APPROXIMATION WITH TENSOR NETWORKS. PART I: APPROXIMATION SPACES.

MAZEN ALI AND ANTHONY NOUY

Centrale Nantes, LMJL UMR CNRS 6629, France

Abstract. We study the approximation of functions by tensor networks (TNs). We show that Lebesgue L^p-spaces in one dimension can be identified with tensor product spaces of arbitrary order through tensorization. We use this tensor product structure to define subsets of L^p of rank-structured functions of finite representation complexity. These subsets are then used to define different approximation classes of tensor networks, associated with different measures of complexity. These approximation classes are shown to be quasi-normed linear spaces. We study some elementary properties and relationships of said spaces.

In part II of this work, we will show that classical smoothness (Besov) spaces are continuously embedded into these approximation classes. We will also show that functions in these approximation classes do not possess any Besov smoothness, unless one restricts the *depth* of the tensor networks.

The results of this work are both an analysis of the approximation spaces of TNs and a study of the expressivity of a particular type of neural networks (NN) – namely feed-forward sum-product networks with sparse architecture. The input variables of this network result from the tensorization step, interpreted as a particular featuring step which can also be implemented with a neural network with a specific architecture. We point out interesting parallels to recent results on the expressivity of rectified linear unit (ReLU) networks – currently one of the most popular type of NNs.

1. Introduction

1.1. **Approximation of Functions.** We present a new perspective and therewith a new tool for approximating one-dimensional real-valued functions $f : \Omega \to \mathbb{R}$ on bounded intervals $\Omega \subset \mathbb{R}$. We focus on the one-dimensional setting to keep the presentation comprehensible, but we intend to address the multi-dimensional setting in a forthcoming part III.

The approximation of general functions by simpler "building blocks" has been a central topic in mathematics for centuries with many arising methods: algebraic polynomials, trigonometric polynomials, splines, wavelets or rational functions are among some of the by now established tools. Recently, more sophisticated tools such as tensor networks (TNs) or neural networks (NNs) have proven to be powerful techniques. Approximation methods find application in various areas: signal processing, data compression, pattern recognition, statistical learning, differential equations, uncertainty quantification, and so on. See [5, 6, 8, 10, 21, 22, 32, 33, 35, 36, 39] for examples.

In the 20th century a deep mathematical theory of approximation has been established. It is by now well understood that approximability properties of a function by more standard tools – such as polynomials or splines – are closely related to its smoothness. Moreover, functions that can be approximated with a certain rate can be grouped to form quasi-normed linear spaces. Varying the approximation rate then generates an entire scale of spaces that turn out to be so-called interpolation spaces. See [13, 14] for more details.

In this work, we address the classical question of one-dimensional function approximation but with a new set of tools relying on tensorization of functions and the use of rank-structured tensor formats (or tensor networks). We analyze the resulting approximation classes. We will show in part II [2] that many known classical spaces of smoothness are embedded in these newly defined approximation

E-mail address: mazen.ali@ec-nantes.fr, anthony.nouy@ec-nantes.fr.

Date: July 2, 2020.

2010 *Mathematics Subject Classification.* 41A65, 41A15, 41A10 (primary); 68T05, 42C40, 65D99 (secondary).

Key words and phrases. Tensor Networks, Tensor Trains, Matrix Product States, Neural Networks, Approximation Spaces, Besov Spaces, direct (Jackson) and inverse (Bernstein) inequalities.

Acknowledgments: The authors acknowledge AIRBUS Group for the financial support with the project AtRandom.

classes. On the other hand, we will also show that these classes are, in a sense, much larger than classical smoothness spaces.

1.2. **Neural Networks.** Our work was partly motivated by current developments in the field of deep learning. Originally developed in [34], artificial neural networks (NNs) were inspired by models in theoretical neurophysiology of the nervous system that constitutes a brain. The intention behind NNs was to construct a mathematical (and ultimately digital) analogue of a biological neural network. The increase in computational power in recent decades has led to many successful applications of NNs in various fields [43]. This in turn has sparked interest in a better mathematical foundation for NNs. One flavor of this research is the analysis of the approximation power of feed-forward NNs. In particular, relevant for this work is the recent paper [23], where the authors analyzed the approximation spaces of deep ReLU and rectified power unit (RePU) networks. In Figure 1, we sketch an example of a feed-forward NN.

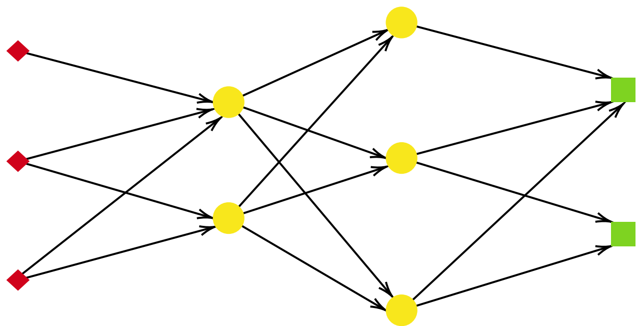


Figure 1. Example of an artificial neural network. On the left we have the *input* nodes marked in red that represent input data to the neural system. The yellow nodes are the *neurons* that perform some simple operations on the input. The edges between the nodes represent synapses or *connections* that transfer (after possibly applying an affine linear transformation) the output of one node into the input of another. The final green nodes are the *output* nodes. In this particular example the number of *layers* L is three, with two *hidden layers*.

Mathematically, a feed-forward NN can be represented by a tuple

$$\Psi = ([T_1, \sigma_1], \dots, [T_L, \sigma_L]),$$

where $T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$ are affine maps; N_l is the number of neurons in layer l, with N_0 being the number of inputs and N_L the number of outputs; the functions $\sigma_l : \mathbb{R}^l \to \mathbb{R}^l$ are typically non-linear and represent the operations performed on data inside the neurons. The functions σ_l are often implemented via a component-wise application of a single non-linear function $\rho : \mathbb{R} \to \mathbb{R}$ referred to as the *activation* function. The realization $\mathcal{R}(\Psi)$ of the NN Ψ is the function

$$\mathcal{R}(\Psi) : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}, \quad \mathcal{R}(\Psi) := \sigma_L \circ T_L \circ \dots \circ \sigma_1 \circ T_1.$$

Before one can proceed with *training* (i.e., estimating) a NN, one usually specifies the *architecture* of the NN: this includes the number of layers L, neurons N_l, connections between the neurons and the non-linearities σ_l. Only after this, one proceeds with training which entails determining the affine maps T_l. This is typically done by minimizing some *loss* or distance functional $\mathcal{J}(\mathcal{R}(\Psi), f)$, where $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ is the target function that we want to approximate, in some sense. If we set $N_0 = N_L = 1$ and, e.g., $\mathcal{J}(\mathcal{R}(\Psi), f) = \|f - \mathcal{R}(\Psi)\|_p$, then we are in the classical setting of one-dimensional approximation[1] (in the L^p-norm).

1.3. **Tensor Networks.** Tensor networks have been studied in parallel in different fields, sometimes under different names: e.g., hierarchical tensor formats in numerical analysis, sum-product networks in machine learning, belief networks in bayesian inference. Tensor networks are commonly applied and studied in condensed matter physics, where understanding phenomena in quantum many-body systems has proven to be a challenging problem, to a large extent due to the sheer amount of dependencies that cannot be simulated even on the most powerful computers (see [38] for a non-technical introduction). In

[1]Though one would typically not train a network by directly minimizing the L^p-norm.

all these fields, a common challenging problem is the approximation of functions of a very large number of variables. This led to the development of tools tailored to so-called *high-dimensional* problems.

The mathematical principle behind tensor networks is to use tensor products. For approximating a d-variate function f, there are several types of tensor formats. The simplest is the so-called r-*term* or *CP* format, where f is approximated as

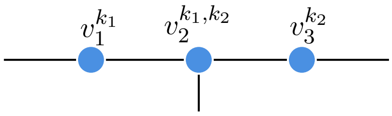
$$f(x_1, \dots, , x_d) \approx \sum_{k=1}^{r} v_1^k(x_1) \cdots v_d^k(x_d). \tag{1.1}$$

If each factor v_ν^k is encoded with N parameters, the total number of parameters is thus dNr, which is linear in the number of variables. The approximation format (1.1) is successful in many applications (chemometrics, inverse problems in signal processing...) but due to a few unfavorable properties (see [24, Chapter 9]), different types of tensor formats are frequently used in numerical approximation. In particular, with the so-called *tensor train* (TT) format or *matrix product state* (MPS), the function f is approximated as

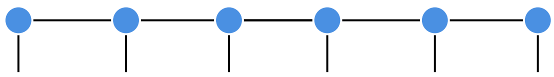
$$f(x_1, \dots, , x_d) \approx \sum_{k_1=1}^{r_1} \dots \sum_{k_{d-1}=1}^{r_{d-1}} v_1^{k_1}(x_1) v_2^{k_1,k_2}(x_2) \dots v_{d-1}^{k_{d-2},k_{d-1}}(x_{d-1}) v_d^{k_{d-1}}(x_d). \tag{1.2}$$

The numbers r_ν are referred to as *multi-linear ranks* or *hierarchical ranks*. The rank r_ν is related to the classical notion of rank for bi-variate functions, by identifying a d-variate function as a function of two complementary groups of variables $(x_1, \dots, x_\nu)$ and $(x_{\nu+1}, \dots, x_d)$. It corresponds to the so-called β-rank r_β, with $\beta = \{1, \dots, \nu\}$. The format in (1.2) is a particular case of *tree-based tensor formats*, or *tree tensor networks* [17, 24], the TT format being associated with a linear dimension partition tree. Numerically, such formats have favorable stability properties with robust algorithms (see [19, 21, 37, 40]). Moreover, the corresponding tensor networks and decompositions have a physical interpretation in the context of entangled many-body systems, see [38, 39].

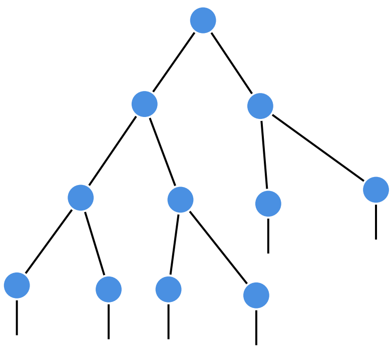
For more general tensor networks, we refer to Figure 2 for graphical representations.



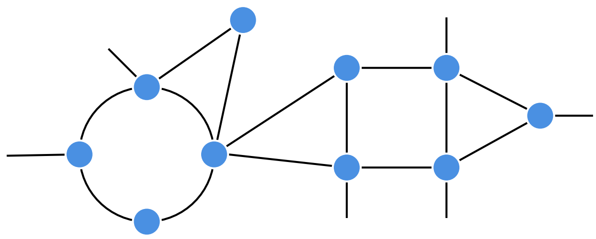
(A) Tensor corresponding to (1.2) with $d = 3$.



(B) General Tensor Train (TT) or Matrix Product State (MPS).



(C) Hierarchical Tucker (HT) or a tree-based format.



(D) General tensor network. Can be seen as an instance of Projected Entangled Pair States (PEPS).

Figure 2. Examples of tensor networks. The vertices in Figure 2 represent the low-dimensional functions in the decomposition, such as $v^1, \dots, v^d$ in (1.2). The edges between the vertices represent summation over an index (*contraction*) between two functions, such as summation over k_ν in (1.2). The free edges represent input variables $x_1, \dots, x_d$ in (1.2).

The specific choice of a tensor network is sometimes suggested by the problem at hand: e.g., in quantum physics by the entanglement/interaction structure of the quantum system that f is to model – see, e.g., [1, 3, 26, 44].

At first glance, it seems that tensor networks are a tool suited only for approximating high-dimensional functions. However, such formats can be applied in any multi-variate setting and this multi-variate setting can be enforced even if $d = 1$ by a "coarse-graining" of an interval in $\mathbb{R}$ allowing to identify a one-dimensional function with a multi-variate function (or tensor). This identification is the *tensorization* of functions which is at the core of the approximation tools considered in this work. It was originally applied for matrices in [41] and later coined as *quantized tensor format* when tensorization is combined with the use of a tensor format.

In high-dimensional approximation, keeping the ranks r_β small relies on the correct choice of the tensor network that "fits" the interaction structure as hinted above. For the approximation of tensorized functions a different type of structure is required. In [20, 41], it was shown that, if f is vector of evaluations of a polynomial on a grid, then choosing the TT format yields hierarchical ranks that are bounded by the degree of the polynomial. Similar statements were shown for trigonometric polynomials and exponential functions. In [29], this fact was utilized to show that a finite element approximation of two-dimensional functions with singularities, where the coefficient vector was stored in a quantized tensor format, automatically recovers an exponential rate of convergence, analogue to that of *hp*-approximation.

This work can be seen as a consolidation and a deeper analysis of approximation of one-dimensional functions using quantized tensor formats. We first show that Lebesgue spaces of p-integrable functions are isometric to tensor product spaces of any order and analyze some basic properties of this identification. We then define and analyze the approximation classes of L^p functions that can be approximated by rank-structured functions in the TT format with a certain rate. In Part II [2], we will show direct and (lack of) inverse embeddings.

1.4. **Tensor vs. Neural Networks.** Recently multiple connections between TNs and NNs have been discovered. In [7], the author exhibits an analogy between the Renormalization Group (RG) – the fundamental physical concept behind TNs – and deep learning, with the scale in RG being akin to depth in NNs. In [31], it was observed that in fact tree tensor networks can be viewed as a specific type of feed-forward NNs with multi-linear functions σ_l, namely sum-product NNs (or arithmetic circuits) [42]. This connection offers intriguing perspectives and it can be exploited both ways: we can use our fundamental understanding of quantum entanglement[2] to measure and design NNs, see [12, 31]. Or we can use NNs to *learn* entanglement and augment tensor networks to better represent highly entangled systems, see [11, 30]. Tensor networks also offer a big choice of well studied and robust numerical algorithms.

In this spirit, our work in part II [2] can be seen as a result on the approximation power of a particular type of NN, where the TT format is a feed-forward sum-product NN and a recurrent neural network architecture. When compared to the results of [23] on approximation classes of RePU networks, we will observe in Part II [2] that both tools achieve optimal approximation order for Sobolev spaces. We also show that TNs (using the TT format) achieve optimal approximation order for Besov spaces on the embedding line (corresponding to non-linear approximation). These statements hold for a tensor network of fixed polynomial degree and *any* smoothness order of the Sobolev/Besov spaces.

On the other hand, TNs are much easier to handle – both analytically and numerically. Moreover, it seems the much simpler architecture of TNs does not sacrifice anything in terms of representing functions of classical smoothness when compared to RePU networks. In particular, both tools are able to recover optimal or close to optimal approximation rates – without being "adapted" to the particular space in question. In other words, all smoothness spaces are included in the same approximation class. This is to be contrasted with more standard approximation tools, such as splines or wavelets, where the approximation class (and thus the approximation method) has to be adapted to the smoothness notion in question. Moreover, both tools will frequently perform better than predicted on specific instances of functions that possess structural features that are not captured by classical smoothness theory[3].

Of course, this is simply to say that both tools do a good job when it comes to classical notions of smoothness. We still expect that NN approximation classes are very different than those of TNs, in an appropriate sense. We also show in Part II [2] that TNs approximation classes (using the TT format) are not embedded into any Besov space – as was shown in [23] for RePU networks.

1.5. **Main Results.** First, we show that any L^p-function f defined on the interval $[0, 1)$ can be identified with a tensor. For a given $b \in \mathbb{N}$ (the base) and $d \in \mathbb{N}$ (the level), we first note that any $x \in [0, 1)$

[2] That is not to say that we have understood quantum entanglement. But an argument can be made that our understanding of entanglement and thus tensor networks offers a different perspective on neural networks.

