |
| 2. The Gauss error function | 15 |
| 3. The quadratic | 15 |
| 4. The hyperbolic tangent and logistic | 15 |
| 5. The “Swish” and “Mish” functions | 16 |
| 6. Hermite polynomials | 16 |
| G. The replica-symmetric capacity with sparsely active inputs | 17 |
| 1. The rectified linear unit | 19 |
| 2. Analytic activation functions | 20 |
| H. Numerical experiments | 22 |
| References | 23 |

Appendix A: Gaussian measures, Hermite polynomials, and weak differentiability

In this appendix, we review relevant background material from the theory of Gaussian measures. Our discussion is a specialization of the more general discussion in Chapter 1 of Bogachev [1] to the one-dimensional case. We merely seek to summarize the relevant definitions and results, and will not attempt to provide rigorous proofs.

We let γ be the standard Gaussian probability measure on \mathbb{R} , which has density $\exp(-x^2/2)/\sqrt{2\pi}$ with respect to Lebesgue measure. We let $\mathcal{L}^2(\gamma)$ be the Lebesgue space of functions on \mathbb{R} that are square-integrable with respect to

* jzavatoneveth@g.harvard.edu

† cpehlevan@seas.harvard.edu

γ , and, for brevity, denote the norm on this space as $\|\cdot\|_\gamma$. The natural orthonormal basis for $\mathcal{L}^2(\gamma)$ is given by the set of Hermite polynomials $\{\text{He}_k\}_{k=0}^\infty$, which can be defined by the formula

$$\text{He}_k(x) = \frac{(-1)^k}{\sqrt{k!}} \exp\left(\frac{x^2}{2}\right) \frac{d^k}{dx^k} \exp\left(-\frac{x^2}{2}\right). \quad (\text{A1})$$

The Hermite polynomials satisfy the recurrence relation

$$\text{He}'_k(x) = \sqrt{k} \text{He}_{k-1}(x) = x \text{He}_k(x) - \sqrt{k+1} \text{He}_{k+1}(x) \quad (\text{A2})$$

for $k \geq 1$, with $\text{He}_0 \equiv 1$. For a given function $g \in \mathcal{L}^2(\gamma)$, we define its Fourier-Hermite coefficients

$$g_k = \int g \text{He}_k d\gamma, \quad (\text{A3})$$

and the Fourier-Hermite series

$$g(x) = \sum_{k=0}^{\infty} g_k \text{He}_k(x), \quad (\text{A4})$$

which is guaranteed to converge in mean-square by the fact that $\|g\|_\gamma^2 = \sum_{k=0}^{\infty} g_k^2$ is finite.

Then, using the recurrence relation $\text{He}'_k(x) = \sqrt{k} \text{He}_{k-1}(x)$ and the fact that $\text{He}'_0 \equiv 0$, we can express the l^{th} weak derivative of g as a formal Fourier-Hermite series

$$g^{(l)}(x) = \sum_{k=l}^{\infty} g_k \sqrt{(k)_l} \text{He}_{k-l}(x), \quad (\text{A5})$$

where $(k)_r = k(k-1)\cdots(k-r+1)$ is the falling factorial. If, for some $r \geq 0$, the sum

$$\|g^{(r)}\|_\gamma^2 = \sum_{k=r}^{\infty} (k)_r g_k^2 \quad (\text{A6})$$

is finite, then the Fourier-Hermite series for g and its weak derivatives up to order r converge in mean-square. The class of functions satisfying this condition is the Sobolev class $\mathcal{H}^r(\gamma)$, which has Sobolev norm

$$\|g\|_{\mathcal{H}^r(\gamma)} = \left(\sum_{l=0}^r \|g^{(l)}\|_\gamma^2 \right)^{1/2}. \quad (\text{A7})$$

Having defined $\mathcal{H}^r(\gamma)$ in terms of Fourier-Hermite expansions, we now connect this definition to a more generic notion of weak differentiability. Let $\mathcal{C}_0^\infty(\mathbb{R})$ be the set of all infinitely-differentiable functions with compact support, and let $p \geq 1$. For a locally integrable function f , we define its weak derivative f' as a locally integrable function that satisfies the integration by parts formula

$$\int_{\mathbb{R}} \phi'(x) f(x) dx = - \int_{\mathbb{R}} \phi(x) f'(x) dx \quad (\text{A8})$$

for every $\phi \in \mathcal{C}_0^\infty(\mathbb{R})$. The subset of functions in $\mathcal{L}^2(\mathbb{R})$ with weak derivatives up to order r of finite \mathcal{L}^2 norm forms the Sobolev class $\mathcal{H}^r(\mathbb{R})$. We can then define the class $\mathcal{H}_{\text{loc}}^r(\mathbb{R})$ as the set of all functions f on \mathbb{R} such that $\phi f \in \mathcal{H}^r(\mathbb{R})$ for all $\phi \in \mathcal{C}_0^\infty(\mathbb{R})$. $\mathcal{H}^r(\gamma)$ coincides with the class of all functions $f \in \mathcal{H}_{\text{loc}}^r(\mathbb{R})$ such that f and its weak derivatives up to order r have finite $\mathcal{L}^2(\gamma)$ norm, and the corresponding weak derivatives coincide as well. In one dimension, the criterion that the $(r-1)^{\text{th}}$ derivative is differentiable almost everywhere and is equal almost everywhere to the Lebesgue integral of its derivative implies the required weak differentiability condition. Furthermore, by Rademacher's theorem, every function that is locally Lipschitz continuous belongs to $\mathcal{H}_{\text{loc}}^1(\mathbb{R})$.

Appendix B: Finite-size effects in treelike committee machines

The treelike committee machine has three relevant scales: the total number of inputs N , the number of hidden units K , and the number of inputs per hidden unit N/K . We note that we implicitly assume that $N \geq K$ throughout,

and ignore whether these three scales are truly integer valued. In our calculations, we consider a limit in which N is first taken to infinity for fixed K (hence N/K tends to infinity), and then K is taken to infinity. In this appendix, we discuss how this limit relates to alternative infinite-size limits, and how finite size effects in each of these scales might affect our results.

For finite N , a phase transition in the Gardner volume is not possible. Instead, the fraction of realizable dichotomies decays smoothly from one to zero with increasing load. A rigorous analysis of this effect for the perceptron with zero margin ($\kappa = 0$) is provided by Cover's theorem [2]. Cover's theorem identifies the critical load α_c for finite N as that for which half of all dichotomies are realizable, and shows that this value is independent of N . As $N \rightarrow \infty$, the decay at α_c becomes infinitely sharp. The combinatorial methods used by Cover are not easily extensible to multilayer networks [2–4], rendering rigorous analysis of finite- N effects more difficult. However, one expects these finite-size effects to be qualitatively similar even in two-layer models [3, 5–7].

Taking the limit $K \rightarrow \infty$ substantially simplified our calculations, as the resulting limiting distribution of the output preactivation is Gaussian. One could systematically study finite- K corrections to this limit by expanding the distribution of the output preactivation in powers of $K^{-1/2}$ as an Edgeworth series [8]. However, studying the behavior of networks with small K is analytically challenging for general choices of activation function, as one must deal directly with the full distributions of the hidden unit activations rather than just their first few cumulants.

For sign activation functions, one can derive a combinatorial expression for the finite- K capacity thanks to the fact that the required distributions are discrete [5, 7, 9]. In particular, the capacity is a monotonically increasing function of K , and scales with $\sqrt{\log K}$ for $K \gg 1$ [3, 5, 7, 9]. Such a simplification is not in general possible for weakly differentiable activation functions, as the resulting hidden unit activation distributions will contain a continuous component provided that the activation function is not almost-everywhere constant.

The fact that small- K output preactivation distributions can have substantial qualitative differences from the $K \rightarrow \infty$ Gaussian limit is illustrated by the simple example of ReLU activation functions. If there are only a few hidden units, a substantial fraction of the total probability mass will be concentrated as a Dirac mass at $-\vartheta$, which is a measure-zero event in the infinite-width limit. Such considerations suggest that the behavior of treelike committee machines with only a few hidden units may differ noticeably from that of infinitely-wide networks.

Finally, given a generative model in which the components of the input patterns are independent and identically distributed with mean zero and unit variance, the mean field theory depends on the number of inputs to each hidden unit N/K only through the fact that the distribution of preactivations is taken to be Gaussian. If N/K were finite, this should be a reasonable approximation provided that this ratio is not too small. Alternatively, one could simply take the input distribution to be Gaussian, which can be technically convenient if one aims to provide rigorous proofs [10]. Heuristically, this suggests that our results should carry over to an alternative thermodynamic limit in which one simultaneously takes $N, K \rightarrow \infty$ with a large but fixed integer ratio $N/K = \mathcal{O}(1)$.

Appendix C: The Gardner volume of the treelike committee machine

In this appendix, we give a detailed account of the computation of the Gardner volume of the treelike committee machine using the replica method. As described in the main text, the treelike committee machine [5–7, 11] is a two-layer neural network with a total of N inputs divided into disjoint groups of N/K among K hidden units, with output

$$y(\mathbf{x}; \{\mathbf{w}_j\}, \mathbf{v}, \vartheta) = \text{sign}(s(\mathbf{x}; \{\mathbf{w}_j\}, \mathbf{v}, \vartheta)) \quad (\text{C1})$$

for

$$s(\mathbf{x}; \{\mathbf{w}_j\}, \mathbf{v}, \vartheta) = \frac{1}{\sqrt{K}} \sum_{j=1}^K v_j g\left(\frac{\mathbf{w}_j \cdot \mathbf{x}_j}{\sqrt{N/K}}\right) - \vartheta, \quad (\text{C2})$$

where $\mathbf{x}_j \in \mathbb{R}^{N/K}$ is the vector of inputs to the j^{th} hidden unit, $\{\mathbf{w}_j \in \mathbb{R}^{N/K}\}_{j=1}^K$ are the hidden unit weight vectors, $\mathbf{v} \in \mathbb{R}^K$ is the fixed readout weight vector, g is the activation function, and $\vartheta \in \mathbb{R}$ is a threshold. We want to characterize the ability of this network to classify a dataset of P independent and identically distributed random examples $\{(\mathbf{x}^\nu, y^\nu)\}_{\nu=1}^P$, where $\mathbf{x}^\nu \in \{-1, 1\}^N$ and $y^\nu \in \{-1, 1\}$, in terms of the Gardner volume [12, 13]

$$Z_{N,P,K} = \int d\rho(\{\mathbf{w}_j\}) \prod_{\nu=1}^P \Theta(y^\nu s(\mathbf{x}^\nu; \{\mathbf{w}_j\}, \mathbf{v}, \vartheta) - \kappa), \quad (\text{C3})$$

where ρ is a measure on the space of hidden unit weights. We will compute the limiting quenched free entropy per weight f in the sequential limit $N, P \rightarrow \infty$, $K \rightarrow \infty$, with load $N/P \rightarrow \alpha \in (0, \infty)$ using the replica trick as

$$f \equiv \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} f_{N, \alpha N, K} = \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}_{\mathbf{x}, y} \log Z_{N, \alpha N, K} = \lim_{n \downarrow 0} \lim_{K \rightarrow \infty} \lim_{N \rightarrow \infty} \frac{1}{nN} \log \mathbb{E}_{\mathbf{x}, y} Z_{N, \alpha N, K}^n, \quad (\text{C4})$$

where $\mathbb{E}_{\mathbf{x}, y}$ denotes expectation over the quenched Bernoulli disorder represented by the dataset.

We take the elements of \mathbf{x}^ν to be independent and identically distributed, with equal probability of being positive or negative. We allow the distribution of y^ν to be asymmetric, with $\mathbb{P}(y^\nu = +1) = 1 - \mathbb{P}(y^\nu = -1) = p$ for some $p \in [0, 1]$. We consider the case of spherical weights [5, 7, 11–13], in which the hidden unit weight vectors are uniformly distributed on the N/K -sphere of radius $(N/K)^{1/2}$. The total volume of weight space, which determines the normalizing constant required to make ρ a probability measure, is then $S_{N/K}^K$, where

$$S_D \equiv \frac{2\pi^{D/2}}{\Gamma(D/2)} D^{(D-1)/2} \quad (\text{C5})$$

is the surface area of the D -dimensional sphere of radius \sqrt{D} . We will assume that $\|\mathbf{v}\|_2 = \sqrt{K}$, but will not initially impose further conditions on the readout weights or threshold. Finally, we assume that $g \in \mathcal{L}^2(\gamma)$.

