Pure and Applied Functional Analysis
Volume 8, Number 2, 2023, 405–428

Yokohama Publishers
ISSN 2189-3764 ONLINE JOURNAL

APPROXIMATION BY TREE TENSOR NETWORKS IN HIGH DIMENSIONS: SOBOLEV AND COMPOSITIONAL FUNCTIONS

MARKUS BACHMAYR*, ANTHONY NOUY, AND REINHOLD SCHNEIDER

Dedicated to Ronald DeVore on the occasion of his 80th birthday

Abstract. This paper is concerned with convergence estimates for fully discrete tree tensor network approximations of high-dimensional functions from several model classes. For functions having standard or mixed Sobolev regularity, new estimates generalizing and refining known results are obtained, based on notions of linear widths of multivariate functions. In the main results of this paper, such techniques are applied to classes of functions with compositional structure, which are known to be particularly suitable for approximation by deep neural networks. As shown here, such functions can also be approximated by tree tensor networks without a curse of dimensionality – however, subject to certain conditions, in particular on the depth of the underlying tree. In addition, a constructive encoding of compositional functions in tree tensor networks is given.

1. Introduction

The performance of standard approximation schemes based on splines or wavelets can be characterized by classical notions of Sobolev or Besov smoothness. In the approximation of functions on high-dimensional domains, such standard methods are too inefficient, which is related to the fact that the associated smoothness classes are too broad: in order to approximate high-dimensional functions with tractable complexity, one needs to exploit more specific features of these functions. This motivates the analysis of more narrow model classes of functions and of their interplay with corresponding approximation algorithms. A classical example are sparse grids, whose performance is characterized by model classes of functions of high-order mixed regularity.

Here, we consider approximation algorithms based on tree tensor networks (or hierarchical tensor formats) [9–11], which are a particular type of low-rank approximation of high-order tensors with favorable numerical properties. We study the performance of such approximations for two types of model classes. On the one hand, we consider a class of functions that can be written as compositions of lower-dimensional component functions. These compositional functions may represent

2020 *Mathematics Subject Classification.* 41A46, 41A63, 41A65.

Key words and phrases. Tree tensor networks, hierarchical tensors, low-rank approximation, linear widths, compositional structure.

*M.B. acknowledges funding by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummern 233630050; 211504053 – TRR 146; SFB 1060.

complex hierarchical decision systems where one agent takes a decision based on the decisions taken by other agents, or complex simulation systems where the inputs of a system are given by the outputs (or states) of other systems [4, 16, 21]; see also the discussion in [20]. This class of functions has been shown by Mhaskar and Poggio [17] to allow for efficient approximations – with a weak dimension-dependence under certain conditions – by deep neural networks. To obtain convergence estimates for tree tensor networks, we develop two techniques based on estimates of linear widths and on a direct constructive encoding of compositions.

On the other hand, to put these convergence results into context, we also revisit the approximation of functions of (mixed) Sobolev regularity by tree tensor networks. By a similar technique based on linear widths, we extend and refine estimates from [22], where this has been considered for periodic functions. In comparison, our results address general non-periodic functions on the unit cube and yield slightly improved asymptotic rates in the storage complexity of approximations. The approximation of Sobolev functions by certain tree tensor networks, including Tucker tensors and tensors trains, has recently also been considered in [12, 13]; there, however, semidiscrete approximation rates in terms of tensor ranks are obtained from singular value estimates, without discretization in the tensor modes. The corresponding complexity estimates are thus not directly comparable to the fully discrete approximations considered here.

The approximations by tree tensor networks that we consider are associated to *dimension trees*, which are assumed to be fixed in advance. An example of such a tree is shown in Figure 1; in general, for a tensor of order d, the set $D = \{1, \dots, d\}$ of modes is recursively subdivided up to the singletons $\{1\}, \dots, \{d\}$. The set of all nodes resulting from this subdivision is then denoted by T. The most common choice here is a binary tree, where each interior node of the tree has two children. A tree tensor network with T-ranks bounded by $r = (r_\alpha)_{\alpha \in T}$ is a multivariate function v that admits for each $\alpha \in T$ a representation $v(x) = \sum_{k=1}^{r_\alpha} v_k^\alpha(x_\alpha) v_k^{\alpha^c}(x_{\alpha^c})$ for some functions v_k^α and $v_k^{\alpha^c}$ of complementary groups of variables x_α and x_{α^c}, $\alpha^c = D \setminus \alpha$. For functions in a Hilbert tensor space equipped with a canonical inner product, such a representation is related to the singular value decomposition of the α-matricization (or α-unfolding) of v, identified with a bivariate function. The approximability of a function by tree tensor networks is therefore related to the decay of singular values of its α-matricizations for each $\alpha \in T$.

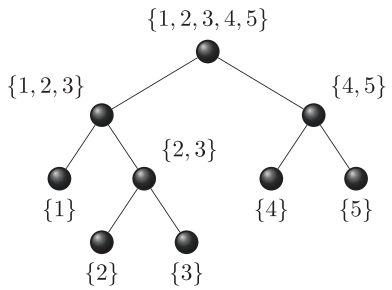

FIGURE 1. Example of a dimension partition tree T over $D = \{1, 2, 3, 4, 5\}$.

The results on approximation of certain compositional functions by neural networks in [17] are also based on the notion of (binary) dimension trees: the class of approximands considered there is comprised of functions that are compositions with a tree structure. For instance, the tree in Figure 1 corresponds to compositions of the form

$$f(x) = f_D\big(f_{\{1,2,3\}}\big(x_1, f_{\{2,3\}}(x_2, x_3)\big), f_{\{4,5\}}(x_4, x_5)\big),$$

where the tree being binary corresponds to composing bivariate functions, and where the constituent functions are assumed to be at least Lipschitz continuous.

The general result from [17] for approximating such compositions with an underlying tree of depth L can be paraphrased as follows: *Assume that f has compositional structure according to a binary dimension tree with L levels, where each component function is Lipschitz continuous with Lipschitz constant $B > 0$ and has s weak derivatives in L^∞. Then for any smooth, non-polynomial activation function, there exists a neural network $\tilde{f}$ such that $\|f - \tilde{f}\|_{L^\infty} \leq \varepsilon$ with $\mathcal{O}(LB^L\varepsilon^{-2/s})$ coefficients.*

Note that since $B \leq 1$ is assumed in [17], the dependence on L is not explicitly mentioned there. The dependence of L on d depends on the tree structure, with the most favorable dependence $L \sim \log d$ for a balanced tree: in this case, B^L is polynomial in d. The proof is based on the following estimate: for functions f, g, h satisfying the above assumptions with approximations $\tilde{f}, \tilde{g}, \tilde{h}$, one has

$$\begin{aligned} \|f(g,h) - \tilde{f}(\tilde{g},\tilde{h})\|_{L^\infty} &\leq \|f(g,h) - f(\tilde{g},\tilde{h})\|_{L^\infty} \\ &\quad + \|f(\tilde{g},\tilde{h}) - \tilde{f}(\tilde{g},\tilde{h})\|_{L^\infty} \\ &\leq B(\|g-\tilde{g}\|_{L^\infty} + \|h-\tilde{h}\|_{L^\infty}) + \|f-\tilde{f}\|_{L^\infty}. \end{aligned} \tag{1.1}$$

Applying this estimate recursively starting from the root of the tree, the bound for the approximation complexity follows, using that each component function can be approximated separately by a neural network with $\mathcal{O}(\varepsilon^{-2/s})$ parameters; the composition of these approximations is then again a neural network.

One of the main results of the present work is that a very similar approximation complexity for this class of compositional functions can be achieved by approximations by tree tensor networks, with error measured in L^2 (for arbitrary s) or L^∞ (with the restriction $s \leq 2$). More specifically, we show that a tree tensor network approximation $\tilde{f}$ (that is, a composition of *multilinear* mappings according to the same binary tree structure as the approximand) can be found such that accuracy ε is achieved with $\mathcal{O}(L^3B^{3L}\varepsilon^{-3/s})$ coefficients, possibly up to terms logarithmic in ε that depend on the particular construction, and up to a constant polynomial in d. In other words, we obtain a very similar dependence on d with d-independent convergence rate for tree tensor network approximations, which are substantially easier to handle numerically than approximations by neural networks. We also show that these tensor approximations of functions with compositional structure as considered here can be constructed explicitly in certain cases. The curse of dimensionality is thus shown to be avoided for tree tensor networks under very similar conditions as for deep neural networks.

The outline of the paper is as follows. In Section 2, we recall the definition of tree tensor networks and provide upper bounds for the best approximation error

of a function in L^2 in terms of linear widths. In Section 3, using these upper bounds based on linear widths, we provide approximation results for functions with (mixed) Sobolev regularity. Finally in Section 4, we consider the approximation of compositional functions by tree tensor networks and discuss the conditions under which the curse of dimensionality is avoided. For the approximation in L^2, our proof is based on estimates of linear widths of compositional functions, while for the approximation in L^∞, we use a constructive proof and provide an explicit encoding of an approximation that achieves the announced convergence rates.

2. Linear widths and tree tensor networks

In this section, we first discuss notions of linear widths in the context of multivariate functions. We then recall the definition of the model class of functions in tree based tensor format (or tree tensor networks), which is interpreted as a particular class of compositional functions. Finally, in the case of square-integrable functions on the unit cube in d dimensions, we deduce upper bounds of the best approximation error in terms of linear widths.

2.1. **Linear widths and singular value decomposition.** We consider functions defined on the unit cube $\mathcal{X} = (0,1)^d$ with $d \geq 2$; other sets $\mathcal{X}$ with Cartesian product structure could be treated in the same manner in what follows, but we restrict ourselves to this special case for simplicity. We denote by $D = \{1, \dots, d\}$ the set of dimensions. Throughout this section, we assume α to be a nonempty strict subset of D, and we define $\alpha^c = D \setminus \alpha$. We set $\mathcal{X}_\alpha = (0,1)^{|\alpha|}$, and for $x = (x_1, \dots, x_d) \in \mathcal{X}$, we write $x_\alpha = (x_\nu)_{\nu \in \alpha} \in \mathcal{X}_\alpha$.

For closed subspaces V of Banach spaces Y, for the error of best approximation of $u \in Y$ by elements of V, we introduce the notation

$$E(u, V)_Y = \inf_{v \in V} \|u - v\|_Y.$$

Recall that the classical Kolmogorov n-width of a compact subset $K \subset Y$ then reads

$$\mathsf{d}_n(K)_Y = \inf_{\dim(V)=n} \sup_{u \in K} E(u, V)_Y,$$

where the infimum is taken over all n-dimensional subspaces $V \subset Y$.

In the following summary of basic notions of related linear widths of multivariate functions, we focus on functions in the tensor product Hilbert space

$$X := L^2(\mathcal{X}) = L^2(\mathcal{X}_1) \otimes \cdots \otimes L^2(\mathcal{X}_d),$$

where we abbreviate $X_\alpha := L^2(\mathcal{X}_\alpha)$. We first note that by the canonical isomorphism $\mathcal{M}_\alpha \colon L^2(\mathcal{X}) \to L^2(\mathcal{X}_{\alpha^c}; X_\alpha)$, any $f \in L^2(\mathcal{X})$ can be isometrically identified with $f^\alpha := \mathcal{M}_\alpha f \in L^2(\mathcal{X}_{\alpha^c}; X_\alpha)$ given by $f^\alpha : x_{\alpha^c} \mapsto f(\cdot, x_{\alpha^c})$. For a given closed subspace $V \subset X_\alpha$, we define the projection

$$\mathcal{P}_V^\alpha f := \mathcal{M}_\alpha^{-1}\Big(\operatorname*{arg\,min}_{g^\alpha \in L^2(\mathcal{X}_{\alpha^c}; V)} \|f^\alpha - g^\alpha\|_{L^2(\mathcal{X}_{\alpha^c}; X_\alpha)} \Big),$$

which amounts to applying the L^2-orthogonal projection onto V to $f^\alpha(x_{\alpha^c})$ for each x_{α^c}. For more details on projections on tensor spaces, see also [18].