[3] A fundamental theory of these structures remains an open question for both tools.

can be uniquely decomposed as

$$x = \sum_{k=1}^{d} i_k b^{-k} + b^{-d} y := t_{b,d}(i_1, \dots, i_d, y),$$

where $(i_1, ..., i_d)$ is the representation of $\lfloor b^d x \rfloor$ in base b and $y = b^d x - \lfloor b^d x \rfloor$. This allows to identify a function with a tensor (or multivariate function)

$$\boldsymbol{f}(i_1, \dots, i_d, y) = f(t_{b,d}(i_1, \dots, i_d, y)) := T_{b,d} f(i_1, \dots, i_d, y),$$

and to define different notions of ranks for a univariate function. A function f can be tensorized at different levels $d \in \mathbb{N}$. We analyze the relation between tensorization maps at different levels, and the relation between the ranks of the corresponding tensors of different orders. When looking at $T_{b,d}$ as a map on $L^p([0,1))$, a main result is given by Theorem 2.15 and Lemma A.1.

Main Result 1.1. *For any $0 < p \leq \infty$, $b \in \mathbb{N}$ ($b \geq 2$) and $d \in \mathbb{N}$, the map $T_{b,d}$ is a linear isometry from $L^p([0,1))$ to the algebraic tensor space $\mathbf{V}_{b,d,L^p} := (\mathbb{R}^b)^{\otimes d} \otimes L^p([0,1))$, where $\mathbf{V}_{b,d,L^p}$ is equipped with a reasonable crossnorm.*

For later use in approximation, we introduce the tensor subspace

$$\mathbf{V}_{b,d,S} := (\mathbb{R}^b)^{\otimes d} \otimes S,$$

where $S \subset L^p([0,1))$ is some finite-dimensional subspace. Then, we can identify $\mathbf{V}_{b,d,S}$ with a finite-dimensional subspace of L^p as

$$V_{b,d,S} := T_{b,d}^{-1} (\mathbf{V}_{b,d,S}) \subset L^p([0,1)).$$

We introduce the crucial assumption that S is closed under b-adic dilation, i.e., for any $f \in S$ and any $k \in \{0, \dots, b-1\}$, $f(b^{-1}(\cdot + k)) \in S$. Under this assumption, which is reminiscent of multi-resolution analysis (MRA), we obtain bounds for multilinear ranks that are related to the dimension of S. Also, under this assumption on S, we obtain the main results given by Propositions 2.23 and 2.24 and Theorem 2.25.

Main Result 1.2. *The spaces $V_{b,d,S}$ form a hierarchy of L^p-subspaces, i.e.*

$$S := V_{b,0,S} \subset V_{b,1,S} \subset V_{b,2,S} \subset \dots,$$

and $V_{b,S} := \bigcup_{d \in \mathbb{N}} V_{b,d,S}$ is a linear space. If we further assume that S contains the constant function one, $V_{b,S}$ is dense in L^p for $1 \leq p < \infty$.

For the approximation of multivariate functions (or tensors), we use the set of tensors in the tensor train (TT) format $\mathcal{TT}_{\boldsymbol{r}} (\mathbf{V}_{b,d,S})$, $\boldsymbol{r} = (r_\nu)_{\nu=1}^{d}$, cf. Section 1.3 and Figure 2b. Given a basis $\{\varphi_k\}_{k=1}^{\dim S}$ of S, a tensor $\boldsymbol{f}$ in $\mathcal{TT}_{\boldsymbol{r}} (\mathbf{V}_{b,d,S})$ can be written as

$$\boldsymbol{f}(i_1, \dots, i_d, y) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \sum_{k_{d+1}}^{\dim S} v_1^{k_1}(i_1) v_2^{k_1,k_2}(i_2) v_3^{k_2,k_3}(i_3) \cdots v_d^{k_{d-1},k_d}(i_d) v_{d+1}^{k_d,k_{d+1}} \varphi_k(y),$$

where the parameters $\mathbf{v} := (v_1, \dots, v_{d+1})$ form a tensor network (a collection of low-order tensors) with

$$\mathbf{v} := (v_1, \dots, v_{d+1}) \in \mathbb{R}^{b \times r_1} \times \mathbb{R}^{b \times r_1 \times r_2} \times \dots \times \mathbb{R}^{b \times r_{d-1} \times r_d} \times \mathbb{R}^{r_d \times \dim S} := \mathcal{P}_{b,d,S,\boldsymbol{r}}.$$

With this we define

$$\Phi_{b,d,S,\boldsymbol{r}} = T_{b,d}^{-1} (\mathcal{TT}_{\boldsymbol{r}} (\mathbf{V}_{b,d,S})) = \{\mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) : \mathbf{v} \in \mathcal{P}_{b,d,S,\boldsymbol{r}}\},$$

where $\mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v})$ is the map which associates to a tensor network $\mathbf{v}$ the function $f = T_{b,d}\boldsymbol{f}$ with $\boldsymbol{f}$ defined as above. Then our approximation tool for univariate functions is defined as

$$\Phi := (\Phi_n)_{n \in \mathbb{N}}, \quad \Phi_n = \{\varphi \in \Phi_{b,d,S,\boldsymbol{r}} : d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d, \mathrm{compl}(\varphi) \leq n\},$$

where $\mathrm{compl}(\varphi)$ is some measure of complexity of a function φ, defined as

$$\mathrm{compl}(\varphi) := \min\{\mathrm{compl}(\mathbf{v}) : \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) = \varphi, d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d\},$$

where the infimum is taken over all tensor networks $\mathbf{v}$ whose realization is the function φ. We introduce three different measures of complexity

$$\begin{aligned} \mathrm{compl}_{\mathcal{N}}(\mathbf{v}) &:= \sum_{\nu=1}^{d} r_\nu, \\ \mathrm{compl}_{\mathcal{C}}(\mathbf{v}) &:= br_1 + b\sum_{k=2}^{d} r_{k-1}r_k + r_d \dim S, \\ \mathrm{compl}_{\mathcal{S}}(\mathbf{v}) &:= \sum_{\nu=1}^{d+1} \|v_\nu\|_{\ell_0}, \end{aligned}$$

where $\|v_\nu\|_{\ell_0}$ is the number of non-zero entries in the tensor v_ν. Consequently, this defines three types of subsets

$$\begin{aligned} \Phi_n^{\mathcal{N}} &:= \left\{\varphi \in \Phi : \ \mathrm{compl}_{\mathcal{N}}(\varphi) \le n\right\}, \\ \Phi_n^{\mathcal{C}} &:= \left\{\varphi \in \Phi : \ \mathrm{compl}_{\mathcal{C}}(\varphi) \le n\right\}, \\ \Phi_n^{\mathcal{S}} &:= \left\{\varphi \in \Phi : \ \mathrm{compl}_{\mathcal{S}}(\varphi) \le n\right\}. \end{aligned}$$

Complexity measures $\mathrm{compl}_{\mathcal{C}}$ and $\mathrm{compl}_{\mathcal{N}}$ are related to the TT-ranks of the tensor $T_{b,d(\varphi)}\varphi$, where $d(\varphi)$ is the minimal d such that $\varphi \in V_{b,d,S}$. The function $\mathrm{compl}_{\mathcal{C}}$ is a natural measure of complexity which corresponds to the dimension of the parameter space. The function $\mathrm{compl}_{\mathcal{S}}$ is also a natural measure of complexity which counts the number of non-zero parameters. When interpreting a tensor network $\mathbf{v}$ as a sum-product neural network, $\mathrm{compl}_{\mathcal{N}}(\mathbf{v})$ corresponds to the number of neurons, $\mathrm{compl}_{\mathcal{C}}(\mathbf{v})$ to the number of weights, and $\mathrm{compl}_{\mathcal{S}}(\mathbf{v})$ the number of non-zero weights (or connections).

We use $\Phi_n \in \left\{\Phi_n^{\mathcal{N}}, \Phi_n^{\mathcal{C}}, \Phi_n^{\mathcal{S}}\right\}$ and the corresponding best approximation error

$$E_n(f)_p := \inf_{\varphi \in \Phi_n} \|f - \varphi\|_p$$

for functions f in $L^p([0,1))$ to define approximation classes

$$A_q^\alpha := A_q^\alpha(L^p, (\Phi_n)_{n\in\mathbb{N}}) := \left\{ f \in L^p([0,1)) : \ \|f\|_{A_q^\alpha} < \infty \right\},$$

for $\alpha > 0$ and $0 < q \le \infty$, where

$$\|f\|_{A_q^\alpha} := \begin{cases} \left(\sum_{n=1}^\infty [n^\alpha E_{n-1}(f)]^q \frac{1}{n}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{n\ge 1} [n^\alpha E_{n-1}(f)], & q = \infty. \end{cases}$$

For the three approximation classes

$$\begin{aligned} N_q^\alpha(X) &:= A_q^\alpha(X, (\Phi_n^{\mathcal{N}})_{n\in\mathbb{N}}), \\ C_q^\alpha(X) &:= A_q^\alpha(X, (\Phi_n^{\mathcal{C}})_{n\in\mathbb{N}}), \\ S_q^\alpha(X) &:= A_q^\alpha(X, (\Phi_n^{\mathcal{S}})_{n\in\mathbb{N}}), \end{aligned}$$

we obtain the main result of this part I given by Theorems 3.17 and 3.19.

Main Result 1.3. *For any $\alpha > 0$, $0 < p \le \infty$ and $0 < q \le \infty$, the classes $N_q^\alpha(L^p)$, $C_q^\alpha(L^p)$ and $S_q^\alpha(L^p)$ are quasi-normed vector spaces and satisfy the continuous embeddings*

$$C_q^\alpha(L^p) \hookrightarrow S_q^\alpha(L^p) \hookrightarrow N_q^\alpha(L^p) \hookrightarrow C_q^{\alpha/2}(L^p).$$

1.6. **Outline.** In Section 2, we discuss how one-dimensional functions can be identified with tensors and analyze some basic properties of this identification. In Section 3, we introduce our approximation tool, briefly review general results from approximation theory, and analyze several approximation classes of rank-structured functions. In particular, we show that these classes are quasi-normed linear spaces. We conclude in Section 4 by a brief discussion on how tensorization can be viewed as a particular featuring step and tensor networks as a particular neural network with features as input variables.

2. Tensorization of Functions

We begin by introducing how one-dimensional functions can be identified with tensors of arbitrary dimension. We then introduce finite-dimensional subspaces of tensorized functions and show that these form a hierarchy of subspaces that are dense in L^p. This will be the basis for our approximation tool in Section 3.

2.1. The Tensorization Map.

Consider one-dimensional functions on the unit interval

$$f : [0, 1) \to \mathbb{R}.$$

We tensorize such functions by encoding the input variable $x \in [0, 1)$ as follows. Let $b \in \mathbb{N}$ be the base and $d \in \mathbb{N}$ the level. We introduce a uniform partition of $[0, 1)$ with b^d intervals $[x_i, x_{i+1})$ with $x_i = b^{-d}i$, $0 \le i \le b^d$. An integer $i \in \{0, \dots, b^d - 1\}$ admits a representation $(i_1, \dots, i_d)$ in base b such that

$$i = \sum_{k=1}^{d} i_k b^{d-k},$$

where $i_k \in \{0, \dots, b - 1\} := I_b$. We define a *conversion map* $t_{b,d}$ from $I_b^d \times [0, 1)$ to $[0, 1)$ defined by

$$t_{b,d}(i_1, \dots, i_d, y) = \sum_{k=1}^{d} i_k b^{-k} + b^{-d} y.$$

For any $x \in [0, 1)$, there exists a unique $(i_1, \dots, i_d, y) \in I_b^d \times [0, 1)$ such that $t_{b,d}(i_1, \dots, i_d, y) = x$, where $(i_1, \dots, i_d)$ is the representation of $\lfloor b^d x \rfloor$ in base b and $y = b^d x - \lfloor b^d x \rfloor$. We therefore deduce the following property.

Lemma 2.1. *The conversion map $t_{b,d}$ defines a linear bijection from the set $I_b^d \times [0, 1)$ to the interval $[0, 1)$, with inverse defined for $x \in [0, 1)$ by*

$$t_{b,d}^{-1}(x) = (\lfloor bx \rfloor, \lfloor b^2 x \rfloor \ \mathit{mod}\ b, \dots, \lfloor b^d x \rfloor \ \mathit{mod}\ b, b^d x - \lfloor b^d x \rfloor).$$

Definition 2.2 (Tensorization Map)**.** *We define the* tensorization map

$$T_{b,d} : \mathbb{R}^{[0,1)} \to \mathbb{R}^{I_b^d \times [0,1)}, \quad f \mapsto f \circ t_{b,d} := \boldsymbol{f}$$

which associates to a function $f \in \mathbb{R}^{[0,1)}$ the multivariate function $\boldsymbol{f} \in \mathbb{R}^{I_b^d \times [0,1)}$ such that

$$\boldsymbol{f}(i_1, \dots, i_d, y) := f(t_{b,d}(i_1, \dots, i_d, y)).$$

From Lemma 2.1, we directly deduce the following property of $T_{b,d}$.

Proposition 2.3. *The tensorization map $T_{b,d}$ is a linear bijection from $\mathbb{R}^{[0,1)}$ to $\mathbb{R}^{I_b^d \times [0,1)}$, with inverse given for $\boldsymbol{f} \in \mathbb{R}^{I_b^d \times [0,1)}$ by $T_{b,d}^{-1} \boldsymbol{f} = \boldsymbol{f} \circ t_{b,d}^{-1}$.*

The space $\mathbb{R}^{I_b^d \times [0,1)}$ can be identified with the algebraic tensor space

$$\mathbf{V}_{b,d} := \mathbb{R}^{I_b^d} \otimes \mathbb{R}^{[0,1)} = \underbrace{\mathbb{R}^{I_b} \otimes \dots \otimes \mathbb{R}^{I_b}}_{d \text{ times}} \otimes \mathbb{R}^{[0,1)} =: (\mathbb{R}^{I_b})^{\otimes d} \otimes \mathbb{R}^{[0,1)},$$

which is the set of functions $\boldsymbol{f}$ defined on $I_b^d \times [0, 1)$ that admit a representation

$$\boldsymbol{f}(i_1, \dots, i_d, y) = \sum_{k=1}^{r} v_1^k(i_1) \dots v_d^k(i_d) g^k(y) := \sum_{k=1}^{r} (v_1^k \otimes \dots \otimes v_d^k \otimes g^k)(i_1, \dots, i_d, y) \tag{2.1}$$

for some $r \in \mathbb{N}$ and for some functions $v_\nu^k \in \mathbb{R}^{I_b}$ and $g^k \in \mathbb{R}^{[0,1)}$, $1 \le k \le r$, $1 \le \nu \le d$. Letting $\{\delta_{j_\nu} : j_\nu \in I_b\}$ be the canonical basis of $\mathbb{R}^{I_b}$, defined by $\delta_{j_\nu}(i_\nu) = \delta_{i_\nu, j_\nu}$, a function $\boldsymbol{f} \in \mathbb{R}^{I_b^d \times [0,1)}$ admits the particular representation

$$\boldsymbol{f} = \sum_{j_1 \in I_b} \dots \sum_{j_d \in I_b} \delta_{j_1} \otimes \dots \otimes \delta_{j_d} \otimes \boldsymbol{f}(j_1, \dots, j_d, \cdot). \tag{2.2}$$

The following result provides an interpretation of the above representation.

Lemma 2.4. *Let $f \in \mathbb{R}^{[0,1)}$ and $\boldsymbol{f} = T_{b,d} f \in \mathbf{V}_{b,d}$. For $(j_1, \ldots, j_d) \in I_b^d$ and $j = \sum_{k=1}^d j_k b^{d-k}$, it holds*

$$T_{b,d}(f \mathbb{1}_{[b^{-d}j, b^{-d}(j+1))}) = \delta_{j_1} \otimes \ldots \otimes \delta_{j_d} \otimes \boldsymbol{f}(j_1, \ldots, j_d, \cdot), \tag{2.3}$$

and

$$\boldsymbol{f}(j_1, \ldots, j_d, \cdot) = f(b^{-d}(j + \cdot)), \tag{2.4}$$

where $f(b^{-d}(j+\cdot))$ is the restriction of f to the interval $[b^{-d}j, b^{-d}(j+1))$ rescaled to $[0,1)$.