Introducing replicas indexed by $a = 1, \dots, n$, we can write the n^{th} quenched moment of the Gardner volume as

$$\mathbb{E}_{\mathbf{x}, y} Z^n = \mathbb{E}_{\mathbf{x}, y} \int \prod_a d\rho(\{\mathbf{w}_j^a\}) \prod_{a, \nu} \Theta \left(y^\nu \left[\frac{1}{\sqrt{K}} \sum_j v_j g \left(\frac{\mathbf{w}_j^a \cdot \mathbf{x}_j^\nu}{\sqrt{N/K}} \right) - \vartheta \right] - \kappa \right). \quad (\text{C6})$$

We observe immediately that the fact that the different patterns are independent and identically distributed implies that

$$\mathbb{E}_{\mathbf{x}, y} Z^n = \int \prod_a d\rho(\{\mathbf{w}_j^a\}) \left[\mathbb{E}_{\mathbf{x}, y} \prod_a \Theta \left(y \left[\frac{1}{\sqrt{K}} \sum_j v_j g \left(\frac{\mathbf{w}_j^a \cdot \mathbf{x}_j}{\sqrt{N/K}} \right) - \vartheta \right] - \kappa \right) \right]^n, \quad (\text{C7})$$

allowing us to simplify our notation by eliminating the pattern index ν . We now consider the local fields

$$h_j^a \equiv \sqrt{\frac{K}{N}} \mathbf{w}_j^a \cdot \mathbf{x}_j, \quad (\text{C8})$$

which have mean zero and covariance

$$\text{cov}(h_j^a, h_l^b) = \delta_{jl} \frac{K}{N} \mathbf{w}_j^a \cdot \mathbf{x}_j. \quad (\text{C9})$$

In this setting, the natural order parameters are the Edwards-Anderson (EA) order parameters [12, 14, 15]

$$q_j^{ab} \equiv \frac{K}{N} \mathbf{w}_j^a \cdot \mathbf{w}_j^b \quad (a \neq b), \quad (\text{C10})$$

which measure the overlap between the weight vectors of each hidden unit in two different replicas. As we have chosen the weight vectors to lie on the sphere, the self-overlap of each hidden unit is fixed to unity, and the EA order parameters are bounded between negative one and one. In terms of the EA order parameters, we have

$$\text{cov}(h_j^a, h_l^b) = \delta_{jl} [\delta_{ab} + q_j^{ab} (1 - \delta_{ab})]. \quad (\text{C11})$$

Then, as each of the local fields is the sum of N/K independent random variables and their covariance is finite, by central limit theorem they converge in distribution as $N \rightarrow \infty$ for any fixed K to a multivariate Gaussian with the same mean and covariance [16, 17]. We note that this limiting result would alternatively follow by inserting Fourier representations of the delta function to enforce the definition of the variables h_j^a , evaluating the averages over the inputs, and expanding the result to lowest order in $1/N$.

We then define the function

$$G_1(\{q_j^{ab}\}) \equiv \frac{1}{n} \log \mathbb{E}_y \mathbb{E}_h \prod_a \Theta \left(y \left[\frac{1}{\sqrt{K}} \sum_j v_j g(h_j^a) - \vartheta \right] - \kappa \right), \quad (\text{C12})$$

where the average \mathbb{E}_h is taken over the q_j^{ab} -dependent Gaussian distribution of the local fields. Introducing Lagrange multipliers \hat{q}_j^{ab} to enforce the definitions of the order parameters q_j^{ab} , we obtain

$$\mathbb{E}_{\mathbf{x},y} Z^n = \int \prod_{b<a,j} \frac{dq_j^{ab} d\hat{q}_j^{ab}}{2\pi i K/N} \exp\left(-\frac{N}{K} \sum_{b<a,j} q_j^{ab} \hat{q}_j^{ab} + Nn\alpha G_1(\{q_j^{ab}\})\right) \int \prod_a d\rho(\{\mathbf{w}_j^a\}) \exp\left(\sum_{b<a,j} \hat{q}_j^{ab} \mathbf{w}_j^a \cdot \mathbf{w}_j^b\right), \quad (\text{C13})$$

where we have, by convention, rescaled the Lagrange multipliers to absorb the factor of K/N in the definition of the order parameters [12]. With the choice that the weight vectors of each branch are uniformly distributed on the N/K -sphere of radius $\sqrt{N/K}$, the integral over the weights expands as

$$\frac{1}{S_{N/K}^K} \int \prod_{a,j} d\mathbf{w}_j^a \left[\prod_{a,j} \delta\left(\|\mathbf{w}_j^a\|^2 - \frac{N}{K}\right) \right] \exp\left(\sum_{b<a,j} \hat{q}_j^{ab} \mathbf{w}_j^a \cdot \mathbf{w}_j^b\right). \quad (\text{C14})$$

To enforce this normalization constraint, we introduce Lagrange multipliers \hat{E}_j^a , allowing us to factor the integrals over the input dimensions of each branch. Furthermore, we note that

$$\lim_{N \rightarrow \infty} \frac{K}{N} \log S_{N/K} = \frac{1 + \log(2\pi)}{2} \quad (\text{C15})$$

for any fixed K . Then, defining the function

$$\begin{aligned} G_2(\{q_j^{ab}\}, \{\hat{q}_j^{ab}\}, \{\hat{E}_j^a\}) &\equiv \frac{1}{2nK} \sum_{a,j} \hat{E}_j^a - \frac{1}{nK} \sum_{b<a,j} q_j^{ab} \hat{q}_j^{ab} - \frac{1 + \log(2\pi)}{2} \\ &+ \frac{1}{nK} \sum_j \log \int \prod_a d\mathbf{w}_j^a \exp\left(-\frac{1}{2} \sum_a \hat{E}_j^a (w_j^a)^2 + \sum_{b<a} \hat{q}_j^{ab} w_j^a w_j^b\right), \end{aligned} \quad (\text{C16})$$

we can write

$$\mathbb{E}_{\mathbf{x},y} Z^n = \int \prod_{a,j} \frac{d\hat{E}_j^a}{4\pi i} \int \prod_{b<a,j} \frac{dq_j^{ab} d\hat{q}_j^{ab}}{2\pi i K/N} \exp\left(Nn \left[\alpha G_1(\{q_j^{ab}\}) + G_2(\{q_j^{ab}\}, \{\hat{q}_j^{ab}\}, \{\hat{E}_j^a\})\right]\right) \quad (\text{C17})$$

in the limit $N \rightarrow \infty$. In this limit, we can evaluate the integrals over the order parameters and the Lagrange multipliers using the method of steepest descent, which yields an expression for the quenched free entropy as

$$f = \lim_{n \downarrow 0} \text{extr}_{\{q_j^{ab}\}, \{\hat{q}_j^{ab}\}, \{\hat{E}_j^a\}} \left\{ \alpha G_1(\{q_j^{ab}\}) + G_2(\{q_j^{ab}\}, \{\hat{q}_j^{ab}\}, \{\hat{E}_j^a\}) \right\}. \quad (\text{C18})$$

We note that the function G_1 represents the energetic contribution to the quenched free entropy, while the function G_2 represents the entropic contribution.

Appendix D: Replica-symmetric solution

In this appendix, we study the quenched free entropy derived in Appendix C using a replica-symmetric ansatz [6, 12–15]. In addition, as we expect the different hidden units to be equivalent to one another after averaging over patterns [5, 7], we make the ansatz that the order parameters are the same across all hidden units. Concretely, we make the ansatz

$$\begin{cases} \hat{E}_j^a = \hat{E} \\ q_j^{ab} = q \\ \hat{q}_j^{ab} = \hat{q} \end{cases} \quad (\text{D1})$$

which substantially simplifies the saddle point equations.

1. The replica-symmetric quenched free entropy

Considering the entropic contribution G_2 , we immediately obtain the simplification

$$G_2 = \frac{1}{2}\hat{E} - \frac{1}{2}(n-1)q\hat{q} + \frac{1}{n} \log \int d^n w \exp\left(-\frac{1}{2}\mathbf{w}^T \mathbf{A} \mathbf{w}\right) - \frac{1 + \log(2\pi)}{2}, \quad (\text{D2})$$

where we have defined the $n \times n$ matrix

$$\mathbf{A} = (\hat{E} + \hat{q})\mathbf{I}_n - \hat{q}\mathbf{1}_n\mathbf{1}_n^T. \quad (\text{D3})$$

Applying the matrix determinant lemma, we have

$$\det \mathbf{A} = (\hat{E} + \hat{q})^n \left(1 - \frac{\hat{q}}{\hat{E} + \hat{q}}n\right), \quad (\text{D4})$$

hence we find that

$$\lim_{n \downarrow 0} \frac{1}{n} \log \int d^n w \exp\left(-\frac{1}{2}\mathbf{w}^T \mathbf{A} \mathbf{w}\right) = \frac{1}{2} \left[\log \frac{2\pi}{\hat{E} + \hat{q}} + \frac{\hat{q}}{\hat{E} + \hat{q}} \right]. \quad (\text{D5})$$

Then, as we would expect, the entropic contribution to the quenched free entropy is the same as for a perceptron [12], yielding

$$\lim_{n \downarrow 0} G_2 = \frac{1}{2} \left[\hat{E} + q\hat{q} + \log \frac{1}{\hat{E} + \hat{q}} + \frac{\hat{q}}{\hat{E} + \hat{q}} - 1 \right]. \quad (\text{D6})$$

As the Lagrange multipliers \hat{E} and \hat{q} appear only in G_2 , they can easily be eliminated from the saddle point equations, yielding

$$\lim_{n \downarrow 0} G_2 = \frac{1}{2} \left[\frac{q}{1-q} + \log(1-q) \right] \quad (\text{D7})$$

at the replica-symmetric saddle point.

We now consider the energetic term G_1 . With the replica- and branch-symmetric ansatz, the covariance matrix of the Gaussian-distributed local fields simplifies to

$$\text{cov}(h_j^a, h_l^b) = \delta_{jl}[\delta_{ab} + q(1 - \delta_{ab})]. \quad (\text{D8})$$

Then, as the local fields h_j are independent, the internal fields

$$s^a \equiv \frac{1}{\sqrt{K}} \sum_j v_j g(h_j^a) - \vartheta \quad (\text{D9})$$

are the sums of K independent random variables, with mean

$$\mu \equiv \mathbb{E}_h s^a = \left[\frac{1}{\sqrt{K}} \sum_j v_j \right] (\mathbb{E}g) - \vartheta \quad (\text{D10})$$

and covariance

$$\text{cov}(s^a, s^b) = \frac{1}{K} \sum_{j,l=1}^K v_j v_l \text{cov}(g(h_j^a), g(h_l^b)) = \frac{1}{K} \sum_{j=1}^K v_j^2 \text{cov}(g(h_j^a), g(h_j^b)) = \text{cov}(g(h^a), g(h^b)) \quad (\text{D11})$$

where we have used the fact the fields are independent and identically distributed across branches and the assumption that $\|\mathbf{v}\|_2^2 = K$. Then, if $\text{cov}(g(h^a), g(h^b))$ is finite, which holds for any $g \in \mathcal{L}^2(\gamma)$, then the classical central limit theorem implies that the internal fields s^a converge in distribution as $K \rightarrow \infty$ to a multivariate Gaussian with the mean and variance given above [16, 17].

Defining the quantity

$$\sigma^2 \equiv \text{var}[g(x) : x \sim \mathcal{N}(0, 1)] = \|g\|_\gamma^2 - (\mathbb{E}g)^2 \quad (\text{D12})$$

and the effective order parameter

$$\tilde{q} = \text{cov} \left[g(x), g(y) : \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} 1 & q \\ q & 1 \end{bmatrix} \right) \right], \quad (\text{D13})$$

we can see from the joint distribution of the local fields h^a that we can write the covariance matrix of the internal fields as

$$\text{cov}(s^a, s^b) = \sigma^2 \delta_{ab} + \tilde{q}(1 - \delta_{ab}). \quad (\text{D14})$$

Then, we can expand the first average in $\exp(nG_1)$ in terms of the joint characteristic function of s^a as

$$\int \prod_a \frac{ds^a d\hat{s}^a}{2\pi} \left[\prod_a \Theta(-s^a - \kappa) \right] \exp \left(i \sum_a s^a \hat{s}^a - \frac{1}{2}(\sigma^2 - \tilde{q}) \sum_a (\hat{s}^a)^2 - \frac{1}{2}\tilde{q} \left[\sum_a \hat{s}^a \right]^2 + i\mu \sum_a \hat{s}^a \right). \quad (\text{D15})$$

To evaluate the remaining integrals, we perform a Hubbard-Stratonovich transformation, which is defined via the integral identity [6]

$$\exp \left(-\frac{1}{2}x^2 \right) = \int d\gamma(z) \exp(-ixz), \quad (\text{D16})$$

to decouple the replicas at the expense of introducing an auxiliary field z . From a statistical point of view, we can see that this has the effect of shifting the mean of s^a from μ to $\mu + \sqrt{\tilde{q}}z$, which yields

$$\int d\gamma(z) \left[\int \frac{ds d\hat{s}}{2\pi} \Theta(-s - \kappa) \exp \left(is\hat{s} - \frac{1}{2}(\sigma^2 - \tilde{q})\hat{s}^2 + i(\mu + \sqrt{\tilde{q}}z)\hat{s} \right) \right]^n \quad (\text{D17})$$

$$= \int d\gamma(z) \left[H \left(\frac{\kappa + \mu + \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right) \right]^n, \quad (\text{D18})$$

where $H(z) = \int_z^\infty d\gamma(x)$ is the Gaussian tail distribution function. Analogously, we can see that the second term in $\exp(nG_1)$ can be written in a similar form, yielding

$$\exp(nG_1) = (1-p) \int d\gamma(z) \left[H \left(\frac{\kappa + \mu + \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right) \right]^n + p \int d\gamma(z) \left[H \left(\frac{\kappa - \mu - \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right) \right]^n. \quad (\text{D19})$$

Applying the identity

$$\mathbb{E} \log x = \lim_{n \downarrow 0} \frac{\log(\mathbb{E}x^n)}{n}, \quad (\text{D20})$$

upon passing to the limit $n \downarrow 0$ we obtain the replica-symmetric free entropy

$$f_{\text{RS}} = \text{extr}_q \left\{ (1-p)\alpha \int d\gamma(z) \log H \left(\frac{\kappa + \mu + \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right) + p\alpha \int d\gamma(z) \log H \left(\frac{\kappa - \mu - \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right) + \frac{1}{2} \left[\frac{q}{1-q} + \log(1-q) \right] \right\}. \quad (\text{D21})$$

