CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 138 **January 11, 2023**

Foundations of Deep Learning: Compositional Sparsity of Computable Functions

Tomaso Poggio
Center for Brains, Minds, and Machines, MIT, Cambridge, MA, USA

Abstract

The claim is that compositional sparsity of the underlying target function, which corresponds to the task to be learned, is the key principle underlying machine learning. Under restrictions of smoothness of the constituent functions, sparsity of the compositional target functions naturally leads to sparse deep networks that allow approximation, optimization and generalization. This is the case of CNNs, in which the known sparse graph of the target function is reflected in the architecture of the network. It is a reasonable conjecture that transformers are able to implement a flexible version of sparsity (selecting which input tokens interact in the MLP layer), through the self-attention layers, when the target function is unknown.

Surprisingly, the assumption of compositional sparsity of the target function is not restrictive in practice, since for computable functions with Lipschitz continuous derivatives *compositional sparsity is equivalent to efficient computability, that is computability in polynomial time*.

This material is based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

1 Introduction

We still do not understand why deep networks work. Until recently this question could have been be rephrased as a question about why CNNs work so well. In the meantime, other architectures, especially transformers, also show amazing performance. Is there a common principle at the core of these successful neural network architectures? In the following, I describe a framework built around a specific principle that, I conjecture, must underlie most of deep learning.

This note is thus about foundations of ML, connecting the avoidance of the curse of dimensionality in the approximations of certain function classes with their computability – establishing *an unusual bridge between computability and approximation*. The main result is that for smooth functions – functions with Lipschitz continuous first derivatives – compositional sparsity is equivalent to e□ciently computability, that is, they can be approximated in polynomial time (we use here e□cient to mean non-exponential). The result implies that all smooth functions are – in "practice" – compositionally sparse. It also says that associated good parametric and constructive approximators are deep sparse RELU networks. In a sense, compositionally sparse functions are a universal function class and sparse RELU networks are the associated class of universal approximators.

2 Sparse functions and computable functions

Let us first define an interesting class of sparse functions, that is *sparse compositional functions* that are the composition of sparse constituent functions. This class is interesting for approximation theory: in fact the assumption of sparse target functions has appeared often in the recent approximation literature (see [1, 2, 3, 4, 5]).

Definition 1. *A sparse compositional function of d variables is a function that can be represented as the composition of no more than $poly(d)$ sparse functions, each depending on $\leq d_0$ variables with $d_0 < d$.*

Let us now provide a specific version for this paper of the definition of a *computable* function. There are various notions of computability on the reals. The simplest is Borel-Turing computability. As shown very recently, they all have problems (see [6, 7]) in the sense that several interesting functions on the reals are not always computable (such as the pseudoinverse). Here we bypass this issues and consider functions that are computable by an appropriate machine. Our focus is whether such functions, that are computable, are or not computable in polynomial time.

Definition 2. *A real-valued continuous function $f : \mathbf{R}^d \to \mathbf{R}^k$ is computable if there is a procedure or algorithm that correctly calculates a parametrized sequence of constructive computable approximations (e.g. by RELU networks) f_n to f, each depending on n parameters, with desired errors $\leq \epsilon_n$ in the sup norm; a special case is when it is computable in polynomial time/space in d.*

As emphasized by H. Mhaskar, just the degree of approximation theorem is too crude a tool to use - one has to add that the approximation must be constructive, based on values of the target function. Otherwise, as shown in [8] where for the first time both dimension independent as well as constructive bounds for the same class of functions are proven), ReLU networks can achieve dimension independent bounds, obviating the need to have deep networks from the point of view of degree of approximation alone.

Definition 2 has a deep motivation in the classical framework of machine learning. In the theory of ML, the first conceptual step is to define parametric approximators of the class of functions to be learned. Examples of approximators are generalized additive models, polynomials and deep RELU network. We want approximators to a class of functions, which should be as large as possible. Furthermore, we want the parametric approximators to be e□ciently computable, to ensure that optimization on the training data is possible. In particular, this means that the number of parameters in the approximators cannot

be exponential in d: in other words, the approximators must avoid the curse of dimensionality. We call this requirement *e ciently computable approximation*, in short *e cient computability*. It is equivalent to requiring that the computation of the approximator by a Turing machine should not have worse complexity than $poly(d)$. Recall that approximation of a generic continuous function using multivariate polynomials of degree k has exponential complexity $O(k^d)$. Notice that evaluation at points of f and even continuity of f are not enough to ensure a meaningful approximation, defined as a convergent sequence of approximating f_n that converges to f[1].

2.1 Equivalence of computability and compositional sparsity

Consider smooth real-valued functions in d variables, that is functions with Lipschitz continuous first derivatives. The MP theorem (see Appendix) shows that such functions are e□ciently computed by deep RELU networks. The converse follows from the observation that an e□ciently computable function can be computed by a Turing machine which computes a composition of sparse functions. The following theorem holds (see Appendix 6.1 for a proof):

Theorem 1. *For computable functions with Lipschitz continuous first derivatives, compositional sparsity is equivalent to $poly(d)$ computability.*

The restriction to computable functions in the theorem is to avoid the problem that several functions on the reals are not computable according to standard definitions [7] (the first, but not only, issue here is that real numbers are not computable). An alternative way that avoids the issue of computability of functions on the reals is to work with Boolean functions throughout, replacing the MP theorem with its Boolean version, as sketched in section 6.7.1.

Without the assumption of smooth target functions, which is equivalent to smooth constituent functions for a compositional function, there is no equivalence between compositional sparsity and computability. The Appendices discuss the situation.

One of the main implications of theorem 1 is that in practice all functions may be approximated by an appropriate neural network without curse of dimensionality. This, in turn, provides theoretical foundations for

1. using deep sparse networks in rather general learning tasks where the parametric approximation is optimized by training on a training set;
2. using sparse tensor representations such as the Hierarchical Tucker format [9, 10] in representing rather generic functions.

3 Learning theory: compositional sparsity leads to order-of-magnitude better generalization

The theorems above imply that deep, sparse RELU networks can be used for training, that is for optimization of the function class wrt given data and a chosen loss function. The optimized network may or may not generalize well. The next question provides some light on this issue, independently of whether the optimization is in the underparametrized or overparametrized case. The latter is more relevant for current usage.

It is possible to prove that sparsity of a network approximating a (sparse) target function reduces its complexity by orders of magnitude. In particular, the following result holds [11]

Theorem 2. *(informal) The Rademacher complexity of a deep overparametrized network is much smaller for convolutional layers with a local kernel than for a dense layers: if the kernel has dimensionality k and the dimensionality of the layer is n, then the contribution of the layer to the Rademacher complexity of the network is $\sqrt{\frac{k}{n}}\|W\|$ instead of $\|W\|$, where $\|W\|$ is the Frobenius norm of the layer weight matrix W.*

[1] Classical analysis suggests that this in turn can be guaranteed by compactness, that is, equicontinuity, that is, uniformly bounded derivatives of the target functions.

In complete analogy with the approximation result the key property here is locality of the convolution kernel ($k << n$) and *not weight sharing*. In a somewhat similar way, I conjecture that lower rank of the weight matrices (dense or convolutional) can improve generalization. Notice that an equivalent result for underparametrized networks follows directly from considerations of VC dimension (see Appendix, section 8 in [12]). The novelty here is to show that sparsity can lead to generalization in the overparametrized case, when weight decay, that is regularization, is present[2].

4 Optimization and open questions

4.1 The sparse graph is known: CNNs

In the underparametrized case, recent work (see for instance [2]) has shown that an optimal tradeo between approximation and generalization error can be achieved, assuming that optimization finds a good minimum. In the much more interesting overparametrized square loss case, generalization depends on solving a sort of *regularized ERM*, that consists of finding minimizers of the empirical risk with zero loss, and then select the one with lowest complexity [14]. Recent work [11] has provided theoretical and empirical evidence that this can be accomplished by SGD provided that the following conditions are satisfied:

1. the sparse function graph of the underlying regression function is assumed to be known and to be reflected in the architecture of the approximating network;
2. the network is overparametrized allowing zero empirical loss;
3. the loss function is the regularized (e.g. with weight decay) square loss (or an exponential loss) function.