Proof. For $x = t_{b,d}(i_1, \ldots, i_d, y)$,

$$\begin{aligned} f(x)\mathbb{1}_{[b^{-d}j, b^{-d}(j+1))}(x) &= \delta_{j_1}(i_1) \ldots \delta_{j_d}(i_d) f(t_{b,d}(i_1, \ldots, i_d, y)) \\ &= \delta_{j_1}(i_1) \ldots \delta_{j_d}(i_d) f(t_{b,d}(j_1, \ldots, j_d, y)) \\ &= \delta_{j_1}(i_1) \ldots \delta_{j_d}(i_d) \boldsymbol{f}(j_1, \ldots, j_d, y). \end{aligned}$$

The property (2.4) simply results from the definition of $\boldsymbol{f}$. □

From Lemma 2.4, we deduce that the representation (2.2) corresponds to the decomposition of $f = T_{b,d}^{-1}(\boldsymbol{f})$ as a superposition of functions with disjoint supports,

$$f(x) = \sum_{j=0}^{b^d-1} f_j(x), \quad f_j(x) = \mathbb{1}_{[b^{-d}j, b^{-d}(j+1))}(x) f(x), \tag{2.5}$$

where f_j is the function supported on the interval $[b^{-d}j, b^{-d}(j+1))$ and equal to f on this interval. Also, Lemma 2.4 yields the following result.

Corollary 2.5. *A function $f \in \mathbb{R}^{[0,1)}$ defined by*

$$f(x) = \begin{cases} g(b^d x - j) & \text{for} \quad x \in [b^{-d}j, b^{-d}(j+1)) \\ 0 & \text{elsewhere} \end{cases}$$

with $g \in \mathbb{R}^{[0,1)}$ and $0 \le j < b^d$ admits a tensorization $T_{b,d} f = \delta_{j_1} \otimes \ldots \otimes \delta_{j_d} \otimes g$, which is an elementary tensor.

We now provide a useful result on compositions of tensorization maps for changing the representation level of a function.

Lemma 2.6. *Let $\bar{d}, d \in \mathbb{N}$ such that $\bar{d} > d$. For any $(i_1, \ldots, i_{\bar{d}}, y) \in I_b^{\bar{d}} \times [0,1)$, it holds*

$$t_{b,\bar{d}}(i_1, \ldots, i_{\bar{d}}, y) = t_{b,d}(i_1, \ldots, i_d, t_{b,\bar{d}-d}(i_{d+1}, \ldots, i_{\bar{d}}, y)),$$

and the operator $T_{b,\bar{d}} \circ T_{b,d}^{-1}$ from $\mathbf{V}_{b,d}$ to $\mathbf{V}_{b,\bar{d}}$ is such that

$$T_{b,\bar{d}} \circ T_{b,d}^{-1} = id_{\{1,\ldots,d\}} \otimes T_{b,\bar{d}-d}$$

where $id_{\{1,\ldots,d\}}$ is the identity operator on $\mathbb{R}^{I_b^d}$ and $T_{b,\bar{d}-d}$ is the tensorization map from $\mathbb{R}^{[0,1)}$ to $\mathbf{V}_{b,\bar{d}-d}$. Also, the operator $T_{b,d} \circ T_{b,\bar{d}}^{-1}$ from $\mathbf{V}_{b,\bar{d}}$ to $\mathbf{V}_{b,d}$ is such that

$$T_{b,d} \circ T_{b,\bar{d}}^{-1} = id_{\{1,\ldots,d\}} \otimes T_{b,\bar{d}-d}^{-1}.$$

Proof. See Appendix A. □

For $d = 0$, we adopt the conventions that $t_{b,0}$ is the identity on $[0,1)$, $T_{b,0}$ is the identity operator on $\mathbb{R}^{[0,1)}$, and $\mathbf{V}_{b,0} = \mathbb{R}^{[0,1)}$.

2.2. **Ranks and Minimal Subspaces.** The minimal integer r such that $\boldsymbol{f} \in \mathbf{V}_{b,d}$ admits a representation of the form (2.1) is the *canonical tensor rank* of $\boldsymbol{f}$ denoted $r(\boldsymbol{f})$. We deduce from the representation (2.2) that

$$r(\boldsymbol{f}) \le b^d.$$

Other notions of ranks can be defined from the classical notion of rank by identifying a tensor with a tensor of order two (through unfolding). Letting $V_\nu := \mathbb{R}^{I_b}$ for $1 \le \nu \le d$, and $V_{d+1} := \mathbb{R}^{[0,1)}$, we have

$$\mathbf{V}_{b,d} = \bigotimes_{\nu=1}^{d+1} V_\nu.$$

Then for any $\beta \subset \{1, \dots, d+1\}$ and its complementary set $\beta^c = \{1, \dots, d+1\} \setminus \beta$, a tensor $\boldsymbol{f} \in \mathbf{V}_{b,d}$ can be identified with an order-two tensor in $\mathbf{V}_\beta \otimes \mathbf{V}_{\beta^c}$, where $\mathbf{V}_\gamma = \bigotimes_{\nu \in \gamma} V_\nu$, called the β-unfolding of $\boldsymbol{f}$. This allows us to define the notion of β-rank.

Definition 2.7 (β-rank)**.** *For $\beta \subset \{1, \dots, d+1\}$, the β-rank of $\boldsymbol{f} \in \mathbf{V}_{b,d}$, denoted $r_\beta(\boldsymbol{f})$, is the minimal integer such that $\boldsymbol{f}$ admits a representation of the form*

$$\boldsymbol{f} = \sum_{k=1}^{r_\beta(\boldsymbol{f})} \boldsymbol{v}_\beta^k \otimes \boldsymbol{v}_{\beta^c}^k, \tag{2.6}$$

where $\boldsymbol{v}_\beta^k \in \mathbf{V}_\beta$ and $\boldsymbol{v}_{\beta^c}^k \in \mathbf{V}_{\beta^c}$.

Since $\mathbf{V}_{b,d}$ is an algebraic tensor space, the β-rank is finite and we have $r_\beta(\boldsymbol{f}) \le r(\boldsymbol{f})$ (though the β-rank can be much smaller). Moreover, we have the following straightforward property

$$r_\beta(\boldsymbol{f}) = r_{\beta^c}(\boldsymbol{f}),$$

and the bound

$$r_\beta(\boldsymbol{f}) \le \min \left\{ \prod_{\nu \in \beta} \dim V_\nu, \prod_{\nu \in \beta^c} \dim V_\nu \right\}, \tag{2.7}$$

which can be useful for small b and either very small or very large $\#\beta$.

Representation (2.6) is not unique but the space spanned by the $\boldsymbol{v}_k^\beta$ is unique and corresponds to the *β-minimal subspace* of $\boldsymbol{f}$.

Definition 2.8 (β-minimal subspace)**.** *For $\beta \subset \{1, \dots, d+1\}$, the β-minimal subspace of $\boldsymbol{f}$, denoted $U_\beta^{\min}(\boldsymbol{f})$, is the smallest subspace $\mathbf{U}_\beta \subset \mathbf{V}_\beta$ such that $\boldsymbol{f} \in \mathbf{U}_\beta \otimes \mathbf{V}_{\beta^c}$, and its dimension is*

$$\dim(U_\beta^{\min}(\boldsymbol{f})) = r_\beta(\boldsymbol{f}).$$

We have the following useful characterization of minimal subspaces from partial evaluations of a tensor.

Lemma 2.9. *For $\beta \subset \{1, \dots, d\}$ and $\boldsymbol{f} \in \mathbf{V}_{b,d}$,*

$$U_{\beta^c}^{\min}(\boldsymbol{f}) = \operatorname{span}\{\boldsymbol{f}(j_\beta, \cdot) : j_\beta \in I_b^{\#\beta}\} \subset \mathbf{V}_{b,d-\#\beta},$$

where $\boldsymbol{f}(j_\beta, \cdot) \in \mathbf{V}_{\beta^c} = \mathbf{V}_{b,d-\#\beta}$ is a partial evaluation of $\boldsymbol{f}$ along dimensions $\nu \in \beta$.

Proof. See Appendix A. □

Next we define a notion of (β, d)-rank for univariate functions in $\mathbb{R}^{[0,1)}$.

Definition 2.10 ((β, d)-rank)**.** *For a function $f \in \mathbb{R}^{[0,1)}$, $d \in \mathbb{N}$ and $\beta \subset \{1, \dots, d+1\}$, we define the (β, d)-rank of f, denoted $r_{\beta,d}(f)$, as the β-rank of its tensorization in $\mathbf{V}_{b,d}$,*

$$r_{\beta,d}(f) = r_\beta(T_{b,d} f).$$

In the rest of this work, we will essentially consider subsets β of the form $\{1, \dots, \nu\}$ or $\{\nu+1, \dots, d+1\}$ for some $\nu \in \{1, \dots, d\}$. For the corresponding β-ranks, we will use the shorthand notations

$$r_\nu(\boldsymbol{f}) := r_{\{1,\dots,\nu\}}(\boldsymbol{f}), \quad r_{\nu,d}(f) = r_{\{1,\dots,\nu\},d}(f).$$

Note that $r_\nu(\boldsymbol{f})$ should not be confused with $r_{\{\nu\}}(\boldsymbol{f})$. The ranks $(r_\nu(\boldsymbol{f}))_{1 \le \nu \le d}$ of a tensor $\boldsymbol{f} \in \mathbf{V}_{b,d}$ have to satisfy some relations, as seen in the next lemma.

Lemma 2.11 (Ranks Admissibility Conditions)**.** *Let $\boldsymbol{f} = \mathbf{V}_{b,d}$. For any set $\beta \subset \{1, \dots, d+1\}$ and any partition $\beta = \gamma \cup \alpha$, we have*

$$r_\beta(\boldsymbol{f}) \le r_\gamma(\boldsymbol{f}) r_\alpha(\boldsymbol{f})$$

and in particular

$$r_{\nu+1}(\boldsymbol{f}) \le b r_\nu(\boldsymbol{f}) \quad \text{and} \quad r_\nu(\boldsymbol{f}) \le b r_{\nu+1}(\boldsymbol{f}), \quad 1 \le \nu \le d-1, \tag{2.8}$$

Proof. See Appendix A. □

A function f admits infinitely many tensorizations of different levels. The following result provides a relation between minimal subspaces.

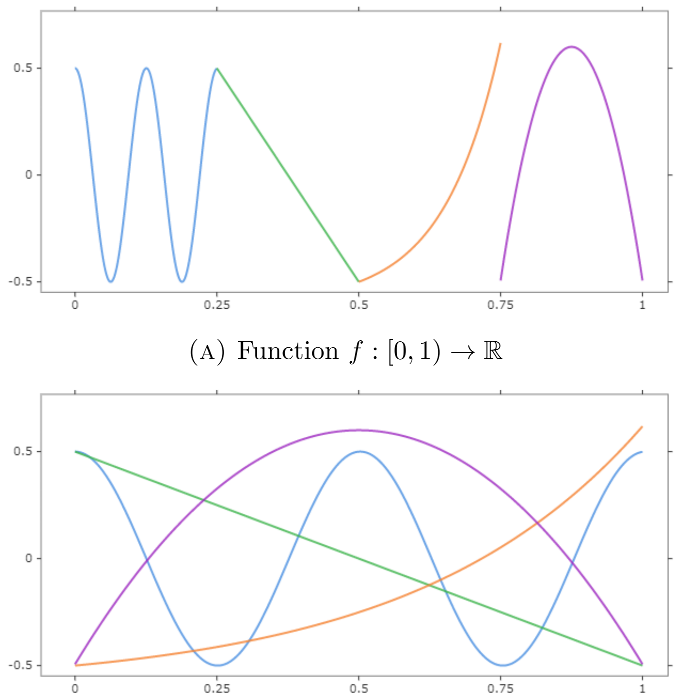
Lemma 2.12. *Consider a function $f \in \mathbb{R}^{[0,1)}$ and its tensorization $\boldsymbol{f}^d = T_{b,d} f$ at level d. For any $1 \le \nu \le d$,*

$$T_{b,d-\nu}^{-1}(U^{\min}_{\{\nu+1,\dots,d+1\}}(\boldsymbol{f}^d)) = \operatorname{span}\{\boldsymbol{f}^\nu(j_1,\dots,j_\nu,\cdot) : (j_1,\dots,j_\nu) \in I_b^\nu\} = U^{\min}_{\{\nu+1\}}(\boldsymbol{f}^\nu),$$

where $\boldsymbol{f}^\nu = T_{b,\nu} f$ is the tensorization of f at level ν.

Proof. See Appendix A. □

For $j = \sum_{k=1}^{\nu} j_k b^{\nu-k}$, since $\boldsymbol{f}^\nu(j_1,\dots,j_\nu,\cdot) = f(b^{-\nu}(j+\cdot))$ is the restriction of f to the interval $[b^{-\nu}j, b^{-\nu}(j+1))$ rescaled to $[0,1)$, Lemma 2.12 provides a simple interpretation of minimal subspace $U^{\min}_{\{\nu+1\}}(\boldsymbol{f}^\nu)$ as the linear span of contiguous pieces of f rescaled to $[0,1)$, see the illustration in Figure 3.



(A) Function $f : [0,1) \to \mathbb{R}$

(B) Partial evaluations $\boldsymbol{f}^\nu(j_1,j_2,\cdot)$ for $(j_1,j_2) \in \{0,1\}^2$.

FIGURE 3. A function $f : [0,1) \to \mathbb{R}$ and partial evaluations of $\boldsymbol{f}^\nu \in \mathbf{V}_{b,d}$ for $b = d = 2$.

Corollary 2.13. *Let $f \in \mathbb{R}^{[0,1)}$ and $d \in \mathbb{N}$. For any $1 \le \nu \le d$,*

$$r_{\nu,d}(f) = r_{\nu,\nu}(f)$$

and

$$r_{\nu,\nu}(f) = \dim \operatorname{span}\{f(b^{-\nu}(j+\cdot)) : 0 \le j \le b^\nu - 1\}.$$

Proof. It holds

$$r_{\nu,d}(f) = r_{\{1,\dots,\nu\}}(\boldsymbol{f}^d) = r_{\{\nu+1,\dots,d+1\}}(\boldsymbol{f}^d) = \dim U^{\min}_{\{\nu+1,\dots,d+1\}}(\boldsymbol{f}^d)$$

and

$$r_{\nu,\nu}(f) = r_{\{1,\dots,\nu\}}(\boldsymbol{f}^\nu) = r_{\nu+1}(\boldsymbol{f}^\nu) = \dim U^{\min}_{\{\nu+1\}}(\boldsymbol{f}^\nu).$$

Lemma 2.12 then implies that $r_{\nu,d}(f) = r_{\nu,\nu}(f)$ and provides the characterization from the linear span of $\boldsymbol{f}^\nu(j_1,\dots,j_\nu,\cdot) = f(b^{-\nu}(j+\cdot))$, with $j = \sum_{k=1}^{\nu} j_k b^{\nu-k}$, which is linearly identified with the restriction $f_{|[b^{-\nu}j, b^{-\nu}(j+1))}$ shifted and rescaled to $[0,1)$. □

2.3. **Measures, Lebesgue and Smoothness Spaces.** We now look at $T_{b,d}$ as a linear map between spaces of measurable functions, by equipping the interval $[0,1)$ with the Lebesgue measure.

Proposition 2.14. *The Lebesgue measure λ on $[0,1)$ is the push-forward measure through the map $t_{b,d}$ of the product measure $\mu_{b,d} := \mu_b^{\otimes d} \otimes \lambda$, where μ_b is the uniform probability measure on I_b. Then the tensorization map $T_{b,d}$ defines a linear isomorphism from the space of measurable functions $\mathbb{R}^{[0,1)}$ to the space of measurable functions $\mathbb{R}^{I_b^d \times [0,1)}$, where $[0,1)$ is equipped with the Lebesgue measure and $I_b^d \times [0,1)$ is equipped with the product measure $\mu_{b,d}$.*

Proof. See Appendix A. $\square$

2.3.1. *Lebesgue Spaces.* For $0 < p \leq \infty$, we consider the Lebesgue space $L^p([0,1))$ of functions defined on $[0,1)$ equipped with its standard (quasi-)norm. Then we consider the algebraic tensor space

$$\mathbf{V}_{b,d,L^p} := \mathbb{R}^{I_b}{}^{\otimes d} \otimes L^p([0,1)) \subset \mathbf{V}_{b,d},$$

which is the space of multivariate functions $\boldsymbol{f}$ on $I_b^d \times [0,1)$ with partial evaluations $\boldsymbol{f}(j_1, \ldots, j_d, \cdot) \in L^p([0,1))$. From hereon we frequently abbreviate $L^p := L^p([0,1))$.