2. The replica-symmetric capacity

The replica-symmetric capacity α_{RS} is determined by the value of α such that $q \uparrow 1$ [5–7, 11–13]. Solving the saddle point equation for q from the replica-symmetric free entropy f_{RS} for α^{-1} , we have

$$\frac{1}{\alpha_{\text{RS}}} = \lim_{q \uparrow 1} \frac{(1-q)^2}{q} \frac{\partial \tilde{q}}{\partial q} \left[(1-p) \int d\gamma(z) \frac{1}{H(c_+)} \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{c_+^2}{2} \right) \frac{1}{(\sigma^2 - \tilde{q})^{3/2}} \left(\kappa + \mu + \frac{z\sigma^2}{\sqrt{\tilde{q}}} \right) + p \int d\gamma(z) \frac{1}{H(c_-)} \frac{1}{\sqrt{2\pi}} \exp \left(-\frac{c_-^2}{2} \right) \frac{1}{(\sigma^2 - \tilde{q})^{3/2}} \left(\kappa - \mu - \frac{z\sigma^2}{\sqrt{\tilde{q}}} \right) \right], \quad (\text{D22})$$

where, for brevity, we write

$$c_{\pm} \equiv \frac{\kappa \pm \mu \pm \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}}. \quad (\text{D23})$$

We can then see that the finiteness of the replica-symmetric critical capacity depends on the analytic properties of \tilde{q} in the limit $q \uparrow 1$. To study the properties of this limit, we make the change of variables $q = 1 - \varepsilon$. We generically expect $\tilde{q} \uparrow \sigma^2$, but the way in which \tilde{q} approaches σ^2 depends on the activation function. As observed by Baldassi *et al.* [11], for $g(x) = \text{sign}(x)$, $\sigma^2 - \tilde{q} \sim \sqrt{\varepsilon}$, while, for $g(x) = \text{ReLU}(x)$, $\sigma^2 - \tilde{q} \sim \varepsilon$. As shown in the main text, the asymptotic scaling $\sigma^2 - \tilde{q} \sim \varepsilon$ holds for all $g \in \mathcal{H}^1(\gamma)$. We thus make the ansatz

$$\tilde{q} \sim \sigma^2 - \beta\varepsilon^\ell \quad (\text{D24})$$

for some parameters $\beta, \ell > 0$. Then, the contribution of the first term in the saddle point equation above to α_{RS}^{-1} is given to leading order as

$$\frac{\varepsilon^{1-\ell/2}}{1-\varepsilon} \frac{\ell(1-p)}{\sqrt{\beta}} \int d\gamma(z) \frac{1}{H(c_+)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{c_+^2}{2}\right) \frac{1}{(\sigma^2 - \tilde{q})^{3/2}} \left(\kappa + \mu + \frac{z\sigma^2}{\sqrt{\sigma^2 - \beta\varepsilon^\ell}} \right), \quad (\text{D25})$$

where we have reparameterized c_+ in terms of ε . In the limit $\varepsilon \downarrow 0$, c_+ tends to $+\infty$ if $z \geq -(\kappa + \mu)/\sigma$ and to $-\infty$ otherwise. Noting that

$$\frac{1}{H(x)} \sim \begin{cases} 1 + (2\pi x^2)^{-1/2} \exp(-x^2/2) [1 - x^{-2} + \mathcal{O}(x^{-4})] & x \ll -1 \\ \sqrt{2\pi} x \exp(x^2/2) [1 + x^{-2} + \mathcal{O}(x^{-4})] & x \gg +1 \end{cases}, \quad (\text{D26})$$

we can see that the only non-vanishing contribution comes from the interval $z \geq -(\kappa + \mu)/\sigma$. Thus, to leading order, this term is given as

$$\frac{\varepsilon^{1-\ell}}{1-\varepsilon} \frac{\ell(1-p)}{\beta} \int_{-(\kappa+\mu)/\sigma}^{\infty} d\gamma(z) \left(\kappa + \mu + z\sqrt{\sigma^2 - \beta\varepsilon^\ell} \right) \left(\kappa + \mu + \frac{z\sigma^2}{\sqrt{\sigma^2 - \beta\varepsilon^\ell}} \right), \quad (\text{D27})$$

which we can further approximate as

$$\ell\varepsilon^{1-\ell} \frac{1-p}{\beta} \int_{-(\kappa+\mu)/\sigma}^{\infty} d\gamma(z) (\kappa + \mu + \sigma z). \quad (\text{D28})$$

By an identical procedure, we can derive the leading-order contribution to the second term, in which case the non-vanishing contribution to the integral comes from $z \leq (\kappa - \mu)/\sigma$, yielding

$$\frac{1}{\alpha_{\text{RS}}} = \lim_{\varepsilon \downarrow 0} \ell\varepsilon^{1-\ell} \left[\frac{1-p}{\beta} \int_{-(\kappa+\mu)/\sigma}^{\infty} d\gamma(z) (\kappa + \mu + \sigma z) + \frac{p}{\beta} \int_{-\infty}^{(\kappa-\mu)/\sigma} d\gamma(z) (\kappa - \mu - \sigma z) \right]. \quad (\text{D29})$$

We note that this result can alternatively be obtained using the method of Engel *et al.* [5], which exploits the properties of the function whose extremum with respect to q defines the free entropy f_{RS} to avoid the need to explicitly compute the saddle point equation.

Thus, we can see that the limit $\varepsilon \downarrow 0$ vanishes for $\ell < 1$, implying divergence of the replica-symmetric capacity. If $\ell \geq 1$, which holds for all functions $g \in \mathcal{H}^1(\gamma)$, the capacity remains finite, and we have $\beta = \|g'\|_\gamma^2$. We note that the boundary case $\ell = 1$, which corresponds to non-zero $\|g'\|_\gamma^2$, is special, as the capacity vanishes if $\ell > 1$. For this class of functions, we therefore obtain

$$\frac{1}{\alpha_{\text{RS}}} = \frac{\sigma^2}{\|g'\|_\gamma^2} \left[(1-p) \int_{-(\kappa+\mu)/\sigma}^{\infty} d\gamma(z) \left(\frac{\kappa + \mu}{\sigma} + z \right)^2 + p \int_{-(\kappa-\mu)/\sigma}^{\infty} d\gamma(z) \left(\frac{\kappa - \mu}{\sigma} + z \right)^2 \right]. \quad (\text{D30})$$

By inspection, we can see that if $p = 1/2$ and the output distribution is symmetric, α_{RS} is maximized by taking $\mu = 0$. If the condition $\mu = 0$ holds, the formula above simplifies to

$$\frac{1}{\alpha_{\text{RS}}} = \frac{\sigma^2}{\|g'\|_\gamma^2} \int_{-\kappa/\sigma}^{\infty} d\gamma(z) (\kappa/\sigma + z)^2, \quad (\text{D31})$$

as given in the main text.

Appendix E: One-step replica-symmetry-breaking solution

In this appendix, we consider a one-step replica-symmetry-breaking (1-RSB) ansatz, in which we divide the n replicas into groups of size m , known as the Parisi parameter, and allow the overlaps between groups to differ from the overlaps within groups [5–7, 11, 14, 15]. Again, we assume that the order parameters are translation-invariant across branches. We let q_0 denote the overlaps between replicas in different groups, and q_1 the overlap between replicas within the same group, with corresponding Lagrange multipliers \hat{q}_0 and \hat{q}_1 .

1. The 1-RSB quenched free entropy

With the 1-RSB ansatz, the entropic contribution G_2 simplifies to

$$G_2 = \frac{1}{2}\hat{E} - \frac{1}{2}(n-m)q_0\hat{q}_0 - \frac{1}{2}(m-1)q_1\hat{q}_1 + \frac{1}{n} \log \int d^n w \exp\left(-\frac{1}{2}\mathbf{w}^T \mathbf{C} \mathbf{w}\right) - \frac{1 + \log(2\pi)}{2}, \quad (\text{E1})$$

where we have defined the $n \times n$ block Toeplitz matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \cdots & \mathbf{B} \\ \mathbf{B} & \mathbf{A} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \mathbf{B} & \cdots & & \mathbf{A} \end{pmatrix}, \quad (\text{E2})$$

where the $m \times m$ blocks are defined as

$$\mathbf{A} = (\hat{E} + \hat{q}_1)\mathbf{I}_m - \hat{q}_1 \mathbf{1}_m \mathbf{1}_m^T \quad (\text{E3})$$

and

$$\mathbf{B} = -\hat{q}_0 \mathbf{1}_m \mathbf{1}_m^T, \quad (\text{E4})$$

respectively. Then, as the integral over \mathbf{w} is Gaussian, it can easily be evaluated, yielding

$$\frac{1}{n} \log \int d^n w \exp\left(-\frac{1}{2}\mathbf{w}^T \mathbf{C} \mathbf{w}\right) = \frac{1}{2} \log(2\pi) - \frac{1}{2n} \log \det \mathbf{C}. \quad (\text{E5})$$

To compute the determinant of \mathbf{C} , we will use a convenient lemma. For n/m a power of two, we have

$$\det \mathbf{C} = \det(\mathbf{A} - \mathbf{B})^{n/m-1} \det(\mathbf{A} + (n/m - 1)\mathbf{B}), \quad (\text{E6})$$

which follows from the identity

$$\det \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B} & \mathbf{A} \end{pmatrix} = \det(\mathbf{A} - \mathbf{B}) \det(\mathbf{A} + \mathbf{B}), \quad (\text{E7})$$

and induction on n/m in powers of two. By the matrix determinant lemma, we have

$$\det(\mathbf{A} - \mathbf{B}) = (\hat{E} + \hat{q}_1)^m \left(1 + m \frac{\hat{q}_0 - \hat{q}_1}{\hat{E} + \hat{q}_1}\right) \quad (\text{E8})$$

and

$$\det(\mathbf{A} + (n/m - 1)\mathbf{B}) = (\hat{E} + \hat{q}_1)^m \left(1 + m \frac{(1 - n/m)\hat{q}_0 - \hat{q}_1}{\hat{E} + \hat{q}_1}\right). \quad (\text{E9})$$

Therefore, for n/m a power of two, we have

$$\frac{1}{n} \log \det \mathbf{C} = \log(\hat{E} + \hat{q}_1) + \frac{n-m}{nm} \log \left(1 + m \frac{\hat{q}_0 - \hat{q}_1}{\hat{E} + \hat{q}_1}\right) + \frac{1}{n} \log \left(1 + \frac{(m-n)\hat{q}_0 - m\hat{q}_1}{\hat{E} + \hat{q}_1}\right). \quad (\text{E10})$$

Assuming the validity of analytic continuation to $n \downarrow 0$, we have

$$\lim_{n \downarrow 0} \frac{1}{n} \log \det \mathbf{C} = \log(\hat{E} + \hat{q}_1) - \frac{\hat{q}_0}{\hat{E} + \hat{q}_1 + m(\hat{q}_0 - \hat{q}_1)} + \frac{1}{m} \log \left(\frac{\hat{E} + \hat{q}_1 + m(\hat{q}_0 - \hat{q}_1)}{\hat{E} + \hat{q}_1} \right). \quad (\text{E11})$$

Therefore, we obtain

$$\begin{aligned} \lim_{n \downarrow 0} G_2 = & \frac{1}{2} \hat{E} + \frac{1}{2} q_1 \hat{q}_1 + \frac{1}{2} m(q_0 \hat{q}_0 - q_1 \hat{q}_1) - \frac{1}{2} \\ & + \frac{1}{2} \left[\log \left(\frac{1}{\hat{E} + \hat{q}_1} \right) + \frac{\hat{q}_0}{\hat{E} + \hat{q}_1 + m(\hat{q}_0 - \hat{q}_1)} + \frac{1}{m} \log \left(\frac{\hat{E} + \hat{q}_1}{\hat{E} + \hat{q}_1 + m(\hat{q}_0 - \hat{q}_1)} \right) \right]. \end{aligned} \quad (\text{E12})$$

We note that this result can alternatively be obtained with substantially more algebra by performing many Hubbard-Stratonovich transformations [5]. As the Lagrange multipliers \hat{E} , \hat{q}_0 , and \hat{q}_1 appear only in G_2 , we can eliminate them from the saddle point equations, yielding

$$\lim_{n \downarrow 0} G_2 = \frac{1}{2} \left[\frac{q_0}{1 - q_1 - m(q_0 - q_1)} + \frac{m-1}{m} \log(1 - q_1) + \frac{1}{m} \log(1 - q_1 - m(q_0 - q_1)) \right] \quad (\text{E13})$$

at the 1-RSB saddle point, which reduces to the replica-symmetric result if we take $q_0 = q_1$.