Thus the conjecture is that this optimization problem can be solved by SGD if the graph of the underlying regression function *is known and takes the form of a compositionally sparse graph*, such as, for instance, a convolutional network.

Empirical evidence suggests that for dense networks that do not reflect the sparse graph the same problem cannot be solved using ℓ_2 minimization. Sparsity must be explicit in the architecture of the network for ℓ_2 minimization to work. Theoretical and empirical evidence points in the same direction: generalization bounds are several orders of magnitudes better for CNNs than for dense networks and close to be non-vacuous for CNNs (and presumably for other sparse networks).

The performance of trained neural networks is robust to harsh levels of pruning[3]. This empirical fact supports the hypothesis that the network should reflect the sparsity of the underlying target function. However, ℓ_2 optimization cannot attain sparsity by itself, since it preserves very small weights that should in fact be zero. Appendix 6.9 is about pruning and related issues. Empirically it seems that the graph of the target function needs to be known approximately. I conjecture that it is su cient that the sparse network contains as a subgraph the target function graph.

The conclusion is that if the sparse graph is known and approximately implemented in the architecture of the network minimization in either ℓ_2 or ℓ_1 should work. A conjecture may be

Conjecture 1. *If the sparse graph of the target function is reflected in the network and zero loss is attained then both ℓ_1 and ℓ_2 minimization lead to solutions with good expected error.*

In addition, it is likely that ℓ_1 minimization – when successful – can lead to pruned networks wrt ℓ_2 optimized networks.

[2] And even without weight decay under appropriate conditions that induce small ρ – which is the product of the Frobenius norm of the weight matrices [13].

[3] Coupled with the ever-growing size of deep learning models, this observation has motivated extensive research on learning sparse models.

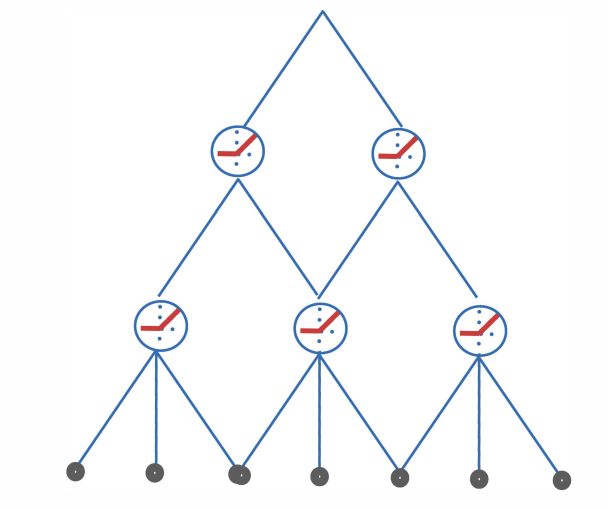

Figure 1: *The network here – similar to a CNN – reflects in a "hardwired" way the sparse compositional function graph of the target function. The function graph is supposed to be known.*

4.2 The sparse graph is unknown

The second part of the argument is about the case of unknown function graph and sparsity constraints in optimization. I propose the conjecture that when the sparse graph structure of the underlying regression function is not known, optimization with sparsity constraints is needed. In particular, two situations should be considered. The first main one is focused on dense networks under sparsity constraints, the second on transformers.

4.2.1 Dense networks optimized under sparsity constraints

For dense networks it is known that a CNN-like inductive bias can be learned from data and through training by using a modified ℓ_1 regularization. Consistent with this empirical finding, pruning of a dense network by using iterative magnitude pruning (IMP) also seems to work.

4.2.2 Self-attention as flexible sparsity

For transformers a key question is: how does self-attention find the sparse set of tokens that are input to a processing node (that is are the variables of a constituent function)? I propose the conjecture that self-attention selects, for each token, the relevant other tokens in the sequence, that is a flexible node of a hardwired CNN network. An equivalent formulation is

Conjecture 2. *Self-attention in a transformer selects a sparse subset of variables (e.g. tokens) for each RELU unit, trying to mimic the compositional sparsity of the underlying target function.*

This conjecture leaves open the interesting question of whether self-attention can deal with all compositionally sparse functions. A more likely possibility is that not all sparse functions are easy to learn by transformers.

The matrices W_Q and W_K that are set during the training time in such a way that $A = QK^T$ – with $Q = xW_Q$, $K = xW_K$ – may be together somewhat similar to a learned Malanhobis distance. In Appendix 6.10 the normalized softmax $H_D(x) = xH(x)$ (with H being a threshold on x) induces sparsity in the selection of "active" connections, preferring only a small number of very similar token – where the similarity is tuned via the learned W_H, W_Q matrices. After the attention step, there is a one-layer dense network on the linear combination of a few tokens – this is very similar to the node of a convolutional network, but with soft-wired connections instead of hard-wired.

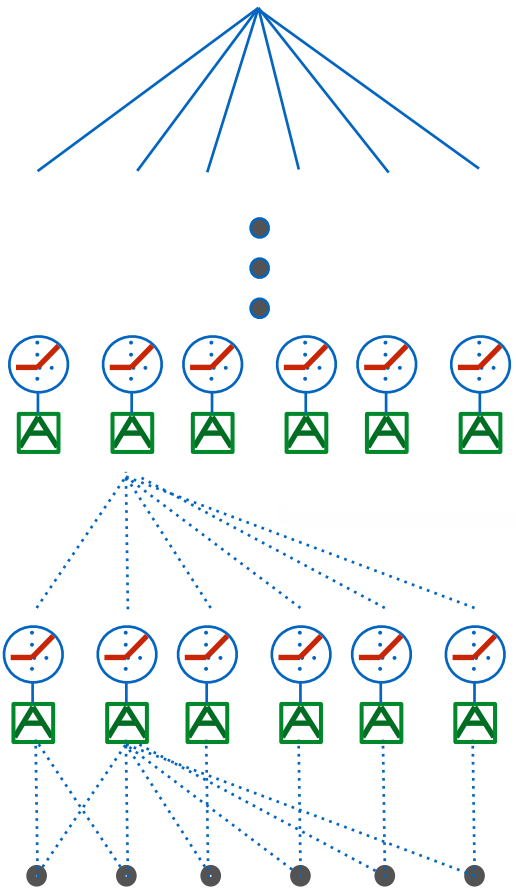

Figure 2: *Here attention (followed by a one-layer RELU network) selects for each input token its connections to other tokens, e ciently instantiating a network that reflects a compositionally sparse function graph. Each input here is a token, that is a vector, such as a patch of an image. The "A" box is the self-attention algorithm; the RELU circle represents a one-layer NN.*

5 Summary

This paper introduces a theoretical framework to explain why deep networks work and what are the properties of di erent architectures.

The *key claim* is about the world, that is about the tasks that networks could try to learn. The claim is that all functions that in practice can be approximated/computed must have a representation with the property of compositional sparsity, that is they can be represented as compositional functions with a function graph comprising constituent functions – each with a bounded, "small" dimensionality. The connection with deep networks depends on a conjecture stating that for functions with bounded first order derivatives *computable approximation is equivalent to compositional sparsity.*

Consider now sparse networks: if each unit in a certain layer of a deep network receives inputs from only a small subset of the units below, the corresponding weight matrix is sparse, with several zero components in each row. Somewhat surprisingly, sparsity of the network is a key property for good *generalization*. An interesting case of this sparsity is represented by convolutional layers with a local kernel. The following property then holds: *the Rademacher complexity of a deep network is much smaller for convolutional deep networks with local kernels, relative to dense networks.* In complete analogy with the approximation result the key property here is locality of the convolution kernel and not weight sharing. In a similar way, lower rank of the weight matrices can improve generalization bounds.

From the point of view of *optimization*, two main cases should be considered: 1) the sparse graph of

the underlying target functions is known, 2) the sparse graph is unknown.