We now introduce an average linear width associated to f and α as

$$\delta_n^\alpha(f) = \inf_{\dim(V)=n} \left(\int_{\mathcal{X}_{\alpha^c}} E\big(f^\alpha(x_{\alpha^c}), V\big)_{X_\alpha}^2 \, dx_{\alpha^c} \right)^{1/2}. \tag{2.1}$$

As we shall now describe, these widths are closely connected to low-rank approximations of f. To this end, we define the compact operator

$$\mathcal{S}_f^\alpha \colon X_{\alpha^c} \to X_\alpha, \ v \mapsto \int_{\mathcal{X}_{\alpha^c}} f^\alpha \, v \, dx_{\alpha^c}.$$

We then define the α-rank of f by

$$\operatorname{rank}_\alpha(f) := \dim \operatorname{Range} \mathcal{S}_f^\alpha,$$

which in general may be infinite. Note that $\operatorname{rank}_\alpha(g) \le n$ implies that g can be written in the form

$$\sum_{k=1}^n u_k(x_\alpha) \, v_k(x_{\alpha^c})$$

with functions $u_k \in X_\alpha$, $v_k \in X_{\alpha^c}$ for $k = 1, \dots, n$.

The operator $\mathcal{S}_f^\alpha$ admits a singular value decomposition (see, e.g., [10, Section 4.4.3]); let $(\sigma_k^\alpha)_{k\ge1}$ be the non-increasing, non-negative sequence of singular values. Then it is easy to see that for each $n \in \mathbb{N}$,

$$\delta_n^\alpha(f) = \min_{\operatorname{rank}_\alpha(v)\le n} \|f - v\|_X = \Big(\sum_{k>n} (\sigma_k^\alpha)^2 \Big)^{1/2};$$

in other words, $\delta_n^\alpha(f)$ is the error of L^2-best approximation of f of α-rank n. Moreover, if $U_n \subset X_\alpha$ is a principal subspace of $\mathcal{S}_f^\alpha$ associated to n largest singular values, then

$$\delta_n^\alpha(f) = \|f - \mathcal{P}_{U_n}^\alpha f\|_X = \min_{\dim(V)=n} \|f - \mathcal{P}_V^\alpha f\|_X,$$

that is, such best approximations of α-rank at most n can be obtained from the singular value decomposition. As a further consequence, note that

$$\delta_n^\alpha = \delta_n^{\alpha^c}, \quad n \in \mathbb{N}. \tag{2.2}$$

2.2. **Tree-based tensor formats.** We next introduce some notions that are fundamental to tree-based tensor formats; for further details, we refer to [5, 10]. Let T be a dimension partition tree over $D = \{1, \dots, d\}$ (see an example on Figure 1). For any node $\alpha \in T$, we denote by $S(\alpha) \subset T$ the set of sons of α, which forms a partition of α. $S(\alpha)$ is either empty or has cardinality $\#S(\alpha) \ge 2$. If $S(\alpha) = \emptyset$, α is called a leaf of T. We let $\mathcal{L}(T)$ be the set of leaves of T and write $\mathcal{I}(T) = T \setminus \mathcal{L}(T)$ for the interior nodes of T.

We let $\operatorname{level}(\alpha)$ be the level of α in T. We use the convention $\operatorname{level}(D) = 0$ and for any $\beta \in T \setminus \{D\}$ such that $\beta \in S(\alpha)$, we define $\operatorname{level}(\beta) = \operatorname{level}(\alpha) + 1$. Also, we define the depth of T as $\operatorname{depth}(T) = \max\{\operatorname{level}(\alpha) : \alpha \in T\}$. We set $T_\ell = \{\alpha \in T : \operatorname{level}(\alpha) = \ell\}$ for $0 \le \ell \le \operatorname{depth}(T)$.

Example 2.1 (Trivial tree)**.** The trivial tree $T = \{D, \{1\}, \dots, \{d\}\}$ has a single interior node D and $\operatorname{depth}(T) = 1$.

Example 2.2 (Linear binary tree)**.** The linear binary tree

$$T = \{\{1\}, \dots, \{d\}, \{1,2\}, \dots, \{1, \dots, d-1\}, D\}$$

satisfies $\operatorname{depth}(T) = d-1$ and $T_\ell = \{\{1, \dots, d-\ell\}, \{d-\ell+1\}\}$ for $1 \le \ell \le d-1$.

Example 2.3 (Balanced binary tree)**.** For a balanced binary tree T, one has $\operatorname{depth}(T) = \lceil \log_2(d) \rceil$. For $\ell \le \operatorname{depth}(T)$, we have $\#T_\ell \le 2^\ell$ and $\#\alpha \le \lceil \frac{d}{2^\ell} \rceil$ for all $\alpha \in T_\ell$.

Let X be a tensor product space of multivariate functions. For a tuple $r = (r_\alpha)_{\alpha \in T}$ (with $r_D = 1$), we define a tree-based tensor format in X as

$$\mathcal{T}_r^T(X) = \big\{ v \in X : \operatorname{rank}_\alpha(v) \le r_\alpha, \alpha \in T \big\}.$$

Tensors satisfying these rank constraints are also known as *hierarchical tensors* [11] or as *tree tensor network states* in quantum physics [23]. Letting $U = U_1 \otimes \dots \otimes U_d$ be a subspace of X, where the U_ν are finite-dimensional subspaces of functions defined on $\mathcal{X}_\nu$, we also define

$$\mathcal{T}_r^T(U) = \{ v \in U : \operatorname{rank}_\alpha(v) \le r_\alpha \} = \mathcal{T}_r^T(X) \cap U.$$

A tuple r is called *admissible* if $\mathcal{T}_r^T(X) \neq \emptyset$.

2.3. **Tree based tensor formats as compositional functions and tensor networks.** For each $\nu \in D$, we let $\{\varphi_i^\nu\}_{i=1}^{n_\nu}$ denote a basis of U_ν, and introduce the map $\varphi^\nu : \mathcal{X}_\nu \to \mathbb{R}^{n_\nu}$ such that $\varphi^\nu(x_\nu) = (\varphi_i^\nu(x_\nu))_{i=1}^{n_\nu}$ for each $x_\nu \in X_\nu$. A function $f \in \mathcal{T}_r^T(U)$ can be parametrized by a set of multilinear functions $\{G^\alpha : \alpha \in T\}$, where $G^\alpha : \bigtimes_{\beta \in S(\alpha)} \mathbb{R}^{r_\beta} \to \mathbb{R}^{r_\alpha}$ for $\alpha \in \mathcal{I}(T)$ is multilinear, and $G^\alpha : \mathbb{R}^{n_\alpha} \to \mathbb{R}^{r_\alpha}$ for $\alpha \in \mathcal{L}(T)$ is linear. To obtain this parameterization, for each $x \in \mathcal{X}$, we set $z_{\{\nu\}} = \varphi^\nu(x_{\{\nu\}})$ for each leaf node $\{\nu\} \in \mathcal{L}(T)$ for $\nu = 1, \dots, d$, and for each interior node $\alpha \in \mathcal{I}(T)$ recursively define the evaluations of compositions

$$z_\alpha = G^\alpha((z_\beta)_{\beta \in S(\alpha)}),$$

and thus obtain

$$f(x) = z_D = G^D((z_\alpha)_{\alpha \in S(D)}).$$

Example 2.4. For the tree T of Figure 1, $f \in \mathcal{T}_r^T(U)$ can be written in the form

$$f(x) = G^D(G^{\{1,2,3\}}(G^{\{1\}}(z_1), G^{\{2,3\}}(G^{\{2\}}(z_2), G^{\{3\}}(z_3))), \\ G^{\{4,5\}}(G^{\{4\}}(z_4), G^{\{5\}}(z_5)))$$

where $z_\nu = \varphi^\nu(x_\nu)$, $1 \le \nu \le d$.

Example 2.5 (Trivial tree and Tucker format)**.** For the trivial tree of Example 2.1, $\mathcal{T}_r^T(U)$ corresponds to the Tucker format and $f \in \mathcal{T}_r^T(U)$ can be written

$$f(x) = G^D(\varphi^1(x_1), \dots, \varphi^d(x_d)).$$

Example 2.6 (Linear tree and tensor train format)**.** For the linear binary tree of Example 2.2, $\mathcal{T}_r^T(U)$ corresponds to the tensor train (TT) Tucker format.

The multilinear functions G^α can be identified with tensors of order $\#S(D)$ for $\alpha = D$, $1 + \#S(\alpha)$ for $\alpha \in \mathcal{I}(T) \setminus \{D\}$ and 2 if $\alpha \in \mathcal{L}(T)$. This yields the interpretation of the tree-based format as a tree tensor network.

The number of parameters (or representation complexity) of an element in $\mathcal{T}_r^T(U)$ is

$$N(T, r, U) = \sum_{\alpha \in \mathcal{I}(T)} r_\alpha \prod_{\beta \in S(\alpha)} r_\beta + \sum_{\nu=1}^{d} r_\nu n_\nu,$$

with $n_\nu = \dim(U_\nu)$. If $r_\alpha \leq R$ for all α and $\dim(U_\nu) \leq n$ for all ν, then

$$N(T, r, U) \leq R^a + (\#T - 1 - d)R^{a+1} + dRn \leq R^a + (d-2)R^{a+1} + dRn,$$

where $a = \max_{\alpha \in \mathcal{I}(T)} \#S(\alpha)$ is the arity of the tree ($a = 2$ for a binary tree, and $a = d$ for a trivial tree).

2.4. **Best approximation error and linear widths.** Let T be a fixed dimension tree and $r = (r_\alpha)_{\alpha \in T}$ be an admissible rank. For any subspace $U \subset X = L^2(\mathcal{X})$, the error of best approximation of a function $f \in X$ by an element of $\mathcal{T}_r^T(U)$ is

$$e_{r,U}^T(f)_X = \inf_{v \in \mathcal{T}_r^T(U)} \|f - v\|_X ,$$

and the error of best approximation of a function $f \in X$ by an element of $\mathcal{T}_r^T(X)$ is

$$e_r^T(f)_X := e_{r,X}^T(f)_X.$$

The following result provides an upper bound of the best approximation error with tree tensor networks in terms of linear widths of f. The argument is similar to the one for the discrete case given in [9].

Proposition 2.7. *Let $f \in X$ and let $r \in \mathbb{N}^{\#T}$ be an admissible rank. Then*

$$e_r^T(f)_X^2 \leq \sum_{\alpha \in T \setminus \{D\}} \big(\delta_r^\alpha(f)\big)^2. \tag{2.3}$$

Furthermore, for any finite-dimensional subspace $U = U_1 \otimes \ldots \otimes U_d$, we have

$$e_{r,U}^T(f)_X^2 \leq \sum_{\nu=1}^{d} \int_{\mathcal{X}_{\nu^c}} E\big(f^\nu(x_{\nu^c}), U_\nu\big)_{X_\nu}^2 \, dx_{\nu^c} + \sum_{\alpha \in A \setminus \{D\}} \big(\delta_r^\alpha(f)\big)^2, \tag{2.4}$$

with $A = \mathcal{I}(T)$ if $\dim(U_\nu) = r_\nu$ for all $1 \leq \nu \leq d$, or $A = T$ otherwise.

Proof. We first show that for any finite-dimensional subspace $U = U_1 \otimes \ldots \otimes U_d$, and any collection of subspaces $V_\alpha \subset X_\alpha$ with $\dim(V_\alpha) = r_\alpha$, $\alpha \in T \setminus \{D\}$, with A as in the hypothesis we have

$$e_r^T(f)_X \leq e_{r,U}^T(f)_X \leq \sum_{\nu=1}^{d} \|f - \mathcal{P}_{U_\nu}^{\{\nu\}} f\|_X^2 + \sum_{\alpha \in A \setminus \{D\}} \|f - \mathcal{P}_{V_\alpha}^\alpha f\|_X^2. \tag{2.5}$$

The result will be proved by constructing a particular approximation $f_r \in \mathcal{T}_r^T(U)$ and by providing an upper bound of $\|f - f_r\|_X^2$. We define the approximation

$$f_r = \mathcal{P}_{L+1} \mathcal{P}_L \ldots \mathcal{P}_1 f,$$

where $L = \mathrm{depth}(T)$, $\mathcal{P}_\ell = \prod_{\alpha\in T_\ell} \mathcal{P}^\alpha_{V_\alpha}$, for $1 \le \ell \le L$, and $\mathcal{P}_{L+1} = \mathcal{P}^{\{1\}}_{U_1} \dots \mathcal{P}^{\{d\}}_{U_d}$. For disjoint subsets α and β, the projections $\mathcal{P}^\alpha_{V_\alpha}$ and $\mathcal{P}^\beta_{V_\beta}$ commute. Therefore, the definition of $\mathcal{P}_\ell$ does not depend on the order of projections $\mathcal{P}^\alpha_{V_\alpha}$, $\alpha \in T_\ell$, for $\ell = 1, \dots, L$.