Theorem 2.15 (Tensorization is an L^p-Isometry)**.** *For any $0 < p \leq \infty$, $T_{b,d}$ is a linear isometry from $L^p([0,1))$ to $\mathbf{V}_{b,d,L^p}$ equipped with the (quasi-)norm $\|\cdot\|_p$ defined by*

$$\|\boldsymbol{f}\|_p^p = \sum_{j_1 \in I_b} \cdots \sum_{j_d \in I_b} b^{-d} \|\boldsymbol{f}(j_1, \ldots, j_d, \cdot)\|_p^p$$

for $p < \infty$, or

$$\|\boldsymbol{f}\|_\infty = \max_{j_1 \in I_b} \ldots \max_{j_d \in I_b} \|\boldsymbol{f}(j_1, \ldots, j_d, \cdot)\|_\infty.$$

Proof. The results follows from Proposition 2.14 and by noting that for $\boldsymbol{f} = T_{b,d} f = f \circ t_{b,d}$,

$$\|f\|_p^p = \int_{[0,1)} |f(x)|^p d\lambda(x) = \int_{I_b^d \times [0,1)} |\boldsymbol{f}(j_1, \ldots, j_d, y)|^p d\mu_{b,d}(j_1, \ldots, j_d, y) = \|\boldsymbol{f}\|_p^p$$

for $p < \infty$, and $\|f\|_\infty = \operatorname{ess\,sup}_x |f(x)| = \operatorname{ess\,sup}_{(j_1, \ldots, j_d, y)} |\boldsymbol{f}(j_1, \ldots, j_d, y)| = \|\boldsymbol{f}\|_\infty$. $\square$

We denote by $\ell^p(I_b)$ the space $\mathbb{R}^{I_b}$ equipped with the (quasi-)norm $\|\cdot\|_{\ell^p}$ defined for $v = (v_k)_{k \in I_b}$ by

$$\|v\|_{\ell^p}^p := b^{-1} \sum_{k=0}^{b-1} |v_k|^p \quad (p < \infty), \quad \|v\|_{\ell^\infty} := \max_{0 \leq k \leq b-1} |v_k|.$$

The space $\mathbf{V}_{b,d,L^p}$ can then be identified with the algebraic tensor space

$$(\ell^p(I_b))^{\otimes d} \otimes L^p([0,1)).$$

and $\|\cdot\|_p$ is a crossnorm, i.e., satisfying for an elementary tensor $v^1 \otimes \ldots \otimes v^{d+1} \in \mathbf{V}_{b,d,L^p}$,

$$\|v^1 \otimes \ldots \otimes v^{d+1}\|_p = \|v^1\|_{\ell^p} \ldots \|v^d\|_{\ell^p} \|v^{d+1}\|_p.$$

and even a reasonable crossnorm for $1 \leq p \leq \infty$ (see Lemma A.1 in the appendix). We let $\{e_k^p\}_{k \in I_b}$ denote the normalized canonical basis of $\ell^p(I_b)$, defined by

$$e_k^p = b^{1/p} \delta_k \text{ for } 0 < p < \infty, \quad \text{and} \quad e_k^\infty = \delta_k \text{ for } p = \infty. \tag{2.9}$$

The tensorization $\boldsymbol{f} = T_{b,d} f$ of a function $f \in L^p([0,1))$ admits a representation

$$\boldsymbol{f} = \sum_{j_1 \in I_b} \cdots \sum_{j_d \in I_b} e_{j_1}^p \otimes \ldots \otimes e_{j_d}^p \otimes f_{j_1, \ldots, j_d}^p, \tag{2.10}$$

with $f_{j_1, \ldots, j_d}^p = b^{-d/p} \boldsymbol{f}(j_1, \ldots, j_d, \cdot)$ for $p < \infty$ and $f_{j_1, \ldots, j_d}^\infty = \boldsymbol{f}(j_1, \ldots, j_d, \cdot)$. The crossnorm property implies that

$$\|e_{j_1}^p \otimes \ldots \otimes e_{j_d}^p \otimes f_{j_1, \ldots, j_d}^p\|_p = \|f_{j_1, \ldots, j_d}^p\|_p,$$

so that Theorem 2.15 implies

$$\|f\|_p = \Big(\sum_{(j_1, \ldots, j_d) \in I_b^d} \|f_{j_1, \ldots, j_d}^p\|_p^p \Big)^{1/p} \quad (p < \infty), \quad \|f\|_\infty = \max_{(j_1, \ldots, j_d) \in I_b^d} \|f_{j_1, \ldots, j_d}^\infty\|_p.$$

2.3.2. *Sobolev Spaces.* Now consider functions f in the Sobolev space $W^{k,p} := W^{k,p}([0,1))$, equipped with the (quasi-)norm

$$\|f\|_{W^{k,p}} = (\|f\|_p^p + |f|_{W^{k,p}}^p)^{1/p} \quad (p < \infty), \quad \|f\|_{W^{k,\infty}} = \max\{\|f\|_p, |f|_{W^{k,\infty}}\},$$

where $|f|_{W^{k,p}}$ is a (quasi-)semi-norm defined by

$$|f|_{W^{k,p}} = \|D^k f\|_p,$$

with $D^k f := f^{(k)}$ the k-th weak derivative of f. Since f and its tensorization $\boldsymbol{f} = T_{b,d} f$ are such that $f(x) = \boldsymbol{f}(j_1, \ldots, j_d, b^d x - j)$ for $x \in [b^d j, b^d(j+1))$ and $j = \sum_{k=1}^d b^{d-k} j_k$, we deduce that

$$D^k f(x) = b^{kd} \frac{\partial^k}{\partial y^k} \boldsymbol{f}(j_1, \ldots, j_d, b^d x - j)$$

for $x \in [b^d j, b^d(j+1))$, that means that D^k can be identified with a rank-one operator over $\mathbf{V}_{b,d,L^p}$,

$$T_{b,d} \circ D^k \circ T_{b,d}^{-1} = id_{\{1,\ldots,d\}} \otimes (b^{kd} D^k).$$

We deduce from Theorem 2.15 that for $f \in W^{k,p}$,

$$|f|_{W^{k,p}} = \|D^k f\|_p = \|T_{b,d}(D^k f)\|_p = b^{kd} \|(id_{\{1,\ldots,d\}} \otimes D^k)\boldsymbol{f}\|_p,$$

with

$$(id_{\{1,\ldots,d\}} \otimes D^k)\boldsymbol{f} = \sum_{j \in I_b^d} \delta_{j_1} \otimes \ldots \otimes \delta_{j_d} \otimes D^k \boldsymbol{f}(j_1, \ldots, j_d, \cdot).$$

Then we deduce that if $f \in W^{k,p}$, $\boldsymbol{f} = T_{b,d} f$ is in the algebraic tensor space

$$\mathbf{V}_{b,d,W^{k,p}} := (\mathbb{R}^{I_b})^{\otimes d} \otimes W^{k,p},$$

and

$$|f|_{W^{k,p}} = b^{kd} \| \sum_{j_1 \in I_b} \cdots \sum_{j_d \in I_b} \delta_{j_1} \otimes \ldots \otimes \delta_{j_d} \otimes D^k \boldsymbol{f}(j_1, \ldots, j_d, \cdot)\|_p.$$

This implies that $T_{b,d} W^{k,p} \subset \mathbf{V}_{b,d,W^{k,p}}$ but $T_{b,d}^{-1}(\mathbf{V}_{b,d,W^{k,p}}) \not\subset W^{k,p}$. In fact, $T_{b,d}^{-1}(\mathbf{V}_{b,d,W^{k,p}}) = W^{k,p}(\mathcal{P}_{b,d})$, the broken Sobolev space associated with the partition $\mathcal{P}_{b,d} = \{[b^d j, b^d(j+1)) : 0 \le j \le b^d - 1\}$. From the above considerations, we deduce

Theorem 2.16. *For any $0 < p \le \infty$ and $k \in \mathbb{N}_0$, $T_{b,d}$ is a linear isometry from the broken Sobolev space $W^{k,p}(\mathcal{P}_{b,d})$ to $\mathbf{V}_{b,d,W^{k,p}}$ equipped with the (quasi-)norm*

$$\|\boldsymbol{f}\|_{W^{k,p}} = (\|\boldsymbol{f}\|_p^p + |\boldsymbol{f}|_{W^{k,p}}^p)^{1/p} \quad (p < \infty), \quad \|\boldsymbol{f}\|_{W^{k,\infty}} = \max\{\|\boldsymbol{f}\|_\infty, |\boldsymbol{f}|_{W^{k,\infty}}\},$$

where $|\cdot|_{W^{k,p}}$ is a (quasi-)semi-norm defined by

$$|\boldsymbol{f}|_{W^{k,p}}^p := b^{d(kp-1)} \sum_{(j_1,\ldots,j_d) \in I_b^d} |\boldsymbol{f}(j_1, \ldots, j_d, \cdot)|_{W^{k,p}}^p$$

for $p < \infty$, and

$$|\boldsymbol{f}|_{W^{k,\infty}}^\infty := b^{dk} \max_{(j_1,\ldots,j_d) \in I_b^d} |\boldsymbol{f}(j_1, \ldots, j_d, \cdot)|_{W^{k,\infty}}.$$

2.3.3. *Besov Spaces.* Let $f \in L^p$, $0 < p \le \infty$ and consider the difference operator

$$\Delta_h : L^p([0,1)) \to L^p([0,1-h)),$$
$$\Delta_h[f](\cdot) := f(\cdot + h) - f(\cdot).$$

For $r = 2, 3, \ldots$, the r-th difference is defined as

$$\Delta_h^r := \Delta_h \circ \Delta_h^{r-1},$$

with $\Delta_h^1 := \Delta_h$. *The r-th modulus of smoothness* is defined as

$$\omega_r(f,t)_p := \sup_{0<h\le t} \|\Delta_h^r[f]\|_p, \quad t > 0. \tag{2.11}$$

Definition 2.17 (Besov Spaces)**.** *For parameters $\alpha > 0$ and $0 < p, q \leq \infty$, define $r := \lfloor \alpha \rfloor + 1$ and the Besov (quasi-)semi-norm as*

$$|f|_{B^\alpha_{p,q}} := \begin{cases} \left(\int_0^1 [t^{-\alpha}\omega_r(f,t)_p]^q \frac{\mathrm{d}t}{t}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{0<t\leq 1} t^{-\alpha}\omega_r(f,t)_p, & q = \infty. \end{cases}$$

The Besov (quasi-)norm is defined as

$$\|f\|_{B^\alpha_{p,q}} := \|f\|_p + |f|_{B^\alpha_{p,q}}.$$

The Besov space *is defined as*

$$B^\alpha_{p,q} := \left\{ f \in L^p : \|f\|_{B^\alpha_{p,q}} < \infty \right\}.$$

As in Section 2.3.2, we would like to compare the Besov space $B^\alpha_{p,q}$ with the algebraic tensor space

$$\mathbf{V}_{b,d,B^\alpha_{p,q}} := (\mathbb{R}^{I_b})^{\otimes d} \otimes B^\alpha_{p,q}.$$

First, we briefly elaborate how the Besov (quasi-)semi-norm scales under affine transformations of the interval. I.e., suppose we are given a function $f : [a,b) \to \mathbb{R}$ with $-\infty < a < b < \infty$ and a transformed $\bar{f}$ such that

$$\bar{f} : [\bar{a}, \bar{b}) \to \mathbb{R}, \quad \bar{x} \mapsto x := \frac{b-a}{\bar{b}-\bar{a}}(\bar{x} - \bar{a}) + a \mapsto f(x) = \bar{f}(\bar{x}),$$

for $-\infty < \bar{a} < \bar{b} < \infty$. Then,

$$\Delta^r_{\bar{h}}[\bar{f}] : [\bar{a}, \bar{b} - r\bar{h}) \to \mathbb{R},$$

$$\Delta^r_{\bar{h}}[\bar{f}](\bar{x}) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} \bar{f}(\bar{x} + k\bar{h}) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x + kh) = \Delta^r_h[f](x)$$

with

$$h := \frac{b-a}{\bar{b}-\bar{a}}\bar{h}.$$

For $0 < p < \infty$, we obtain for the modulus of smoothness

$$\omega_r(\bar{f}, \bar{t})^p_p = \frac{\bar{b}-\bar{a}}{b-a}\omega_r(f,t)^p_p, \quad t := \frac{b-a}{\bar{b}-\bar{a}}\bar{t},$$

and for $p = \infty$, $\omega_r(\bar{f}, \bar{t})_\infty = \omega_r(f,t)_\infty$. Finally, for the Besov (quasi-)semi-norm this implies

$$|\bar{f}|_{B^\alpha_{p,q}} = \left(\frac{\bar{b}-\bar{a}}{b-a}\right)^{1/p-\alpha} |f|_{B^\alpha_{p,q}}, \quad 0 < q \leq \infty,\ 0 < p < \infty,$$

$$|\bar{f}|_{B^\alpha_{p,q}} = \left(\frac{\bar{b}-\bar{a}}{b-a}\right)^{-\alpha} |f|_{B^\alpha_{p,q}}, \qquad 0 < q \leq \infty,\ p = \infty.$$

With this scaling at hand, for $p < \infty$, what remains is "adding up" Besov (quasi-)norms of partial evaluations $\boldsymbol{f}(j_1, \ldots, j_d, \cdot)$. The modulus of smoothness from (2.11) is not suitable for this task. Instead, we can use an equivalent measure of smoothness via the *averaged modulus of smoothness* (see [14, §5 of Chapter 6 and §5 of Chapter 12])

$$\mathrm{w}_r(f,t)^p_p := \frac{1}{t}\int_0^t \|\Delta^r_h[f]\|^p_p \,\mathrm{d}h, \quad 0 < p < \infty.$$

With this definition of Besov norm, we can define a (quasi-)semi-norm

$$|f|_{B^\alpha_{p,q}} = \left(\int_0^1 [t^{-\alpha}\mathrm{w}_r(f,t)_p]^q \frac{\mathrm{d}t}{t}\right)^{1/q}, \quad 0 < q < \infty,$$

which is equivalent to the former one and therefore results in the same Besov space $B^\alpha_{p,q}$. Expanding the right-hand-side and interchanging the order of integration allows us to add up the contributions to $|f|_{B^\alpha_{p,q}}$ over the intervals $[b^d j, b^d(j+1))$, provided that $q = p$. However, note that this is not the same as summing over $|\boldsymbol{f}(j_1, \ldots, j_d, \cdot)|_{B^\alpha_{p,q}}$, since the latter necessarily omits the contributions of $|\Delta^r_h[f]|$ across the right boundaries of the intervals $[b^d j, b^d(j+1))$.

Example 2.18. *Consider the function*

$$f(x) := \begin{cases} 1, & 0 \le x \le 1/2, \\ 0, & \text{otherwise.} \end{cases}$$

Take $0 < \alpha < 1$ and $r = 1$ in Definition 2.17. The first difference is then

$$\Delta_h[f](x) = \begin{cases} 1, & 1/2 - h < x \le 1/2, \\ 0, & \text{otherwise.} \end{cases}$$

For $0 < p < \infty$,

$$\|\Delta_h[f]\|_p^p = h,$$

and for $p = \infty$,

$$\|\Delta_h[f]\|_\infty = 1.$$

Thus, for the ordinary modulus of smoothness we obtain

$$\omega_r(f,t)_p = t^{1/p}, \quad 0 < p < \infty,$$
$$\omega_r(f,t)_\infty = 1.$$

Inserting this into Definition 2.17, we see that $f \in B^\alpha_{p,q}$ if and only if $p \neq \infty$ and $0 < \alpha < 1/p$. In this case $0 < |f|_{B^\alpha_{p,q}} < \infty$.

On the other hand, for $b = 2$ and $d = 1$, the partial evaluations of the tensorization $T_{2,1}f$ are the constant functions 0 and 1. Thus, any Besov semi-norm of these partial evaluations is 0 and consequently the sum as well. We see that, unlike in Theorem 2.16, even if a function f has Besov regularity, the Besov norm of f is in general not equivalent to the sum of the Besov norms of partial evaluations.