1. In the overparametrized square loss case, generalization depends on solving a sort of *regularized ERM*, that consists of finding minimizers of the empirical risk with zero loss, while selecting the one with lowest complexity. Recent work has provided theoretical and empirical evidence that this can be accomplished by SGD (with norm regularization under the square loss or without regularization under an exponential loss) with weight decay in the overparametrized case when the network architecture reflects the sparse graph of the target function. This implies that this optimization problem can be solved if the graph of the underlying regression function *is known and takes the form of a compositionally sparse graph, such as, for instance, a convolutional network.*

   Empirical (and perhaps theoretical) evidence shows that for dense networks the same problem cannot be solved using ℓ_2 minimization. Sparsity must be explicit in the architecture of the network for ℓ_2 minimization to work.

2. The second part of the theory is about the case of unknown function graph and sparsity constraints in optimization. In particular, two situations should be considered. The main one is focused on transformers, the second on dense networks under sparsity constraints.

   For transformers the conjecture is that the self-attention layer finds the sparse graph structure of the underlying regression function. I will show that the stages of self-attention and MLP with normalization and residual connections can be seen as a sparsification step followed by a one-layer MLP.

   For dense networks it is known that a CNN-like inductive bias can be learned from data and through training by using a modified ℓ_1 regularization. Consistent with this empirical finding, pruning of a dense network by using iterative magnitude pruning (IMP) also works.

Summary Why do deep networks work as well as they do? The answer I propose here is that certain deep architectures – such as CNNs and transformers – exploit a general property of all e□ciently computable smooth functions: their compositional sparsity.

Acknowledgments I thanks Fabio Anselmi, Sophie Langer, Tomer Galanti, Akshay Rangamani, Shimon Ullman, Yaim Cooper, Gitta Kutyniok, and the Compositional Sparsity (CoSp) Collaboration (Santosh Vempala, Hrushikesh Mhaskar, Eran Malach, Seth Lloyd) for illuminating discussions. This material is based upon work supported by the Center for Minds, Brains and Machines (CBMM), funded by NSF STC award CCF-1231216. This research was also sponsored by grants from the National Science Foundation (NSF-0640097, NSF-0827427), AFSOR-THRL (FA8650-05-C-7262) and Lockeed-Martin.

References

[1] Wolfgang Dahmen. Compositional sparsity, approximation classes, and parametric transport equations, 2022.

[2] Michael Kohler and Sophie Langer. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function". *The Annals of Statistics*, 48(4):1906 – 1910, 2020.

[3] Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020.

[4] Markus Bachmayr, Anthony Nouy, and Reinhold Schneider.

[5] Gitta Kutyniok. Discussion of: "Nonparametric regression using deep neural networks with ReLU activation function". *The Annals of Statistics*, 48(4):1902 – 1905, 2020.

[6] Alexander Bastounis, Anders C Hansen, and Verner Vlačić. The extended smale's 9th problem – on computational barriers and paradoxes in estimation, regularisation, computer-assisted proofs and learning, 2021.

[7] Holger Boche, Adalbert Fono, and Gitta Kutyniok. Limitations of deep learning for inverse problems on digital hardware, 2022.

[8] H. Mhaskar. Dimension independent bounds for general shallow networks. *Neural Networks*, 2020.

[9] Wolfgang Hackbusch and Stephan Kühn. A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications*, 15:706–722, 2009.

[10] Lars Grasedyck. Hierarchical Singular Value Decomposition of Tensors. *SIAM J. Matrix Anal. Appl.*, (31,4):2029–2054, 2010.

[11] M. Xu, A. Rangamani, A.and Banburski, Q.and Galanti Liao, T., and T. Poggio. Dynamics and neural collapse in deep classifiers trained with the square loss. CBMM Memo 117, CBMM, MIT, 2022.

[12] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. Theory I: Why and when can deep - but not shallow - networks avoid the curse of dimensionality. Technical report, CBMM Memo No. 058, MIT Center for Brains, Minds and Machines, 2016.

[13] Mengjia Xu, Akshay Rangamani, Qianli Liao, Tomer Galanti, and Tomaso Poggio. Dynamics in deep classifiers trained with the square loss: normalization, low rank, neural collapse and generalization bounds. *Research*, 2023.

[14] Eran Malach and Tomaso Poggio. Compositional locality and optimization. *CBMM Memo*, 2023.

[15] A. N. Kolmogorov. On the representation of continuous functions of several variables by superposition of continuous functions of one variable and addition. *Dokl. Akad. Nauk SSSR*, 114:953–956, 1957.

[16] V.I. Arnol'd. On functions of three variables. *Dokl. Akad. Nauk SSSR*, 114:679–681, 1957.

[17] J.P. Kahane. Sur le theoreme de superposition de Kolmogorov. *Journal of Approximation Theory*, 13:229–234, 1975.

[18] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta Numerica*, 8:143–195, 1999.

[19] GG Lorentz. Approximation of functions, athena series. *Selected Topics in Mathematics*, 1966.

[20] A. G. Vitushkin and G.M. Henkin. Linear superposition of functions. *Russian Math. Surveys*, 22:77–125, 1967.

[21] A. G. Vitushkin. On Hilbert's thirteenth problem. *Dokl. Akad. Nauk SSSR*, 95:701–704, 1954.

[22] H.N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, pages 829– 848, 2016.

[23] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova. Nonlinear approximation and (deep) relu networks. *arXiv e-prints*, page arXiv:1905.02199, May 2019.

[24] Hrushikesh Narhar Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80, 1993.

[25] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: a tensor analysis. *CoRR*, abs/1509.0500, 2015.

[26] Behnam Neyshabur. Towards learning convolutions from scratch. *CoRR*, abs/2007.13657, 2020.

[27] Franco Pellegrini and Giulio Biroli. Sifting out the features by pruning: Are convolutional networks the winning lottery ticket of fully connected ones? *CoRR*, abs/2104.13343, 2021.

[28] Stéphane d'Ascoli, Levent Sagun, Joan Bruna, and Giulio Biroli. Finding the needle in the haystack with convolutions: on the benefits of architectural bias, 2019.

[29] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision, 2021.

[30] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *CoRR*, abs/1712.06541, 2017.

[31] Patrick Rebeschini. Algorithmic foundations of learning lecture 3: Rademacher complexity, 2020.

[32] M. Xu, T. Poggio, and T. Galanti. Complexity bounds for sparse networks. *CBMM memo 1XX*, 2022.

6 Appendices

6.1 Proof of Theorem 1

Consider smooth real-valued functions in d variables, that is functions with Lipschitz continuous first derivatives. The MP theorem (see Appendix 6.5) shows that the compositionally sparsity of computable functions implies their e□cient computability by deep RELU networks, assuming computability of the latter.

For the converse assume that a function is e□ciently computable by a Turing machine. This means that the function can be represented by the composition of at most a polynomial number of sparse functions, each corresponding to the basic read-write step in a Turing machine.

6.2 Compositionality and sparsity

6.2.1 All continuous function are compositionally (non-smooth) sparse

An obvious question is how "big" is the class of compositionally sparse functions wrt the class of all continuous functions (all functions are trivially compositional, since every function can be composed with the identity function). An answer for continuous functions was given by the solution of Hilbert's thirteen problem due to Kolmogorov and Arnold [15, 16, 17]: every continuous functions can be represented *exactly* as a compositions of $poly(d)$ functions of one variable, that is as the composition of sparse functions.

Theorem 3. *All continuous functions are compositionally sparse, that is they have an exact representation in terms of sparse non-smooth constituent functions.*

In this representation, the constituent functions are very non-smooth, that is $s = 0$. Appendix 6.4 has more information. This fact implies that the Kolmogorov-Arnold representation is not e□ciently computable in the sense that continuous functions cannot be approximated with non-exponential rates.

6.2.2 Compositional S-sparsity implies computable approximation

Let us first define smooth sparse functions.

Definition 3. *A S-sparse (e.g. smoothly sparse) compositional function of d variables is a sparse compositional function with constituent functions that have bounded first derivatives.*

With this definition, we can then reformulate the MP theorem (see Appendix) as

Theorem 4. *Compositionally S-sparse functions are e ciently computable.*

6.2.3 Computable approximation is not equivalent to S-sparsity

Networks with non-smooth RELU can approximate arbitrarily well S-sparse functions. They are compositionally sparse but not smooth. Thus one can imagine, in a theorem such as d'Arzela'-Ascoli, that there is a sequence of sparse compositional f_n – provided by RELU networks – converging to a smooth compositionally sparse f without the f_n being smooth themselves! CAN THIS BE TRUE?