We now consider the energetic contribution G_1 . As in the replica-symmetric case, the central limit theorem implies that the internal fields

$$s^a \equiv \frac{1}{\sqrt{K}} \sum_j v_j g(h_j^a) - \vartheta \quad (\text{E14})$$

converge in distribution to a Gaussian as $K \rightarrow \infty$. Their mean μ is the same as before, but now their covariance is given by the block Toeplitz matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{A} & \mathbf{B} & \cdots & \mathbf{B} \\ \mathbf{B} & \mathbf{A} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \\ \mathbf{B} & \cdots & & \mathbf{A} \end{pmatrix}, \quad (\text{E15})$$

with $m \times m$ blocks

$$\mathbf{A} = (\sigma^2 - \tilde{q}_1) \mathbf{I}_m + \tilde{q}_1 \mathbf{1}_m \mathbf{1}_m^T \quad (\text{E16})$$

and

$$\mathbf{B} = \tilde{q}_0 \mathbf{1}_m \mathbf{1}_m^T, \quad (\text{E17})$$

where $\sigma^2 = \text{var}[g(h)]$ as before and the effective order parameter now takes two values

$$\tilde{q}_j = \text{cov} \left[g(x), g(y) : \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N} \left(0, \begin{bmatrix} 1 & q_j \\ q_j & 1 \end{bmatrix} \right) \right], \quad j = 1, 2. \quad (\text{E18})$$

We now want to understand the structure of the joint characteristic function of the replicated internal fields, in terms of which we will expand G_1 , such that we can decouple replicas by performing appropriate Hubbard-Stratonovich transformations. Introducing Lagrange multipliers $\hat{\mathbf{s}}$ and indexing blocks by Greek superscripts, we have

$$\hat{\mathbf{s}} \cdot \mathbf{C} \hat{\mathbf{s}} = \sum_{\lambda=1}^{n/m} \hat{\mathbf{s}}^\lambda \cdot \mathbf{A} \hat{\mathbf{s}}^\lambda + \sum_{\nu \neq \lambda} \hat{\mathbf{s}}^\nu \cdot \mathbf{B} \hat{\mathbf{s}}^\lambda = (\sigma^2 - \tilde{q}_1) (\hat{\mathbf{s}} \cdot \hat{\mathbf{s}}) + (\tilde{q}_1 - \tilde{q}_0) \sum_{\lambda=1}^{n/m} (\mathbf{1}_m \cdot \hat{\mathbf{s}}^\lambda)^2 + \tilde{q}_0 (\mathbf{1}_n \cdot \hat{\mathbf{s}})^2. \quad (\text{E19})$$

Then, we can see that we will need to perform one Hubbard-Stratonovich transformation to decouple the $(\mathbf{1}_n \cdot \hat{\mathbf{s}})^2$ term at the expense of introducing an auxiliary field z_0 , which has the effect of shifting the mean of s^a from μ to $\mu + \sqrt{\tilde{q}_0} z_0$. To decouple the $(\mathbf{1}_m \cdot \hat{\mathbf{s}}^\lambda)^2$ terms, we introduce n/m auxiliary fields z_1^λ , which further shifts the mean of s^a from $\mu + \sqrt{\tilde{q}_0} z_0$ to $\mu + \sqrt{\tilde{q}_0} z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0} z_1^\lambda$. Then, recognizing that the contribution of each replica within a block

to the integral over the corresponding z_1^λ is identical, and that the contribution of each block to the integral over z_0 is in turn identical, the first average in $\exp(nG_1)$ is given as

$$\int d\gamma(z_0) \left\{ \int d\gamma(z_1) \left[\int \frac{ds d\hat{s}}{2\pi} \Theta(-s - \kappa) \exp \left(is\hat{s} - \frac{1}{2}(\sigma^2 - \tilde{q}_1)\hat{s}^2 + i(\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)\hat{s} \right) \right]^m \right\}^{n/m} \quad (\text{E20})$$

$$= \int d\gamma(z_0) \left\{ \int d\gamma(z_1) \left[H \left(\frac{\kappa + (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^m \right\}^{n/m}. \quad (\text{E21})$$

Analogously, we can see that the second term in $\exp(nG_1)$ can be written in a similar form, yielding

$$\begin{aligned} \exp(nG_1) &= (1-p) \int d\gamma(z_0) \left\{ \int d\gamma(z_1) \left[H \left(\frac{\kappa + (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^m \right\}^{n/m} \\ &\quad + p \int d\gamma(z_0) \left\{ \int d\gamma(z_1) \left[H \left(\frac{\kappa - (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^m \right\}^{n/m}. \end{aligned} \quad (\text{E22})$$

Therefore, passing to the limit $n \downarrow 0$, we obtain the 1-RSB saddle point free entropy

$$\begin{aligned} f_{1\text{-RSB}} &= \text{extr}_{q_0, q_1, m} \left\{ \frac{1}{m} (1-p)\alpha \int d\gamma(z_0) \log \int d\gamma(z_1) \left[H \left(\frac{\kappa + (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^m \right. \\ &\quad + \frac{1}{m} p\alpha \int d\gamma(z_0) \log \int d\gamma(z_1) \left[H \left(\frac{\kappa - (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^m \\ &\quad \left. + \frac{1}{2} \left[\frac{q_0}{1 - q_1 - m(q_0 - q_1)} + \frac{m-1}{m} \log(1 - q_1) + \frac{1}{m} \log(1 - q_1 - m(q_0 - q_1)) \right] \right\}. \end{aligned} \quad (\text{E23})$$

2. The 1-RSB capacity

To determine the capacity under the 1-RSB ansatz, we need to find the value of α such that $q_1 \uparrow 1$. In this limit, we expect $m \downarrow 0$ such that the non-negative quantity

$$r \equiv \frac{m}{1 - q_1} \quad (\text{E24})$$

remains finite [5–7, 11, 14]. We thus re-parameterize the saddle point equations by writing $q_1 = 1 - \varepsilon$ and $m = r\varepsilon$, which yields

$$\begin{aligned} f_{1\text{-RSB}} &= \text{extr}_{q_0, \varepsilon, r} \left\{ \frac{1}{r} (1-p)\alpha \int d\gamma(z_0) \log \int d\gamma(z_1) \left[H \left(\frac{\kappa + (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^{r\varepsilon} \right. \\ &\quad + \frac{1}{r} p\alpha \int d\gamma(z_0) \log \int d\gamma(z_1) \left[H \left(\frac{\kappa - (\mu + \sqrt{\tilde{q}_0}z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0}z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^{r\varepsilon} \\ &\quad \left. + \frac{1}{2} \left[\frac{q_0}{1 + r(1 - q_0 - \varepsilon)} + \varepsilon \log(\varepsilon) + \frac{1}{r} \log(1 + r(1 - q_0 - \varepsilon)) \right] \right\}, \end{aligned} \quad (\text{E25})$$

where \tilde{q}_1 is now a function of ε alone.

To derive a formula for the 1-RSB critical capacity, we follow the method used by Engel *et al.* [5]. This method starts by observing that the quantity inside the curly braces above must vanish in the limit $\varepsilon \downarrow 0$ in order for the extremum with respect to ε to be well-defined in this limit. This condition gives an implicit expression for $\alpha_{1\text{-RSB}}$ as

$$0 = \min_{q_0, r} \left\{ \frac{q_0}{1 + r(1 - q_0)} + \frac{1}{r} \log(1 + r(1 - q_0)) - \frac{2}{r} \alpha_{1\text{-RSB}} \psi(q_0, r) \right\}, \quad (\text{E26})$$

where

$$\psi(q_0, r; \kappa) \equiv - \lim_{\varepsilon \downarrow 0} \left\{ (1-p) \int d\gamma(z_0) \log \int d\gamma(z_1) [H(c_+)]^{r\varepsilon} + p \int d\gamma(z_0) \log \int d\gamma(z_1) [H(c_-)]^{r\varepsilon} \right\}, \quad (\text{E27})$$

and, for brevity, we write

$$c_{\pm} = \frac{\kappa \pm (\mu + \sqrt{\tilde{q}_0} z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0} z_1)}{\sqrt{\sigma^2 - \tilde{q}_1}}. \quad (\text{E28})$$

As $\psi \geq 0$ for all q_0 , r , and κ , we can explicitly express the capacity as

$$\alpha_{1\text{-RSB}}(\kappa) = \min_{q_0, r} \left\{ \frac{1}{2\psi(q_0, r; \kappa)} \left[\frac{r q_0}{1 + r(1 - q_0)} + \log(1 + r(1 - q_0)) \right] \right\}. \quad (\text{E29})$$

We note that one could obtain the same formula for α as a function of the saddle-point values of q_0 and r by solving the saddle-point equation for ε for α and taking the limit $\varepsilon \downarrow 0$.

We must now evaluate the limit $\varepsilon \downarrow 0$ in the definition of ψ . Following our analysis of the RS critical capacity, we focus on the symmetric case $\mu = 0$, and make the ansatz that $\tilde{q}_0 \sim \sigma^2 - \beta \varepsilon^\ell$ for some $\beta, \ell > 0$. The assumption of symmetry yields the simplification

$$\psi(q_0, r; \kappa) = -\lim_{\varepsilon \downarrow 0} \int d\gamma(z_0) \log \int d\gamma(z_1) \left[H \left(\frac{\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0} z_1}{\sqrt{\sigma^2 - \tilde{q}_1}} \right) \right]^{r\varepsilon}. \quad (\text{E30})$$

Expanding the argument of H to leading order in ε , we have

$$\frac{\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\tilde{q}_1 - \tilde{q}_0} z_1}{\sqrt{\sigma^2 - \tilde{q}_1}} \sim \frac{\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\sigma^2 - \tilde{q}_0} z_1}{\sqrt{\beta \varepsilon^\ell}}, \quad (\text{E31})$$

hence the argument of H tends to $+\infty$ for $z_1 \geq -(\kappa + \sqrt{\tilde{q}_0} z_0)/\sqrt{\sigma^2 - \tilde{q}_0}$ and to $-\infty$ otherwise. Noting that

$$H(x) \sim \begin{cases} 1 - (2\pi x^2)^{-1/2} \exp(-x^2/2) [1 - x^{-2} + \mathcal{O}(x^{-4})] & x \ll -1 \\ (2\pi x^2)^{-1/2} \exp(-x^2/2) [1 - x^{-2} + \mathcal{O}(x^{-4})] & x \gg +1 \end{cases}, \quad (\text{E32})$$

we can then write the argument of the logarithm in the definition of ψ to leading order in ε as

$$\begin{aligned} & \int_{-\infty}^{-Q} d\gamma(z_1) \left[1 - \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\beta \varepsilon^{\ell/2}}}{\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\sigma^2 - \tilde{q}_0} z_1} \exp \left(-\frac{(\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\sigma^2 - \tilde{q}_0} z_1)^2}{2\beta \varepsilon^\ell} \right) \right]^{r\varepsilon} \\ & + \int_{-Q}^{\infty} d\gamma(z_1) \left[\frac{1}{\sqrt{2\pi}} \frac{\sqrt{\beta \varepsilon^{\ell/2}}}{\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\sigma^2 - \tilde{q}_0} z_1} \exp \left(-\frac{(\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\sigma^2 - \tilde{q}_0} z_1)^2}{2\beta \varepsilon^\ell} \right) \right]^{r\varepsilon}, \end{aligned} \quad (\text{E33})$$

where we have defined the function

$$Q \equiv \frac{\kappa + \sqrt{\tilde{q}_0} z_0}{\sqrt{\sigma^2 - \tilde{q}_0}} \quad (\text{E34})$$

for brevity. Using the continuity of the logarithm and passing to the limit $\varepsilon \downarrow 0$, this simplifies to

$$\int_{-\infty}^{-Q} d\gamma(z_1) + \int_{-Q}^{\infty} d\gamma(z_1) \lim_{\varepsilon \downarrow 0} \exp \left(-\frac{r(\kappa + \sqrt{\tilde{q}_0} z_0 + \sqrt{\sigma^2 - \tilde{q}_0} z_1)^2}{2\beta} \varepsilon^{1-\ell} \right) \quad (\text{E35})$$

for any $\ell > 0$. If $\ell < 1$, the remaining limit tends to unity, hence the argument of the logarithm tends to unity and ψ vanishes, resulting in a divergent 1-RSB capacity. If $g \in \mathcal{H}^1(\gamma)$, we have $\ell \geq 1$ and $\beta = \|g'\|_\gamma^2$, hence the 1-RSB capacity, like the RS capacity, remains finite. For functions of this class, evaluating the integrals over z_1 , we find that

$$\psi(q_0, r; \kappa) = -\int d\gamma(z_0) \log \left[H(Q) + R \exp \left(-\frac{1}{2} \frac{r(\kappa + \sqrt{\tilde{q}_0} z_0)^2}{\|g'\|_\gamma^2 + r(\sigma^2 - \tilde{q}_0)} \right) H(-Q) \right], \quad (\text{E36})$$

where Q is given as above and we have defined

$$R \equiv \sqrt{\frac{\|g'\|_\gamma^2}{\|g'\|_\gamma^2 + r(\sigma^2 - \tilde{q}_0)}} \quad (\text{E37})$$

for brevity.

