On the Expressive Power of Deep Polynomial Neural Networks

We thank the reviewers for their positive and useful comments. One shared concern among the reviewers seems to be that our study of the *exact* functional space and its dimension might not be directly helpful for understanding the practical ability of a network to represent functions *approximately*. This criticism may be prompted in part by our claim that "we do not emphasize approximation properties, but rather the study of the functions that can be expressed exactly using a network" (Sec.1.1). While it is true that we do not directly address the issue of approximation, our theory also suggests a geometric framework for characterizing the class of functions that can be approximated well by a neural network. In particular, when the functional variety is not filling, one can naturally consider functions that are "close" to the variety (nearness could be formalized by considering tubular neighborhoods of the variety). Given that our functional varieties are generalizations of families of low-rank tensors, this perspective might lead to quantitative results in terms of generalized versions of SVD (applied for example to tensor flattenings). In fact, motivated by the reviewers' questions, we plan to think about computational methods for "projecting" a polynomial on a functional variety. In a broader sense, we believe that a theory of "exact expressivity" of neural networks should have greater explanatory power compared to a purely approximation-based analysis. The fact that the former perspective has received less attention could be due to the lack of appropriate tools for addressing it; we hope that our algebraic framework can open the door to new work on this topic. If the paper is accepted, we will mention these directions, and remove the aforementioned sentence from Sec.1.1.

Reviewer 1.

- *"I'd ask the authors to respond to my above question about approximation."* See our detailed answer above.

Reviewer 2.

- *"it is difficult to see how this measure translates to actual measures of interest to expressiveness analysis [...] the naive dimension bound provided matches exactly with other works on memorization, i.e., it is proportional to the number of parameters in a network."* In addition to our general answer above, we believe that the algebraic framework may be used to derive new approximation bounds. For example, while it is true that our naive dimension bound is proportional to the number of parameters, the existence of "asymptotic bottlenecks" (Theorem 19) shows that in many situations these two quantities are very different (in the presence of a bottleneck, even if widths grow arbitrarily, the functional dimension stays bounded): our theory detects this discrepancy, and the corresponding effect on expressivity.
- *"it appears that even for very small input spaces (e.g., 28x28 MNIST images) and squared activations the required minimal widths are already infeasible"* We agree that the filling conditions are unlikely to be satisfied in practice, and for this reason we believe that learning in real-life architectures takes place in a non-filling regime. Still, qualitative distinctions between filling and non-filling architectures are theoretically important, and consistent with numerous existing results showing that learning is easier in the infinite-width setting. Furthermore, we believe that refined notions of filling will help bring our theory closer to practice. Specifically, one could consider "empirical filling", *i.e.*, whether given any sample set of a fixed size, there exist weights that perfectly interpolate the data. Also, a notion of "relative filling" could help compare different architectures, *i.e.*, when two architectures give the same functional space. We will remark in the paper how these notions could be studied in future work.
- *"are the ranks computed to give exact answers [...], or are you using the standard floating-point numerical methods that are only estimates? I suggest the authors emphasize this aspect in the paper."* Our computations are based on finite field arithmetic, using a large prime number to define the base field. As we will clarify in the paper, all filling dimensions are provably correct over $\mathbb{R}$ while all other computed dimensions are correct over $\mathbb{R}$ with very high probability.

Reviewer 3.

-*"It would have been nicer to see more investigation of whether such networks can be trained to learn polynomials"* The focus of this paper was on expressivity, and we only briefly touched upon optimization/learning in Sec.2.3. We are currently working on developing the connection between filling architectures and optimization further. For example, we conjecture that the absence of non-global local minima in the landscape of deep polynomial networks is actually equivalent to the condition that the functional space is filling. If true, this property would vastly generalize known results on shallow quadratic networks. We are also investigating how refined notions of filling (see above) may yield favorable optimization properties. We believe however that these issues fall outside the scope of the current paper.
- *"Another point is that there may be advantages to use activations like sigmoid"* In general, we agree that there may be benefits in using non-polynomial activations, as universal approximation theorems require them. However, we are unsure about the practical difference between using polynomial activations and ReLU or sigmoids (folklore seems to say that the choice of the activation is not so important). For example, there are optimization methods that involve expressing a non-smooth function (*e.g.*, absolute value) as a limit of polynomial functions ("η-trick"). Spelling out the connection between polynomial networks and sigmoid or ReLU networks is an important issue that we will look into.
- *"it would be better to list the main results upfront"* We thank the reviewer for this useful suggestion. Unfortunately, for most of our results, a precise formulation of the statement requires notions and terminology that are only introduced in Sec.2. However, if the paper is accepted, we will include informal versions of our main results at the end of Sec.1.