6.3 Are computable functions compositionally sparse?

Let us state an obvious conjecture.

Conjecture 3. *Functions with an e ciently computable approximation are compositionally sparse.*

If the above were true the story would take the following form

Conjecture 4. *Compositional S-sparse functions have an e ciently computable approximation; functions with an e ciently computable approximation are compositionally sparse.*

This would mean that the class of computable functions is larger than the class of functions for which we can guarantee e cient approximation.

6.3.1 Remarks

- The curse of dimensionality bound is tight, see Pinkus [18] comments about Maiorov's results.
- All functions have a compositional representation which is not unique, since, in general, a function admits more than one compositional graph representation.
- A definition of sparse compositional functions must include an implicit or explicit constraint on the number of nodes in the associated DAG, that is on the number of constituent functions. In the specific example of a binary tree graph the depth of the graph increases only logarithmically as the dimensionality d increases.
- The curse of dimensionality holds not only for real-valued continuous functions but also for Boolean functions. A specific constructions of relevant Boolean functions is discussed in the Appendix (see 6.8.1 and [12]).
- Because of theorem 8 (see also [12]) all compositionally S-sparse, continuous, real-valued functions can be approximated by a Boolean compositionally sparse function.
- Computable by a Turing machine usually doesn't assume polynomial time/space complexity, but the term *e ciently computable* used in this paper implies polynomial time/space requirements.
- All compositionally S-sparse functions are e ciently computable by a Turing machine and admit an e cient approximator in terms of a deep RELU network with an architecture reflecting the sparse function graph.
- E cient approximation can be thought of as the computation by a Turing machine of a Boolean function. Such a Boolean function is the composition of the Boolean functions that approximate the constituent functions. The complexity of the resulting function is at most $O(poly(d))$.
- Observe that a non-sparse continuous function $f : \mathcal{R}^{1000} \to \mathcal{N}$ requires a memory of up to $> 10^{1000}$ bits, larger than the number of protons in the Universe, which is in the order of 10^{80}. A classical example is a dense polynomial in d dimensions of arbitrarily high degree.
- There are (compositional) functions with constituent functions that are not smooth and are e ciently computable. An example of such functions are the Boolean functions that are ultimately used in a Turing machine to represent (and approximate) a continuous smooth function. Another example of (compositional) functions with constituent functions that are not smooth are some of the functions in the Takagi class (see Appendix). Furthermore, deep RELU networks are compositionally sparse but non smooth functions.
- Stability of the approximation in probability wrt perturbations of the inputs seems an important requirement for any function and its approximation. This seems to imply smoothness of the constituent functions when they are continuous.

6.4 Kolmogorov's theorem[15]

Theorem 5. *(Kolmogorov, 1957; see also [19]). There exist fixed (universal) increasing continuous functions $h_{pq}(x)$, on $I = [0, 1]$ so that each continuous function f on I^d can be written in the form*

$$f(x_1, \ldots, x_d) = \sum_{q=1}^{2d+1} g_q \left(\sum_{p=1}^{d} h_{pq}(x_p) \right),$$

where g_q are properly chosen continuous functions of one variable.

This result asserts that every multivariate continuous function can be represented by the superposition of a small number of univariate continuous functions. In terms of networks this means that every continuous function of many variables can be computed by a network with two hidden layers, whose hidden units compute continuous functions (the functions g_q and h_{pq}).

The interpretation of Kolmogorov's theorem in term of networks is very appealing: the representation of a function requires a fixed number of nodes, polynomially increasing with the dimension of the input space. Unfortunately, these results are somewhat pathological and their practical implications are very limited. The problem lies in the inner functions of Kolmogorov's formula: although they are continuous, theorems of Vitushkin and Henkin [20] prove that they must be highly non-smooth. One could ask if it is possible to find a superposition scheme in which the functions involved are smooth. The answer is negative, even for two variable functions, and was given by [21] with the following theorem:

Theorem 6. *(Vitushkin 1954). There are $r(r = 1, 2, \ldots)$ times continuously di erentiable functions of $n \geq 2$ variables, not representable by superposition of r times continuously di erentiable functions of less than n variables; there are r times continuously di erentiable functions of two variables that are not representable by sums and continuously di erentiable functions of one variable.*

6.5 MP theorem

An interesting, specific pair (function class, approximator) can be given in terms of the class of compositional smooth functions and of deep RELU networks. In fact, the following theorem by Mhaskar and Poggio [22] shows that functions that are *compositionally S-sparse* can be approximated arbitrarily well by deep, sparse RELU networks with $poly(d)$ parameters. The motivation for the result was the classical curse of dimensionality: an upper bound on the number of parameters needed for approximation of a continuous function supported on a compact domain of $\mathcal{R}^d$ is $W = \mathcal{O}(\epsilon^{-\frac{d}{s}})$, where ϵ is the approximation error and s – the number of bounded derivatives – is a a measure of smoothness of the function with $s \geq 1$. The curse can be avoided by shallow or deep networks if s is large and in particular if s grows with d. The curse can also be avoided by deep networks, but not by shallow ones, if the function is *compositionally S-sparse*, that is if the function graph is such that each constituent function has low e□ective dimensionality[4].

Theorem 7. *Let $\mathcal{G}$ be a DAG, n be the number of source nodes, and for each $v \in V$, let d_v be the number of in-edges of v. Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be a compositional $\mathcal{G}$-function, where each of the constituent functions is in the Sobolev space $W_{m_v}^{d_v}$. Consider shallow and deep networks with infinitely smooth activation function. Then deep networks - with an associated graph that corresponds to the graph of f – avoid the curse of dimensionality in approximating f for increasing n, whereas shallow networks cannot directly avoid the curse. In particular, the complexity of the best approximating shallow network is exponential in n*

$$N_s = \mathcal{O}\left(\epsilon^{-\frac{n}{m}}\right),$$

where $m = \min_{v \in V} m_v$, while the complexity of the deep network is

$$N_d = \mathcal{O}\left(\sum_{\eta \in V} \epsilon^{-d_v/m_v} \right).$$

tpKill

[4] I use the term *compositional sparsity* following [1] instead of another equivalent term we used earlier: *hierarchical compositionality*.

6.6 Tagaki class of functions (from [23])

Here are examples of functions F that are well approximated by ReLU networks. For the most part, these functions cannot be well approximated by standard approximation families. Let us recall that functions of the form

$$F = \sum_{k\geq 1} t^k g\left(\psi^{\circ k}\right), \quad |t| < 1, \tag{1}$$

with $\psi : [0,1] \to [0,1]$ and $g : [0,1] \to \mathbb{R}$, provide primary examples of self similar functions and dynamical systems. If $g \in \Upsilon^{W_1,\ell}$ and $\psi \in \Upsilon^{W_2,\ell}$, with $W_1 + W_2 = W$, Proposition 4.4 in [23] implies that the partial sum $S_m := \sum_{k=1}^m t^k g\left(\psi^{\circ k}\right)$ belongs to $\bar{\Upsilon}^{W,\ell(m+1)} \subset \bar{\Upsilon}_{\ell(m+1)}$. Therefore, in this case, the function F defined via 1 is approximated by the partial sum S_m with exponential accuracy by ReLU networks, that is

$$\sigma_{\ell(m+1)}(F, \underline{\underline{\Upsilon}})_{C[0,1]} \leq Ct^{m+1}, \quad |t| < 1.$$

Now, we consider a special class of functions. For this purpose, we recall that the hat function $H \in \Upsilon^{2,1}$ and its k-fold composition $H^{\circ k} := H \circ H \circ \cdots \circ H$, according to the composition property belongs to $\underline{\Upsilon}^{2,k}$. The collection of all such functions is called the Takagi class. It contains a number of interesting and important examples.