To gain some understanding of the behavior of the 1-RSB capacity, we exploit the fact that it is defined as a minimization problem to derive upper bounds by fixing the value of the inter-block overlap q_0 . Trivially, by taking $q_0 \uparrow 1$ we recover the RS result and the bound $\alpha_{1\text{-RSB}} \leq \alpha_{\text{RS}}$. If we instead fix $q_0 = 0$, the problem simplifies dramatically because \tilde{q}_0 vanishes. Denoting this family of candidate capacities by $\alpha_{1\text{-RSB}_0}$, we have

$$\alpha_{1\text{-RSB}_0}(\kappa) = \min_{r \geq 0} \left\{ \frac{\log(1+r)}{2\psi(0, r; \kappa)} \right\}, \quad (\text{E38})$$

where

$$\psi(0, r; \kappa) = -\log \left[H\left(\frac{\kappa}{\sigma}\right) + \sqrt{\frac{\|g'\|_\gamma^2}{\|g'\|_\gamma^2 + \sigma^2 r}} \exp\left(-\frac{1}{2} \frac{\kappa^2 r}{\|g'\|_\gamma^2 + \sigma^2 r}\right) H\left(-\frac{\kappa}{\sigma} \sqrt{\frac{\|g'\|_\gamma^2}{\|g'\|_\gamma^2 + \sigma^2 r}}\right) \right]. \quad (\text{E39})$$

Evaluating this expression at $\kappa = 0$ and noting that $\alpha_{\text{RS}} = 2\|g'\|_\gamma^2/\sigma^2$, we have

$$\alpha_{1\text{-RSB}_0} = \min_{s \geq 0} u(s), \quad (\text{E40})$$

where we have re-expressed the optimization problem in terms of $s \equiv r/\alpha_{\text{RS}}$ and defined the function

$$u(s) \equiv \frac{1}{2} \frac{\log(1 + \alpha_{\text{RS}} s)}{\log(2) - \log(1 + 1/\sqrt{1 + 2s})}. \quad (\text{E41})$$

For all $s > -1/\max\{2, \alpha_{\text{RS}}\}$, $u(s)$ is a continuously differentiable transcendental function of s , with $u(s=0) = \alpha_{\text{RS}}$. For all $0 < \alpha_{\text{RS}} \leq 5/2$, $u'(s)$ is positive for all $s > 0$, hence it is minimized at the boundary. For $\alpha_{\text{RS}} > 5/2$, $u'(s)$ vanishes for some positive s , and the minimum is less than α_{RS} . To obtain an asymptotic bound on the 1-RSB capacity for large α_{RS} , we can use the fact that

$$\alpha_{1\text{-RSB}_0} \leq u(1) = \frac{\log(1 + \alpha_{\text{RS}})}{2\log(3 - \sqrt{3})} \quad (\text{E42})$$

for all α_{RS} to obtain the asymptotic

$$\alpha_{1\text{-RSB}} = \mathcal{O}(\log \alpha_{\text{RS}}). \quad (\text{E43})$$

In summary, the 1-RSB and RS ansätze yield the same conditions on the activation function for the capacity to remain finite in the infinite-width limit. For a given activation function, we can in principle determine $\alpha_{1\text{-RSB}}$ numerically by solving an explicit two-dimensional minimization problem over $q_0 \in [0, 1]$ and $r_0 \in [0, \infty)$, though we do not obtain a simple closed-form solution like that for α_{RS} . Unlike in the calculation of the RS capacity, we cannot in this case avoid the need to compute the effective order parameter $\tilde{q}_0(q_0)$ for generic values of q_0 . We note that expression for the 1-RSB capacity with $g(x) = \text{ReLU}(x)$ from [11] is equivalent in functional form to that presented here. We observe that the first-order conditions on q_0 and r resulting from this minimization are precisely the saddle-point equations for those order parameters. However, Engel *et al.* [5]'s prescription for expressing the 1-RSB capacity as a minimization problem is advantageous relative to simply solving the saddle point equations as it allows one to derive relatively tractable upper bounds. In particular, the $q_0 = 0$ family of candidate solutions yields a tighter upper bound on $\alpha_{1\text{-RSB}}$ than α_{RS} itself, showing that $\alpha_{1\text{-RSB}}$ can grow at most logarithmically with α_{RS} .

Appendix F: Computation of the capacity for common activation functions

In this appendix, we provide details of the computation of the RS and 1-RSB capacities for several commonly-used activation functions. For these examples, we illustrate this minimization problem by plotting its landscape in Figure S1. For comparison purposes, we include the landscape for linear activation functions, for which RSB does not occur [12–14, 18]. Numerical computation of the 1-RSB capacity was performed in MATLAB 9.6. Integrals with respect to Gaussian measure were estimated using 20-point Gauss-Hermite quadrature [19], and minimization was performed using the interior-point solver `fmincon` [20]. These results were then checked against computations performed in MATHEMATICA 12.1 using the numerical integrator `NIntegrate` and minimizer `NMinimize` with 100 digits of internal precision. These methods were also used to generate and cross-check the contour plots in Figure S1.

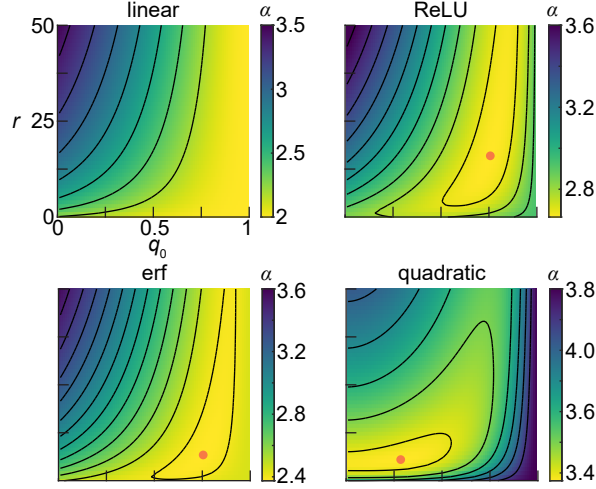


FIG. S1. The landscape of the function whose minimum determines the 1-RSB capacity as a function of the inter-block overlap q_0 and the rescaled Parisi parameter $r \equiv m/(1 - q_1)$ for several example activation functions. In each panel, the value of this function is shown in false color, with the location of minimum indicated by an orange dot.

1. The rectified linear unit

The rectified linear unit $\text{ReLU}(x) \equiv \max\{0, x\}$ is the most commonly-used activation function in modern machine learning applications [21, 22], and has weak derivative $\text{ReLU}'(x) = \Theta(x)$. With our conventions, the Hermite expansion of ReLU is given as

$$\text{ReLU}(x) = \frac{1}{\sqrt{2\pi}} + \frac{1}{2} \text{He}_1(x) + \frac{1}{\sqrt{2\pi}} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{2^k (2k-1)k!} \sqrt{(2k)!} \text{He}_{2k}(x). \quad (\text{F1})$$

By direct computation of the required integrals, we have

$$\sigma^2 = \frac{1}{2} - \frac{1}{2\pi} = \frac{\pi - 1}{2\pi} \quad (\text{F2})$$

and

$$\|\text{ReLU}'\|_{\gamma}^2 = \|\Theta\|_{\gamma}^2 = H(0) = \frac{1}{2}, \quad (\text{F3})$$

hence we recover the result of Baldassi *et al.* [11] that

$$\alpha_{\text{RS}}(\kappa = 0) = \frac{2\pi}{\pi - 1} \simeq 2.93388. \quad (\text{F4})$$

For ReLU, we can express $\tilde{q}(q)$ in closed form by direct integration or by summation of the series expansion resulting from the function's Hermite expansion as

$$\tilde{q}(q) = \frac{q}{4} + \frac{q \arcsin(q) + \sqrt{1 - q^2} - 1}{2\pi}. \quad (\text{F5})$$

Using this formula, we obtain the estimate

$$\alpha_{1\text{-RSB}}(\kappa = 0) \simeq 2.66428 \quad (\text{F6})$$

at $(q_0^*, r^*) \simeq (0.75716, 16.63737)$. This result is consistent with the upper bound $\alpha_{1\text{-RSB}} \lesssim 2.85021$ resulting from the $q_0 = 0$ family of candidate capacities.

This estimate of the 1-RSB capacity of a network with ReLU activations is consistent with the $\alpha_{1\text{-RSB}} \simeq 2.6643$ reported by Baldassi *et al.* [11] in an update to their Letter. Previous versions of their work reported an erroneous value of $\alpha_{1\text{-RSB}} \simeq 2.92$, which does not agree with our estimate of $\alpha_{1\text{-RSB}}$ and exceeds the $q_0 = 0$ bound. To estimate the 1-RSB capacity, they solved the saddle-point equations for q_0 and r numerically rather than directly minimizing over $\alpha_{1\text{-RSB}}$. After the appearance of our work in preprint form, they found that incorrect initialization had allowed the solver to converge on the RS saddle point. Their revised estimate is consistent with ours.

2. The Gauss error function

The Gauss error function

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z dx \exp(-x^2) = 1 - 2H(\sqrt{2}z) \quad (\text{F7})$$

is the most analytically convenient of the commonly-used sigmoidal activation functions. It has the Hermite expansion

$$\operatorname{erf}(x) = \frac{2}{\sqrt{3\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{3^k (2k+1)k!} \sqrt{(2k+1)!} \operatorname{He}_{2k+1}(x), \quad (\text{F8})$$

which allows us to easily obtain the closed-form expressions

$$\tilde{q}(q) = \frac{4}{3\pi} \sum_{k=0}^{\infty} \frac{(2k+1)!}{3^{2k} (k!)^2 (2k+1)^2} q^{2k+1} = \frac{2}{\pi} \arcsin\left(\frac{2}{3}q\right) \quad (\text{F9})$$

and

$$\sigma^2 = \tilde{q}(1) = \frac{2}{\pi} \arcsin\left(\frac{2}{3}\right). \quad (\text{F10})$$

Similarly, we can easily work out that

$$\|\operatorname{erf}'\|_{\gamma}^2 = \int d\gamma(x) \left[\frac{2}{\sqrt{\pi}} \exp(-x^2) \right]^2 = \frac{4}{\sqrt{5}\pi}. \quad (\text{F11})$$

This yields

$$\alpha_{\text{RS}}(\kappa = 0) = \frac{4}{\sqrt{5} \arcsin(2/3)} \simeq 2.45140 \quad (\text{F12})$$

and

$$\alpha_{1\text{-RSB}}(\kappa = 0) \simeq 2.37500, \quad (\text{F13})$$

with $(q_0^*, r^*) \simeq (0.75463, 7.75682)$. This is consistent with our upper bounds for $\alpha_{1\text{-RSB}}$, which in this case simply reduce to the RS capacity as it is less than 2.5.

3. The quadratic

In neuroscientific studies of two-layer network models, expansive activations such as a quadratic are sometimes considered [23]. With $g(x) = x^2$, we can trivially work out that $\tilde{q}(q) = 2q^2$, hence we have $\sigma^2 = 2$ and $\|g'\|_{\gamma}^2 = \|2x\|_{\gamma}^2 = 4$. Thus, we find that

$$\alpha_{\text{RS}}(\kappa = 0) = 4 \quad (\text{F14})$$

and

$$\alpha_{1\text{-RSB}}(\kappa = 0) \simeq 3.37466, \quad (\text{F15})$$

with $(q_0^*, r^*) \simeq (0.28452, 6.39299)$. This result is consistent with the $q_0 = 0$ bound of $\alpha_{1\text{-RSB}} \lesssim 3.38100$.

4. The hyperbolic tangent and logistic

Though it is less analytically convenient than the error function, the hyperbolic tangent and logistic sigmoid $g(x) = (\tanh(x) + 1)/2$ are more commonly used in practical applications [21]. We can numerically evaluate the required integrals, yielding $\|\tanh\|_{\gamma}^2 \simeq 0.39429$ and $\|\tanh'\|_{\gamma}^2 \simeq 0.46440$ and the estimate

$$\alpha_{\text{RS}}(\kappa = 0) \simeq 2.35561. \quad (\text{F16})$$

As the RS capacity is scale- and shift-invariant, the RS capacity of the logistic function is the same as that of the hyperbolic tangent.

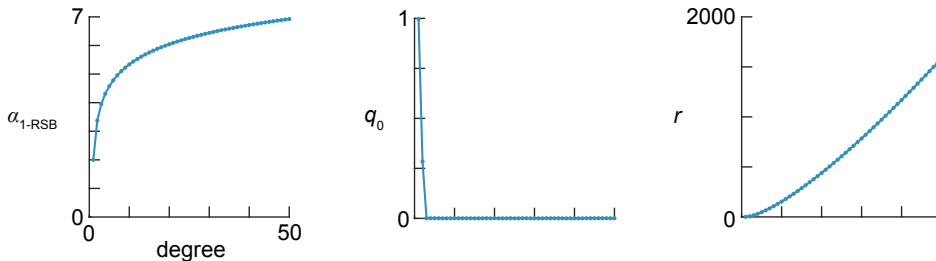


FIG. S2. 1-RSB solutions for Hermite polynomial activation functions of varying degree. In each panel, the abscissa is the degree of the polynomial. The leftmost panel shows the 1-RSB capacity $\alpha_{1\text{-RSB}}$, the middle panel the saddle-point value of the inter-block overlap q_0 , and the rightmost panel the saddle-point value of the rescaled Parisi parameter r .