Proposition 2.19. *Let $0 < p = q \le \infty$ and $\alpha > 0$. Let $B^\alpha_{p,p}$ be equipped with the (quasi-)norm associated with the modulus of smoothness when $p = \infty$ or the averaged modulus of smoothness when $p < \infty$. Then, we equip the tensor space $\mathbf{V}_{b,d,B^\alpha_{p,p}}$ with the (quasi-)norm*

$$\|\boldsymbol{f}\|_{B^\alpha_{p,p}} := (\|\boldsymbol{f}\|_p^p + |\boldsymbol{f}|^p_{B^\alpha_{p,p}})^{1/p} \quad (p < \infty), \quad \|\boldsymbol{f}\|_{B^\alpha_{\infty,\infty}} := \max\{\|\boldsymbol{f}\|_\infty, |\boldsymbol{f}|_{B^\alpha_{\infty,\infty}}\},$$

where $|\cdot|_{B^\alpha_{p,p}}$ is a (quasi-)semi-norm defined by

$$|\boldsymbol{f}|^p_{B^\alpha_{p,p}} := b^{d(\alpha p - 1)} \sum_{(j_1,\ldots,j_d) \in I_b^d} |\boldsymbol{f}(j_1,\ldots,j_d,\cdot)|^p_{B^\alpha_{p,p}},$$

for $p < \infty$, and

$$|\boldsymbol{f}|_{B^\alpha_{\infty,\infty}} := b^{d\alpha} \max_{(j_1,\ldots,j_d)\in I_b^d} |\boldsymbol{f}(j_1,\ldots,j_d,\cdot)|_{B^\alpha_{\infty,\infty}}$$

Then, $T_{b,d}(B^\alpha_{p,p}) \hookrightarrow \mathbf{V}_{b,d,B^\alpha_{p,p}}$ with

$$|f|_{B^\alpha_{p,p}} \ge |T_{b,d}(f)|_{B^\alpha_{p,p}}.$$

2.4. **Tensor Subspaces and Corresponding Function Spaces.** For a linear space of functions $S \subset \mathbb{R}^{[0,1)}$, we define the tensor subspace

$$\mathbf{V}_{b,d,S} := (\mathbb{R}^{I_b})^{\otimes d} \otimes S \subset \mathbf{V}_{b,d},$$

and the corresponding linear subspace of functions in $\mathbb{R}^{[0,1)}$,

$$V_{b,d,S} = T^{-1}_{b,d}(\mathbf{V}_{b,d,S}).$$

In the majority of this work we will be using finite-dimensional subspaces S for approximation. In particular, we will frequently use $S = \mathbb{P}_m$ where $\mathbb{P}_m$ is the space of polynomials of degree up to $m \in \mathbb{N}_0$. In this case we use the shorthand notation

$$V_{b,d,m} := V_{b,d,\mathbb{P}_m}.$$

The tensorization $\boldsymbol{f} = T_{b,d}(f)$ of a function $f \in V_{b,d,S}$ admits a representation (2.2) with functions $\boldsymbol{f}(j_1,\ldots,j_d,\cdot) := f_{j_1,\ldots,j_d}$ in S. For $x = t_{b,d}(j_1,\ldots,j_d,y)$ in the interval $[x_j, x_{j+1})$, with $j = \sum_{k=1}^b j_k b^{d-k}$, we have $f(x) = f_{j_1,\ldots,j_d}(y) = f_{j_1,\ldots,j_d}(b^d x - j)$. Therefore, the functions $f \in V_{b,d,S}$ have restrictions on intervals $[x_j, x_{j+1})$ that are obtained by shifting and scaling functions in S. In particular, the space

$V_{b,d,m}$ corresponds to the space of piecewise polynomials of degree m over the uniform partition of $[0,1)$ with b^d intervals.

For considering functions with variable levels $d \in \mathbb{N}$, we introduce the set

$$V_{b,S} := \bigcup_{d\in\mathbb{N}} V_{b,d,S}.$$

It is straight-forward to see that, in general,

$$V_{b,d,S} \not\subset V_{b,\bar{d},S}$$

for $\bar{d} < d$. E.g., take $S = \mathbb{P}_m$ and let $f \in V_{b,d,m}$ be a piece-wise polynomial but discontinuous function. Then, clearly f does not have to be a polynomial over the intervals

$$[kb^{-\bar{d}}, (k+1)b^{-\bar{d}}), \quad 0 \le k \le b^{-\bar{d}} - 1,$$

for $\bar{d} < d$. The same holds for the other inclusion, as the following example demonstrates.

Example 2.20. *Consider the one-dimensional subspace $S := \operatorname{span}\{\cos(2\pi\cdot)\}$. A function $0 \neq f \in V_{b,d,S}$ is thus a piece-wise cosine. Take for simplicity $b = 2$, $d = 0$ (i.e., $V_{2,0,S} = S$) and $\bar{d} = 1$. Then, $f \notin V_{2,1,S}$ due to $\operatorname{span}\{\cos(2\pi\cdot)\} \not\supset \operatorname{span}\{\cos(\pi\cdot), \cos(\pi+\pi\cdot)\}$, since cosines of different frequencies are linearly independent. The same reasoning can be applied to any $b \ge 2$ and $d, \bar{d} \in \mathbb{N}$ with $d < \bar{d}$.*

This motivates the following definition that is reminiscent of multi-resolution analysis (MRA).

Definition 2.21 (Closed under b-adic dilation)**.** *We say that a linear space S is closed under b-adic dilation if for any $f \in S$ and any $k \in \{0, \dots, b-1\}$,*

$$f(b^{-1}(\cdot + k)) \in S.$$

Lemma 2.22. *If S is closed under b-adic dilation, then for all $f \in S$,*

$$f(b^{-d}(\cdot + k)) \in S$$

for all $d \in \mathbb{N}$ and $k \in \{0, \dots, b^d - 1\}$.

Proof. See Appendix A. □

Important examples of spaces S that satisfy the above property include spaces of polynomials and MRAs. The closedness of S under b-adic dilation implies a hierarchy between spaces $V_{b,d,S}$ with different levels, and provides $V_{b,S}$ with a linear space structure.

Proposition 2.23. *If S is closed under b-adic dilation, then*

$$S := V_{b,0,S} \subset V_{b,1,S} \subset \ldots \subset V_{b,d,S} \subset \ldots.$$

Proof. See Appendix A. □

Proposition 2.24 ($V_{b,S}$ is a linear space)**.** *If S is closed under b-adic dilation, then $V_{b,S}$ is a linear space.*

Proof. See Appendix A. □

If $S \subset L^p([0,1))$, then $V_{b,S}$ is clearly a subspace of $L^p([0,1))$. However, it is not difficult to see that, in general, $V_{b,S}$ is not a closed subspace of $L^p([0,1))$. On the other hand, we have the following density result.

Theorem 2.25 ($V_{b,S}$ dense in L^p)**.** *Let $1 \le p < \infty$. If $S \subset L^p([0,1))$ and S contains the constant function one, then $V_{b,S} = \bigcup_{d\in\mathbb{N}} V_{b,d,S}$ is dense in $L^p([0,1))$.*

Proof. See Appendix A. □

Now we provide bounds for ranks of functions in $V_{b,S}$, directly deduced from (2.7).

Lemma 2.26. *For $f \in V_{b,d,S}$ and any $\beta \subset \{1, \dots, d\}$,*

$$r_{\beta,d}(f) \le \min\{b^{\#\beta}, b^{d-\#\beta} \dim S\}.$$

In particular, for all $1 \le \nu \le d$,

$$r_{\nu,d}(f) \le \min\{b^{\nu}, b^{d-\nu} \dim S\}.$$

In the case where S is closed under b-adic dilation, we can obtain sharper bounds for ranks.

Lemma 2.27. *Let S be closed under b-adic dilation.*

(i) If $f \in S$, then for any $d \in \mathbb{N}$, $f \in V_{b,d,S}$ and we have

$$r_{\nu,d}(f) \leq \min\{b^\nu, \dim S\}, \quad 1 \leq \nu \leq d.$$

(ii) If $f \in V_{b,d,S}$, then for any $\bar{d} \geq d$, $f \in V_{b,\bar{d},S}$ and we have

$$r_{\nu,\bar{d}}(f) = r_{\nu,d}(f) \leq \min\left\{b^\nu, b^{d-\nu} \dim S\right\}, \quad 1 \leq \nu \leq d,$$

$$r_{\nu,\bar{d}}(f) \leq \min\{b^\nu, \dim S\}, \quad d < \nu \leq \bar{d}. \tag{2.12}$$

Proof. See Appendix A. □

Remark 2.28. *Lemma 2.27 shows that the ranks are independent of the representation level of a function φ, so that we will frequently suppress this dependence and simply note $r_{\nu,d}(\varphi) = r_\nu(\varphi)$ for any d such that $\varphi \in V_{b,d,S}$.*

We end this section by introducing projection operators based on local projection. Let $\mathcal{I}_S$ be a linear projection operator from $L^p([0,1))$ to a finite-dimensional space S. Then, define the linear operator $\mathcal{I}_{b,d,S}$ on $L^p([0,1))$ defined for $f \in L^p([0,1))$ by

$$(\mathcal{I}_{b,d,S} f)(b^{-d}(j+\cdot)) = \mathcal{I}_S(f(b^{-d}(j+\cdot))), \quad 0 \leq j < b^d. \tag{2.13}$$

Lemma 2.29 (Local projection)**.** *The operator $\mathcal{I}_{b,d,S}$ is a linear operator from $L^p([0,1))$ to $V_{b,d,S}$ and satisfies*

$$T_{b,d} \circ \mathcal{I}_{b,d,S} \circ T_{b,d}^{-1} = id_{\{1,\ldots,d\}} \otimes \mathcal{I}_S. \tag{2.14}$$

Proof. See Appendix A. □

We now provide a result on the ranks of projections.

Lemma 2.30 (Local projection ranks)**.** *For any $f \in L^p$, $\mathcal{I}_{b,d,S} f \in V_{b,d,S}$ satisfies*

$$r_{\nu,d}(\mathcal{I}_{b,d,S} f) \leq r_{\nu,d}(f), \quad 1 \leq \nu \leq d.$$

Proof. Lemma 2.29 implies that $T_{b,d} \circ \mathcal{I}_{b,d,S} \circ T_{b,d}^{-1}$ is a rank one operator. Since a rank-one operator can not increase β-ranks, we have for all $1 \leq \nu \leq d$

$$r_{\nu,d}(\mathcal{I}_{b,d,S}(f)) = r_\nu(T_{b,d} \circ \mathcal{I}_{b,d,S} \circ T_{b,d}^{-1} \boldsymbol{f}) = r_\nu((id_{\{1,\ldots,d\}} \otimes \mathcal{I}_S)\boldsymbol{f}) \leq r_\nu(\boldsymbol{f}) = r_{\nu,d}(f).$$

□

3. Tensor Networks and Their Approximation spaces

In this section, we begin by describing particular tensor formats, namely tree tensor networks that will constitute our approximation tool. We then briefly review classical approximation spaces (see [14]). We conclude by introducing different measures of complexity of tree tensor networks and analyze the resulting approximation classes.

3.1. **Tree Tensor Networks and The Tensor Train Format.**

Let S be a finite-dimensional subspace of $\mathbb{R}^{[0,1)}$. A tensor format in the tensor space $\mathbf{V}_{b,d,S} = (\mathbb{R}^{I_b})^{\otimes d} \otimes S$ is defined as a set of tensors with β-ranks bounded by some integers r_β, for a certain collection A of subsets $\beta \subset \{1, \ldots, d+1\}$,

$$\mathcal{T}_{\boldsymbol{r}}^A(\mathbf{V}_{b,d,S}) = \{\boldsymbol{f} \in \mathbf{V}_{b,d,S} : r_\beta(\boldsymbol{f}) \leq r_\beta, \beta \in A\}.$$

When A is a dimension partition tree (or a subset of such a tree), the resulting format is called a hierarchical or tree-based tensor format [17, 25]. A tensor $\boldsymbol{f} \in \mathcal{T}_{\boldsymbol{r}}^A(\mathbf{V}_{b,d,S})$ in a tree-based tensor format admits a parametrization in terms of a collection of low-order tensors v_β, $\beta \in A$. Hence, the interpretation as a tree tensor network (see [37, Section 4]).

Remark 3.1. *Tree tensor networks are convolutional feedforward neural networks with non-linear feature maps, product pooling, a number of layers equal to the depth of the dimension partition tree and a number of neurons equal to the sum of ranks r_β.*

For the most part we will work with the tensor train format with the exception of a few remarks. This format considers the collection of subsets $A = \{\{1\}, \{1,2\}, \ldots, \{1,\ldots,d\}\}$, which is a subset of a linear dimension partition tree.

Definition 3.2 (Tensor Train Format)**.** *The set*[4] *of tensors in* $\mathbf{V}_{b,d}$ *in tensor train (TT) format with ranks at most* $\boldsymbol{r} := (r_\nu)_{\nu=1}^d$ *is defined as*

$$\mathcal{TT}_{\boldsymbol{r}}(\mathbf{V}_{b,d,S}) := \{\boldsymbol{f} \in \mathbf{V}_{b,d,S} : \; r_\nu(\boldsymbol{f}) \le r_\nu, \; 1 \le \nu \le d\},$$

where we have used the shorthand notation $r_\nu(\boldsymbol{f}) := r_{\{1,\ldots,\nu\}}(\boldsymbol{f})$*. This defines a set of univariate functions*

$$\Phi_{b,d,S,\boldsymbol{r}} = T_{b,d}^{-1}(\mathcal{TT}_{\boldsymbol{r}}(\mathbf{V}_{b,d,S})) = \{f \in V_{b,d,S} : r_\nu(f) \le r_\nu, 1 \le \nu \le d\},$$

where $r_\nu(f) := r_{\nu,d}(f)$*, that we further call the tensor train format for univariate functions.*

Letting $\{\varphi_k\}_{k=1}^{\dim S}$ be a basis of S, a tensor $\boldsymbol{f} \in \mathcal{TT}_{\boldsymbol{r}}(\mathbf{V}_{b,d,S})$ admits a representation

$$\boldsymbol{f}(i_1, \ldots, i_d, y) = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \sum_{k=1}^{\dim S} v_1^{k_1}(i_1) v_2^{k_1,k_2}(i_2) \cdots v_d^{k_{d-1},k_d}(i_d) v_{d+1}^{k_d,k} \varphi_k(y), \tag{3.1}$$

with parameters $v_1 \in \mathbb{R}^{b \times r_1}$, $v_\nu \in \mathbb{R}^{b \times r_{\nu-1} \times r_\nu}$, $2 \le \nu \le d$, and $v_{d+1} \in \mathbb{R}^{r_d \times \dim S}$ forming a tree tensor network

$$\mathbf{v} = (v_1, \ldots, v_{d+1}) \in \mathcal{P}_{b,d,S,\boldsymbol{r}} := \mathbb{R}^{b \times r_1} \times \mathbb{R}^{b \times r_1 \times r_2} \times \ldots \times \mathbb{R}^{b \times r_{d-1} \times r_d} \times \mathbb{R}^{r_d \times \dim S}.$$

The format $\mathcal{TT}_{\boldsymbol{r}}(\mathbf{V}_{b,d,S})$ then corresponds to the image of the space of tree tensor networks $\mathcal{P}_{b,d,S,\boldsymbol{r}}$ through the map

$$R_{b,d,S,\boldsymbol{r}} : \mathcal{P}_{b,d,S,\boldsymbol{r}} \to \mathcal{TT}_{\boldsymbol{r}}(\mathbf{V}_{b,d,S}) \subset \mathbf{V}_{b,d,S}$$

such that for $\mathbf{v} = (v_1, \ldots, v_{d+1}) \in \mathcal{P}_{b,d,S,\boldsymbol{r}}$, the tensor $\boldsymbol{f} = R_{b,d,S,\boldsymbol{r}}(\mathbf{v})$ is defined by (3.1). The set of functions $\Phi_{b,d,S,\boldsymbol{r}}$ in the tensor train format can be parametrized as follows:

$$\Phi_{b,d,S,\boldsymbol{r}} = \{\varphi = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) : \mathbf{v} \in \mathcal{P}_{b,d,S,\boldsymbol{r}}\}, \quad \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) = T_{b,d}^{-1} \circ R_{b,d,S,\boldsymbol{r}}.$$

With an abuse of terminology, we call tensor networks both the set of tensors $\mathbf{v}$ and the corresponding function $\varphi = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v})$. The representation complexity of $\boldsymbol{f} = R_{b,d,S,\boldsymbol{r}}(\mathbf{v}) \in \mathcal{TT}_{\boldsymbol{r}}(V_{b,d,S})$ is

$$\mathcal{C}(b, d, S, \boldsymbol{r}) := \dim(\mathcal{P}_{b,d,S,\boldsymbol{r}}) = br_1 + b\sum_{\nu=2}^{d} r_{\nu-1} r_\nu + r_d \dim S. \tag{3.2}$$

Remark 3.3 (Re-Ordering Variables in the TT Format)**.** *We chose in Definition 2.2 to order the input variables of the tensorized function* $\boldsymbol{f}$ *such that* $y \in [0,1)$ *is in the last position. This specific choice allows the interpretation of partial evaluations of* $\{1,\ldots,\nu\}$*-unfoldings as contiguous pieces of* $f = T_{b,d}^{-1}(\boldsymbol{f})$ *(see Lemma 2.12 and the discussion thereafter). Alternatively, we could have chosen the ordering* $(y, i_1, \ldots, i_d) \mapsto \boldsymbol{f}(y, i_1, \ldots, i_d)$*, and defined the TT-format and TT-ranks correspondingly. Essentially this is the same as considering a different tensor format, see discussion above. Many of the results of part II [2] remain the same. In particular, the order of magnitude of the rank bounds and therefore the resulting direct and inverse estimates would not change. However, this re-ordering may lead to slightly smaller rank bounds as in Remark 3.9 or slightly larger rank bounds as in [2, Remark 4.6].*

3.2. **General Approximation Spaces.** Approximation spaces have been extensively studied in the second half of the last century. They provide a systematic way of classifying functions that can be approximated with a certain rate. Moreover, they have intriguing connections to smoothness and interpolation spaces (see [13]) and thus provide a complete characterization of approximation properties. We briefly review some fundamentals that we require for the rest of this work. For details we refer to [13, 14].

