Activation function dependence of the storage capacity of treelike neural networks

Jacob A. Zavatone-Veth and Cengiz Pehlevan

Phys. Rev. E **103**, L020301 — Published 19 February 2021

DOI: 10.1103/PhysRevE.103.L020301

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*

(Dated: February 9, 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.

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 (Figure 1a) [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

$$y(\mathbf{x}) = \mathrm{sign}(s(\mathbf{x})) \quad \text{for} \tag{1}$$

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

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 dataset of P examples subject to constraints on the hidden unit weights imposed by a probability measure ρ, we define

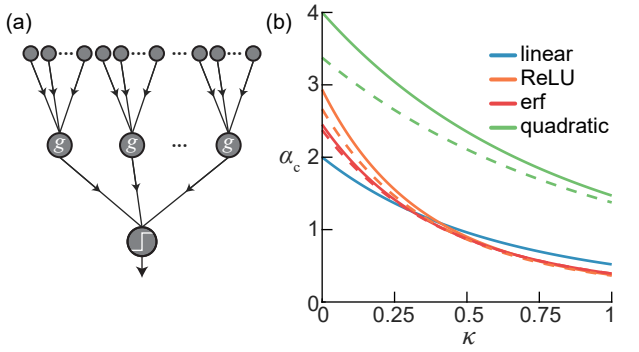

FIG. 1. Pattern storage in treelike committee machines. (a) Network architecture. (b) Capacity α_c as a function of margin κ for several common activation functions. Solid and dashed lines indicate estimates of the capacity under replica-symmetric and one-step replica-symmetry-breaking ansätze, respectively.

* jzavatoneveth@g.harvard.edu
† cpehlevan@seas.harvard.edu

the Gardner volume [16, 17]

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

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 dataset in which the components of the inputs and the target outputs are independent and identically distributed as $x^\mu_{jk} = \pm 1$ 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 $N, P \to \infty$ with load $\alpha \equiv P/N = \mathcal{O}(1)$ and then take $K \to \infty$ [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 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 20$^{\text{th}}$ century, showing that the capacity diverges as $K \to \infty$ [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 $K \to \infty$ 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 datasets 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 non-rigorous interchange of limits to write

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

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

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

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 [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 non-vanishing. For the special cases of $\text{sign}(x)$ 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, [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]: $\varphi(x, y; q) = \varphi(x)\varphi(y) \sum_{k=0}^{\infty} q^k \operatorname{He}_k(x) \operatorname{He}_k(y)$, 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 $\tilde{q}(\varepsilon) + g_0^2 = \sum_{l=0}^{\infty} \frac{(-\varepsilon)^l}{l!} \sum_{k=l}^{\infty} (k)_l g_k^2$, where $(k)_l = k(k-1)\cdots(k-l+1)$ 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 $g^{(l)}(x) = \sum_{k=l}^{\infty} g_k \sqrt{(k)_l} \operatorname{He}_{k-l}(x)$, which follow from the recurrence relation $\operatorname{He}_k'(x) = \sqrt{k} \operatorname{He}_{k-1}(x)$ [29]. Therefore, $\tilde{q}$ admits a formal power series expansion in ε as

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

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 non-vanishing 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