5. The “Swish” and “Mish” functions

Recent experimental works on deep neural networks have proposed various smooth, non-monotonic functions as alternatives to the rectifier unit. One proposal is the product of a logistic function and a linear function, termed “Swish” [22, 24]:

$$\text{swish}(x; \beta) = \frac{x}{1 + \exp(-\beta x)}, \quad (\text{F17})$$

where β is a positive parameter. Conventionally, β is either fixed to unity or treated as a trainable weight, and yields the limiting behavior $\lim_{\beta \downarrow 0} \text{swish}(x; \beta) = x$, $\lim_{\beta \rightarrow \infty} \text{swish}(x; \beta) = \text{ReLU}(x)$. We can see that α_{RS} is a monotone increasing function of β , which tends to the perceptron result $\alpha_{\text{RS}}(\kappa = 0) = 2$ as $\beta \downarrow 0$ and the ReLU result $\alpha_{\text{RS}}(\kappa = 0) = 2\pi/(\pi - 1)$ as $\beta \rightarrow \infty$. With $\beta = 1$, we have $\|\text{swish}(\cdot; 1)\|_{\gamma}^2 \simeq 0.31308$ and $\|\text{swish}'(\cdot; 1)\|_{\gamma}^2 \simeq 0.37948$, and the estimate

$$\alpha_{\text{RS}}(\kappa = 0) \simeq 2.42416. \quad (\text{F18})$$

Another alternative to ReLU is the “Mish” function, defined as [24]

$$\text{mish}(x) = x \tanh \log(1 + \exp(x)), \quad (\text{F19})$$

for which we have $\|\text{mish}\|_{\gamma}^2 \simeq 0.47908$ and $\|\text{mish}'\|_{\gamma}^2 \simeq 0.39455$, and the estimate

$$\alpha_{\text{RS}}(\kappa = 0) \simeq 2.42852. \quad (\text{F20})$$

6. Hermite polynomials

To illustrate how slowly the 1-RSB capacity grows as a function of the RS capacity, we consider Hermite polynomial activations, i.e.

$$g_k(x) = \text{He}_k(x) \quad (\text{F21})$$

for $k > 0$. For $\text{He}_k(x)$, we of course have $\sigma_k^2 = 1$, $\|\text{He}_k'\|_{\gamma}^2 = k$, and $\tilde{q}_k(q) = q^k$. Thus, the RS capacity at zero margin is simply

$$\alpha_{\text{RS}}(k) = 2k. \quad (\text{F22})$$

As we have the simple expression $\tilde{q}(q) = q^k$ for any k , we can numerically estimate the 1-RSB capacity as a function of degree, yielding the results shown in S2. Importantly, the 1-RSB capacity grows far slower than linearly with k ; in particular, it scales roughly as $\log k$ for $k \gg 1$. For this class of activation functions, the saddle point value of the inter-block overlap q_0 is nearly zero for $k \geq 3$, hence the $q_0 = 0$ upper bound is quite close to the actual estimated value of $\alpha_{1\text{-RSB}}$. We observe that the saddle-point value for the rescaled Parisi parameter r scales roughly as a power law for large k .

Appendix G: The replica-symmetric capacity with sparsely active inputs

In this appendix, we generalize our previous calculation of the replica-symmetric capacity to input distributions with sparsely active input units. Following Gardner [12], we consider a distribution with $\mathbb{P}(x_{jk}^\mu = +1) = (1+r)/2$, where $r = \mathbb{E}x_{jk}^\mu \in [-1, 1]$ is the constrained magnetization of the input patterns. We will focus on the limit in which inputs are very sparse ($r \uparrow 1$), but allow the output distribution to potentially be symmetric (that is, we do not assume that $\mathbb{P}(y^\mu = +1) = p$ is a function of r , as studied by Gardner [12]). We note that, as shown by Gardner [12], it is possible to store more than $\mathcal{O}(N)$ patterns in the limit in which both the input and target output distributions are infinitely sparse. However, such finely tuned matching is not generally realistic in supervised learning tasks. We will follow our previous calculation of the replica-symmetric calculation, noting only where we must make adjustments to account for the non-zero average input magnetization.

The local fields

$$h_j^a \equiv \sqrt{\frac{K}{N}} \mathbf{w}_j^a \cdot \mathbf{x}_j \quad (\text{G1})$$

now have non-zero mean

$$\mathbb{E}_{\mathbf{x}} h_j^a = r \sqrt{\frac{K}{N}} \sum_{k=1}^{N/K} w_{jk}^a \quad (\text{G2})$$

and covariance

$$\text{cov}_{\mathbf{x}}(h_j^a, h_l^b) = (1-r^2) \delta_{kl} \frac{K}{N} \mathbf{w}_j^a \cdot \mathbf{w}_l^b. \quad (\text{G3})$$

We can then see that, in addition to the Edwards-Anderson order parameters $q_j^{ab} \equiv (K/N) \mathbf{w}_j^a \cdot \mathbf{w}_j^b$, we will need to introduce local magnetizations

$$m_j^a \equiv \sqrt{\frac{K}{N}} \sum_{k=1}^{N/K} w_{jk}^a. \quad (\text{G4})$$

As noted by Gardner [12], the effect of the corresponding Lagrange multiplier on the entropic contribution to the saddle-point free entropy can be neglected in the limit $N \rightarrow \infty$, as it is suppressed relative to the contribution from the Lagrange multipliers corresponding to the EA order parameters by a factor of $\sqrt{K/N}$. Thus, the local magnetization affects the free entropy only through its appearance in the energetic term.

Under a replica- and branch-symmetric ansatz, the defining moments of the local fields are given as

$$\mathbb{E} h_j^a = r m \quad (\text{G5})$$

and

$$\text{cov}(h_j^a, h_l^b) = (1-r^2) \delta_{kl} [\delta_{ab} + q(1-\delta_{ab})]. \quad (\text{G6})$$

Then, by comparison to our previous results, we can see that we can map all of our reasoning onto the sparse case provided that we take expectations with respect to the modified distribution. In particular, we will have effective order parameters

$$\tilde{m}(m, r) = \mathbb{E} \left[g(x) : x \sim \mathcal{N}(rm, 1-r^2) \right], \quad (\text{G7})$$

$$\sigma^2(m, r) = \text{var} \left[g(x) : x \sim \mathcal{N}(rm, 1-r^2) \right], \quad (\text{G8})$$

and

$$\tilde{q}(q, m, r) = \text{cov} \left[g(x), g(y) : \begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N} \left(rm \begin{bmatrix} 1 \\ 1 \end{bmatrix}, (1-r^2) \begin{bmatrix} 1 & q \\ q & 1 \end{bmatrix} \right) \right]; \quad (\text{G9})$$

the corresponding quantities in our original calculation are the $r = 0$ special case of these expressions. Defining the average output preactivation

$$\mu \equiv \tilde{m} \frac{1}{\sqrt{K}} \sum_{j=1}^K v_j - \vartheta, \quad (\text{G10})$$

the appropriate generalization of the energetic term in the $K \rightarrow \infty$ limit from our previous calculation is therefore

$$G_1 = (1-p) \int d\gamma(z) \log H \left(\frac{\kappa + \mu + \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right) + p \int d\gamma(z) \log H \left(\frac{\kappa - \mu - \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \right), \quad (\text{G11})$$

where in our previous calculation μ was a constant.

The replica-symmetric free entropy is then given by

$$f_{\text{RS}} = \text{extr}_{q,m} \left\{ \alpha G_1(q, m) + G_2(q) \right\}, \quad (\text{G12})$$

where, as noted above, the entropic term remains unchanged by the introduction of a non-zero magnetization, and is given as

$$G_2(q) = \frac{1}{2} \left[\frac{q}{1-q} + \log(1-q) \right]. \quad (\text{G13})$$

Under suitable smoothness conditions on the effective order parameters, we can then extract the RS capacity as

$$\frac{1}{\alpha_{\text{RS}}} = \lim_{q \uparrow 1} \frac{2(1-q)^2}{q} \frac{\partial G_1}{\partial q} \Big|_{m=m^*}, \quad (\text{G14})$$

where m^* is the solution to the equation

$$\lim_{q \uparrow 1} \frac{\partial G_1}{\partial m} = 0. \quad (\text{G15})$$

Until this point, we have ignored the question of how the local magnetization m should scale with K , which affects how μ scales with K . In particular, divergence of μ corresponds to a trivial committee machine that almost surely predicts the same class for all inputs. In our previous calculation, we handled the scaling of μ post hoc, as for $r = 0$ it is a constant and the subsequent expressions have sensible $|\mu| \rightarrow \infty$ limits. Here, however, μ is a function of the order parameter m . As we are interested in the case in which one changes the sparsity of the input distribution while keeping the output distribution fixed, we will demand that $\mu = \mathcal{O}(1)$. We note that solutions with diverging μ may be optimal if one considers a target distribution that is always either positive or negative, but this is not in general the case. This constraint matches that considered by Gardner [12] in her analysis of the perceptron, where she demanded that the total magnetization

$$M = \frac{1}{\sqrt{N}} \sum_{k=1}^N w_k = \frac{1}{\sqrt{K}} \sum_{j=1}^K m_j = \sqrt{K} m \quad (\text{G16})$$

remain $\mathcal{O}(1)$, corresponding to the scaling $m = \mathcal{O}(K^{-1/2})$.

As we assume that $\sum_{j=1}^K v_j^2 = K$, we should have $|v_j| = \mathcal{O}(1)$, hence $\sum_{j=1}^K v_j = \mathcal{O}(K)$ provided that the readout weights do not exactly sum to zero. We note that this is precisely the case for the classic committee machine readout $v_j = 1$. As the zero-sum case results in $\mu = -\vartheta$ independent of m , we choose to exclude it as it is in this sense trivial. We then define the $\mathcal{O}(1)$ quantity

$$\bar{v} \equiv \frac{1}{K} \sum_{j=1}^K v_j, \quad (\text{G17})$$

in terms of which we have $\mu = \sqrt{K} \bar{v} \tilde{m} - \vartheta$. Following our previous calculation, we choose the threshold such that it cancels the constant component of $\sqrt{K} \bar{v} \tilde{m}$, i.e. we set $\vartheta = \sqrt{K} \bar{v} \tilde{m}_0$, where we have defined $\tilde{m}_0 \equiv \tilde{m}(0, r)$. This

prevents μ from trivially diverging due to the addition of a constant offset to the activations. With these choices, we have $\mu = \sqrt{K\bar{v}}(\tilde{m} - \tilde{m}_0)$. Therefore, to have $\mu = \mathcal{O}(1)$, we must have $\tilde{m} - \tilde{m}_0 = \mathcal{O}(K^{-1/2})$.

We can now write down the $K \rightarrow \infty$ limit of the saddle point equation for m . Defining

$$c_{\pm} \equiv \frac{\kappa \mp \mu \mp \sqrt{\tilde{q}}z}{\sqrt{\sigma^2 - \tilde{q}}} \quad (\text{G18})$$

for brevity, the saddle-point equation for m prior to taking the $q \uparrow 1$ limit is

$$\begin{aligned} 0 = (1-p) \int d\gamma(z) \frac{\sqrt{\sigma^2 - \tilde{q}}\phi(c_-)}{H(c_-)} \left[\sqrt{K\bar{v}} \frac{\partial \tilde{m}}{\partial m} - \frac{\kappa + \mu + \sqrt{\tilde{q}}z}{2(\sigma^2 - \tilde{q})} \frac{\partial \sigma^2}{\partial m} + \frac{\kappa + \mu + \sigma^2 \tilde{q}^{-1/2} z}{2(\sigma^2 - \tilde{q})} \frac{\partial \tilde{q}}{\partial m} \right] \\ + p \int d\gamma(z) \frac{\sqrt{\sigma^2 - \tilde{q}}\phi(c_+)}{H(c_+)} \left[-\sqrt{K\bar{v}} \frac{\partial \tilde{m}}{\partial m} - \frac{\kappa - \mu - \sqrt{\tilde{q}}z}{2(\sigma^2 - \tilde{q})} \frac{\partial \sigma^2}{\partial m} + \frac{\kappa - \mu - \sigma^2 \tilde{q}^{-1/2} z}{2(\sigma^2 - \tilde{q})} \frac{\partial \tilde{q}}{\partial m} \right]. \end{aligned} \quad (\text{G19})$$

All K -dependence in this equation is contained in the $\sqrt{K\bar{v}}$ and μ terms. Demanding that $\mu = \mathcal{O}(1)$, it simplifies to

$$(1-p) \int d\gamma(z) \frac{\sqrt{\sigma^2 - \tilde{q}}\phi(c_-)}{H(c_-)} \frac{\partial \tilde{m}}{\partial m} = p \int d\gamma(z) \frac{\sqrt{\sigma^2 - \tilde{q}}\phi(c_+)}{H(c_+)} \frac{\partial \tilde{m}}{\partial m} \quad (\text{G20})$$

in the $K \rightarrow \infty$ limit.

We now have the necessary ingredients to compute the RS capacity in the limit of sparse inputs. We first would like to gain some general understanding of whether the conditions for the RS capacity to remain finite are the same in the sparse limit, assuming that p does not tend to unity with r . As in the non-sparse case, this depends on whether the limit

$$\beta(m, r) = \lim_{q \uparrow 1} \frac{\sigma^2(m, r) - \tilde{q}(q, m, r)}{1 - q} \quad (\text{G21})$$

is finite. We assume that r is not exactly equal to one, and that m is bounded. Then, all effective order parameters are the non-sparse ($r = 0$) order parameters of a network with a transformed activation function

$$\bar{g}(x) = g(\sqrt{1 - r^2}x + rm). \quad (\text{G22})$$

If $g \in \mathcal{L}^2(\gamma)$, then \bar{g} must also be in $\mathcal{L}^2(\gamma)$, which follows from the fact that g must be exponentially bounded at infinity. The function \bar{g} is weakly differentiable if and only if g is weakly differentiable, which is easy to see based on the fact that the linear transformation $x \mapsto \sqrt{1 - r^2}x + rm$ is smooth and invertible. If g is weakly differentiable, then the chain rule for the weak derivative implies that

$$\bar{g}'(x) = \sqrt{1 - r^2}g'(\sqrt{1 - r^2}x + rm). \quad (\text{G23})$$

Furthermore, if g' has finite $\mathcal{L}^2(\gamma)$ norm, then the reasoning applied to \bar{g} above implies that \bar{g}' also has finite norm. Thus, by applying our arguments from the non-sparse case to \bar{g} , we find that β is finite if the weak derivative of g exists and has finite norm. Therefore, changing the sparsity of the input distribution does not change whether or not the capacity diverges in the infinite-width limit.