For instance the Takagi function

$$T := \sum_{k\geq 1} 2^{-k} H^{\circ k}$$

can be approximated with exponential accuracy by ReLU networks with roughly $W^2 m$ parameters. However, T is nowhere di□erentiable and so it has very little smoothness in the classical sense. This means that all of the traditional methods of approximation will fail miserably to approximate it. Note that the function T has self similarity, in that it satisfies a simple refinement equation. Other functions in the Takagi class do not satisfy a Lipschitz condition of any order and yet they can be approximated to exponential accuracy by RELU networks. Many functions from the Takagi class are fractals, in the sense that the Hausdor□ dimension of their graph is strictly greater than one. The main point to draw from these examples is that the approximation classes corresponding to RELU networks contain many functions which are not smooth in any classical sense.

6.7 Boolean functions

One of the most important tools for theoretical computer scientists for the study of computable functions is the study of Boolean functions, that is functions of n Boolean variables. A key tool here is the Fourier transform over the Abelian group $\mathcal{Z}_2^n$. This is known as Fourier analysis over the Boolean cube $\{-1,1\}^n$. The Fourier expansion of a Boolean function $f : \{-1,1\}^n \to \{-1,1\}$ or even a real-valued Boolean function $f : \{-1,1\}^n \to [-1,1]$ is its representation as a real polynomial, which is multilinear because of the Boolean nature of its variables. Thus for Boolean functions their Fourier representation is identical to their polynomial representation. Unlike functions of real variables, the full finite Fourier expansion is exact, instead of an approximation. There is no need to distingush between trigonometric and real polynomials. Most of the properties of standard harmonic analysis are otherwise preserved, including Parseval theorem. The terms in the expansion correspond to the various monomials; the low order ones are parity functions over small subsets of the variables and correspond to low degrees and low frequencies in the case of polynomial and Fourier approximations, respectively, for functions of real variables.

6.7.1 *Boolean Functions and Sparsity*

The curse of dimensionality holds not only for real-valued continuous functions but also for Boolean functions (see discussion in [12]). The following theorem states that compositionally sparse Boolean functions avoid the curse.

Theorem 8. *Let $\mathcal{G}$ be a DAG, n be the number of source nodes, and for each $v \in V$, let d_v be the number of in-edges of v. Let $f : \{1,-1\}^n \mapsto \mathbb{R}$ be a compositional $\mathcal{G}$-function, where each of the constituent functions is a function g in d_v Boolean variables, $g : \{1,-1\}^d \mapsto \{1,-1\}$. Consider shallow and deep networks with a RELU activation functions or a hard threshold. Then deep networks - with an associated graph that corresponds*

to the graph of f - can avoid the curse of dimensionality in approximating f for increasing d, since the number of required parameters is $\propto \mathcal{O}(poly(d))(max_v d_v)$.

The converse results from the observation that the Fourier representation of a Boolean function in d variables can have up to N non-zero monomials where $N = \binom{2d}{d} = \frac{2d!}{d!d!}$. Thus $N > 2^d$. Clearly a Boolean functions with all non-zero monomials is not e□ciently computable. In fact the following holds

Theorem 9. *All e ciently computable Boolean functions are compositionally sparse, that is they can be represented as the composition of a $\leq poly(d)$ number of constituent functions with a bounded "small" dimensionality.*

Combining the previous two theorems we obtain the following

Theorem 10. *For Boolean functions, e ciently computable is equivalent to compositionally sparse.*

6.8 Spline approximations, Boolean functions and tensors

Consider the case of a multivariate smooth function $f : [0, 1]^d \to \mathbf{R}$. Suppose to discretize it by a set of piecewise constant splines and their tensor products[5]. Each coordinate is e□ciently replaced by n boolean variables. This results in a d-dimensional table with $N = n^d$ entries. This in turn corresponds to a Boolean function $f : \{0, 1\}^N \to \mathbf{R}$.

- Every smooth function f can be approximated by an epsilon-close binary function f_B. Binarization of $f : \mathbf{R}^n \to \mathbf{R}$ is done by using k partitions for each variable x_i and indicator functions. Thus $f \mapsto f_B : \{0, 1\}^{kn} \to \mathbf{R}$ and $sup|f - f_B| \leq \epsilon$, with ϵ depending on k and bounded Df.
- f_B can be written as a polynomial (a Walsh decomposition) $f_B \approx p_B$. It is always possible to associate a p_b to any f, given ϵ.
- One can think about tensors in terms of d-dimensional tables. The framework of hierarchical decompositions of tensors – in particular the *Hierarchical Tucker format* – is closely connected to our notion of compositionality. Interestingly, the hierarchical Tucker decomposition has been the subject of recent papers on Deep Learning (for instance see [25]). This work, as well more classical papers [10], does not characterize directly the class of functions for which these decompositions are e□ective approximations. Notice that tensor decompositions *assume* that the sum of polynomial functions of order d is sparse (see eq. at top of page 2030 of [10]). Our results provide a rigorous grounding in terms of approximation theory for papers on tensors related to deep learning. There is obviously a wealth of interesting connections with approximation theory that should be explored.

6.8.1 *On multivariate function approximation*

Consider a smooth multivariate continuous function $f : [0, 1]^d \to \mathbf{R}$ discretized by tensor basis functions:

$$\phi_{(i_1,...,i_d)}(x_1, ..., x_d) := \prod_{\mu=1}^{d} \phi_{i_\mu}(x_\mu), \tag{2}$$

with $\phi_{i_\mu} : [0, 1] \to \mathbf{R}, 1 \leq i_\mu \leq n_\mu, 1 \leq \mu \leq d$
to provide

$$f(x_1, ..., x_d) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} c(i_1, ..., i_d)\phi(i_1, ..., i_d)(x_1, ..., x_d). \tag{3}$$

The one-dimensional basis functions could be polynomials (as above), indicator functions, polynomials, wavelets, or other sets of basis functions. The total number N of basis functions scales exponentially in d as $N = \prod_{\mu=1}^{d} n_\mu$ for a fixed smoothness class m (it scales as $\frac{d}{m}$).

[5] An argument similar to the one below for polynomials was used by Mhaskar[24] to show that a multivariate tensor product spline can be synthesized exactly using a deep network with activation function $(x_+)^2$; with Yarotsky's theorem, this can be translated into deep networks with ReLU functions without incurring in saturation phenomena.

We can regard neural networks as implementing some form of this general approximation scheme. The problem is that the type of operations available in the networks are limited. In particular, most of the networks do not include the product operation (apart from "sum-product" networks also called "algebraic circuits") which is needed for the straightforward implementation of the tensor product approximation described above. Equivalent implementations can be achieved however. We describe nest how networks with a univariate ReLU nonlinearity may perform multivariate function approximation with a polynomial basis and with a spline basis respectively. The first result is known and we give it for completeness. The second is simple but new.

Neural Networks: polynomial viewpoint One of the choices listed above leads to polynomial basis functions. The standard approach to prove degree of approximations uses polynomials. It can be summarized in three steps:

1. Let us denote with $\mathcal{H}_k$ the linear space of homogeneous polynomials of degree k in $\mathbf{R}^n$ and with $P_k = \bigcup_{s=0}^{k} \mathcal{H}_s$ the linear space of polynomials of degree at most k in n variables. Set $r = \binom{n-1+k}{k} = dim\mathcal{H}_k$ and denote by π_k the space of univariate polynomials of degree at most k. We recall that the number of monomials in a polynomial in d variables with total degree $\leq N$ is $\binom{d+N}{d}$ and can be written as a linear combination of the same number of terms of the form $((w, x) + b)^N$.

   We first prove that

$$P_k(x) = span(((w^i, x))^s : i = 1, \cdots, r,\ s = 1, \cdots, k \tag{4}$$

   and thus, with, $p_i \in \pi_k$,

$$P_k(x) = \sum_{i=1}^{r} p_i((w_i, x)). \tag{5}$$

   Notice that the e□ective r, as compared with the theoretical r which is of the order $r \approx k^n$, is closely related to the *separation rank* of a tensor. Also notice that a polynomial of degree k in n variables can be represented exactly by a network with $r = k^n$ units.

2. Second, we prove that each univariate polynomial can be approximated on any finite interval from

$$\mathcal{N}(\sigma) = span\{\sigma(\lambda t - \theta)\}, \lambda, \theta \in \mathbf{R} \tag{6}$$

   in an appropriate norm.