Let X be a quasi-normed linear space, $\Phi_n \subset X$ subsets of X for $n \in \mathbb{N}_0$ and $\Phi := (\Phi_n)_{n \in \mathbb{N}_0}$ an approximation tool. Define the best approximation error

$$E_n(f) := E(f, \Phi_n) := \inf_{\varphi \in \Phi_n} \|f - \varphi\|_X.$$

With this we define approximation classes as

[4] It is in fact a manifold, see [16, 18, 27].

Definition 3.4 (Approximation Classes). *For any $f \in X$ and $\alpha > 0$, define the quantity*

$$\|f\|_{A_q^\alpha} := \begin{cases} \left(\sum_{n=1}^\infty [n^\alpha E_{n-1}(f)]^q \frac{1}{n}\right)^{1/q}, & 0 < q < \infty, \\ \sup_{n \geq 1} [n^\alpha E_{n-1}(f)], & q = \infty. \end{cases}$$

The approximation classes A_q^α of $\Phi = (\Phi_n)_{n \in \mathbb{N}_0}$ are defined by

$$A_q^\alpha := A_q^\alpha(X) := A_q^\alpha(X, \Phi) := \left\{ f \in X : \|f\|_{A_q^\alpha} < \infty \right\}.$$

These approximation classes have many useful properties if we further assume that Φ_n satisfy the following criteria for any $n \in \mathbb{N}_0$.

(P1) $0 \in \Phi_n$, $\Phi_0 = \{0\}$.
(P2) $\Phi_n \subset \Phi_{n+1}$.
(P3) $a\Phi_n = \Phi_n$ for any $a \in \mathbb{R} \setminus \{0\}$.
(P4) $\Phi_n + \Phi_n \subset \Phi_{cn}$ for some $c := c(\Phi)$.
(P5) $\bigcup_{n \in \mathbb{N}_0} \Phi_n$ is dense in X.
(P6) Φ_n is proximinal in X, i.e. each $f \in X$ has a best approximation in Φ_n.

Additionally, properties (P1) – (P6) will be frequently combined with the so-called *direct* or *Jackson* inequality

$$E_n(f) \leq C n^{-r_{\mathrm{J}}} |f|_Y, \quad \forall f \in Y, \tag{3.3}$$

for a semi-normed vector space Y and some parameter $r_{\mathrm{J}} > 0$, and the *inverse* or *Bernstein* inequality

$$|\varphi|_Y \leq C n^{r_{\mathrm{B}}} \|\varphi\|_X, \quad \forall \varphi \in \Phi_n, \tag{3.4}$$

for some parameter $r_{\mathrm{B}} > 0$.

The implications of (P1) – (P6) about the properties of A_q^α are as follows

- (P1)+(P3)+(P4) $\Rightarrow$ A_q^α is a linear space with a quasi-norm.
- (P1)+(P3)+(P4) $\Rightarrow$ A_q^α satisfies the *direct* or *Jackson* inequality

  $$E_n(f) \leq C n^{-r_{\mathrm{J}}} \|f\|_{A_q^\alpha}, \quad \forall f \in A_q^\alpha,$$

  for $r_{\mathrm{J}} = \alpha$.
- (P1)+(P2)+(P3)+(P4) $\Rightarrow$ A_q^α satisfies the *inverse* or *Bernstein* inequality

  $$\|\varphi\|_{A_q^\alpha} \leq C n^{r_{\mathrm{B}}} \|\varphi\|_X, \quad \forall \varphi \in \Phi_n,$$

  for $r_{\mathrm{B}} = \alpha$.

The other properties (P5) and (P6) are required for characterizing approximation spaces by interpolation spaces, see [14]. Specifically, (P1) – (P4) together with a Jackson estimate as in (3.3) are required to prove so-called *direct embeddings*: a range of smoothness spaces is continuously embedded into A_q^α. While (P1) – (P6) together with a Bernstein estimate (3.4) are required for *inverse embeddings*: A_q^α is continuously embedded into smoothness spaces. We will see in Part II [2] that, in general, for approximation spaces of tensor networks no inverse estimates[5] are possible, since these spaces are "too large". Therefore, properties (P5) – (P6) are not essential, while (P5) is typically true for any type of reasonable approximation tool[6].

We have the continuous embeddings

$$A_q^\alpha \hookrightarrow A_{\bar{q}}^\beta, \quad \text{if } \alpha > \beta \quad \text{or if } \alpha = \beta \text{ and } q \leq \bar{q}.$$

We will see that, while most properties are easy to satisfy, property (P4) will be the most critical one. In essence (P4) is a restriction on the non-linearity of the sets Φ_n, with $c(\Phi) = 1$ being satisfied by linear subspaces.

[5]The same was shown for RePU networks in [23].
[6]Think of *universality theorems* for neural networks which hold for tensor networks as well as we will see shortly.

3.3. **Measures of Complexity.** We consider as an approximation tool Φ the collection of tensor networks $\Phi_{b,d,S,\boldsymbol{r}}$ associated with different levels and ranks,

$$\Phi := (\Phi_{b,d,S,\boldsymbol{r}})_{d\in\mathbb{N},\boldsymbol{r}\in\mathbb{N}^d},$$

and define the sets of functions Φ_n as

$$\Phi_n := \left\{\varphi \in \Phi_{b,d,S,\boldsymbol{r}} : d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d, \mathrm{compl}(\varphi) \le n\right\}, \tag{3.5}$$

where $\mathrm{compl}(\varphi)$ is some measure of complexity of a function φ. The approximation classes of tensor networks depend on the chosen measure of complexity. We propose different measures of complexity and discuss the critical property (P4). A function $\varphi \in \Phi$ may admit representations at different levels. We set

$$d(\varphi) = \min\{d : \varphi \in V_{b,d,S}\}$$

to be the minimal representation level of φ, and $\boldsymbol{r}(\varphi) = (r_\nu(\varphi))_{\nu=1}^{d(\varphi)}$ be the corresponding ranks. Measures of complexity may be based on a measure of complexity $\mathrm{compl}(\mathbf{v})$ of tensor networks $\mathbf{v}$ such that $\varphi = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v})$. Then, we would define

$$\Phi_n := \left\{\varphi = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) \in \Phi_{b,d,S,\boldsymbol{r}} : \mathbf{v} \in \mathcal{P}_{b,d,S,\boldsymbol{r}}, d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d, \mathrm{compl}(\mathbf{v}) \le n\right\}, \tag{3.6}$$

which is equivalent to the definition (3.5) if we let

$$\mathrm{compl}(\varphi) := \min\{\mathrm{compl}(\mathbf{v}) : \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) = \varphi, d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d\}, \tag{3.7}$$

where the minimum is taken over all possible representations of φ.

3.3.1. *Complexity Measure: Maximum Rank.* In many high-dimensional approximation problems it is common to consider the maximum rank as an indicator of complexity (see, e.g., [4]). By this analogy we consider for $\varphi \in \Phi$,

$$\mathrm{compl}(\varphi) := bdr_{\max}^2(\varphi) + r_{\max}(\varphi)\dim S, \quad r_{\max}(\varphi) = \max\{r_\nu(\varphi) : 1 \le \nu \le d(\varphi)\}. \tag{3.8}$$

This complexity measure does not satisfy (P4).

Proposition 3.5 ((P4) not satisfied by the complexity measure based on $r_{\max}$)**.** *Let S be closed under b-adic dilation and assume $\dim S < \infty$. Then, with Φ_n as defined in (3.5) with the measure of complexity (3.8),*

(i) There exists no constant $c \in \mathbb{R}$ such that

$$\Phi_n + \Phi_n \subset \Phi_{cn}.$$

(ii) There exists a constant $c > 1$ such that

$$\Phi_n + \Phi_n \subset \Phi_{cn^2}.$$

Proof. See Appendix B. □

3.3.2. *Complexity Measure: Sum of Ranks.* For a neural network, a natural measure of complexity is the number of neurons. By analogy (see Remark 3.1), we can define a complexity measure equal to the sum of ranks

$$\mathrm{compl}(\varphi) := \mathrm{compl}_{\mathcal{N}}(\varphi) := \sum_{\nu=1}^{d(\varphi)} r_\nu(\varphi), \tag{3.9}$$

and the corresponding set

$$\Phi_n^{\mathcal{N}} := \left\{\varphi \in \Phi_{b,d,S,\boldsymbol{r}} : d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d, \mathrm{compl}_{\mathcal{N}}(\varphi) \le n\right\}. \tag{3.10}$$

This complexity measure can be equivalently defined by (3.7) with $\mathrm{compl}_{\mathcal{N}}(\mathbf{v}) = \sum_{\nu=1}^{d} r_\nu$ for $\mathbf{v} \in \mathcal{P}_{b,d,S,\boldsymbol{r}}$.

Lemma 3.6 ($\Phi_n^{\mathcal{N}}$ satisfies (P4))**.** *Let S be closed under b-adic dilation and $\dim S < \infty$. Then, the set $\Phi_n^{\mathcal{N}}$ as defined in (3.10) satisfies (P4) with $c = 2 + \dim S$.*

Proof. See Appendix B. □

3.3.3. *Complexity Measure: Representation Complexity.* A straight-forward choice for the complexity measure is the number of parameters required for representing φ as in (3.2), i.e.,

$$\operatorname{compl}(\varphi) := \operatorname{compl}_{\mathcal{C}}(\varphi) := \mathcal{C}(b, d(\varphi), S, \boldsymbol{r}(\varphi)) = br_1(\varphi) + b\sum_{k=2}^{d(\varphi)} r_{k-1}(\varphi)r_k(\varphi) + r_d(\varphi)\dim S, \tag{3.11}$$

and the corresponding set is defined as

$$\Phi_n^{\mathcal{C}} := \left\{\varphi \in \Phi_{b,d,S,\boldsymbol{r}} : d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d, \operatorname{compl}_{\mathcal{C}}(\varphi) \leq n\right\}. \tag{3.12}$$

This complexity measure can be equivalently defined by (3.7) with $\operatorname{compl}_{\mathcal{C}}(\mathbf{v}) = \mathcal{C}(b, d, S, \boldsymbol{r})$ for $\mathbf{v} \in \mathcal{P}_{b,d,S,\boldsymbol{r}}$.

Remark 3.7. *When interpreting tensor networks as neural networks (see Remark 3.1), the complexity measure* $\operatorname{compl}_{\mathcal{C}}$ *is equivalent to the number of weights for a fully connected neural network with r_ν neurons in layer ν.*

We can show the set $\Phi_n^{\mathcal{C}}$ satisfies (P4) with the help of Lemmas 2.11 and 2.27.

Lemma 3.8 ($\Phi_n^{\mathcal{C}}$ satisfies (P4))**.** *Let S be closed under b-adic dilation and* $\dim S < \infty$*. Then, the set $\Phi_n^{\mathcal{C}}$ as defined in* (3.12) *satisfies (P4) with $c = c(b, \dim S) > 1$.*

Proof. See Appendix B. □

Remark 3.9 (Re-Ordering Input Variables)**.** *In the proof, we have used the property* (2.12) *from Lemma 2.11. As mentioned in Remark 3.3, we could consider a different ordering of the input variables $(y, i_1, \ldots, i_d) \mapsto \boldsymbol{f}(y, i_1, \ldots, i_d)$, and the corresponding TT-format. This would change* (2.12) *to*

$$r_{\nu,\bar{d}}(f) = 1, \quad d + 1 \leq \nu \leq \bar{d}.$$

We still require S to be closed under b-adic dilation to ensure $f \in V_{b,\bar{d},S}$.

Remark 3.10 (ℓ^2-norm of Ranks)**.** *We also considered defining the complexity measure as a ℓ^2-norm of the tuple of ranks*

$$\operatorname{compl}(\varphi) := b\sum_{k=1}^{d(\varphi)} r_k(\varphi)^2 + r_d(\varphi)\dim S.$$

This definition satisfies (P4) as well with analogous results as for the complexity measure $\operatorname{compl}_{\mathcal{C}}$ *for direct and inverse embeddings. The ℓ^2-norm of ranks is less sensitive to rank-anisotropy than the representation complexity* $\operatorname{compl}_{\mathcal{C}}(\varphi)$*. Note that both complexity measures reflect the cost of representing a function with tensor networks, not the cost of performing arithmetic operations, where frequently an additional power of r is required (e.g., $\sim r^3$ or higher).*

3.3.4. *Complexity Measure: Sparse Representation Complexity.* Finally, for a function $\varphi = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) \in \Phi_{b,d,S,\boldsymbol{r}}$, we consider a complexity measure that takes into account the sparsity of the tensors $\mathbf{v} = (v_1, \ldots, v_{d+1})$,

$$\operatorname{compl}(\mathbf{v}) = \operatorname{compl}_{\mathcal{S}}(\mathbf{v}) := \sum_{\nu=1}^{d+1} \|v_\nu\|_{\ell_0}, \tag{3.13}$$

where $\|v_\nu\|_{\ell_0}$ is the number of non-zero entries in the tensor v_ν. By analogy with neural networks (see Remark 3.1), this corresponds to the number of non-zero weights for sparsely connected neural networks. We define the corresponding set as

$$\Phi_n^{\mathcal{S}} := \left\{\varphi \in \Phi_{b,d,S,\boldsymbol{r}} : d \in \mathbb{N}, \boldsymbol{r} \in \mathbb{N}^d, \operatorname{compl}_{\mathcal{S}}(\varphi) \leq n\right\}. \tag{3.14}$$

We can show the set $\Phi_n^{\mathcal{S}}$ satisfies (P4). For that, we need the following two lemmas.