Let us first prove that $f_r \in \mathcal{T}^T_r(U)$. We clearly have $f_r \in U$. Then we note that for any function g and any pair $\alpha, \beta \in T$ such that $\beta \subset \alpha$ or $\beta \subset \alpha^c$, we have $\mathrm{rank}_\alpha(\mathcal{P}_{V_\beta} g) \le \mathrm{rank}_\alpha(g)$ for any subspace V_β in X_β. Then for $\alpha \in T$ with level ℓ, since the projections $\mathcal{P}_{\ell'}$ with $\ell' > \ell$ only involve projections $\mathcal{P}^\beta_{V_\beta}$ with $\beta \subset \alpha$ or $\beta \subset \alpha^c$, we have $\mathrm{rank}_\alpha(f_r) \le \mathrm{rank}_\alpha(\mathcal{P}_\ell \dots \mathcal{P}_1 f) = \mathrm{rank}_\alpha(\mathcal{P}^\alpha_{V_\alpha} g) \le r_\alpha$, where $g = \prod_{\beta\in T_\ell, \beta\neq\alpha} \mathcal{P}^\beta_{V_\beta} \mathcal{P}_{\ell-1} \dots \mathcal{P}_1 f$. This proves that $\mathrm{rank}_\alpha(f_r) \le r_\alpha$ for all $\alpha \in T$, which implies $f_r \in \mathcal{T}^T_r(X)$. We therefore deduce that $f_r \in \mathcal{T}^T_r(X) \cap U = \mathcal{T}^T_r(U)$.

Now let us provide the desired upper bound for $\|f - f_r\|_X$. For clarity, we let $\|\cdot\| = \|\cdot\|_X$. Using the properties of orthogonal projections, we have

$$\begin{aligned}\|f - f_r\|^2 &= \|f - \mathcal{P}_{L+1} \dots \mathcal{P}_1 f\|^2 \\ &= \|f - \mathcal{P}_{L+1} f\|^2 + \|\mathcal{P}_{L+1}(f - \mathcal{P}_{L-1} \dots \mathcal{P}_1 f)\|^2 \\ &\le \|f - \mathcal{P}_{L+1} f\|^2 + \|f - \mathcal{P}_L \dots \mathcal{P}_1 f\|^2\end{aligned}$$

Repeating the above arguments, we obtain $\|f - f_r\|^2 \le \sum_{1\le\ell\le L+1} \|f - \mathcal{P}_\ell f\|^2$. For $1 \le \ell \le L$, we have $\|f - \mathcal{P}_\ell f\|^2 = \|f - \prod_{\alpha\in T_\ell} \mathcal{P}^\alpha_{V_\alpha} f\|^2 \le \sum_{\alpha\in T_\ell} \|f - \mathcal{P}^\alpha_{V_\alpha} f\|^2$, which provides the desired bound for the general case. In the case where $\dim(U_\nu) = r_\nu$ for $1 \le \nu \le d$, the result is deduced from the above result by choosing $V_\nu = U_\nu$ for $1 \le \nu \le d$ in the definition of f_r, and by defining $\mathcal{P}_{L+1} = \mathrm{id}$.

Now (2.3) follows from (2.5) by taking the infimum over spaces U_α and V_α, and (2.4) follows from (2.5) by taking the infimum over spaces V_α. □

3. Approximation of functions in Sobolev spaces

In this section, we consider the approximation of functions in Sobolev spaces on $\mathcal{X} = (0,1)^d$ using tree tensor networks: on the one hand, the standard fractional Sobolev spaces $H^s(\mathcal{X})$ for $s > 0$, and on the other hand, the mixed Sobolev spaces $H^s_{\mathrm{mix}}(\mathcal{X})$, which can be characterized as tensor products $H^s_{\mathrm{mix}}(\mathcal{X}) = H^s(0,1) \otimes \dots \otimes H^s(0,1)$ with the canonical cross norm. Assuming a dimension tree T for D, we again write $\mathcal{X}_\alpha = (0,1)^{|\alpha|}$ for $\alpha \in T$ and abbreviate $H^s_\alpha = H^s(\mathcal{X}_\alpha)$ and $H^s_{\alpha,\mathrm{mix}} = H^s_{\mathrm{mix}}(\mathcal{X}_\alpha)$. In addition, for such α, we set

$$d_\alpha = \min\{\#\alpha, d - \#\alpha\}\,.$$

3.1. Sobolev spaces.

We first recall a standard result on Kolmogorov widths of Sobolev balls (see, e.g., [19, Chapter VII]). Here and in what follows, we denote by $B_1(X)$ the unit ball of a given normed space X.

Theorem 3.1. *Let $I = (0,1)^m$. Then*

$$\mathsf{d}_n\big(B_1(H^s(I))\big)_{L^2} \le R n^{-s/m},$$

where $R > 0$ is independent of n but depends on s and on m.

It is well known that there exist approximation tools, such as splines or wavelets, that achieve the rate of convergence given by the Kolmogorov widths, which is even the optimal rate achievable by nonlinear manifold approximation [6]. In other words, there exists a sequence of n-dimensional spaces $V_n \subset L^2(I)$ such that for all $f \in H^s(I)$,

$$E(f, V_n)_{X_\alpha} \leq M n^{-s/m} \|f\|_{H^s}$$

where $M \geq R$ is a constant independent of f and n. For $f \in H^s(\mathcal{X})$, $s > 0$, from this bound we deduce the following estimate on the average linear widths of f defined in (2.1).

Proposition 3.2. *Let $f \in H^s(\mathcal{X})$, $s > 0$. For any $\alpha \in T \setminus \{D\}$, we have*

$$\delta_n^\alpha(f) \leq C n^{-s/d_\alpha} \|f\|_{H^s}$$

where C is independent of r and f, but depends on s and may depend exponentially on d_α.

Proof. Let $f \in H^s(\mathcal{X})$. Since $f^\alpha(x_{\alpha^c}) \in H^s_\alpha$ for almost all x_{α^c}, for x_{α^c} such that $f^\alpha(x_{\alpha^c}) \neq 0$ we have

$$E(f^\alpha(x_{\alpha^c}), V_n)_{L^2_\alpha} = E\big(\|f^\alpha(x_{\alpha^c})\|^{-1}_{H^s_\alpha} f^\alpha(x_{\alpha^c}), V_n\big)_{X_\alpha} \|f^\alpha(x_{\alpha^c})\|_{H^s_\alpha},$$

as well as $E(f^\alpha(x_{\alpha^c}), V_n)_{X_\alpha} = 0$ otherwise. Thus

$$\begin{aligned}
\delta_n^\alpha(f) &= \inf_{\dim(V_n)=n} \left(\int_{\mathcal{X}_{\alpha^c}} E(f^\alpha(x_{\alpha^c}), V_n)^2_{X_\alpha} dx_{\alpha^c} \right)^{1/2} \\
&\leq \inf_{\dim(V_n)=n} \operatorname*{ess\,sup}_{x_{\alpha^c}} E\big(\|f^\alpha(x_{\alpha^c})\|^{-1}_{H^s_\alpha} f^\alpha(x_{\alpha^c}), V_n\big)_{X_\alpha} \|f\|_{L^2(\mathcal{X}_{\alpha^c}; H^s_\alpha)} \\
&\leq \mathsf{d}_n\big(B_1(H^s_\alpha)\big)_{X_\alpha} \|f\|_{H^s},
\end{aligned}$$

where we have used $\|f^\alpha\|_{L^2(\mathcal{X}_{\alpha^c}; H^s_\alpha)} \leq \|f\|_{H^s}$. By Theorem 3.1, we have $\mathsf{d}_n\big(B_1(H^s_\alpha)\big)_{X_\alpha} \leq C_{\#\alpha} n^{-s/\#\alpha}$ with $C_{\#\alpha}$ independent of f and n. The statement now follows with (2.2). □

For each $\nu \in D$, we introduce a sequence of spaces U_{ν,n_ν} with dimension n_ν (such as splines or wavelets) such that for all $u \in H^s_\nu$,

$$E(u, U_{\nu,n_\nu})_{X_\nu} \leq M n_\nu^{-s} \|u\|_{H^s_\nu}, \tag{3.1}$$

which implies

$$\int_{\mathcal{X}_{\nu^c}} E\big(f^\nu(x_{\nu^c}), U_{\nu,n_\nu}\big)^2_{X_\nu} dx_{\nu^c} \leq M^2 n_\nu^{-2s} \|f\|^2_{H^s}. \tag{3.2}$$

Then we let $U_n = U_{1,n_1} \otimes \ldots \otimes U_{d,n_d}$. Now, we can deduce an approximation result for the approximation of functions in Sobolev spaces using tree tensor networks.

Theorem 3.3. *Let $f \in H^s(\mathcal{X})$ and $0 < \varepsilon < 1$, and let $N(f, \varepsilon, d)$ be the minimal complexity $N(T, r, U_n)$ such that*

$$e^T_{r,U_n}(f)_X \leq \varepsilon \|f\|_{H^s}.$$

For any dimension partition tree T, there exists a constant C depending on d such that

$$N(f, \varepsilon, d) \leq C \varepsilon^{-d/s}.$$

Proof. From Proposition 2.7 and Proposition 3.2, we deduce that if

$$r_\alpha \geq \varepsilon^{-d_\alpha/s}(C_{d_\alpha}\sqrt{\#T-1})^{d_\alpha/s}$$

for each interior node $\alpha \in \mathcal{I}(T)\setminus\{D\}$, and $r_\nu = n_\nu \geq \varepsilon^{-1/s}(M\sqrt{\#T-1})^{1/s}$ for all $1\leq \nu \leq d$, then

$$e^T_{r,U_n}(f)_X \leq \varepsilon \|f\|_{H^s}.$$

The minimal values of ranks such that the above conditions hold are such that $r_\alpha := r_\alpha(\varepsilon) \sim \varepsilon^{-d_\alpha/s}$, $\alpha \in T\setminus\{D\}$, with constants depending on d, M and s. Then recalling that $N(f,r,U_n) = \sum_{\nu=1}^d r_\nu^2 + \sum_{\alpha\in\mathcal{I}(T)} r_\alpha \prod_{\beta\in S(\alpha)} r_\beta$, we have

$$N(f,\varepsilon,d) \lesssim d\varepsilon^{-2/s} + \varepsilon^{-(\sum_{\alpha\in S(D)} d_\alpha)/s} + \sum_{\alpha\in\mathcal{I}(T)\setminus\{D\}} \varepsilon^{-(d_\alpha+\sum_{\beta\in S(\alpha)} d_\beta)/s}.$$

We note that $\sum_{\alpha\in S(D)} d_\alpha \leq \sum_{\alpha\in S(D)} \#\alpha = d$. Then consider $\alpha \in \mathcal{I}(T)\setminus\{D\}$. If $d_\alpha = \#\alpha$, we have $\#\alpha \leq d/2$ and $d_\alpha + \sum_{\beta\in S(\alpha)} d_\beta \leq \#\alpha + \sum_{\beta\in S(\alpha)} \#\beta = 2\#\alpha \leq d$. Otherwise, $d_\alpha = \#\alpha^c$, and we have $d_\alpha + \sum_{\beta\in S(\alpha)} d_\beta \leq \#\alpha^c + \sum_{\beta\in S(\alpha)} \#\beta = \#\alpha^c + \#\alpha = d$. Then for any tree, we have $N(f,\varepsilon,d) \lesssim d\varepsilon^{-2/s} + (\#T-d)\varepsilon^{-d/s}$. □

An important observation is that for any dimension partition tree T the complexity $N(f,\varepsilon,d)$ scales as $\varepsilon^{-d/s}$, the optimal rate deduced from nonlinear manifold widths of Sobolev balls [6]. For Sobolev spaces, a shallow network associated with a trivial tree with depth one (Tucker format) has a similar performance as deep tensor networks associated with binary trees.