3. The last step is to use classical results about approximation by polynomials of functions in a Sobolev space:

$$E(\mathcal{B}_p^m; P_k; L_p) \leq Ck^{-m} \tag{7}$$

   where $\mathcal{B}_p^m$ is the Sobolev space of functions supported on the unit ball in $\mathbf{R}^n$.

The key step from the point of view of possible implementations by a deep neural network with ReLU units is step number 2. A univariate polynomial can be synthesized – in principle – via the linear combination of ReLUs units as follows. The limit of the linear combination $\frac{\sigma((a+h)x+b)-\sigma(ax+b)}{h}$ contains the monomial x (assuming the derivative of σ is nonzero). In a similar way one shows that the set of shifted and dilated ridge functions has the following property. Consider for $c_i, b_i, \lambda_i \in \mathbf{R}$ the space of univariate functions

$$\mathcal{N}_r(\sigma) = \left\{ \sum_{i=1}^{r} c_i \sigma(\lambda_i x - b_i) \right\}. \tag{8}$$

The following (see Propositions 3.6 and 3.8 in [18]) holds

Lemma 1. *If $\sigma \in \mathcal{C}()$ is not a polynomial and $\sigma \in C^\infty$, the closure of $\mathcal{N}$ contains the linear space of algebraic polynomial of degree at most $r-1$.*

Since $r \approx k^n$ and thus $k \approx r^{1/n}$ equation 7 gives

$$E(\mathcal{B}_p^m; P_k; L_p) \leq Cr^{-\frac{m}{n}}. \tag{9}$$

Neural Networks: splines viewpoint Another choice of basis functions for discretization consists of splines. In particular, we focus for simplicity on indicator functions on partitions of $[0,1]$, that is piecewise constant splines. Another attractive choice are Haar basis functions. If we focus on the binary case, section 6.8.1 tells the full story that does not need to be repeated here. We just add a note on establishing a partition.

Suppose that $a = x_1 < x_2 \cdots < x_m = b$ are given points, and set Δx the maximum separation between any two points.

- If $f \in C[a,b]$ then for every $\epsilon > 0$ there is a $\delta > 0$ such that if $\Delta x < \delta$, then $|f(x) - Sf(x)| < \epsilon$ for all $x \in [a,b]$, where Sf is the spline interpolant of f.
- if $f \in^2 [a,b]$ then for all $x \in [a,b]$

$$|f(x) - Sf(x)| \leq \frac{1}{8}(\Delta x)^2 max_{a \leq z \leq b} |f''(z)|$$

The first part of the Proposition states that piecewise linear interpolation of a continuous function converges to the function when the distance between the data points goes to zero. More specifically, given a tolerance, we can make the error less than the tolerance by choosing Δx su□ciently small. The second part gives an upper bound for the error in case the function is smooth, which in this case means that f and its first two derivatives are continuous.

Boolean functions and curse of dimensionality The classical curse of dimensionality result is based on polynomial approximation. Because of the n-width result other approaches to approximation cannot yield better rates than polynomial approximation. It is, however, interesting to consider other kinds of approximation that may better capture what deep neural network with the ReLU activation functions implement in practice.

A network with non-smooth ReLU activation functions can approximate any continuous function. A weakness of this results wrt to other ones is that it is valid in the L_2 norm but not in the sup norm. This weakness does not matter in practice since a discretization of real number, say, by using 64 bits floating point representation, will make the class of functions a finite class for which the result is valid also in the L_∞ norm. The logic of the argument is simple:

- Consider the constituent functions of a compositional function with a function graph given by a binary tree, that is functions of two variables such as $g(x_1, x_2)$. Assume that g is Lipschitz with Lipschitz constant L. Then for any ϵ it is possible to set a partition of x_1, x_2 on the unit square that allows piecewise constant approximation of g with accuracy at least ϵ in the sup norm.
- We show then that a multilayer network of ReLU units can compute the required partitions in the L_2 norm and perform piecewise constant approximation of g.

Notice that partitions of two variables x and y can in principle be chosen in advance yielding a finite set of points $0 =: x_0 < x_1 < \cdots < x_k := 1$ and an identical set $0 =: y_0 < y_1 < \cdots < y_k := 1$. In the extreme, there may be as little as one partition – the binary case. In practice, the partitions can be assumed to be set by the architecture of the network and optimized during learning. The simple way to choose partitions is to choose an interval on a regular grid. The other way is an irregular grid optimized to the local smoothness of the function. This is the di□erence between fixed-knots splines and free-knots splines.

I describe next a specific construction.

Here is how a linear combination of ReLUs creates a unit that is active if $x_1 \leq x \leq x_2$ and $y_0 \leq y \leq y_1$. Since the ReLU activation t_+ is a basis for piecewise linear splines, an approximation to an indicator function (taking the value 1 or 0, with knots at x_1, $x_1 + \eta$, x_2 $x_2 + \eta$,) for the interval between x_1 and

x_2 can be synthesized using at most 4 units in one layer. A similar set of units creates an approximate indicator function for the second input y. A set of 3 ReLU's can then perform a min operations between the x and the y indicator functions, thus creating an indicator function in two dimensions.

In greater detail, the argument is as follows: For any $\epsilon > 0$, $0 \leq x_0 < x_1 < 1$, it is easy to construct an ReLU network R_{x_0,x_1} with 4 units as described above so that

$$\|\chi_{[x_0,x_1)} - R\|_{L^2[0,1]} \leq \epsilon.$$

We define another ReLU network with two inputs and 3 units by

$$\begin{aligned} \phi(x_1,x_2) &:= (x_1)_+ - (-x_1)_+ - (x_1 - x_2)_+ = \min(x_1,x_2) \\ &= \frac{x_1+x_2}{2} + \frac{|x_1-x_2|}{2}. \end{aligned}$$

Then, with $I = [x_0,x_1) \times [y_0,y_1)$, we define a two layered network with 11 units total by

$$\Phi_I(x,y) = \phi(R_{x_0,x_1}(x), R_{y_0,y_1}(y)).$$

Then it is not di□cult to deduce that

$$\begin{aligned} \|\chi_I - \Phi_I\|^2_{L^2([0,1]^2)} &= \int_0^1 \int_0^1 \\ &|\min(\chi_{[x_0,x_1)}(x), \chi_{[y_0,y_1)}(y)) - \\ &\min(R_{x_0,x_1}(x), R_{y_0,y_1}(y))|^2 dxdy \leq c\epsilon^2. \end{aligned}$$

Notice that in this case dimensionality is $n = 2$; notice that in general the number of units is proportional to k^n which is of the same order as $\binom{n+k}{k}$ which is the number of parameters in a polynomial in n variables of degree k. The layers we described compute the entries in the 2D table corresponding to the bivariate function g. One node in the graph (there are $n-1$ nodes in a binary tree with n inputs) contains $O(k^2)$ units; the total number of units in the network is $(n-1)O(k^2)$. This construction leads to the following result (for the special case of a compositionally sparse function with a binary tree function graph):

Lemma 2. *Compositional functions on the unit cube with an associated binary tree graph structure and constituent functions that are Lipschitz can be approximated by a deep network of ReLU units within accuracy ϵ in the L_2 norm with a number of units in the order of $O((n-1)L\epsilon^{-2})$, where L is the max of the Lipschitz constants among the constituent functions.*

Of course, in the case of machine numbers – the integers – we can think of zero as a very small positive number. In this case, the symmetric di□erence ratio $((x+\epsilon)_+ - (x-\epsilon)_+)/(2\epsilon)$ is the hard threshold sigmoidal function if ϵ is less than this smallest positive number. So, we have the indicator function exactly as long as we stay away from 0. From here, one can construct a deep network as usual.

Notice that the number of partitions in each of two variables that are input to each node in the graph is $k = \frac{L}{\epsilon}$ where L is the Lipschitz constant associated with the function g approximated by the node. Here the role of smoothness is clear: the smaller L is, the smaller is the number of variables in the approximating Boolean function. Notice that if $g \in W_1^2$, that is g has bounded first derivatives, then g is Lipschitz. However, *higher order smoothness* beyond the bound on the first derivative *cannot be exploited by the network* because of the non-smooth activation function[6].