Lemma 3.11. *Assume S is closed under b-adic dilation and* $\dim S < \infty$*. Let $\varphi = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) \in \Phi_{b,d,S,\boldsymbol{r}}$ with $\boldsymbol{r} = (r_\nu)_{\nu=1}^d$. For $\bar{d} > d$, there exists a representation $\varphi = \mathcal{R}_{b,\bar{d},S,\overline{\boldsymbol{r}}}(\overline{\mathbf{v}}) \in \Phi_{b,\bar{d},S,\overline{\boldsymbol{r}}}$ with $\overline{\boldsymbol{r}} = (\bar{r}_\nu)_{\nu=1}^{\bar{d}}$ such that $\bar{r}_\nu = r_\nu$ for $1 \leq \nu \leq d$ and $\bar{r}_\nu \leq \max\{\dim S, b\}\dim S$ for $d < \nu \leq \bar{d}$, and*

$$\operatorname{compl}_{\mathcal{S}}(\overline{\mathbf{v}}) \leq b\operatorname{compl}_{\mathcal{S}}(\mathbf{v}) + (\bar{d} - d)b^2(\dim S)^3.$$

Proof. See Appendix B. □

Lemma 3.12 (Sum of Sparse Representations). *Let* $\varphi_A = \mathcal{R}_{b,d,S,\boldsymbol{r}^A}(\mathbf{v}_A) \in \Phi_{b,d,S,\boldsymbol{r}^A}$ *and* $\varphi_B = \mathcal{R}_{b,d,S,\boldsymbol{r}^B}(\mathbf{v}_B) \in \Phi_{b,d,S,\boldsymbol{r}^B}$. *Then,* $\varphi_A + \varphi_B$ *admits a representation* $\varphi_A + \varphi_B = \mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v}) \in \Phi_{b,d,S,\boldsymbol{r}}$ *with* $r_\nu = r_\nu^A + r_\nu^B$ *for* $1 \leq \nu \leq d$, *and*

$$\mathrm{compl}_{\mathcal{S}}(\varphi_A + \varphi_B) \leq \mathrm{compl}_{\mathcal{S}}(\mathbf{v}) \leq \mathrm{compl}_{\mathcal{S}}(\mathbf{v}_A) + \mathrm{compl}_{\mathcal{S}}(\mathbf{v}_B).$$

Proof. See Appendix B. □

Lemma 3.13 ($\Phi_n^{\mathcal{S}}$ satisfies (P4)). *Let* S *be closed under* b*-adic dilation and* $\dim S < \infty$. *Then, the set* $\Phi_n^{\mathcal{S}}$ *as defined in* (3.14) *satisfies (P4) with* $c = b + 1 + b^2(\dim S)^3$.

Proof. Let $\varphi_A,\ \varphi_B \in \Phi_n^{\mathcal{S}}$ with $\varphi_A = \mathcal{R}_{b,d_A,S,\boldsymbol{r}^A}(\mathbf{v}_A) \in \Phi_{b,d_A,S,\boldsymbol{r}^A}$ and $\varphi_B = \mathcal{R}_{b,d_B,S,\boldsymbol{r}^B}(\mathbf{v}_B) \in \Phi_{b,d_B,S,\boldsymbol{r}^B}$ and w.l.o.g. $d_A \leq d_B$. From Lemmas 3.11 and 3.12, we know that $\varphi_A + \varphi_B$ admits a representation $\varphi_A + \varphi_B = \mathcal{R}_{b,d_B,S,\boldsymbol{r}}(\mathbf{v})$ at level d_B with

$$\begin{aligned}\mathrm{compl}_{\mathcal{S}}(\mathbf{v}) &\leq b\,\mathrm{compl}_{\mathcal{S}}(\mathbf{v}_A) + \mathrm{compl}_{\mathcal{S}}(\mathbf{v}_B) + (d_B - d_A)b^2(\dim S)^3\\ &\leq (b + 1 + b^2(\dim S)^3)n,\end{aligned}$$

which ends the proof. □

3.3.5. *Necessity of (P4).* We could consider replacing n with n^2 in (P4), i.e.,

$$\Phi_n + \Phi_n \subset \Phi_{cn^2}. \tag{3.15}$$

This implies that A_q^α as defined in Definition 3.4 is no longer a vector space. The statements about Jackson and Bernstein inequalities as well as the relation to interpolation and smoothness spaces is no longer valid as well.

One could try to recover the linearity of A_q^α by modifying Definition 3.4. In Definition 3.4 we measure *algebraic* decay of $E_n(f)$. Algebraic decay is compatible with (P4) that in turn ensures A_q^α is a vector space. We could reverse this by asking: what type of decay behavior is "compatible" with (3.15) in the sense that the corresponding approximation class would be a linear space? We can introduce a growth function $\gamma : \mathbb{N} \to \mathbb{R}^+$ with $\lim_n \gamma(n) = \infty$ and define an approximation class A_∞^γ of $\Phi = (\Phi_n)_{n\in\mathbb{N}_0}$ as

$$A_\infty^\gamma := \left\{ f \in X :\ \sup_{n\geq 1} \gamma(n) E_{n-1}(f) < \infty \right\}.$$

With some elementary computations one can deduce that if the growth function is of the form

$$\gamma(n) := 1 + \ln(n),$$

then (3.15) implies A_∞^γ is closed under addition. However, functions in A_∞^γ have too slowly decaying errors for any practical purposes such that we do not intend to analyze this space further.

We could instead ask what form of (P4) would be compatible with a growth function such as

$$\gamma(n) := \exp(an^\alpha),$$

for some $a > 0$ and $\alpha > 0$, i.e., classes of functions with exponentially decaying errors. In this case we would have to require $c = 1$ in (P4), i.e.,

$$\Phi_n + \Phi_n \subset \Phi_n,$$

in other words, Φ_n is a linear space.

These considerations suggest that preserving (P4) in its original form is necessary to exploit the full potential of classical approximation theory while preserving some flexibility in defining Φ_n. Thus, we only consider definitions of compl$(\cdot)$ that preserve (P4).

3.4. **Approximation Spaces of Tensor Networks.** We denote by $\Phi_n^{\mathcal{N}}$, $\Phi_n^{\mathcal{C}}$ and $\Phi_n^{\mathcal{S}}$ the approximation set Φ_n asscociated with the measures of complexity $\mathrm{compl}_{\mathcal{N}}$, $\mathrm{compl}_{\mathcal{C}}$ and $\mathrm{compl}_{\mathcal{S}}$ respectively. Then, we define three different families of approximation classes

$$N_q^\alpha(X) := A_q^\alpha(X, (\Phi_n^{\mathcal{N}})_{n\in\mathbb{N}}), \tag{3.16}$$

$$C_q^\alpha(X) := A_q^\alpha(X, (\Phi_n^{\mathcal{C}})_{n\in\mathbb{N}}), \tag{3.17}$$

$$S_q^\alpha(X) := A_q^\alpha(X, (\Phi_n^{\mathcal{S}})_{n\in\mathbb{N}}), \tag{3.18}$$

with $\alpha > 0$ and $0 < q \leq \infty$. Below, we will show that these approximation classes are in fact approximation spaces and we will then compare these spaces.

3.4.1. *Approximation Classes are Approximation Spaces.* We proceed with checking if $\Phi_n^{\mathcal{N}}$, $\Phi_n^{\mathcal{C}}$ and $\Phi_n^{\mathcal{S}}$ satisfy properties (P1)–(P6). In particular, satisfying (P1)–(P4) will imply that the corresponding approximation classes are quasi-normed Banach spaces. The only property – other than (P4) – that is not obvious, is (P6). This is addressed in the following Lemma for $\Phi_n^{\mathcal{N}}$ and $\Phi_n^{\mathcal{C}}$.

Lemma 3.14 ($\Phi_n^{\mathcal{N}}$ and $\Phi_n^{\mathcal{C}}$ satisfy (P6))**.** *Let $1 < p < \infty$ and let $S \subset L^p$ be a closed subspace. Then, $\Phi_n^{\mathcal{N}}$ and $\Phi_n^{\mathcal{C}}$ are proximinal in L^p for any $n \in \mathbb{N}$. Moreover, if S is finite-dimensional, the above holds for $1 \leq p \leq \infty$.*

Proof. See Appendix B. □

As the following example shows, we cannot in general guarantee (P6) for $\Phi_n^{\mathcal{S}}$.

Example 3.15. *Suppose $b \geq 3$ and $\dim S \geq 3$. Take two linearly independent vectors $v, w \in \mathbb{R}^b$ and $f, g \in S$. For any $N \in \mathbb{N}$, set*

$$\varphi_N := (w + Nv) \otimes (v + \frac{1}{N} w) \otimes f + v \otimes v \otimes (g - Nf),$$

and

$$\varphi := v \otimes v \otimes g + v \otimes w \otimes f + w \otimes v \otimes f.$$

Then, the following holds (see [24, Proposition 9.10 and Remark 12.4]).

(i) For the canonical tensor rank, we have $r(\varphi_N) = 2$ for any $N \in \mathbb{N}$ and $r(\varphi) = 3$.
(ii) As we will see in Lemma 3.22, $\varphi_N \in \Phi^{\mathcal{S}}_{6b+2\dim S}$ for any $N \in \mathbb{N}$ and $\varphi \in \Phi^{\mathcal{S}}_{9b+3\dim S}$. Moreover, this complexity is minimal for both functions.
(iii) For $N \to \infty$, $\varphi_N \to \varphi$ in any norm.

In other words, $E^{\mathcal{S}}_{6b+2\dim S}(\varphi) = 0$, even though $\varphi \notin \Phi^{\mathcal{S}}_{6b+2\dim S}$.

Remark 3.16. *(P6) is required for showing that the approximation spaces A_q^α are continuously embedded into interpolation spaces, see [14, Chapter 7, Theorem 9.3]. Since inverse embeddings for N_q^α, C_q^α and S_q^α hold only in very restricted cases, property (P6) is not essential for our work and the majority of our results. As a side note, (P6) does not hold for ReLU or RePU networks as was discussed in [23].*

We now derive the main result of this section.

Theorem 3.17 (Properties of $\Phi_n^{\mathcal{N}}$, $\Phi_n^{\mathcal{C}}$ and $\Phi_n^{\mathcal{S}}$)**.** *Let $0 < p \leq \infty$, $S \subset L^p$ be a closed subspace that is also closed under b-adic dilation and $\dim S < \infty$. Then,*

(i) $\Phi_n^{\mathcal{N}}$ and $\Phi_n^{\mathcal{C}}$ and $\Phi_n^{\mathcal{S}}$ satisfy (P1) – (P4).
(ii) If $1 \leq p \leq \infty$, then $\Phi_n^{\mathcal{N}}$, $\Phi_n^{\mathcal{C}}$ additionally satisfy (P6).
(iii) If $1 \leq p < \infty$ and if S contains the constant function one, $\Phi_n^{\mathcal{N}}$, $\Phi_n^{\mathcal{C}}$ and $\Phi_n^{\mathcal{S}}$ additionally satisfy (P5).

Proof. (P1) – (P3) are obvious and (P4) follows from Lemmas 3.6, 3.8 and 3.13, that yields (i). (iii) follows from the fact that

$$\bigcup_{n\in\mathbb{N}} \Phi_n = \bigcup_{d\in\mathbb{N}} \bigcup_{\boldsymbol{r}\in\mathbb{N}^d} \Phi_{b,d,S,\boldsymbol{r}} = \bigcup_{d\in\mathbb{N}} V_{b,d,S} = V_{b,S},$$

and from Theorem 2.25. Finally, (ii) follows from Lemma 3.14. □

Theorem 3.17 (i) implies that the approximation classes $N_q^\alpha(L^p)$, $C_q^\alpha(L^p)$ and $S_q^\alpha(X)$ are quasi-normed vector spaces that satisfy the Jackson and Bernstein inequalities.

3.4.2. *Comparing Approximation Spaces.* For comparing approximation spaces $N_q^\alpha(L^p)$, $C_q^\alpha(L^p)$ and $S_q^\alpha(L^p)$, we first provide some relations between the sets $\Phi_n^{\mathcal{N}}$, $\Phi_n^{\mathcal{C}}$ and $\Phi_n^{\mathcal{S}}$.

Proposition 3.18. *For any $n \in \mathbb{N}$,*

$$\Phi_n^{\mathcal{C}} \subset \Phi_n^{\mathcal{S}} \subset \Phi_n^{\mathcal{N}} \subset \Phi^{\mathcal{C}}_{b\dim S + bn^2}.$$

Proof. See Appendix B. □

From Proposition 3.18, we obtain[7]

[7] Compare to similar results obtained for RePU networks in [23, Section 3.4].

Theorem 3.19. *For any $\alpha > 0$, $0 < p \leq \infty$ and $0 < q \leq \infty$, the classes $N_q^\alpha(L^p)$, $C_q^\alpha(L^p)$ and $S_q^\alpha(L^p)$ satisfy the continuous embeddings*

$$C_q^\alpha(L^p) \hookrightarrow S_q^\alpha(L^p) \hookrightarrow N_q^\alpha(L^p) \hookrightarrow C_q^{\alpha/2}(L^p).$$

3.5. **About The Canonical Tensor Format.** We conclude by comparing tensor networks with the canonical tensor format

$$\mathcal{T}_r(V_{b,d,S}) = \{\boldsymbol{f} \in \mathbf{V}_{b,d,S} : r(\boldsymbol{f}) \leq r\},$$

which is the set of tensors that admit a representation

$$\boldsymbol{f}(i_1, \ldots, i_d, y) = \sum_{k=1}^{r} w_1^k(i_1) \ldots w_d^k(i_d) g_{d+1}^k(y), \quad g_{d+1}^k(y) = \sum_{q=1}^{\dim S} w_{d+1}^{q,k} \varphi_q(y),$$

with $w_\nu \in \mathbb{R}^{b \times r}$ for $1 \leq \nu \leq d$ and $w_{d+1} \in \mathbb{R}^{\dim S \times r}$. The canonical tensor format can be interpreted as a shallow sum-product neural network (or arithmetic circuit), see [12].

We let $R_{b,d,S,r}$ be the map from $(\mathbb{R}^{b\times r})^d \times \mathbb{R}^{\dim S \times r} := \mathcal{P}_{b,d,S,r}$ to $\mathbf{V}_{b,d,S}$ which associates to a set of tensors $(w_1, \ldots, w_{d+1})$ the tensor $\boldsymbol{f} = R_{b,d,S,r}(w_1, \ldots, w_{d+1})$ as defined above. We introduce the sets of functions

$$\Phi_{b,d,S,r} = T_{b,d}^{-1} \mathcal{T}_r(V_{b,d,S}),$$

which can be parametrized as follows:

$$\Phi_{b,d,S,r} = \{\varphi = \mathcal{R}_{b,d,S,r}(\mathbf{w}) : \mathbf{w} \in \mathcal{P}_{b,d,S,r}\}, \quad \mathcal{R}_{b,d,S,r} = T_{b,d}^{-1} \circ R_{b,d,S,r}.$$

For $\varphi \in V_{b,S}$, we let

$$r(\varphi) = \min\{r(\boldsymbol{f}) : \boldsymbol{f} \in \mathbf{V}_{b,d(\varphi),S}, \varphi = T_{b,d}^{-1}(\boldsymbol{f})\}.$$

We introduce as a natural complexity measure the representation complexity

$$\mathrm{compl}_{\mathcal{R}}(\varphi) = bd(\varphi) r(\varphi) + r(\varphi) \dim S,$$

define the sets

$$\Phi_n^{\mathcal{R}} = \{\varphi \in \Phi_{b,d,S,r} : d \in \mathbb{N}, r \in \mathbb{N}, \mathrm{compl}_{\mathcal{R}}(\varphi) \leq n\},$$

and consider the corresponding approximation classes

$$R_q^\alpha(L^p) = A_q^\alpha(L^p, (\Phi_n^{\mathcal{R}})_{n \in \mathbb{N}}),$$

with $\alpha > 0$ and $0 < q \leq \infty$. We start by showing that $\Phi_n^{\mathcal{R}}$ satisfies (P1)-(P3) and $(P5)$ (under some assumptions), but not (P4).