3.2. **Mixed Sobolev spaces.** We recall a standard result on Kolmogorov widths of balls of mixed Sobolev spaces (see e.g. [25]).

Theorem 3.4. *Let $I = (0,1)^m$. For any $s > 0$, there exists $R > 0$ such that for all $n \in \mathbb{N}$,*

$$\mathsf{d}_n\big(B_1(H^s_{\mathrm{mix}}(I))\big)_{L^2} \leq Rn^{-s}\log(n)^{s(m-1)},$$

where R depends on s and on m.

The above result yields the following estimate of the average linear widths of f.

Proposition 3.5. *For $f \in H^s_{\mathrm{mix}}(\mathcal{X})$ and $\alpha \in T\setminus\{D\}$, we have*

$$\delta^\alpha_n(f) \leq Cn^{-s}\log(n)^{s(d_\alpha-1)}\|f\|_{H^s_{\mathrm{mix}}}$$

and with a constant $C > 0$ that is independent of f and n, but depends on s and may depend exponentially on d_α.

Proof. Let $f \in H^s_{\mathrm{mix}}(\mathcal{X})$. Using $\|f^\alpha\|_{L^2(\mathcal{X}_{\alpha^c};H^s_{\alpha,\mathrm{mix}})} \leq \|f\|_{H^s_{\mathrm{mix}}}$ to argue as in the proof of Proposition 3.2, we obtain

$$\delta^\alpha_n(f) \leq \mathsf{d}_n\big(B_1(H^s_{\alpha,\mathrm{mix}})\big)_{X_\alpha}\|f\|_{H^s_{\mathrm{mix}}}.$$

By Theorem 3.4,

$$\mathsf{d}_n\big(B_1(H^s_{\alpha,\mathrm{mix}})\big)_{X_\alpha} \leq C_{\#\alpha}n^{-s}\log(n)^{s(\#\alpha-1)}$$

with $C_{\#\alpha}$ independent of f and n. The statement follows with (2.2). □

Another bound is obtained in the next proposition by exploiting results on hyperbolic cross approximation [14] (see also [8]). Related conversions from hyperbolic cross approximations to tensor formats have also been considered in [10, §7.6] and [22].

Proposition 3.6. *For $f \in H^s_{\mathrm{mix}}(\mathcal{X})$ and $\alpha \in T \setminus \{D\}$, we have*

$$\delta_n^\alpha(f) \leq C_d\, n^{-2s} \log(n)^{2s(d-2)} \|f\|_{H^s_{\mathrm{mix}}}$$

where $C_d > 0$ is independent of f and n, but depends on s and may depend exponentially on d.

Proof. We rely on results on m-term approximation from [14]. We consider the tensor product wavelet system $\{\phi_j\}_{j\in\mathcal{I}}$ from [14, Section 3.2], where $\mathcal{I} \subset \mathbb{N}^d \times \mathbb{Z}^d$ and where for $(l,k) \in \mathcal{I}$, $\phi_{l,k}(x) = \varphi_{l_1,k_1}(x_1) \dots \varphi_{l_d,k_d}(x_d)$ with φ_{l_ν,k_ν} a one-dimensional wavelet system. Consider $f \in H^s_{\mathrm{mix}}$ with $\|f\|_{H^s_{\mathrm{mix}}} = 1$, where H^s_{mix} coincides with the Lizorkin-Triebel space $S^s_{2,2}F$ (see definition in [14, Section 3.1]). It admits an expansion

$$f = \sum_{j\in\mathcal{I}} c_j(f)\phi_j,$$

with a sequence of coefficients $(c_j(f))_{j\in\mathcal{I}}$ in the sequence space $s^s_{2,2}f(\mathcal{I})$ defined in [14, Definition 3.2]. Then consider the multi-index set

$$\mathcal{I}_L = \{(l,k) \in \mathcal{I} : |l|_1 \leq L\},$$

which is an hyperbolic cross with cardinality $\#\mathcal{I}_L \sim L^{d-1}2^L$ (see [14, Remark 5.7]). Then from [14, Proposition 5.6] and the fact that $\|f\|_{H^s_{\mathrm{mix}}} \sim \|(c_j(f))_{j\in\mathcal{I}}\|_{s^s_{2,2}f}$, we have that the approximation

$$f_L = \sum_{j\in\mathcal{I}_L} c_j\phi_j$$

satisfies $\|f - f_L\|_p \lesssim 2^{-Ls}$.

We let $L_\alpha((l,k)) = \sum_{\nu\in\alpha} l_\nu$, $L_{\alpha^c}((l,k)) = |l|_1 - L_\alpha((l,k))$, and define the sets of multi-indices

$$\mathcal{I}_L^{\leq} = \{j \in \mathcal{I}_L : L_\alpha(j) \leq L_{\alpha^c}(j)\} \quad \text{and} \quad \mathcal{I}_L^{>} = \{j \in \mathcal{I}_L : L_\alpha(j) > L_{\alpha^c}(j)\}.$$

We decompose

$$f_L = f_L^{\leq} + f_L^{>}, \quad f_L^{\leq} = \sum_{j\in\mathcal{I}_L^{\leq}} c_j\phi_j, \quad f_L^{>} = \sum_{j\in\mathcal{I}_L^{>}} c_j\phi_j.$$

Defining $\mathcal{I}^S_{L,\beta} = \{j_\beta : (j_\beta, j_{\beta^c}) \in \mathcal{I}^S_L\}$, with $S \in \{\leq, >\}$ and $\beta \in \{\alpha, \alpha^c\}$, we have

$$f_L^S = \sum_{j_\beta \in \mathcal{I}^S_{L,\beta}} \phi_{j_\beta}(x_\beta)\psi^{S,\beta}_{j_\beta}(x_{\beta^c}),$$

with

$$\psi^{S,\beta}_{j_\beta}(x_{\beta^c}) = \sum_{j_{\beta^c} : (j_\beta, j_{\beta^c}) \in \mathcal{I}^S_L} c_{(j_\beta, j_{\beta^c})}\phi_{j_{\beta^c}}(x_{\beta^c}),$$

so that $\operatorname{rank}_\beta(f_L^S) \le \#\mathcal{I}_{L,\beta}^S$. It follows that

$$\begin{aligned}\operatorname{rank}_\alpha(f_L) &\le \operatorname{rank}_\alpha(f_L^{\le}) + \operatorname{rank}_\alpha(f_L^{>}) \\ &= \operatorname{rank}_\alpha(f_L^{\le}) + \operatorname{rank}_{\alpha^c}(f_L^{>}) \\ &\le \#\mathcal{I}_{L,\alpha}^{\le} + \#\mathcal{I}_{L,\alpha^c}^{>}.\end{aligned}$$

We observe that for all $j \in \mathcal{I}_L^{\le}$, $L \ge L_\alpha(j) + L_{\alpha^c}(j) \ge 2L_\alpha(j)$, and therefore

$$\#\mathcal{I}_{L,\alpha}^{\le} \le \#\{(l_\alpha, k_\alpha) : (l_\alpha, l_{\alpha^c}, k_\alpha, k_{\alpha^c}) \in \mathcal{I}, |l_\alpha|_1 \le L/2\} \lesssim 2^{L/2}(L/2)^{\#\alpha - 1}.$$

Also, for all $j \in \mathcal{I}_L^{<}$, $L > 2L_{\alpha^c}(j)$, and therefore

$$\#\mathcal{I}_{L,\alpha^c}^{>} \lesssim 2^{L/2}(L/2)^{\#\alpha - 1}.$$

Since $\max\{\#\alpha, \#\alpha^c\} \le d-1$, we finally deduce $\operatorname{rank}_\alpha(f_L) \le C2^{L/2}(L/2)^{d-2}$ for some constant C. This implies

$$\delta_n^\alpha(f) \lesssim 2^{-Ls} = x^{-2s}$$

for $n \ge C2^{L/2}(L/2)^{d-2} := Cx\log_2(x)^{d-2}$. A solution to $x\log_2(x)^a = t$, for $t \ge 0$, is given by $x = e^{aW_0(t^{1/a}\log(2)/a)}$, where W_0 is the principal branch of the Lambert function. Then for $a = d-2$ and $n \ge Ct$, we have

$$\delta_n^\alpha(f) \lesssim e^{-2saW_0(t^{1/a}\log(2)/a)}.$$

Using [15, Theorem 2.1], we have that for all $t \ge e$, $W_0(t) \ge \log(t) - \log(\log(t))$, which implies $e^{W_0(t)} \ge t\log(t)^{-1}$. Therefore

$$\begin{aligned}\delta_n^\alpha(f) &\lesssim ((n/C)^{1/a}\log(2)a^{-1}\log((n/C)^{1/a}\log(2)a^{-1})^{-1})^{-2sa} \\ &\le C_{a,s}n^{-2s}\log(n)^{2sa}\end{aligned}$$

for $n \ge C(ae/\log(2))^{d-2}$, with a constant $C_{a,s}$ depending on a and s, which completes the proof. □

The above result provides a better rate in n^{-2s} (instead of n^{-s}) but slightly worse exponent of the $\log(n)$ term. In the following, we will only exploit the result of Proposition 3.6.

Remark 3.7. Based on [14], analogous results can be obtained for L^p-weighted widths with H^s_{mix} replaced by the Lizorkin-Triebel space $S^s_{p,2}F$ for $1 \le p < \infty$.

For each $\nu \in D$, we introduce a sequence of spaces U_{ν,n_ν} with dimension n_ν (e.g. trigonometric polynomials or wavelets) such that for all $u \in H^s(\mathcal{X}_\nu)$, the error $E(u, U_{\nu,n_\nu})_{L^2_\nu}$ satisfies (3.1), which implies

$$\int_{\mathcal{X}_{\nu^c}} E\big(f^\nu(x_{\nu^c}), U_{\nu,n_\nu}\big)_{X_\nu}^2 \, dx_{\nu^c} \le M^2 n_\nu^{-2s}\|f\|^2_{H^s_{\mathrm{mix}}}.$$

Then we let $U_n = U_{1,n_1} \otimes \ldots \otimes U_{d,n_d}$. Now, we can state an approximation result for the approximation of functions in mixed Sobolev spaces using tree tensor networks.

Theorem 3.8. *Let $f \in H^s_{\mathrm{mix}}(\mathcal{X})$. Let $0 < \varepsilon < 1$. We denote by $N(f, \varepsilon, d)$ the complexity $N(T, r, U_n)$ sufficient to achieve a relative error ε for the approximation of f in the format $\mathcal{T}_r^T(U_n)$. There exists a constant C_d, which may depend exponentially on d, such that*

(i) *if T is a trivial tree with depth one,*

$$N(f, \varepsilon, d) \leq C_d \varepsilon^{-d/(2s)} \log(\varepsilon^{-1})^{d(d-2)},$$

(ii) *and if T is a binary tree,*

$$N(f, \varepsilon, d) \leq C_d \varepsilon^{-3/(2s)} \log(\varepsilon^{-1})^{3(d-2)}.$$

Proof. From [22, Lemma 1], we know that for some function $c(\varepsilon, d)$ such that $c(\varepsilon, d) \to 1$ as $\varepsilon \to 0$ and $c(\varepsilon, d) \to \infty$ super-exponentially with d_α, the condition

$$r_\alpha \geq c(\varepsilon, d)(C\sqrt{d + \#T - 1})^{1/(2s)} s^{-d+2} \varepsilon^{-1/(2s)} \log(\varepsilon^{-1})^{d-2} \tag{3.3}$$

implies

$$C r_\alpha^{-2s} \log(r_\alpha)^{2s(d-2)} \leq \varepsilon / \sqrt{d + \#T - 1}.$$

Then using Proposition 2.7 and Proposition 3.6, we have that if $n_\nu \geq \varepsilon^{-1/s}(M\sqrt{d + \#T - 1})^{1/s}$ for all $1 \leq \nu \leq d$, and r_α satisfies (3.3) for each node $\alpha \in T$, then

$$e^T_{r,U_n}(f)_X \leq \varepsilon \|f\|_{H^s_{\mathrm{mix}}}.$$

The minimal values of ranks and dimensions n_ν such that the above conditions hold are such that $r_\alpha := r_\alpha(\varepsilon) \sim \varepsilon^{-1/(2s)} \log(\varepsilon^{-1})^{d-2}$, $\alpha \in T \setminus \{D\}$, and $n_\nu := n_\nu(\varepsilon) \sim \varepsilon^{-1/s}$, $1 \leq \nu \leq d$, with constants depending on d, M and s. Then recalling that $N(f, r, U_n) = \sum_{\nu=1}^d r_\nu n_\nu + \sum_{\alpha \in \mathcal{I}(T)} r_\alpha \prod_{\beta \in S(\alpha)} r_\beta$, we have

$$\begin{aligned} N(f, \varepsilon, d) \lesssim\ & d\varepsilon^{-3/(2s)} \log(\varepsilon^{-1})^{d-2} + \varepsilon^{-a/(2s)} \log(\varepsilon^{-1})^{a(d-2)} \\ & + (\#T - d - 1)\varepsilon^{-(a+1)/(2s)} \log(\varepsilon^{-1})^{(a+1)(d-2)}, \end{aligned}$$

where $a = \max_{\alpha \in \mathcal{I}(T)} \#\alpha$ is the arity of the tree T. □

Remark 3.9. From the above result, we can make the following observations.