We *conjecture that the construction above that performs piecewise constant approximation is qualitatively similar to what deep networks may represent after training*. Notice that the partitions we used correspond to a uniform grid, set a priori, depending on global properties of the function, such as a Lipschitz bound.

6.9 Pruning

Empirically it seems that dense networks cannot learn convolution under L_2 minimization but can under L_1 minimization. In particular, the possibility of learning CNN-like inductive bias from data and

[6] In the case of univariate approximation on the interval $[-1,1]$, piecewise linear functions with inter-knot spacing h gives an accuracy of $(h^2/2)M$, where M is the max absolute value of f''. So, a higher derivative does lead to better approximation : we need $\sqrt{2M/\epsilon}$ units to give an approximation of ϵ. This is a saturation though. Even higher smoothness does not help.

through training was investigated in [26]. It was shown that training using a modified L_1 regularization is a way to induce local masks for visual tasks. Consistent with this finding, pruning of a dense network by using iterative magnitude pruning (IMP) on FCNs trained on a low resolution version of ImageNet uncovers (see [27]) sub-networks characterized by local connectivity, especially in the first hidden layer, and masks leading to local features with patterns very reminiscent of the ones of trained CNNs[7].

This is similar to the following empirical result: enforcing sparsity during training leads to structures characterized by locality. [28] studies the role of CNN-like inductive biases by embedding convolutional architectures within the general FCN class. It shows that enforcing CNN-like features in an FCN can improve performance even beyond that of its CNN counterpart. Finally, [29] show that by considering a particular multilayer perceptron architecture, called MLP-mixer, some of the CNN features can be learned from scratch using a large training dataset.

6.10 Transformers

6.10.1 K, Q, V

$X \in \mathcal{R}^{T,d_{in}}$; $Q = XW_Q$ with $W_Q \in \mathcal{R}^{d_{in},d_k}$; $K = XW_K$ with $W_K \in \mathcal{R}^{d_{in},d_k}$; $V = XW_V$ with $W_V \in \mathcal{R}^{d_{in},d_{out}}$

Lemma 3. *The matrix $XW_QW_K^TX^T \in \mathcal{R}^{T,T}$ can be a RIP matrix with appropriate choices of W_K, W_Q.*

Notice that the standard formulation of the transformer layers can be written as

$$y = x + MLP(LayerNorm(x + Attention(LayerNorm(x)))$$

.

This implies that the sparsity function and the nonlinear association are intertwined – together one stage in a multistage architecture. This may be ideal to represent compositional sparsity. There is however a formulation with similar empirical performance (see PALM paper) which can be written as

$$y = x + MLP(LayerNorm(x)) + Attention(LayerNorm(x))$$

.

7 Transformers as associative memories

Consider $AX = Y$. The best A is given by

$$A = YX^T(XX^T)^{-1}. \tag{10}$$

If $(XX^T)^{-1} \approx I$ – which happens for noiselike X – then $A = YX^T$ implying $Ax = YX^Tx$. Typically the dimensionality of the columns of X is large to allow for the noiselike property (and sparsity).

Transformers transform input matrices into output matrices of the same dimensionality for instance a German sentence into a French one. In other words, functions implemented by self-attention map from $R^{T,d}$ to itself, so that instances from this function class can be composed. This is important for compositionality in *compositional sparsity*. It is also important in the use of transformers a sequence of associations from an input x' to an output $x"$ which is then used for another association. x' could be a sentence with a missing word and $x"$ its completion.

The idea of associative memory is consistent with the interpretation of the self-attention layer as a learned, di□erentiable lookup table. The Q, K, and V are described as "queries," "keys," and "values" respectively, which seem to invoke such an interpretation. Consider only one attentional head. Each object or token x_i has a query $Q(x_i)$ that it will use to test "compatibility" with the key $K(x_j)$ of each object x_j. Compatibility of x_i with x_j is defined by the inner product $Q(x_i), K(x_j)$; if this inner product is high, then x_i 's query matches x_j key and so we look up x_js value $V(x_j)$. We construct then a soft lookup of values compatible with x_i's key: we sum up the value of each object x_j proportional to the compatibility of x_i with x_j.

[7] Deeper layers are made up of these local features with larger receptive fields hinting at the hierarchical structure found in CNNs. Pruning induces locality also beyond the first hidden layer. Their remark "These results highlight the role of the task in shaping the properties of the network obtained by pruning: only for the task that the network can e□ciently learn, and not just memorize, local features emerge..." is consistent with our hypothesis of compositional sparsity of the underlying task, in this case a visual task.

8 Generalization bounds for Sparse Networks

8.1 Convolutional networks with local kernel

The classical bounds are for generic deep networks. In such a general case, ρ in those bounds is the product of the Frobenius norms of all the weight matrices. For convolutional networks the weight matrices are Toeplitz matrices. This gives large bounds.

Here we show that the bound on the Rademacher complexity can be reduced by exploiting two typical properties of CNNs: a) the locality of the convolutional kernels and b) shared weights. They allow us to use only the norm of the kernels in the calculation of ρ_k instead of the norm of the corresponding Toeplitz matrix. In this section we give an outline of the results with more precise statements and proofs to be published later.

We start by considering the simple situation of non-overlapping convolutional patches. In other words, the stride of the convolution is equal to the size of the kernel in each layer. This means that in the associated Toeplitz matrix the non-zero components in each row do not overlap with the non-zero components of the row above or the one below. In other words, if K is the number of patches, ℓ is the size of each patch and $x \in \mathbb{R}^d$, then $d = K\ell$. Notice that the standard bounds give a Rademacher complexity proportional to the product of the Frobenius norms of each weight matrix $\|W\|$ time the norm of $\|x\|$, where $\|W\| \propto \sqrt{k}M$, where M is the norm of the kernel.

In [11] we describe generic bounds on the Rademacher complexity of deep neural networks. In these cases, ρ measures the product of the Frobenius norms of the network's weight matrices in each layer. For convolutional networks, however, the operation in each layer is computed with a kernel, described by the vector w, that acts on each patch of the input separately. Therefore, a convolutional layer is represented by a Toeplitz matrix W, whose blocks are each given by w. In this section (from [11]) we provide an informal analysis of the Rademacher complexity, showing that it can be reduced by exploiting mainly the first one of the two properties of convolutional layers: (a) the locality of the convolutional kernels that is the sparsity of the associated Toeplitz matrix and (b) weight sharing. These properties allow us to bound the Rademacher complexity by taking the products of the norms of the kernel w instead of the norm of the associated Toeplitz matrix W. Here we outline the results with more precise statements and proofs to be published separately.

We consider the case of 1-dimensional convolutional networks with non-overlapping patches and one channel per layer. For simplicity, we assume that the input of the network lies in $\mathbb{R}^d$, with $d = 2^L$ and the stride and the kernel of each layer are 2. The analysis can be easily extended to kernels of di□erent sizes. This means that the network $h(x)$ can be represented as a binary tree, where the output neuron is computed as $W^L \cdot \sigma(v_1^L(x), v_2^L(x))$, $v_1^L(x) = W^{L-1} \cdot \sigma(v_1^{L-1}(x), v_2^{L-1}(x))$ and $v_2^L(x) = W^{L-1} \cdot \sigma(v_3^{L-1}(x), v_4^{L-1}(x))$ and so on. This means that we can write the i'th row of the Toeplitz matrix of the l'th layer $(0, \ldots, 0, -W^l-, 0 \ldots, 0)$, where W^l appears on the $2^i - 1$ and 2^i coordinates. We define a set $\mathcal{H}$ of neural networks of this form, where each layer is followed by a ReLU activation function and $\prod_{l=1}^{L} W^l \leq \rho$.