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

where

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

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 $\alpha_{\mathrm{G}}(\kappa)$ 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 $K \to \infty$ 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 $\alpha_{\mathrm{RS}} = 2\pi/(\pi - 1) \simeq 2.93388$. Considering sigmoidal activations, we find that $\alpha_{\mathrm{RS}} = 2 \arcsin(2/3)/\pi \simeq 2.45140$ for the error function, while $\alpha_{\mathrm{RS}} \simeq 2.35561$ for the hyperbolic tangent and the logistic. As an example of a non-monotonic activation function, we consider a quadratic, which yields $\alpha_{\mathrm{RS}} = 4$. We plot the RS capacity as a function of margin for these activation functions in Figure 1b, 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 non-convex, 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 inter-group overlap q_0 and intra-group 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 $\alpha_{\text{1-RSB}}$ 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 $\alpha_{\text{1-RSB}} \simeq 2.66428$ at $(q_0^*, r^*) \simeq (0.75716, 16.6374)$, which is consistent with the result of Baldassi *et al.* [11] (see [33]). For erf, we obtain $\alpha_{\text{1-RSB}} \simeq 2.37500$ at $(q_0^*, r^*) \simeq (0.75463, 7.75682)$. Finally, for the quadratic, we have $\alpha_{\text{1-RSB}} \simeq 3.37466$ at $(q_0^*, r^*) \simeq (0.28452, 6.39299)$. In Figure 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 $\alpha_{\text{1-RSB}}$ by considering candidate solutions with fixed values of the inter-block overlap q_0. From $q_0 \uparrow 1$, we have $\alpha_{\text{1-RSB}} \leq \alpha_{\mathrm{RS}}$. As shown in the Supplemental Material [22], we can also

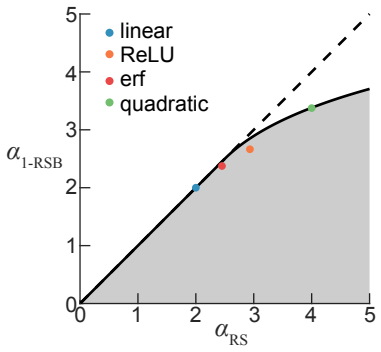

FIG. 2. The accessible region in (α_{RS}, $\alpha_{\text{1-RSB}}$)-space defined by the $q_0 = 0$ bound. The allowed region is shaded in gray, and the locations of the four example activation functions for which we estimate $\alpha_{\text{1-RSB}}$ are indicated by colored dots.

obtain an upper bound for $\alpha_{\text{1-RSB}}$ 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 $\alpha_{\text{1-RSB}} = \mathcal{O}(\log \alpha_{\text{RS}})$. The $q_0 = 0$ bound allows us to define an accessible region in (α_{RS}, $\alpha_{\text{1-RSB}}$)-space, as illustrated in Figure 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—sub-logarithmically—with degree. This result is sensible given the oscillatory nature of high-degree Hermite polynomials, which one expects to yield a highly non-convex 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 [16]'s 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: non-smooth 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 work, 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.

Acknowledgements—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, Mathematics of control, signals and systems **2**, 303 (1989).
[2] K. Hornik, M. Stinchcombe, H. White, *et al.*, Multilayer feedforward networks are universal approximators., Neural networks **2**, 359 (1989).
[3] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Understanding deep learning requires rethinking generalization, in *5th Int. Conf. on Learning Representations (ICLR 2017)* (2016) arXiv: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 *Advances in neural information processing systems* (2016) pp. 3360–3368, arXiv:1606.05340.
[5] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, Nature **521**, 436 (2015).
[6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (MIT press, 2016).
[7] 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.
[8] L. Zdeborová, Understanding deep learning is also a job for physicists, Nature Physics **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.
[10] P. Ramachandran, B. Zoph, and Q. V. Le, Searching for activation functions, arXiv preprint arXiv:1710.05941 (2017), arXiv: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, Physical Review Letters **123**, 170602 (2019), arXiv:1907.07578.
[12] E. Barkai, D. Hansel, and H. Sompolinsky, Broken symmetries in multilayered perceptrons, Physical Review A **45**, 4146 (1992).
[13] 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).
[14] A. Engel and C. Van den Broeck, *Statistical mechanics of learning* (Cambridge University Press, 2001).
[15] A. Mozeika, B. Li, and D. Saad, Space of functions computed by deep-layered machines, Physical Review Letters **125**, 168301 (2020), arXiv:2004.08930.
[16] E. Gardner, The space of interactions in neural network models, Journal of Physics A: Mathematical and General **21**, 257 (1988).
[17] E. Gardner and B. Derrida, Optimal storage properties of neural network models, Journal of Physics A: Mathematical and General **21**, 271 (1988).
[18] M. Talagrand, *Spin glasses: a challenge for mathematicians: cavity and mean field models*, Vol. 46 (Springer Science & Business Media, 2003).
[19] 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.
[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. Papoutsi, Illuminating dendritic function with computational models, Nature Reviews Neuroscience **21**, 303 (2020).
[22] See the Supplemental Material 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 *Advances in neural information processing systems* (2018) pp. 8571–8580, arXiv:1806.07572.
[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* (2020) pp. 8135–8145, arXiv:2002.02561.
[25] G. Mitchison and R. Durbin, Bounds on the learning capacity of some multi-layer networks, Biological Cybernetics **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, 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 Publishing