(i) For a trivial tree (Tucker format), we have a complexity $\varepsilon^{-d/(2s)}$, up to a logarithmic factor, which compared to the result of Theorem 3.3 for H^s-regularity represents a deterioration by a factor two in the rate. In other words, the extra regularity of functions in H^s_{mix} compared to those of H^s is not exploited by shallow tensor networks.

(ii) For binary trees, we observe a significant gain, going from a complexity in $\varepsilon^{-d/s}$ for H^s to a complexity in $\varepsilon^{-3/(2s)} \log(\varepsilon^{-1})^{3(d-2)}$ for H^s_{mix}. The result is similar to the one obtained in [22] for functions on the torus using results on bilinear approximation from [24]. Specifically, under our present assumptions, [22, Thm. 2] yields a complexity of order $\varepsilon^{-(3+1/(2s))/(2s)} \log(\varepsilon^{-1})^{(1+1/(2s))(d-2)}$ for H^s_{mix}. In other words, in addition to our treatment of non-periodic functions, we also obtain an improvement over the result of [22] concerning the leading algebraic rate, at the price of a slight deterioration in the logarithmic factors for large s .

(iii) We note, however, that deep tensor networks (associated with binary trees) do not achieve the optimal rate in $\varepsilon^{-1/s}$ (up to logarithmic factors) obtained from lower bounds of linear widths of mixed Sobolev balls [27], and reached by hyperbolic cross approximation [8, 25].

Remark 3.10. The rate in $\varepsilon^{-1/s}$ (up to logarithmic terms) could be obtained by tree tensor networks by further exploiting sparsity in the tensors, and by using a measure of complexity N counting the number of nonzero entries. In particular, this rate can be achieved with a trivial tree and a tensor C^D having a sparsity pattern based on hyperbolic crosses. We refer the reader to [1–3] for the analysis of approximation classes of tensor networks with sparsity.

4. Approximation of compositional functions

We have seen in Section 2 that tree tensor networks are a particular class of compositional functions, where the functions that are composed are vector-valued multilinear functions. In this section, we consider the approximation with tree tensor networks of a particular class of compositional functions (also considered in [17]) where the functions that are composed are real-valued functions with Sobolev regularity. In this section, we consider a set $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$, with $\mathcal{X}_\nu = I^\nu$ a bounded and closed interval, equipped with the uniform measure. The results can be easily extended to the case of more general measures.

4.1. **A class $\mathcal{F}_s^T$ of compositional functions.** We let T be a given dimension partition tree over $D = \{1, \ldots, d\}$. We consider the model class $\mathcal{F}^T$ of compositional functions $f : \mathcal{X} \to \mathbb{R}$ of the form

$$f(x) = f_D((g_\alpha(x_\alpha))_{\alpha \in S(D)})$$

where $g_\alpha : \mathcal{X}_\alpha \to I^\alpha \subset \mathbb{R}$ and f_D is a multivariate function with values in $\mathbb{R} =: I^D$, where $g_\alpha(x_\alpha) = x_\alpha$ for $\alpha \in \mathcal{L}(T)$, and for $\alpha \in \mathcal{I}(T)$,

$$g_\alpha(x_\alpha) = f_\alpha((g_\beta(x_\beta))_{\beta \in S(\alpha)})$$

where $g_\beta : \mathcal{X}_\beta \to I^\beta \subset \mathbb{R}$ and f_α is a multivariate function. The function f is completely determined by the set $\mathbf{f} = \{f_\alpha\}_{\alpha \in \mathcal{I}(T)}$ of multivariate functions

$$f_\alpha : \bigtimes_{\beta \in S(\alpha)} I^\beta \to I^\alpha, \quad \alpha \in \mathcal{I}(T).$$

Note that for $\alpha \in \mathcal{I}(T) \setminus \{D\}$, we can take $I^\alpha = [-\|f_\alpha\|_{L^\infty}, \|f_\alpha\|_{L^\infty}]$.

Example 4.1. For the dimension tree T of Figure 1, the function f admits the representation

$$f(x) = f_{\{1,2,3,4,5\}}(f_{\{1,2,3\}}(x_1, f_{\{2,3\}}(x_2, x_3)), f_{\{4,5\}}(x_4, x_5)).$$

We write $\mathcal{I}_\ell(T) = \{\alpha \in \mathcal{I}(T) \colon \mathrm{level}(\alpha) = \ell\}$ for the set of interior nodes with level ℓ, and we define $\mathcal{V}_\ell = \mathcal{I}_\ell(T) \cup \{\alpha \in \mathcal{L}(T) \colon \mathrm{level}(\alpha) \leq \ell\}$, which additionally includes leaf nodes on levels up to ℓ. This means that for each ℓ and $x \in \mathcal{X}$, $(x_\alpha)_{\alpha \in \mathcal{V}_\ell}$

corresponds to a partition of the set of entries of x. For each ℓ and $\alpha \in \mathcal{V}_\ell$, we now set

$$z_\alpha = \begin{cases} f_\alpha\big((x_\beta)_{\beta\in S(\alpha)}\big), & \alpha \in \mathcal{I}_\ell(T), \\ x_\alpha, & \text{otherwise.} \end{cases}$$

With this notation, we define the compositions on level ℓ of a given function $G_\ell \colon \bigtimes_{\beta\in\mathcal{V}_{\ell+1}} I^\beta \to I^D$ with all f_α, $\alpha \in \mathcal{I}_{\ell+1}(T)$, by

$$G_\ell \circ_\ell (f_\alpha)_{\alpha\in\mathcal{I}_{\ell+1}(T)} = \big((x_\alpha)_{\alpha\in\mathcal{V}_{\ell+1}} \mapsto G_\ell\big((z_\alpha)_{\alpha\in\mathcal{V}_{\ell+1}}\big)\big)\,. \tag{4.1}$$

Starting with $\mathcal{C}_0(\mathbf{f}) = f_D$, we now recursively define the compositions of all f_α up to a given level ℓ in T by

$$\mathcal{C}_\ell(\mathbf{f}) = \mathcal{C}_{\ell-1}\big(\mathbf{f}\big) \circ_{\ell-1} (f_\alpha)_{\alpha\in\mathcal{I}_\ell(T)}.$$

We denote by $\mathcal{C}$ the map which associates to the entire set of functions $\{f_\alpha\}_{\alpha\in\mathcal{I}(T)}$ the compositional function $f \in \mathcal{F}^T$,

$$f = \mathcal{C}(\mathbf{f}) = \mathcal{C}_{\mathrm{depth}(T)-1}(\mathbf{f}).$$

We now obtain a restriction to functions with parameters f_α having Sobolev regularity $W^{s,\infty}$ with $s \in \mathbb{N}$ by introducing

$$\mathcal{F}_s^T = \{f = \mathcal{C}(\mathbf{f}) : f_\alpha \in W^{s,\infty}, \alpha \in \mathcal{I}(T)\}.$$

Next, we introduce a subset of $\mathcal{F}_s^T$ where the norms of parameters f_α are controlled. For a given $B = (B_1, \ldots, B_s) \geq (1, \ldots, 1)$, we define

$$\begin{aligned}\mathcal{F}_{s,B}^T = \big\{ f \in \mathcal{F}^T : \|f_D\|_{W^{s,\infty}} \leq 1, \\ \|D^{k_\alpha} f_\alpha\|_{L^\infty} \leq B_{|k_\alpha|} \text{ for } 1 \leq |k_\alpha| \leq s \text{ and } \alpha \in \mathcal{I}(T)\setminus\{D\}\big\},\end{aligned}$$

where k_α is a multi-index in $\mathbb{N}^{\#S(\alpha)}$. For any $f \in \mathcal{F}_{s,B}^T$, we have $\|D^{k_\alpha} f_\alpha\|_{L^\infty} \leq B_1$ for $|k_\alpha| \leq 1$. Using the chain rule, we can prove that $\|f\|_{W^{s,\infty}} \leq p_{T,s,B}$, with $p_{T,s,B}$ depending on the tree T, on s and on B.

4.2. Approximation of functions in $\mathcal{F}_s^T$ using tree tensor networks.

4.2.1. *An approach based on linear widths.* Let T be a dimension tree over $D = \{1, \ldots, d\}$ and consider a compositional function $f \in \mathcal{F}_{s,B}^T$.

Lemma 4.2. *Let $f \in \mathcal{F}_{s,B}^T$. For $\alpha \in T \setminus \{D\}$,*

$$f(x) = F_\alpha(g_\alpha(x_\alpha), x_{\alpha^c}),$$

with $F_\alpha : I^\alpha \times \mathcal{X}_{\alpha^c} \to \mathbb{R}$ such that for any fixed x_{α^c}, $F_\alpha(\cdot, x_{\alpha^c}) \in S_{\mathrm{level}(\alpha)}^{s,\infty,B}$ with

$$\begin{aligned}S_\ell^{s,\infty,B} = \{h = h_1 \circ \ldots \circ h_\ell : \|h_1\|_{W^{s,\infty}} \leq 1, \\ \|D^j h_i\|_{L^\infty} \leq B_j, 0 \leq j \leq s, 2 \leq i \leq \ell\}.\end{aligned}$$

Proof. Let $\alpha \in T \setminus \{D\}$. Let $P(\alpha) \in T$ be the parent node of α in T (such that $\alpha \in S(P(\alpha))$), let $B(\alpha) \subset T$ be the set of brothers of α (such that $B(P(\alpha)) \cup \{\alpha\} = S(\alpha)$), and let A_α be the ancestors of α, which is of cardinality $\#A_\alpha = \mathrm{level}(\alpha)$. We let $F_\alpha : I_\alpha \times \mathcal{X}_{\alpha^c} \to \mathbb{R}$ be the function such that

$$f(x) = F_\alpha(y_\alpha, x_{\alpha^c}) \quad \text{with} \quad y_\alpha = g_\alpha(x_\alpha).$$

Letting $\ell = \text{level}(\alpha)$, and letting $A_\alpha = \{\beta_1, \dots, \beta_\ell\}$ be the ancestors of α ordered by increasing level (that is, $D = \beta_1 \supset \dots \supset \beta_\ell = P(\alpha)$), the function F_α admits the representation

$$F_\alpha(t, x_{\alpha^c}) = f_{\beta_1}(f_{\beta_2}(\dots f_{\beta_\ell}(t, (y_\beta)_{\beta\in B(\alpha)})\dots), (y_\beta)_{\beta\in B(\beta_2)}),$$

with $y_\beta = g_\beta(x_\beta)$. Therefore, for a fixed x_{α^c}, the function $F_\alpha(\cdot, x_{\alpha^c}) : I_\alpha \to \mathbb{R}$ can be written as $F_\alpha(t, x_{\alpha^c}) = h_1 \circ \dots \circ h_\ell(t)$, where $h_i = f_{\beta_i}(\cdot, (y_\beta)_{\beta\in B(\beta_{i+1})}) : I_{\beta_{i+1}} \to I_{\beta_i}$ satisfies $h_i \in W^{s,\infty}$. We have $h_1 = f_D(\cdot, (y_\beta)_{\beta\in B(\beta_2)})$, so that $\|h_1\|_{W^{s,\infty}} \le \|f_D\|_{W^{s,\infty}} \le 1$. And for $2 \le i \le \ell$ and $1 \le j \le s$, $\|D^j h_i\|_{L^\infty} = \|D^{(j,0)} f_{\beta_i}\|_{L^\infty}$, with $(j, 0) \in \mathbb{N}^{\#S(\beta_i)}$, so that $\|D^j h_i\|_{L^\infty} \le B_j$. □

We consider the sets of partial evaluations

$$K_\alpha(f) = \big\{f(\cdot, x_{\alpha^c}) : x_{\alpha^c} \in \mathcal{X}_{\alpha^c}\big\} = \{F_\alpha(g_\alpha(\cdot), x_{\alpha^c}) : x_{\alpha^c} \in \mathcal{X}_{\alpha^c}\},$$

for which we have the following width estimate.