We now would like to characterize how the RS capacity of networks with weakly-differentiable activation functions behaves in the sparse limit. However, directly applying the methods that allowed us to evaluate the $q \uparrow 1$ limit in the non-sparse case is more challenging, as we would need to compute the Fourier-Hermite coefficients of the transformed activation \bar{g} for all r and m . Instead, we will focus on two simple cases: ReLU and analytic activations with non-vanishing derivative at the origin. These cases cover most activation functions commonly used in neural networks, and provide some insight into how sparsity can affect the capacity.

1. The rectified linear unit

For $g(x) = \text{ReLU}(x)$, we can compute the effective local magnetization in closed form, yielding

$$\tilde{m}(m, r) = \sqrt{\frac{1 - r^2}{2\pi}} \exp\left(-\frac{(mr)^2}{2(1 - r^2)}\right) + \frac{1}{2}mr \left[1 + \text{erf}\left(\frac{mr}{\sqrt{2(1 - r^2)}}\right)\right]. \quad (\text{G24})$$

We note that the exponential term tends to zero as $r \uparrow 1$ uniformly in m , while the second term tends to m as $r \uparrow 1$ if $m > 0$, and to zero if $m \leq 0$.

To obtain an $\mathcal{O}(1)$ value for $\sqrt{K}(\tilde{m}(m, r) - \tilde{m}_0)$, we adopt the scaling

$$m = \frac{t}{\sqrt{K}} \quad (\text{G25})$$

for $t = \mathcal{O}(1)$. This yields

$$\lim_{K \rightarrow \infty} \sqrt{K}(\tilde{m}(m, r) - \tilde{m}_0) = \frac{1}{2}rt \quad (\text{G26})$$

for any t and all $|r| < 1$; other scalings will not yield an $\mathcal{O}(1)$ value. Similarly, we find that

$$\lim_{K \rightarrow \infty} \frac{\partial \tilde{m}}{\partial m} = \frac{1}{2}r. \quad (\text{G27})$$

Using the continuity of the effective order parameters in m and the fact that ReLU is a positive-homogeneous function, we have

$$\lim_{K \rightarrow \infty} \sigma^2(m, r) = \sigma^2(m = 0, r) = (1 - r^2)\sigma^2(m = 0, r = 0) = (1 - r^2)\frac{\pi - 1}{2\pi} \quad (\text{G28})$$

and

$$\lim_{K \rightarrow \infty} \tilde{q}(q, m, r) = \tilde{q}(q, m = 0, r) = (1 - r^2)\tilde{q}(q, m = 0, r = 0) \quad (\text{G29})$$

for any $|r| < 1$. This implies that

$$\beta(m = 0, r) = (1 - r^2)\beta(m = 0, r = 0) = \frac{1}{2}(1 - r^2). \quad (\text{G30})$$

Then, by the same reasoning we used to take the limit $q \uparrow 1$ to obtain the RS capacity in the non-sparse case, we find that the saddle-point equation for t becomes

$$(1 - p) u_1 \left(\frac{\kappa + \bar{v}rt/2}{\sqrt{\sigma^2(1 - r^2)}} \right) = p u_1 \left(\frac{\kappa - \bar{v}rt/2}{\sqrt{\sigma^2(1 - r^2)}} \right), \quad (\text{G31})$$

where by a minor abuse of notation we now write $\sigma^2 = \sigma^2(m = 0, r = 0)$, and we define

$$u_n(x) \equiv \int_{-x}^{\infty} d\gamma(z) (x + z)^n. \quad (\text{G32})$$

for brevity. Similarly, the re-arranged saddle point equation for q yields

$$\alpha_{\text{RS}} = \frac{\pi}{\pi - 1} \left[(1 - p) u_2 \left(\frac{\kappa + \bar{v}rt/2}{\sqrt{\sigma^2(1 - r^2)}} \right) + p u_2 \left(\frac{\kappa - \bar{v}rt/2}{\sqrt{\sigma^2(1 - r^2)}} \right) \right]^{-1}. \quad (\text{G33})$$

For $p = 1/2$, the fact that u_1 is a monotonically increasing function implies that we must have $t^* = 0$ for any κ , yielding an RS capacity of

$$\alpha_{\text{RS}} = \frac{\pi}{\pi - 1} \left[u_2 \left(\frac{\kappa}{\sqrt{\sigma^2(1 - r^2)}} \right) \right]^{-1}. \quad (\text{G34})$$

2. Analytic activation functions

We now consider analytic activation functions with non-zero derivative at the origin. As for ReLU, we intuitively expect that the required scaling for μ to remain $\mathcal{O}(1)$ is $m = t/\sqrt{K}$ for $t = \mathcal{O}(1)$. This intuition may be made more

concrete by considering the small- m expansion of \tilde{m} in the limit $r \uparrow 1$:

$$\tilde{m}(m, r) = \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi(1-r^2)}} \exp\left(-\frac{(x-rm)^2}{2(1-r^2)}\right) g(x) \quad (\text{G35})$$

$$= \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi(1-r^2)}} \exp\left(-\frac{x^2}{2(1-r^2)}\right) \left[1 + \frac{rxm}{1-r^2} + \mathcal{O}(m^2)\right] g(x) \quad (\text{G36})$$

$$= \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi(1-r^2)}} \exp\left(-\frac{x^2}{2(1-r^2)}\right) g(x) \\ + rm \int_{-\infty}^{\infty} \frac{dx}{\sqrt{2\pi(1-r^2)}} \exp\left(-\frac{x^2}{2(1-r^2)}\right) g'(x) + \mathcal{O}(m^2) \quad (\text{G37})$$

$$\rightarrow g(0) + mg'(0) + \mathcal{O}(m^2), \quad (\text{G38})$$

where we have used the formula for Gaussian integration by parts to obtain the third line, and taken the limit $r \uparrow 1$ on the fourth.

We will proceed under the assumption that r is close enough to unity such that we can expand $\sigma^2(r)$ as a power series in $1-r^2$ by interchanging the expectation with the series expansion of $g(x)$ about the origin. We note that this is justified for sufficiently small $1-r^2$ by the assumption of analyticity. Then, by continuity, we have

$$\lim_{K \rightarrow \infty} \sqrt{K}(\tilde{m}(t/\sqrt{K}, r) - \tilde{m}_0) = g'(0)t + \mathcal{O}(1-r^2), \quad (\text{G39})$$

$$\lim_{K \rightarrow \infty} \sigma^2(m, r) = \sigma^2(m=0, r) \equiv \sigma^2(r) = [g'(0)]^2(1-r^2) + \mathcal{O}[(1-r^2)^2], \quad (\text{G40})$$

and

$$\lim_{K \rightarrow \infty} \tilde{q}(q, m, r) = \tilde{q}(q, m=0, r) \equiv \tilde{q}(q, r) = [g'(0)]^2(1-r^2)q + \mathcal{O}[q^2, (1-r^2)^2], \quad (\text{G41})$$

which yields

$$\beta(r) = \lim_{q \uparrow 1} \frac{\sigma^2(r) - \tilde{q}(q, r)}{1-q} = [g'(0)]^2(1-r^2) + \mathcal{O}[(1-r^2)^2]. \quad (\text{G42})$$

Applying our previous results, we obtain the saddle-point equation for t and the expression for the capacity to leading order in $1-r^2$ as

$$(1-p)u_1 \left(\frac{\kappa + \bar{v}g'(0)t^*}{\sqrt{[g'(0)]^2(1-r^2)}} \right) = pu_1 \left(\frac{\kappa - \bar{v}g'(0)t^*}{\sqrt{[g'(0)]^2(1-r^2)}} \right) \quad (\text{G43})$$

and

$$\frac{1}{\alpha_{\text{RS}}} = (1-p)u_2 \left(\frac{\kappa + \bar{v}g'(0)t^*}{\sqrt{[g'(0)]^2(1-r^2)}} \right) + pu_2 \left(\frac{\kappa - \bar{v}g'(0)t^*}{\sqrt{[g'(0)]^2(1-r^2)}} \right), \quad (\text{G44})$$

respectively.

As for ReLU, if $p = 1/2$, the saddle-point equation has solution $t^* = 0$, which yields

$$\alpha_{\text{RS}} = \left[u_2 \left(\frac{\kappa}{\sqrt{[g'(0)]^2(1-r^2)}} \right) \right]^{-1}. \quad (\text{G45})$$

Comparison of these results reveals an interesting point. With ReLU activation functions, the zero-margin RS capacity remains $2\pi/(\pi-1)$ in the sparse limit. In contrast, the zero-margin RS capacity for an analytic activation function of this type approaches that of the perceptron in this limit. For either, the capacity vanishes at any non-zero margin as $\alpha_{\text{RS}} \sim 1-r^2$, as $u_2 \sim x^2$ for $x \gg 1$.

The difference in the capacities in the sparse limit for ReLU or analytic activation functions is easy to justify intuitively. For $p = 1/2$, one expects the local magnetization to vanish such that the distribution of the output preactivation is symmetric, like that of the target output. Then, the remaining effect of sparsity is the $1-r^2$ variance

of the hidden unit preactivations. For ReLU, this simply corresponds to an overall scaling of the output preactivation by $\sqrt{1-r^2}$. For analytic activation functions with non-zero derivative at the origin, we expect terms of quadratic order and higher to be negligible if $1-r^2$ is small enough such that the preactivations are concentrated very near to the origin. This leaves, approximately, a perceptron with an overall scaling factor of $\sqrt{1-r^2}g'(0)$. Then, as the zero-margin capacity is scale-invariant, the zero-margin capacity for ReLU should remain the same as in the non-sparse case, while that for analytic activation functions of this class should approach that of the perceptron. In either case, one expects the capacity at non-zero margins to vanish as the output preactivation concentrates in some $\mathcal{O}(\sqrt{1-r^2})$ neighborhood of zero. Thus, one can obtain the capacities calculated above via heuristic arguments.

Appendix H: Numerical experiments

The question of how to confront our theory with empirical data raises an important issue in the study of deep networks: the questions of the existence and learnability of solutions to a classification task need not be equivalent [25–27]. The Gardner volume seeks to quantify the existence of solutions, agnostic to how the weights might be determined [5–7, 12, 13]. For the perceptron, one can prove that the eponymous learning algorithm will find solutions whenever they exist [27]. However, there do not exist learning algorithms with corresponding convergence guarantees for deep networks [21, 26, 28, 29]. In the absence of rigorous guarantees, one cannot be sure that a particular learning algorithm will find the solutions which the Gardner volume aims to count. Thus, there exists an important distinction between theories that study the Gardner volume and those that study the storage capacity of networks subject to particular learning rules [12, 13, 27, 30].

With these considerations in mind, it is important to note that theories of the Gardner volume can be falsified using particular learning algorithms. Concretely, the capacity computed using these methods constitutes a non-rigorous upper bound on the true capacity. Therefore, it is possible to falsify such theories by showing empirically or analytically that a particular learning algorithm can find solutions at loads higher than this predicted bound. However, if one seeks to test the main prediction of our theory—that of diverging or finite capacity in the infinite-width limit—one encounters an important problem: networks with activation functions that are not weakly differentiable are not amenable to optimization via commonly-used gradient-based techniques [21, 26]. Instead, one must use ad hoc algorithms developed for particular activation functions.

For “classical” treelike committee machines with sign activation functions and all readout weights equal to unity, the most commonly-used learning algorithms are variants on an algorithm known as least action learning (LAL) [3, 5, 6, 27, 31]. LAL is a greedy heuristic extension of the perceptron learning algorithm: if a training example is classified incorrectly, the perceptron learning rule is applied to the hidden unit with preactivation closest to the threshold among those that “voted” for the incorrect class. Engel *et al.* [5] found that the empirical capacity of a slight variant of LAL appeared to be around 2 for committee machines of widths 3, 5, and 7, failing to increase with width as predicted by their analysis of the Gardner volume. In particular, the 1-RSB estimate of the capacity with three branches is approximately 3. More recently, Baldassi *et al.* [31] showed that a version of LAL that operates batchwise can find solutions at loads approaching the 1-RSB estimate, but they only considered the three-branched case. Because the maximum relative change in the output preactivation scales as $1/K$, one expects the speed of learning with LAL to become extremely slow in wide networks. Furthermore, the theoretical capacity with sign activation functions diverges extremely slowly with width, scaling only as $\sqrt{\log K}$ [9].