Theorem 11. *Let $\mathcal{H}$ be the set of binary-tree structured neural networks over $\mathbb{R}^d$, with $d = 2^L$ for some natural number L. Let $X = \{x_1, \ldots, x_m\} \subset \mathbb{R}^d$ be a set of samples. Then,*

$$\mathcal{R}_X(\mathcal{H}) \leq \frac{2^L \rho \sqrt{\sum_{i=1}^m \|x_i\|^2}}{m} \tag{11}$$

Proof sketch. First we rewrite the Rademacher complexity in the following manner:

$$\begin{aligned} \mathcal{R}_X(\mathcal{H}) &= \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} |\frac{1}{m} \sum_{i=1}^m \epsilon_i \cdot h(x_i)| \\ &= \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{m} |\sum_{i=1}^m \epsilon_i \cdot W^L \cdot \sigma(v_1(x), v_2(x))| \\ &= \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{m} \sqrt{|\sum_{i=1}^m \epsilon_i \cdot W^L \cdot \sigma(v_1(x), v_2(x))|^2} \end{aligned} \tag{12}$$

Next, by the proof of Lem. 1 in [30], we obtain that

$$\begin{aligned}\mathcal{R}_X(\mathcal{H}) &\leq 2\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{m}\sqrt{\|W^L\|^2 \cdot \|\sum_{i=1}^{m}\epsilon_i(v_1(x), v_2(x))\|^2} \\ &= \mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{m}\sqrt{\|W^L\|^2 \cdot \sum_{j=1}^{2}\|\sum_{i=1}^{m}\epsilon_i v_j(x_i)\|^2}\end{aligned} \tag{13}$$

By applying this peeling process L times, we obtain the following inequality:

$$\begin{aligned}\mathcal{R}_X(\mathcal{H}) &\leq 2^{L-1}\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{m}\sqrt{\prod_{l=1}^{L}\|W^l\|^2 \cdot \sum_{j=1}^{d}\|\sum_{i=1}^{m}\epsilon_i x_{ij}\|^2} \\ &= 2^{L-1}\mathbb{E}_\epsilon \sup_{h\in\mathcal{H}} \frac{1}{m}\sqrt{\prod_{l=1}^{L}\|W^l\|^2 \cdot \|\sum_{i=1}^{m}\epsilon_i x_i\|^2} \\ &\leq \frac{2^{L-1}\rho\mathbb{E}_\epsilon\|\sum_{i=1}^{m}\epsilon_i x_i\|}{m} \\ &\leq \frac{2^{L-1}\rho\sqrt{\sum_{i=1}^{m}\|x_i\|^2}}{m}\end{aligned} \tag{14}$$

where the factor 2^{L-1} is obtained because the last layer is linear (see [31]). We note that a better bound can achieved when using the reduction introduced in [30] which would give a factor of $\sqrt{2\log(2)L}+1$ instead of 2^{L-1}. □

One-layer convolutional classifier Consider a ReLU convolutional classifier with k patches. $\hat{\mathcal{R}}_m$, in the standard bounds would be

$$\hat{\mathcal{R}}_m \leq BX$$

where B is the Frobenius norm of the Toeplitz matrix with k rows, each row consisting of the kernel w. Thus $B = \sqrt{K}\|w\|$ and $X = \|x\|$.

Our calculation gives with x^1 representing the first patch of x and x^K the last one:

$$\hat{\mathcal{R}}_m \leq \sqrt{\|w\|^2\|x^1 + \cdots + x^K\|^2} = \sqrt{\|w\|^2\|x\|^2} = \|w\|\|x\|.$$

instead of the general bound usually referred which is

$$\hat{\mathcal{R}}_m \leq \|W\|\|x\| = \sqrt{k}\|w\|\|x\|$$

Multi-layer convolutional classifier The Rademacher complexity of a feed-forward neural network can be bounded recursively by considering each layer at a time. A bound that can be used for the recursion is given by the following proposition (see [31, 30]), that expresses the Rademacher complexities at the outputs of one layer in terms of the outputs at the previous layers.

Lemma 4. *Let $\mathcal{H}$ be a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be the ReLU function which is 1-Lipschitz and define $\mathcal{H}' := \left\{x \in \mathbb{R}^d \to \sigma\left(\sum_{j=1}^{k} w_j h_j(x)\right) \in \mathbb{R} : \|w\|_2 \leq M, h_1, \ldots, h_k \in \mathcal{H}\right\}$. Then, for any $x_1, \ldots, x_m \in \mathbb{R}^d$*

$$\mathcal{R}(\mathcal{H}' \circ \{x_1, \ldots, x_m\}) \leq 2M\mathcal{R}(\mathcal{H} \circ \{x_1, \ldots, x_m\}).$$

We apply now the Lemma to the class of L-depth ReLU real-valued CNN, with each layer's kernel w_d with norm at most M_d.

Theorem 12. *(informal) The Rademacher complexity of a convolutional deep net with RELUs in all d layers but the last linear one and with non-overlapping convolutional patches can be bounded as*

$$\hat{\mathcal{R}_m}(\mathcal{H}_d) \leq (\sqrt{2\log(2)L}+1)\prod_{j=1}^{L} M_j\|x\| \tag{15}$$

Proof sketch. Each $h_k^d \in \mathcal{H}_\lceil$ ($k = 1, \cdots, Q$ is a ReLU classifier inputs from patch j of the layer below. Patch k in layer $d-1$ can be written as a vector v_k consisting of ℓ classifiers $v_k = h_{k\cdot 1}^{d-1}, h_{k\cdot 2}^{d-1}, \cdots, h_{k\cdot \ell}^{d-1}$. Then $h_k^d = \sigma(w \cdot v_k)$. Notice that because of our assumption of non-overlapping patches the number of units in layer $d-1$ is ℓ times the number of units in layer d. Then

$$\hat{\mathcal{R}}_m(\mathcal{H}_d) = \mathbb{E}_\epsilon \sup_{h_i \in \mathcal{H}_d} \frac{1}{m} \sum_{i=1}^m \epsilon_i h_i = \mathbb{E}_\epsilon \sup_{h_i \in \mathcal{H}_{d-1} w: \|w\| \le M} \frac{1}{m} \sum_{i=1}^m \epsilon_i w \cdot (\sum_k v_k), \tag{16}$$

can be upper bounded as follows

$$\hat{\mathcal{R}}_m(\mathcal{H}_d) \le 2M_d \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}_d} \|\frac{1}{m} \sum_{i=1}^m \epsilon_i (\sum_k (v_k)_i)\| \le \frac{1}{m} \sqrt{(w \cdot (\sum_k v_k)^2} = \frac{1}{m} \sqrt{(w \cdot (\sum_k v_k)^2} = 2M_d \hat{\mathcal{R}}_m(\mathcal{H}_{d-1}), \tag{17}$$

because $(\sum_k v_k)^2 = \sum_k v_k^2$ since the various patches are zero-mean and uncorrelated.
Continuing the peeling we obtain

$$\mathcal{R}_m(\mathcal{H}_L) \le 2^{L-1} \hat{M_L} \cdot M_{L-1} \cdots M_1 \|x\|, \tag{18}$$

where the factor 2^{L-1} is obtained because the last layer is linear (see [31]). To this result we can further apply the reduction used by [30] to finally obtain the result. □

One ends up with a bound scaling as the product of the norms of the kernel at each layer. The constants may change depending on the architecture, the number of patches, the size of the patches and their overlap.

Thus one ends up with a bound scaling as the product of the norms of the kernel at each layer. The constants may change depending on the architecture, the number of patches, the size of the patches and their overlap.

This special non-overlapping case can be extended to the general convolutional case. In fact a proof of the following conjecture will be provided in [32]

Conjecture 5. *If a convolutional layer has overlaps among its patches then the bound*

$$\mathcal{R}_m(\mathcal{H}_L) \le 2^{L-1} \hat{M_L} \cdot M_{L-1} \cdots M_1 \|x\|$$

holds with $\|x\|$ replaced by

$$\|x\| \sqrt{\frac{K}{K-O}},$$

where K is the size of the kernel (number of components) and O is the size of the overlap.

Sketch proof Call P the number of patches and O the overlap. With no overlap then $PK = D$ where D is the dimensionality of the input to the layer. In general $P = \frac{D-O}{K-O}$. It follows that a layer with the most overlap can add at most $< \|x\|\sqrt{K}$ to the bound. Notice that we assume that each component of x_i averaged across i will have norm $\sqrt{\frac{1}{d}}$.