Lemma 4.3. *For $f \in \mathcal{F}^T_{s,B}$ and any $\alpha \in T \setminus \{D\}$,*

$$\mathsf{d}_k(K_\alpha(f))_{L^\infty(\mathcal{X}_\alpha)} \le \mathsf{d}_k\big(S^{s,\infty,B}_{\text{level}(\alpha)}\big)_{L^\infty(I^\alpha)}.$$

Proof of Lemma 4.3. Recall that $K_\alpha(f) = \{F_\alpha(g_\alpha(\cdot), x_{\alpha^c}) : x_{\alpha^c} \in \mathcal{X}_{\alpha^c}\}$, with $g_\alpha : \mathcal{X}_\alpha \to I_\alpha$. Therefore

$$\begin{aligned}\mathsf{d}_k(K_\alpha(f))_{L^\infty(\mathcal{X}_\alpha)} &= \inf_{\dim(V_\alpha)=k} \sup_{x_{\alpha^c}} \inf_{v\in V_\alpha} \|F_\alpha(g_\alpha(\cdot), x_{\alpha^c}) - v(\cdot)\|_{L^\infty(\mathcal{X}_\alpha)} \\ &\le \inf_{\dim(W)=k} \sup_{x_{\alpha^c}} \inf_{w\in W} \|F_\alpha(g_\alpha(\cdot), x_{\alpha^c}) - w(g_\alpha(\cdot))\|_{L^\infty(\mathcal{X}_\alpha)},\end{aligned}$$

where the inequality has been obtained by restricting the minimization over k-dimensional subspaces $V_\alpha = \{w(g_\alpha(\cdot)) : w \in W\}$, with W a k-dimensional subspace of functions defined on I_α. Then, introducing $K_1(F_\alpha) = \{F_\alpha(\cdot, x_{\alpha^c}) : x_{\alpha^c} \in \mathcal{X}_{\alpha^c}\} \subset L^\infty(I_\alpha)$, we have

$$\begin{aligned}\mathsf{d}_k(K_\alpha(f))_{L^\infty(\mathcal{X}_\alpha)} &\le \inf_{\dim(W)=k} \sup_{h\in K_1(F_\alpha)} \inf_{w\in W} \|h(g_\alpha(\cdot)) - w(g_\alpha(\cdot))\|_{L^\infty(\mathcal{X}_\alpha)} \\ &\le \inf_{\dim(W)=k} \sup_{h\in K_1(F_\alpha)} \inf_{w\in W} \|h - w\|_{L^\infty(I_\alpha)} \\ &= \mathsf{d}_k(K_1(F_\alpha))_{L^\infty(I_\alpha)}.\end{aligned}$$

The result now follows from the fact that $K_1(F_\alpha) \subset S^{s,\infty,B}_{\text{level}(\alpha)}$. □

Lemma 4.4. *For $h = h_1 \circ \dots \circ h_\ell \in S^{s,\infty,B}_\ell$, we have*

$$\|h\|_{W^{s,\infty}} \le C(B, s, \ell),$$

where $C(B, 1, \ell) = B_1^\ell$, $C(B, 2, \ell) = \ell B_1^{2\ell-2} B_2$, and more generally, for any $s \ge 1$,

$$C(B, s, \ell) = (C\ell)^{s-1} B_1^{s(\ell-1)} B_\star^s$$

with $B_\star = \max_{1\le j\le s} B_j$ and $C = \max\{1, s-1\}$.

Proof. We first note that $\|h\|_{L^\infty} \leq \|h_1\|_{L^\infty} \leq 1$. Let $h_{>i} = h_{i+1} \circ \ldots \circ h_\ell$ and $g_{k,i} = h_i^{(k)} \circ h_{>i}$, which is such that $\|g_{k,i}\|_{L^\infty} \leq B_k$. We have $h' = \prod_{i=1}^{\ell} g_{1,i}$ and therefore $\|h'\|_{L^\infty} \leq B_1^\ell = B_1^{\ell-1} B_\star$. Then we have $h'' = \sum_{i=1}^{\ell} g'_{1,i} \prod_{1\leq j\leq \ell, j\neq i} g_{1,j}$ with $g'_{1,i} = g_{2,i} \prod_{i<j\leq \ell} g_{1,j}$. As a consequence, $\|h''\|_{L^\infty} \leq \sum_{i=1}^{\ell} B_2 B_1^{2\ell-i-1} \leq \ell B_2 B_1^{2(\ell-1)}$. We next prove by induction that for $s \geq 2$, $f^{(s)}$ can be written as

$$f^{(s)} = \sum_{\alpha\in\Gamma} \prod_{i=1}^{n_1^\alpha} g_{1,p_{1,i}^\alpha} \cdots \prod_{i=1}^{n_s^\alpha} g_{s,p_{s,i}^\alpha}$$

for some index set Γ and $p_{k,i}^\alpha \in \{1,\ldots,\ell\}$, and where

$$\#\Gamma \leq (s-1)!\ell^{s-1}, \quad n_1^\alpha \leq s(\ell-1), \quad n_2^\alpha + \ldots + n_s^\alpha \leq s-1.$$

This is true for $s = 2$. Assuming it is true for some $s \geq 2$, we deduce from the expression of $f^{(s)}$ that

$$\begin{aligned} f^{(s+1)} &= \sum_{\alpha\in\Gamma} \sum_{q=1}^{s} \Big(\sum_{i=1}^{n_q^\alpha} g'_{q,p_{q,i}^\alpha} \prod_{1\leq j\leq n_q^\alpha, j\neq i} g_{q,p_{q,j}^\alpha}\Big)\Big(\prod_{1\leq r\leq s, r\neq q} \prod_{i=1}^{n_r^\alpha} g_{r,p_{r,i}^\alpha}\Big) \\ &= \sum_{\alpha\in\tilde{\Gamma}} \prod_{i=1}^{\tilde{n}_1^\alpha} g_{1,\tilde{p}_{1,i}^\alpha} \cdots \prod_{i=1}^{\tilde{n}_s^\alpha} g_{s,\tilde{p}_{s,i}^\alpha}, \end{aligned}$$

where

$$\begin{aligned} \#\tilde{\Gamma} = \sum_{\alpha\in\Gamma} \sum_{q=1}^{s} n_q^\alpha &\leq \#\Gamma(s(\ell-1)+s-1) \leq s!\ell^s, \\ \tilde{n}_1^\alpha &\leq \max_{1\leq q\leq s} \max_{\alpha\in\Gamma} n_q^\alpha + \ell - 1 \leq (s+1)(\ell-1), \\ \sum_{k=2}^{s+1} \tilde{n}_k^\alpha &\leq \max_{\alpha\in\Gamma} \sum_{k=2}^{s} n_k^\alpha + 1 \leq s. \end{aligned}$$

We finally deduce that

$$\|f^{(s)}\|_{L^\infty} \leq \sum_{\alpha\in\Gamma} B_1^{s(\ell-1)} B_\star^{s-1} \leq (s-1)!\ell^{s-1} B_1^{s(\ell-1)} B_\star^{s-1},$$

from which we obtain $\|f^{(s)}\|_{L^\infty} \leq (C\ell)^{s-1} B_1^{s(\ell-1)} B_\star^s$ with $C = s-1$, concluding the proof. □

From Lemmas 4.3 and 4.4 and Theorem 3.1, we directly obtain the following result.

Lemma 4.5. *For $f \in \mathcal{F}_{s,B}^T$ and any node $\alpha \in T \setminus \{D\}$ with level ℓ_α,*

$$\mathsf{d}_n(K_\alpha(f))_{L^\infty(\mathcal{X}_\alpha)} \leq RC(B,s,\ell_\alpha) n^{-s}, \quad n \in \mathbb{N},$$

with a constant R not depending on ℓ_α, B and d.

For each $\nu \in D$, we introduce a sequence of spaces U_{ν,n_ν} with dimension n_ν (e.g. splines or wavelets) such that for all $u \in H^s(\mathcal{X}_\nu)$,

$$E(u, U_{\nu,n_\nu})_{X_\nu} \leq M n_\nu^{-s} \|u\|_{H^s}, \tag{4.2}$$

with $M \geq R$, with R the constant from Lemma 4.5.

Lemma 4.6. *Let $f \in \mathcal{F}^T_{s,B}$ and $\nu \in D$. For all $u \in K_{\{\nu\}}(f)$, with $\ell_\nu = \mathrm{level}(\{\nu\})$, we have*

$$E(u, U_{\nu,n_\nu})_{X_\nu} \leq MC(B, s, \ell_\nu) n_\nu^{-s}$$

with M a constant not depending on d.

Proof. The set of partial evaluations $K_{\{\nu\}}(f)$ satisfies $K_{\{\nu\}}(f) \subset H^s(\mathcal{X}_\nu)$, so that (4.2) holds for all $u \in K_{\{\nu\}}(f)$. Also $K_{\{\nu\}}(f) \subset S^{s,\infty,B}_{\ell_\nu}$. Therefore, from Lemma 4.4, we have $\|u\|_{H^s} \leq C(B, s, \ell_\nu)$ for all $u \in K_{\{\nu\}}(f)$, which completes the proof. □

Proposition 4.7. *Let $f \in \mathcal{F}^T_{s,B}$. For an admissible rank $r \in \mathbb{N}^{\#T}$ and $U_r = U_{1,r_1} \otimes \ldots \otimes U_{d,r_d}$, we have*

$$e^T_{r,U_r}(f)^2_X \leq \sum_{\alpha \in T \setminus D} \big(MC(B, s, \mathrm{level}(\alpha))\big)^2 r_\alpha^{-2s}.$$

Proof. This follows from Proposition 2.7, the bound $\delta^\alpha_n(f) \leq \mathtt{d}_n(K_\alpha(f))_{L^\infty(\mathcal{X}_\alpha)}$ which holds since $\mathrm{meas}(\mathcal{X}) = 1$, Lemma 4.5, and Lemma 4.6. □

4.2.2. *A constructive approach using uniform approximations.* For each $\alpha \in \mathcal{I}(T)$, let $(\mathcal{Q}^\alpha_N)_{N \in \mathbb{N}^{\#S(\alpha)}}$ be a family of linear operators mapping $C(\bigtimes_{\beta \in S(\alpha)} I^\beta)$ to a finite-dimensional tensor subspace spanned by product basis functions,

$$\mathcal{Q}^\alpha_N \colon C\Big(\bigtimes_{\beta \in S(\alpha)} I^\beta\Big) \to \bigotimes_{\beta \in S(\alpha)} U_{\beta,N_\beta}$$

with

$$U_{\beta,N_\beta} = \mathrm{span}\{\varphi^\beta_{N_\beta,i} \colon i = 1, \ldots, N_\beta\}, \quad \beta \in S(\alpha).$$

We assume these operators to have the properties

$$\|\mathcal{Q}^\alpha_N g\|_{L^\infty} \leq \|g\|_{L^\infty} \tag{4.3}$$

and for all $s \in (0, s^*]$ with $s^* \in (0, \infty]$, $\min N := \min_{\beta \in S(\alpha)} N_\beta$,

$$\|g - \mathcal{Q}^\alpha_N g\|_{L^\infty} \leq Q_{\#S(\alpha)} \big(\min N\big)^{-s} \|g\|_{W^{s,\infty}}. \tag{4.4}$$

Here $Q_{\#S(\alpha)} > 0$ is independent of N and g, but may depend on $\#S(\alpha)$, where $Q_{\#S(\alpha)} \leq Q_a$ for a $Q_a > 0$ whenever $\#S(\alpha) \leq a$. The operators $\mathcal{Q}^\alpha_N$ are thus required to be non-expansive and provide approximations in L^∞-norm converging at optimal rate up to some maximum order.