Company, 1987).
[29] V. I. Bogachev, *Gaussian measures* (American Mathematical Society, 1998).
[30] W. Kibble, An extension of a theorem of Mehler's on Hermite polynomials, Mathematical Proceedings of the Cambridge Philosophical Society **41**, 12 (1945).
[31] Y. L. Tong, *The multivariate normal distribution* (Springer Science & Business Media, 2012).
[32] M. Shcherbina and B. Tirozzi, Rigorous solution of the Gardner problem, Communications in Mathematical Physics **234**, 383 (2003), arXiv:math-ph/0112003.
[33] In the published version of their Letter, Baldassi *et al.* [11] reported a value of $\alpha_{\text{1-RSB}} \simeq 2.92$. After the appearance of our work in preprint form, they found that this result was incorrect [22]; their revised estimate of $\alpha_{\text{1-RSB}} \simeq 2.6643$ 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* (2019) pp. 816–827, arXiv:1809.07742.
[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, Journal of Statistical Mechanics: Theory and Experiment **2019**, 124023 (2019), arXiv:1806.05451.
[36] A. Knoblauch, G. Palm, and F. T. Sommer, Memory capacities for synaptic and structural plasticity, Neural Computation **22**, 289 (2010).
[37] B. Willmore and D. J. Tolhurst, Characterizing the sparseness of neural codes, Network: Computation in Neural Systems **12**, 255 (2001).
[38] T. Shinzato and Y. Kabashima, Perceptron capacity revisited: classification ability for correlated patterns, Journal of Physics A: Mathematical and Theoretical **41**, 324013 (2008), arXiv:0712.4050.
[39] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Statistical learning theory of structured data, Phys. Rev. E **102**, 032119 (2020), arXiv:2005.10002.
[40] S. Goldt, M. Mézard, F. Krzakala, and L. Zdeborová, Modelling the influence of data structure on learning in neural networks, Physical Review X **10**, 041044 (2020), arXiv:1909.11500.
[41] D. Breuer, M. Timme, and R.-M. Memmesheimer, Statistical physics of neural systems with nonadditive dendritic coupling, Physical Review X **4**, 011053 (2014), arXiv:1507.03881.
[42] A. Gidon, T. A. Zolnik, P. Fidzinski, F. Bolduan, A. Papoutsi, P. Poirazi, M. Holtkamp, I. Vida, and M. E. Larkum, Dendritic action potentials and computation in human layer 2/3 cortical neurons, Science **367**, 83 (2020).
[43] A. Payeur, J.-C. Béïque, and R. Naud, Classes of dendritic information processing, Current Opinion in Neurobiology **58**, 78 (2019).
[44] 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.
[45] R. Urbanczik, Storage capacity of the fully-connected committee machine, Journal of Physics A: Mathematical and General **30**, L387 (1997).
[46] J. E. Kolassa, *Series approximation methods in statistics*, Vol. 88 (Springer Science & Business Media, 2006).
[47] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Vol. 55 (US Government printing office, 1948).
[48] 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).
[49] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation* (CRC Press, 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, 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.
[51] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).