We implemented the batch-LAL algorithm proposed in Baldassi *et al.* [31] in MATLAB 9.6 and, following the system size considered in that work, trained a treelike committee machine with sign activation functions with $N = 990$ inputs and $K = 3, 5, 9, 15, 33,$ or 99 hidden units to classify a randomly-generated dataset. We used a learning rate of $\eta = 0.005$ and a batch size of 128 [31], allowing a maximum of 10^5 epochs with early stopping if zero classification error was reached. As shown in Figure S3, the empirical capacities saturate at around 3 rather than increasing with width. This apparent ceiling likely results from the imposed maximum number of training epochs, as we find that the number of epochs required to achieve vanishing error increases superexponentially as the load approaches three from below. Compute time is therefore an important limiting factor in determining the true algorithmic capacity of batch-LAL; these simulations required more than fourteen days of compute time over 32 cores of an HPC node to complete. In short, the batch-LAL algorithm behaves in our hands in much the same way as the variant of LAL proposed by Engel *et al.* [5] did in 1992: if one imposes reasonable constraints the runtime of the algorithm, then one does not observe substantial increases in the empirical capacity with increasing hidden layer width.

To study treelike committee machines with weakly differentiable activation functions, we train networks via minimization of the hinge loss. The hinge loss is commonly used for training maximum margin binary classifiers, notably support vector machines [26], and can be optimized using the subgradient methods commonly applied to ReLU networks in contemporary machine learning [21, 26]. We used this method to train treelike committee machines with $N = 1000$ and $K = 10, 50, 100,$ or 200 with ReLU, erf, or quadratic activation functions. We chose the total number

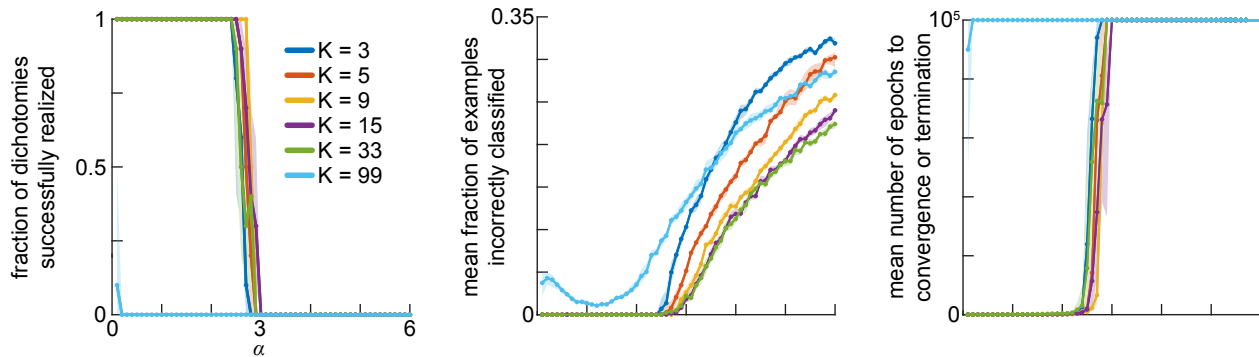


FIG. S3. Training treelike committee machines with sign function activations to classify random datasets using batch-LAL. The total number of inputs is $N = 990$ throughout, and the abscissa in each panel is the load $\alpha = P/N$. In all panels, solid lines indicate the average over 10 realizations, and shaded patches indicate 95% confidence intervals of the mean computed via the bias-corrected and accelerated bootstrap method. In the left panel, the ordinate shows the fraction of the 10 realizations at each load for which zero classification error was reached. The center panel shows the mean fraction of examples classified correctly at the end of training at each load. Finally, the right panel shows the mean number of epochs at which training was terminated, either due to early stopping or the fixed threshold.

of inputs to be a comparable finite size to that of our LAL simulations while being easily divisible among even numbers of hidden units such that we could set half of the readout weights to $+1$ and the remainder to -1 , thus satisfying our constraint on the weights and threshold. We implemented the optimization using TENSORFLOW 2.0 [32] in PYTHON 3.8 using the ADAM [33] optimizer with default parameters and a batch size of 32. As shown in Figure S4, we find much the same phenomena in this case as we did for LAL: the empirical capacity appears to be limited chiefly by the maximum number of training epochs allowed. Here, we fixed the maximum number of training epochs to 5,000; each of the 12 simulations reported in Figure S4 required between five and seven days of compute time on one NVIDIA Tesla V100 GPU of an HPC node. In all cases, we find empirical capacities that are less than those predicted at 1-RSB, with committee machines with error function activations and 50 hidden units coming the closest to achieving the predicted capacity. Therefore, these experiments fail to falsify our theory.

-
- [1] V. I. Bogachev, *Gaussian measures* (American Mathematical Society, 1998).
 - [2] T. M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE transactions on electronic computers*, 326 (1965).
 - [3] G. Mitchison and R. Durbin, Bounds on the learning capacity of some multi-layer networks, *Biological Cybernetics* **60**, 345 (1989).
 - [4] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Statistical learning theory of structured data, *Phys. Rev. E* **102**, 032119 (2020), [arXiv:2005.10002](https://arxiv.org/abs/2005.10002).
 - [5] A. Engel, H. Köhler, F. Tschepke, H. Vollmayr, and A. Zippelius, Storage capacity and learning algorithms for two-layer neural networks, *Physical Review A* **45**, 7590 (1992).
 - [6] A. Engel and C. Van den Broeck, *Statistical mechanics of learning* (Cambridge University Press, 2001).
 - [7] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, *Physical Review A* **45**, 4146 (1992).
 - [8] J. E. Kolassa, *Series approximation methods in statistics*, Vol. 88 (Springer Science & Business Media, 2006).
 - [9] R. Monasson and R. Zecchina, Weight space structure and internal representations: a direct approach to learning and generalization in multilayer neural networks, *Physical Review Letters* **75**, 2432 (1995), [arXiv:cond-mat/9501082](https://arxiv.org/abs/cond-mat/9501082).
 - [10] J. Ding and N. Sun, Capacity lower bound for the Ising perceptron, in *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (2019) pp. 816–827, [arXiv:1809.07742](https://arxiv.org/abs/1809.07742).
 - [11] 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, *Physical Review Letters* **123**, 170602 (2019), [arXiv:1907.07578](https://arxiv.org/abs/1907.07578).
 - [12] E. Gardner, The space of interactions in neural network models, *Journal of Physics A: Mathematical and General* **21**, 257 (1988).
 - [13] E. Gardner and B. Derrida, Optimal storage properties of neural network models, *Journal of Physics A: Mathematical and General* **21**, 271 (1988).
 - [14] M. Talagrand, *Spin glasses: a challenge for mathematicians: cavity and mean field models*, Vol. 46 (Springer Science & Business Media, 2003).

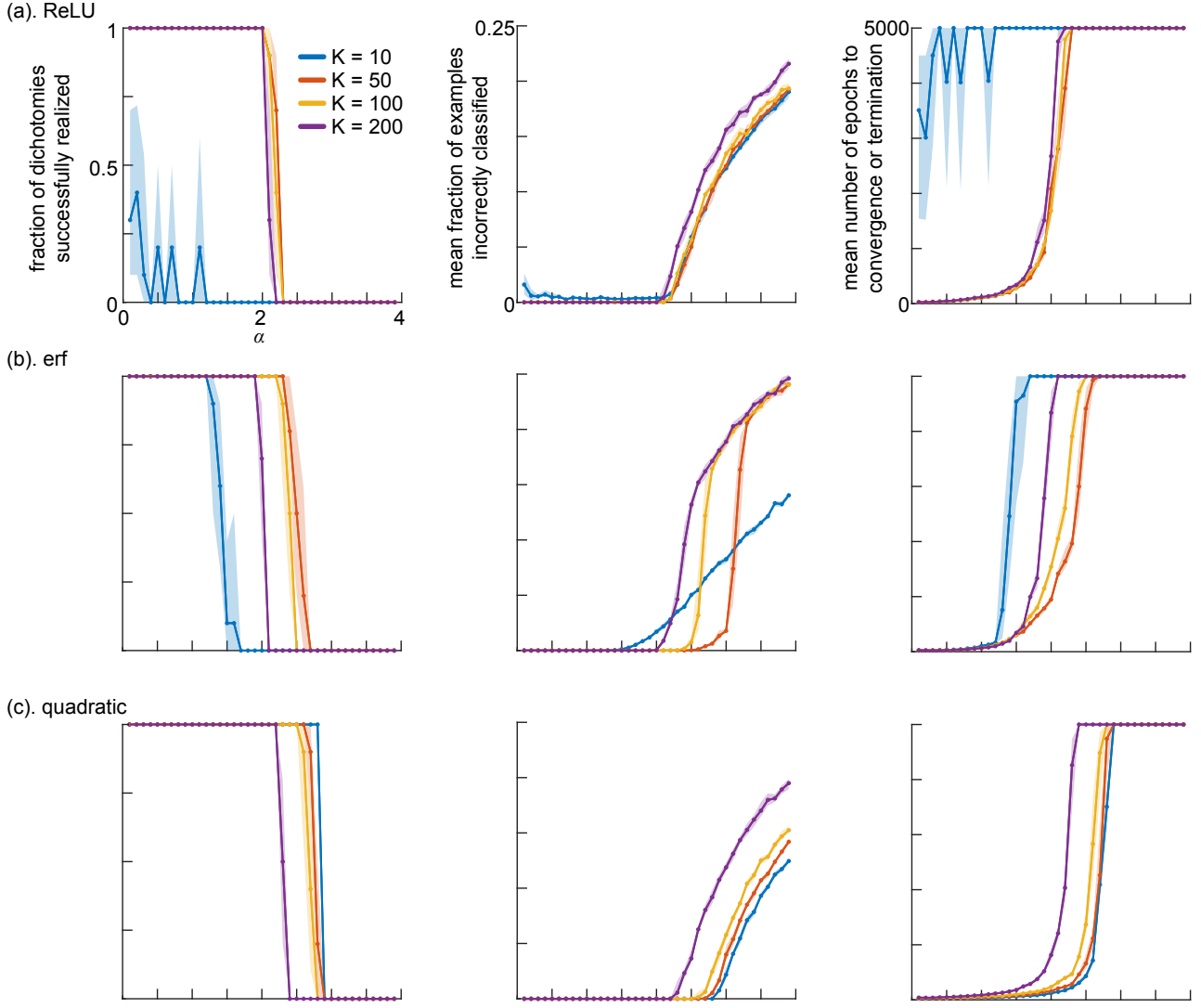


FIG. S4. Training treelike committee machines with weakly-differentiable activation functions using stochastic gradient descent on the hinge loss. The total number of inputs is $N = 1000$ throughout, and the abscissa in each panel is the load $\alpha = P/N$. In all panels, solid lines indicate the average over 10 realizations, and shaded patches indicate 95% confidence intervals of the mean computed via the bias-corrected and accelerated bootstrap method. In the left panel, the ordinate shows the fraction of the 10 realizations at each load for which zero classification error was reached. The center panel shows the mean fraction of examples classified correctly at the end of training at each load. Finally, the right panel shows the mean number of epochs at which training was terminated, either due to early stopping or the fixed threshold. Sub-figures (a), (b), and (c) show results for ReLU, erf, and quadratic activation functions, respectively.

- [15] M. Mézard, G. Parisi, and M. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, Vol. 9 (World Scientific Publishing Company, 1987).
- [16] D. Pollard, *A user's guide to measure theoretic probability*, Vol. 8 (Cambridge University Press, 2002).
- [17] Y. L. Tong, *The multivariate normal distribution* (Springer Science & Business Media, 2012).
- [18] M. Shcherbina and B. Tirozzi, Rigorous solution of the Gardner problem, *Communications in Mathematical Physics* **234**, 383 (2003), [arXiv:math-ph/0112003](https://arxiv.org/abs/math-ph/0112003).
- [19] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Vol. 55 (US Government printing office, 1948).
- [20] R. H. Byrd, J. C. Gilbert, and J. Nocedal, A trust region method based on interior point techniques for nonlinear programming, *Mathematical Programming* **89**, 149 (2000).

- [21] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature* **521**, 436 (2015).
- [22] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions, arXiv preprint arXiv:1710.05941 (2017), [arXiv:1710.05941](https://arxiv.org/abs/1710.05941).
- [23] P. Poirazi, T. Brannon, and B. W. Mel, Pyramidal neuron as two-layer neural network, *Neuron* **37**, 989 (2003).
- [24] A. Panigrahi, A. Shetty, and N. Goyal, Effect of activation functions on the training of overparametrized neural nets, in *International Conference on Learning Representations* (2020) [arXiv:1908.05660](https://arxiv.org/abs/1908.05660).
- [25] E. Malach and S. Shalev-Shwartz, Is deeper better only when shallow is good?, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett (2019) pp. 6429–6438, [arXiv:1903.03488](https://arxiv.org/abs/1903.03488).
- [26] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).
- [27] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (CRC Press, 1991).
- [28] A. Jacot, F. Gabriel, and C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, in *Advances in neural information processing systems* (2018) pp. 8571–8580, [arXiv:1806.07572](https://arxiv.org/abs/1806.07572).
- [29] M. Belkin, D. Hsu, S. Ma, and S. Mandal, Reconciling modern machine-learning practice and the classical bias–variance trade-off, *Proceedings of the National Academy of Sciences* **116**, 15849 (2019), [arXiv:1812.11118](https://arxiv.org/abs/1812.11118).
- [30] A. Knoblauch, G. Palm, and F. T. Sommer, Memory capacities for synaptic and structural plasticity, *Neural Computation* **22**, 289 (2010).
- [31] C. Baldassi, F. Pittorino, and R. Zecchina, Shaping the learning landscape in neural networks around wide flat minima, *Proceedings of the National Academy of Sciences* **117**, 161 (2020), [arXiv:1905.07833](https://arxiv.org/abs/1905.07833).
- [32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.
- [33] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).