Example 4.8. The operators $\mathcal{Q}^\alpha_N$ can be chosen as piecewise constant interpolation on a uniform partition into N_β subintervals in the coordinate β, in which case (4.4) holds for $s \in (0, 1]$; or piecewise linear interpolation with $s \in (0, 2]$.

In general, $\mathcal{Q}_N^\alpha g$ is of the form

$$\mathcal{Q}_N^\alpha g = \sum_{i_1,\dots,i_a} c^\alpha_{i_1,\dots,i_a}(g)\, \varphi^{\beta_1}_{N_{\beta_1},i_1} \otimes \cdots \otimes \varphi^{\beta_a}_{N_{\beta_a},i_a}, \quad S(\alpha) = \{\beta_1,\dots,\beta_a\}$$

with coefficients $c^\alpha_{i_1,\dots,i_a}(g) \in \mathbb{R}$.

For $f \in \mathcal{F}^T_{s,B}$ with $f = \mathcal{C}(\mathbf{f})$, for given tuples of positive integers $N_\alpha \in \mathbb{N}^{\#S(\alpha)}$, $\alpha \in T$, we define $\tilde{f}_\ell$ for $\ell = 0,\dots,\operatorname{depth}(T)-1$ recursively as follows:

$$\tilde{f}_0 = \mathcal{Q}^D_{N_D} f_D,$$

and for $\ell > 0$, with $T_\ell = \{\alpha \in T : \operatorname{level}(\alpha) = \ell\}$,

$$\tilde{f}_\ell = \Big(\bigotimes_{\alpha\in\mathcal{V}_\ell} \mathcal{Q}^\alpha_{N_\alpha}\Big)\big(\tilde{f}_{\ell-1} \circ_{\ell-1} (f_\alpha)_{\alpha\in\mathcal{I}_\ell}\big).$$

We set $\tilde{f} = \tilde{f}_{L-1}$ with $L = \operatorname{depth}(T)$.

Lemma 4.9. *For $f \in \mathcal{F}^T_{s,B}$,*

$$\|f - \tilde{f}\|_{L^\infty} \le \sum_{\alpha\in\mathcal{I}(T)} Q_{\#S(\alpha)} C\big(B,s,\operatorname{level}(\alpha)\big)\big(\min N_\alpha\big)^{-s}. \tag{4.5}$$

Proof. Let $f_\ell = \mathcal{C}_\ell(\mathbf{f})$, that is, f_ℓ are the compositions of the functions f_α up to level ℓ without approximations, so that $f = f_{L-1}$. We set

$$\mathcal{Q}_\ell = \bigotimes_{\alpha\in\mathcal{I}_\ell} \mathcal{Q}^\alpha_{N_\alpha} \otimes \bigotimes_{\alpha\in\mathcal{V}_\ell\setminus\mathcal{I}_\ell} \mathrm{id}_\alpha$$

and note that $\|\mathcal{Q}_\ell\| \le 1$ by (4.3) and that by the triangle inequality, for any h,

$$\|h - \mathcal{Q}_\ell h\|_\infty \le \sum_{\alpha\in\mathcal{I}_\ell} \|h - (\mathcal{Q}^\alpha_{N_\alpha} \otimes \mathrm{id}_{\alpha^c})h\|_\infty.$$

Since $f = f_{L-2} \circ_{L-2} (f_\alpha)_{\alpha\in\mathcal{I}_{L-1}(T)}$, combining the above and (4.4) with Lemma 4.4 we obtain

$$\begin{aligned}
\|f - \tilde{f}\|_{L^\infty} &\le \|f - \mathcal{Q}_{L-1} f\|_{L^\infty} \\
&\quad + \|\mathcal{Q}_{L-1}(f_{L-2} \circ_{L-2} (f_\alpha)_{\alpha\in\mathcal{I}_{L-1}(T)}) \\
&\qquad - \mathcal{Q}_{L-1}(\tilde{f}_{L-2} \circ_{L-2} (f_\alpha)_{\alpha\in\mathcal{I}_{L-1}(T)})\|_{L^\infty} \\
&\le \sum_{\alpha\in\mathcal{I}_{L-1}} \|(I - (\mathcal{Q}^\alpha_{N_\alpha} \otimes \mathrm{id}_{\alpha^c}))f\|_{L^\infty} + \|\mathcal{Q}_{L-1}\|\|f_{L-2} - \tilde{f}_{L-2}\|_{L^\infty} \\
&\le \sum_{\alpha\in\mathcal{I}_{L-1}} C(B,s,\operatorname{level}(\alpha)) Q_{\#S(\alpha)}\big(\min N_\alpha\big)^{-s} \\
&\qquad + \|f_{L-2} - \tilde{f}_{L-2}\|_{L^\infty}.
\end{aligned}$$

Applying the same argument to $\ell < L-1$ starting with $f_{L-2} - \tilde{f}_{L-2}$, we recursively obtain

$$\|f - \tilde{f}\|_{L^\infty} \le \sum_{\ell=0}^{L-1} \sum_{\alpha\in\mathcal{I}_\ell} Q_{\#S(\alpha)} C(B,s,\operatorname{level}(\alpha))\big(\min N_\alpha\big)^{-s},$$

which completes the proof. □

From the above, we deduce a result on the approximation with tree tensor networks in L^∞-norm.

Proposition 4.10. *Let $f \in \mathcal{F}^T_{s,B}$ with $s > 0$, and let (4.3) and (4.4) hold for this s. For an admissible rank $r \in \mathbb{N}^{\#T}$ and $U_r = U_{1,r_1} \otimes \ldots \otimes U_{d,r_d}$, we have*

$$e^T_{r,U_r}(f)_{L^\infty} \leq \sum_{\alpha \in T \setminus \{D\}} Q_a C(B,s,\mathrm{level}(\alpha))\, r_\alpha^{-s}, \tag{4.6}$$

with a the arity of T.

Proof. We let $N_\alpha = (r_\beta)_{\beta \in S(\alpha)}$ for all $\alpha \in \mathcal{I}(T)$ and $\tilde{f}$ be the corresponding approximation defined above, which is such that $\tilde{f} \in U_r$ and $\mathrm{rank}_\alpha(\tilde{f}) \leq r_\alpha$ for each $\alpha \in T \setminus \{D\}$. Therefore, $e^T_{r,U_r}(f)_{L^\infty} \leq \|f - \tilde{f}\|_{L^\infty}$ and the result follows from Lemma 4.9 and the fact that for each $\alpha \in \mathcal{I}(T)$, $Q_{\#S(\alpha)} \leq Q_a$, $\big(\min N_\alpha\big)^{-s} = \max_{\beta \in S(\alpha)} r_\beta^{-s} \leq \sum_{\beta \in S(\alpha)} r_\beta^{-s}$ and for each $\beta \in S(\alpha)$, $C(B,s,\mathrm{level}(\alpha)) \leq C(B,s,\mathrm{level}(\beta))$. ☐

Remark 4.11. Similar results can still be obtained when the assumptions (4.3) and (4.4) are relaxed. One example is for each $\alpha \in T$ to choose $\mathcal{Q}^\alpha_N$ as the Lagrangian interpolation operator on $\bigtimes_{\beta \in S(\alpha)} I^\beta$ corresponding to interpolation in Chebyshev points $\{x^\beta_1, \ldots, x^\beta_{N_\beta}\}$ on each I^β; that is, if $S(\alpha)$ contains only interior nodes in the tree, $\mathcal{Q}^\alpha_N$ acts on $g \in C(\bigtimes_{\beta \in S(\alpha)} I^\beta)$ as

$$\mathcal{Q}^\alpha_N g = \sum_{i_1,\ldots,i_a} g(x^{\beta_1}_{i_1}, \ldots, x^{\beta_a}_{i_a})\, \varphi^{\beta_1}_{N_{\beta_1},i_1} \otimes \cdots \otimes \varphi^{\beta_a}_{N_{\beta_a},i_a}, \quad S(\alpha) = \{\beta_1, \ldots, \beta_a\}$$

where $\varphi^\beta_{N,i}$, $i = 1, \ldots, N$ are the Lagrange basis polynomials for the Chebyshev points on I^β. For the Lebesgue constant Λ_N, we have

$$\Lambda_N \leq \prod_{\beta \in S(\alpha)} \left(\frac{2}{\pi} \log(N_\beta + 1) + 1 \right).$$

Recall that $\|\mathcal{Q}^\alpha_N g\|_{L^\infty} \leq \Lambda_N \|g\|_{L^\infty}$ and by Lebesgue's lemma,

$$\begin{aligned} \|g - \mathcal{Q}^\alpha_N g\|_{L^\infty} &\leq (1 + \Lambda_N) \min_{p \in \Pi_N} \|g - p\|_{L^\infty} \\ &\lesssim \big(\min N\big)^{-s} \left(\prod_{\beta \in S(\alpha)} (1 + \log N_\beta) \right) \|g\|_{W^{s,\infty}}, \end{aligned}$$

with $\Pi_N = \bigotimes_{\beta \in S(\alpha)} U_{\beta,N_\beta}$. Thus (4.3) and (4.4) both hold only up to an additional logarithmic factor. This leads to additional factors in $\log(r_\alpha)^a$ on the right in (4.6).

For a given $r = (r_\alpha)_{\alpha \in T}$, we let $N_\alpha = (r_\beta)_{\beta \in S(\alpha)}$ for each $\alpha \in \mathcal{I}(T)$ and $\tilde{f}$ the corresponding approximation. For $\alpha \in \mathcal{I}(T)$, the component tensor of the tree network representation of $\tilde{f}$ at node α is explicitly given for $\alpha \neq D$ by

$$A^\alpha_{j,i_1,\ldots,i_{\#S(\alpha)}} := c^\alpha_{i_1,\ldots,i_{\#S(\alpha)}}\big(\varphi^\alpha_{r_\alpha,j} \circ f_\alpha\big)$$

for $j \in \{1, \ldots, r_\alpha\}$, $i_\beta \in \{1, \ldots, r_\beta\}$, $\beta \in S(\alpha)$, or for $\alpha = D$ by

$$A^D_{i_1,\ldots,i_a} := c^D_{i_1,\ldots,i_{\#S(D)}}\big(f_D\big), \quad i_\beta \in \{1, \ldots, r_\beta\}, \beta \in S(D).$$

Example 4.12. Let $D = \{1,2,3,4\}$ and let T be the corresponding balanced tree with arity $a = 2$. Then we have the explicit tensor representation

$$\begin{aligned}\tilde{f} &= \sum_{i_{12}=1}^{r_{12}} \sum_{i_{34}=1}^{r_{34}} \sum_{i_1=1}^{r_1} \sum_{i_2=1}^{r_2} \sum_{i_3=1}^{r_3} \sum_{i_4=1}^{r_4} c^D_{i_{12},i_{34}}(f_D)\, c^{\{1,2\}}_{i_1,i_2}(\varphi^{\{1,2\}}_{r_{12},i_{12}} \circ f_{\{1,2\}}) \\ &\qquad \times c^{\{3,4\}}_{i_3,i_4}(\varphi^{\{3,4\}}_{r_{34},i_{34}} \circ f_{\{3,4\}})\, \varphi^{\{1\}}_{r_1,i_1} \otimes \varphi^{\{2\}}_{r_2,i_2} \otimes \varphi^{\{3\}}_{r_3 i_3} \otimes \varphi^{\{4\}}_{r_4,i_4} \\ &= \sum_{\substack{i_{12},i_{34} \\ i_1,i_2,i_3,i_4}} A^D_{i_{12},i_{34}} A^{\{1,2\}}_{i_{12},i_1,i_2} A^{\{3,4\}}_{i_{34},i_3,i_4}\, \varphi^{\{1\}}_{r_1,i_1} \otimes \varphi^{\{2\}}_{r_2,,i_2} \otimes \varphi^{\{3\}}_{r_3,i_3} \otimes \varphi^{\{4\}}_{r_4,i_4}.\end{aligned}$$

With the particular choice of $\mathcal{Q}^\alpha_N$ as piecewise constant approximation, with each $\varphi^\beta_{r_\beta,j}$ the characteristic function of a subinterval of I^β, the entries of the tensors A^α of order three have a simple interpretation: their nonzero entries correspond exactly to parallelepipeds in the chosen three-dimensional product grid that intersect the graph of the bivariate function f_α; in other words, these entries mark a "voxel approximation" of the graph of f_α.

4.2.3. *Approximation complexity estimates.* We are now ready to state the main result on the approximation of compositional functions from $\mathcal{F}^T_{s,B}$ by tree tensor networks.

Theorem 4.13. *Let $f \in \mathcal{F}^T_{s,B}$ with $s \in \mathbb{N}$. For $\epsilon > 0$, we denote by $N(f,\varepsilon,d)$ the complexity $N(T,r,U_n)$ sufficient to achieve an error ε for the approximation of f in the format $\mathcal{T}^T_r(U_n)$, with error measured in L^2 for arbitrary s or in L^∞ for $s \le 2$. Let $a = \max_{\alpha\in\mathcal{I}(T)} \#S(\alpha)$ be the arity of the tree and $L = \operatorname{depth}(T)$. Then we have the following estimates:*

(i) *For a trivial tree T with arity $a = d$ and depth $L = 1$,*

$$N(f,\varepsilon,d) \le C_d\, \varepsilon^{-d/s}.$$

with a constant C_d depending super-exponentially on d but not depending on ε.

(ii) *For a tree with arity a independent of d,*

$$N(f,\varepsilon,d) \le C_d\, L^{a+1} \varepsilon^{-(a+1)/s} B_1^{(a+1)L} B_\star^{a+1} \tag{4.7}$$

with a constant C_d depending polynomially on d but not depending on ε.

Proof. From Proposition 4.7 and Proposition 4.10, we have that

$$e^T_{r,U_n}(f)_{L^p} \le \sum_{\alpha\in T\setminus\{D\}} MC(B,s,\operatorname{level}(\alpha))\, r_\alpha^{-s},$$

for $p = 2$ and arbitrary s with a constant M independent of d, and for $p = \infty$ and $s = 1$ or 2 and a constant M depending on the arity a. If the ranks r_α are such that

$$r_\alpha \ge \varepsilon^{-1/s} (\#T - 1)^{1/s} M^{1/s} C(B,s,\ell_\alpha)^{1/s}, \tag{4.8}$$

then $e^T_{r,U_r}(f)_{L^p} \le \varepsilon$. We know that $C(B,s,\ell) \le (C\ell)^{s-1} B_1^{s(\ell-1)} B_\star^s$ by Lemma 4.4. Therefore, from condition (4.8), we deduce the sufficient condition on r_α to achieve

an error ε is

$$r_\alpha \geq M^{1/s}\varepsilon^{-1/s}(C\ell_\alpha)^{1-1/s}B_1^{\ell_\alpha-1}B_\star(\#T-1)^{1/s}.$$

Letting $r_\alpha := r_\alpha(\varepsilon) \sim M^{1/s}\varepsilon^{-1/s}(C\ell_\alpha)^{1-1/s}B_1^{\ell_\alpha-1}B_\star(\#T-1)^{1/(2s)}$ be the minimal ranks satisfying the above condition, we have $N(f,\varepsilon,d) \leq N(T,r,U_r)$, which yields

$$\begin{aligned} N(f,\varepsilon,d) \lesssim\ & M^{a/s}\varepsilon^{-a/s}(C)^{a-a/s}B_\star^a(\#T-1)^{a/s} \\ & + \sum_{\alpha\in\mathcal{I}(T)\setminus\{D\}} M^{(a+1)/s}\varepsilon^{-(a+1)/s}(C\ell_\alpha)^{1-1/s}(C(\ell_\alpha+1))^{a-a/s} \\ & \qquad \times B_1^{(a+1)\ell_\alpha-1}B_\star^{a+1}(\#T-1)^{(a+1)/s} \\ & \qquad + \sum_{\nu=1}^{d} M^{2/s}\varepsilon^{-2/s}(C\ell_\nu)^{2-2/s}B_1^{2\ell_\nu-2}B_\star^2(\#T-1)^{2/s}. \end{aligned}$$

Noting that $\#T \leq 2d-1$, $\ell_\alpha \leq L-1$ for $\alpha \in \mathcal{I}(T)$ and $\ell_\alpha \leq L$ for $\alpha \in \mathcal{L}(T)$, and letting $a = \max_{\alpha\in\mathcal{I}(T)} \#S(\alpha)$ be the arity of the tree and $L = \operatorname{depth}(T)$, we obtain

$$\begin{aligned} N(f,\varepsilon,d) \leq\ & M^{a/s}\varepsilon^{-a/s}(C)^{a(1-1/s)}B_\star^a(\#T-1)^{a/s} \\ & + (\#\mathcal{I}(T)-1)M^{(a+1)/s}\varepsilon^{-(a+1)/s}(CL)^{(a+1)(1-1/s)} \\ & \qquad \times B_1^{(a+1)(L-1)-1}B_\star^{a+1}(2d)^{(a+1)/s} \\ & + dM^{2/s}\varepsilon^{-2/s}(CL)^{2(1-1/s)}B_1^{2L-2}B_\star^2(2d)^{2/s}. \end{aligned}$$

In particular, for a trivial tree T with arity d and depth 1,

$$\begin{aligned} N(f,\varepsilon,d) \leq\ & M^{d/s}\varepsilon^{-d/s}C^{d(1-1/s)}B_\star^d d^{d/s} \\ & + dM^{2/s}\varepsilon^{-2/s}(CL)^{2-2/s}B_\star^2(2d)^{2/s}, \end{aligned}$$

whereas for a tree with arity a independent of d,

$$N(f,\varepsilon,d) \leq \beta d^{1+(a+1)/s}\varepsilon^{-(a+1)/s}L^{a+1}B_1^{(a+1)L}B_\star^{a+1}$$

with a constant β independent of ε and d. □

Remark 4.14. The following observations can be made:

(i) As expected, we observe that for a trivial tree, a shallow tensor network (Tucker format) does not exploit more than the Sobolev regularity H^s of the function and suffers from the curse of dimensionality.

(ii) In the case where the tree has arity a independent of d, we observe in (4.7) that the complexity is exponential in the depth L through the term $B_1^{(a+1)L}$, which depends on the bound B_1 on the first derivatives of functions f_α.

(iii) An important observation is that if a is independent of d and $B_1 = 1$ (that is, the functions f_α are 1-Lipschitz), then there is no exponential dependence on L and the complexity depends algebraically on d and ε^{-1}. That means that tree tensor networks do not present the curse of dimensionality for functions in $\mathcal{F}_{s,B}^T$.

(iv) When $B_1 > 1$ and a independent of d, tree tensor networks may or may not suffer from the curse of dimensionality for functions in $\mathcal{F}_{s,B}^T$, depending on the dependence of L in d.

(v) For binary trees with $a = 2$,

$$N(f, \varepsilon, d) \leq C_d L^3 \varepsilon^{-3/s} B_1^{3L} B_\star^3,$$

where $L \leq \lceil \log_2(d) \rceil$ for a balanced binary tree, and $L = d - 1$ for a linear binary tree. For a linear tree, we observe a complexity exponential in d of the form B_1^{3d}. However, for a balanced tree, the dependence on d is only algebraic with respect to d of the form $B_1^{3\log_2(d)}$. This means that the approximation complexity may depend exponentially on d for any tree with depth depending polynomially on d, in particular for linear trees, but remains algebraic in d for balanced trees, or more generally for any tree with a depth L depending logarithmically on d.

References

[1] M. Ali and A. Nouy, *Approximation with tensor networks. part I: Approximation spaces*, arXiv e-prints, arxiv:2007.00118, 2020.

[2] M. Ali and A. Nouy, *Approximation with tensor networks. part II: Approximation rates for smoothness classes*, arXiv e-prints, arxiv:2007.00128, 2020.

[3] M. Ali and A. Nouy, *Approximation with tensor networks. part III: Multivariate approximation*, arXiv preprint arXiv:2101.11932, 2021.

[4] S. Amaral, D. Allaire and K. Willcox, *A decomposition-based approach to uncertainty analysis of feed-forward multicomponent systems*, International Journal for Numerical Methods in Engineering **100** (2014), 982–1005.

[5] M. Bachmayr, R. Schneider and A. Uschmajew, *Tensor networks and hierarchical tensors for the solution of high-dimensional partial differential equations*, Foundations of Computational Mathematics **16** (2016), 1423–1472.

[6] R. A. DeVore, *Nonlinear Approximation*, Acta Numerica **7** (1998), 51–150.

[7] R. A. DeVore, R. Howard and C. Micchelli, *Optimal nonlinear approximation*, Manuscripta Mathematica **63** (1989), 469–478.

[8] D. Dũng, V. Temlyakov and T. Ullrich, *Hyperbolic Cross Approximation*, Springer, 2018.

[9] L. Grasedyck, *Hierarchical singular value decomposition of tensors*, SIAM J. Matrix Anal. Appl. **31** (2010), 2029–2054.

[10] W. Hackbusch, *Tensor Spaces and Numerical Tensor Calculus*, vol. 56. Springer Nature, 2019.

[11] W. Hackbusch and S. Kühn, *A New Scheme for the Tensor Representation*, Journal of Fourier Analysis and Applications **15** (2009), 706–722.

[12] M. Griebel and H. Harbrecht, *Analysis of tensor approximation schemes for continuous functions*. Foundations of Computational Mathematics, DOI: 10.1007/s10208-021-09544-6, 2021.

[13] M. Griebel, H. Harbrecht and R. Schneider, *Low-rank approximation of continuous functions in Sobolev spaces with dominating mixed smoothness*, arXiv preprint arXiv:2203.04100, 2022.

[14] M. Hansen and W. Sickel, *Best m-term approximation and Sobolev–Besov spaces of dominating mixed smoothness—the case of compact embeddings*, Constructive Approximation **36** (2012), 1–51.

[15] A. Hoorfar and M. Hassani, *Inequalities on the Lambert w function and hyperpower function*, J. Inequalities in Pure and Applied Math. **9** (2008).

[16] S. Marque-Pucheu, G. Perrin and J. Garnier, *Efficient sequential experimental design for surrogate modeling of nested codes*, ESAIM: Probability and Statistics **23** (2019), 245–270.

[17] H. N. Mhaskar and T. Poggio, *Deep vs. shallow networks: An approximation theory perspective*, Analysis and Applications **14** (2016), 829–848.

[18] A. Nouy, *Higher-order principal component analysis for the approximation of tensors in tree-based low-rank formats*, Numerische Mathematik **141** (2019), 743–789.

[19] A. Pinkus, *N-widths in Approximation Theory*, vol. 7, Springer Science & Business Media, 2012.

[20] T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda and Q. Liao, *Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review*, International Journal of Automation and Computing **14** (2017), 503–519.
[21] F. Sanson, O. Le Maitre and P. M. Congedo, *Systems of Gaussian process models for directed chains of solvers*, Computer Methods in Applied Mechanics and Engineering **352** (2019), 32–55.
[22] R. Schneider and A. Uschmajew, *Approximation rates for the hierarchical tensor format in periodic Sobolev spaces*, Journal of Complexity **30** (2014), 56–71.
[23] S. Szalay, M. Pfeffer, V. Murg, G. Barcza, F. Verstraete, R. Schneider and Ö. Legeza, *Tensor product methods and entanglement optimization for ab initio quantum chemistry*, International Journal of Quantum Chemistry **115** (2015), 1342–1391.
[24] V. Temlyakov, *Estimates of best bilinear approximations of periodic functions*, in: Proc. Steklov Inst. Math., 1989, pp. 275–293.
[25] V. Temlyakov, *Approximations of functions with bounded mixed derivative*, Trudy Matematicheskogo Instituta imeni VA Steklova **178** (1986), 3–113.
[26] V. Temlyakov, *Bilinear approximation and applications*, Trudy Matematicheskogo Instituta imeni VA Steklova **187** (1989), 191–215.
[27] V. Temlyakov, *Multivariate Approximation*, vol. 32, Cambridge University Press, 2018.

Manuscript received December 5 2021
revised July 26 2022

M. Bachmayr
Institut für Geometrie und Praktische Mathematik, RWTH Aachen University, Germany
E-mail address: bachmayr@igpm.rwth-aachen.de

A. Nouy
Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray, CNRS UMR 6629, France
E-mail address: anthony.nouy@ec-nantes.fr

R. Schneider
Technische Universität Berlin, Germany
E-mail address: schneidr@math.tu-berlin.de