Lemma 3.20 ($\Phi_n^{\mathcal{R}}$ satisfies (P1)-(P3) and $(P5)$)**.** *Let $1 \leq p \leq \infty$ and $S \subset L^p$ be a finite-dimensional space. Then $\Phi_n^{\mathcal{R}}$ satisfies (P1)-(P3). Moreover, if S contains the constant function one, $\Phi_n^{\mathcal{R}}$ satisfies (P5) for $1 \leq p < \infty$.*

Proof. (P1)-(P3) are obvious. (P5) follows from the fact that

$$\bigcup_{n \in \mathbb{N}} \Phi_n^{\mathcal{R}} = \bigcup_{d \in \mathbb{N}} \bigcup_{r \in \mathbb{N}} \Phi_{b,d,S,r} = \bigcup_{d \in \mathbb{N}} V_{b,d,S} = V_{b,S},$$

and from Theorem 2.25. ☐

Lemma 3.21 ($\Phi_n^{\mathcal{R}}$ does not satisfy (P4))**.** *Let $1 \leq p \leq \infty$ and $S \subset L^p$ be a finite-dimensional subspace which is closed under b-adic dilation and such that $r(T_{b,d}(\varphi)) = 1$ for any $\varphi \in S$ and $d \in \mathbb{N}$. Then, $\Phi_n^{\mathcal{R}}$ satisfies*

(i) $\Phi_n^{\mathcal{R}} + \Phi_n^{\mathcal{R}} \subset \Phi_{3n^2}^{\mathcal{R}}$,

(ii) there exists no constant $c > 1$ such that $\Phi_n^{\mathcal{R}} + \Phi_n^{\mathcal{R}} \subset \Phi_{cn}^{\mathcal{R}}$.

Proof. See Appendix B. ☐

Lemma 3.22. *For any $n \in \mathbb{N}$, it holds $\Phi_n^{\mathcal{R}} \subset \Phi_n^{\mathcal{S}}$.*

Proof. See Appendix B. ☐

Corollary 3.23. *For any $\alpha > 0$ and $0 < q \leq \infty$,*

$$R_q^\alpha(L^p) \subset S_q^\alpha(L^p).$$

4. Tensor Networks as Neural Networks – The Role of Tensorization

The tensorization of functions is a milestone allowing the use of tensor networks for the approximation of multivariate functions. In this section, we interpret tensorization as a non-standard and powerful featuring step which can be encoded in a neural network with a non-classical architecture. Then, we discuss the role of this particular featuring.

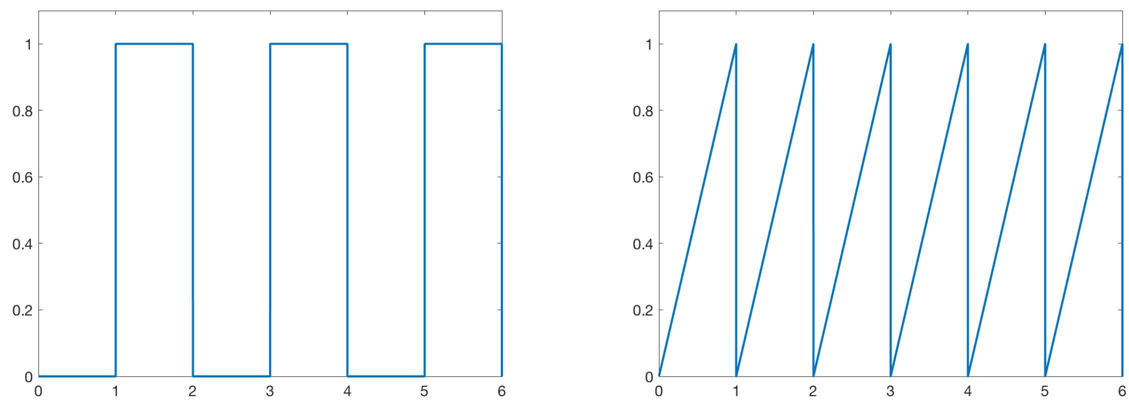
4.1. **Tensorization as Featuring.** When applying $t_{b,d}^{-1}$ to the input variable x, we create $d+1$ new variables $(i_1, ..., i_d, y)$ defined by

$$i_\nu = \sigma(b^\nu x), \quad \sigma(t) = \lfloor t \rfloor \, mod \, b,$$

$1 \leq \nu \leq d$, and

$$y = \tilde{\sigma}(b^d x), \quad \tilde{\sigma}(t) = t - \lfloor t \rfloor,$$

see Figure 4 for a graphical representation of functions σ and $\tilde{\sigma}$. Then for each $1 \leq \nu \leq d$, we create



(A) σ (B) $\tilde{\sigma}$

Figure 4. Functions σ and $\tilde{\sigma}$

b features $\delta_{j_\nu}(i_\nu)$, $0 \leq j_\nu \leq b-1$, and we also create $m+1$ features $\varphi_k(y) = y^k$ from the variable y (or other features for S different from $\mathbb{P}_m$).[8] Figure 5 provides an illustration of these features and of products of these features. Finally, tensorization can be seen as a featuring step with a featuring map

$$\Phi : [0,1) \to \mathbb{R}^{b^d(m+1)}$$

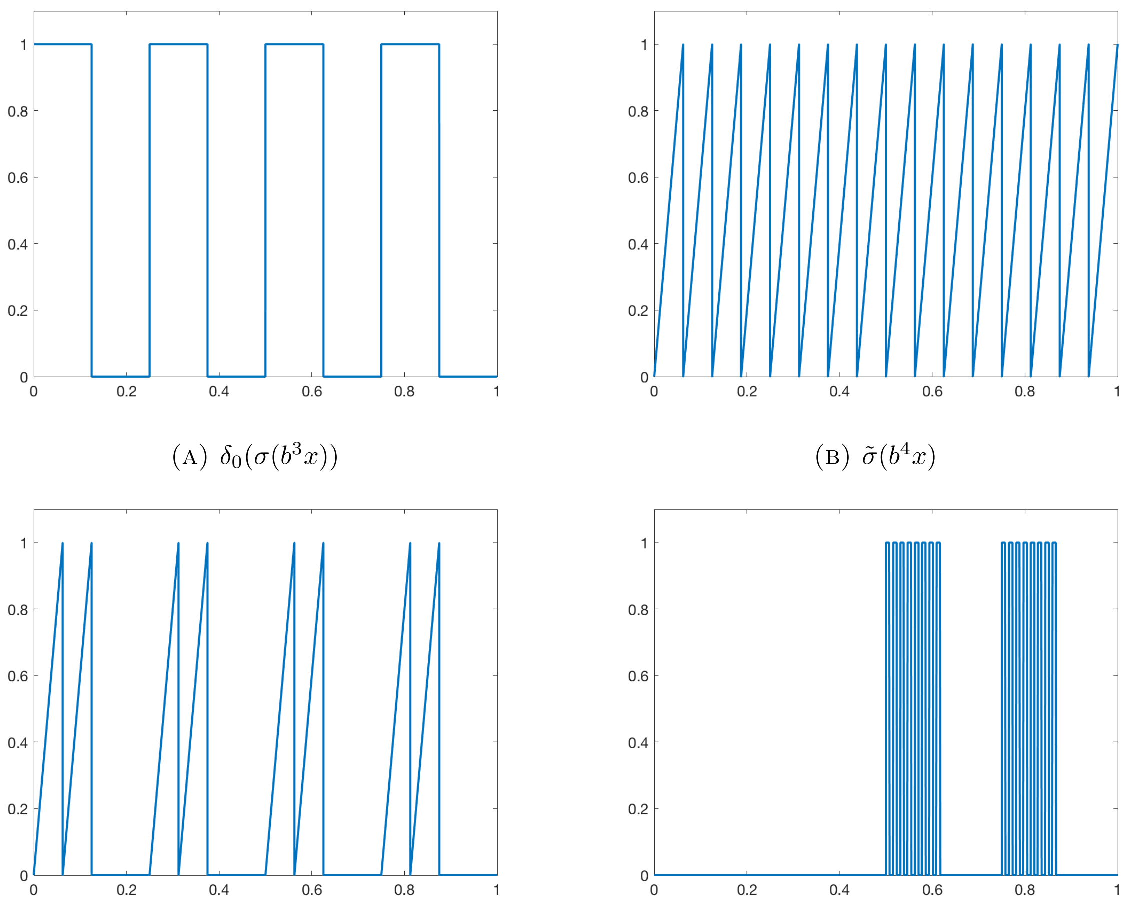
which maps $x \in [0,1)$ to a $(d+1)$-order tensor

$$\Phi(x)_{j_1,\dots,j_{d+1}} = \delta_{j_1}(\sigma(bx)) \dots \delta_{j_d}(\sigma(b^d x)) \tilde{\sigma}(b^d x)^{j_{d+1}}.$$

A function $\varphi \in V_{b,d,m}$ is then represented by $\varphi(x) = \sum_j \Phi(x)_j a_j$, where a is a $(d+1)$-order tensor with entries associated with the $b^d(m+1)$ features. When considering for a a full tensor (not rank-structured), it results in a linear approximation tool which is equivalent to spline approximation. Note that functions represented on Figures 5c and 5d are obtained by summing many features $\Phi(x)_{j_1,\dots,j_{d+1}}$. However, these functions, which have rank-one tensorizations, can be represented with a rank-one tensor a in the feature tensor space, and thus can be encoded with very low complexity within our nonlinear approximation tool.

Increasing d means considering more and more features, and is equivalent to refining the discretisation. At this point, tensorization is an interpretation of a univariate function as a multivariate function, but it is also an alternative way to look at discretization.

[8]For $m = 0$, the extra variable y is not exploited. For $m = 1$, we only consider the variable y and for $m > 1$, we exploit more from this variable.



(A) $\delta_0(\sigma(b^3x))$ (B) $\tilde{\sigma}(b^4x)$

(C) $\delta_0(b^3x)\tilde{\sigma}(b^4x)$ (D) $\delta_1(\sigma(bx))\delta_0(\sigma(b^3x))\delta_0(b^7x)$

FIGURE 5. Representation of some features and their products for $b = 2$.

4.2. **Encoding Tensorization with a Neural Network.** The tensorization step can be encoded as a three-layer feedforward neural network with a single intput x and $b^d(m+1)$ outputs corresponding to the entries of $\Phi(x)$. The variables $(i_1, ..., i_d, y)$ can be seen as the output of a first layer with $d+1$ neurons, which corresponds to the application of a linear map $x \mapsto (b, b^2, \dots, b^d, b^d)x$ to the input variable x, followed by a component-wise application of the activation function σ (for the first d neurons) or $\tilde{\sigma}$ (for the last neuron). Noting that $\delta_{j_\nu}(i_\nu) = \mathbb{1}_{[0,1)}(i_\nu - j_\nu)$, the variable $\delta_{j_\nu}(i_\nu)$ can be seen as the output of a neuron which applies the activation function $t \mapsto \mathbb{1}_{[0,1)}(t)$ to a shift of the input i_ν. The variables y^k correspond to the ouputs of a classical power unit with input y. Therefore, the resulting variables $\delta_{j_\nu}(i_\nu)$ $(1 \le \nu \le d,\ 0 \le j_\nu \le b-1)$ and y^k $(0 \le k \le m)$ are the outputs of a second layer with $bd + (m+1)$ neurons, using the activation function $t \mapsto \mathbb{1}_{[0,1)}(t)$ (for the first bd neurons) or classical power activation functions $y \mapsto y^k$ for the last $m+1$ neurons. The third layer corresponds to a product pooling layer which creates new variables that are products of $d+1$ variables $\{\delta_{j_1}(i_1), \dots, \delta_{j_d}(i_d), \varphi_{j_{d+1}}(y)\}$.

The approximation tool considered in this work then corresponds to neural networks with three first layers implementing the particular featuring map Φ followed by a tensor network or sum-product network (with recurrent network architecture). Note that, if instead of using a sum-product network we take a simple linear combination of the outputs of the third layer, we end up with a neural network implementing classical splines with degree m and b^d knots.

4.3. **The Role of Tensorization.** Another featuring (which is rather straight-forward) would have consisted in taking new variables (or features) $x_{j,k} = (b^d x - j)^k \mathbb{1}_{I_j}(x) = x^k \mathbb{1}_{[0,1)}(b^d x - j)$, $0 \le k \le m$, $0 \le j < b^d$, where I_j is the interval $[b^d j, b^d(j+1))$. This also leads to $(m+1)b^d$ features. This featuring step can be encoded with a two-layer neural network, the first layer with b^d neurons implementing affine transformations $x \mapsto b^d x - j$ followed by the application of the activation function $t \mapsto t\mathbb{1}_{[0,1)}(t)$, and a second layer with $b^d(m+1)$ neurons that apply classical power activation functions to the outputs of the first layer. When considering a simple linear combination of the outputs of this two-layer neural network, we also end up with classical fixed knot spline approximation.

Both featuring (or tensorization) methods can be encoded with feed-forward neural networks, and both methods lead to a linear feature space corresponding to classical spline approximation. One may ask what is the interest of using the very specific feature map Φ? In fact, the use of the particular feature map Φ, which is related to multi-resolution analysis, allows to further exploit sparsity or low-rankness of the tensor when approximating functions from smoothness spaces, and probably other classes of functions. It is well known that the approximation class of splines of degree $m \geq r-1$ is the Sobolev space $W^{r,p}$. Therefore, whatever the featuring used, the approximation class of the resulting linear approximation tool (taking linear combinations of the features) is the Sobolev space $W^{m+1,p}$. We will see in Part II [2] that (near-)optimal performance is achieved by the proposed approximation tool for a large range of smoothness spaces for any fixed m (including $m=0$), at the price of letting d grow (or equivalently the depth of the tensor networks) to capture higher regularity of functions. When working with a fixed m, exploiting low-rank structures will then be crucial.

This reveals that the power of the approximation tool considered in this work comes from the combination of a particular featuring step (the tensorization step) and the use of tensor networks.

References

[1] M. Ali. Low-rank approximability and entropy area laws for ground states of unbounded hamiltonians, 2019.
[2] M. Ali and A. Nouy. Approximation with tensor networks. Part II: Approximation rates for smoothness classes, 2020.
[3] I. Arad, A. Kitaev, Z. Landau, and U. Vazirani. An area law and sub-exponential algorithm for 1d systems, 2013.
[4] M. Bachmayr and W. Dahmen. Adaptive near-optimal rank tensor approximation for high-dimensional operator equations. *Foundations of Computational Mathematics*, 15(4):839–898, 2015.
[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(6):1423–1472, Apr. 2016.
[6] P. L. Bartlett and M. Anthony. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press, 2009.
[7] C. Bény. Deep learning and the renormalization group, 2013.
[8] H. Boche, R. Calderbank, G. Kutyniok, and J. Vybíral, editors. *Compressed Sensing and its Applications.* Springer International Publishing, 2015.
[9] V. I. Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007.
[10] D. Braess. *Finite Elements.* Cambridge University Press, 2007.
[11] G. Carleo and M. Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, feb 2017.
[12] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In V. Feldman, A. Rakhlin, and O. Shamir, editors, *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, volume 49 of *JMLR Workshop and Conference Proceedings*, pages 698–728. JMLR.org, 2016.
[13] R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, jan 1998.
[14] R. A. DeVore and G. G. Lorentz. *Constructive Approximation.* Springer-Verlag Berlin Heidelberg, 1993.
[15] A. Falcó and W. Hackbusch. On minimal subspaces in tensor representations. *Foundations of Computational Mathematics*, 12:765–803, 2012.
[16] A. Falco, W. Hackbusch, and A. Nouy. Geometric structures in tensor representations (final release), 2015.
[17] A. Falcó, W. Hackbusch, and A. Nouy. Tree-based tensor formats. *SeMA Journal*, Oct 2018.
[18] A. Falcó, W. Hackbusch, and A. Nouy. On the dirac–frenkel variational principle on tensor banach spaces. *Foundations of Computational Mathematics*, 19(1):159–204, Feb 2019.
[19] L. Grasedyck. Hierarchical singular value decomposition of tensors. *SIAM Journal on Matrix Analysis and Applications*, 31(4):2029–2054, 2010.

[20] L. Grasedyck. Polynomial Approximation in Hierarchical Tucker Format by Vector-Tensorization. Technical report, Institut für Geometrie und Prakitsche Mathematik, RWTH Aachen, 2010.

[21] E. Grelier, A. Nouy, and M. Chevreuil. Learning with tree-based tensor formats. 2018.

[22] E. Grelier, A. Nouy, and R. Lebrun. Learning high-dimensional probability distributions using tree tensor networks. *ArXiv e-prints*, 2019.

[23] R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks. 2019.

[24] W. Hackbusch. *Tensor Spaces and Numerical Tensor Calculus*. Springer Berlin Heidelberg, 2012.

[25] W. Hackbusch and S. Kuhn. A New Scheme for the Tensor Representation. *Journal of Fourier analysis and applications*, 15(5):706–722, 2009.

[26] M. B. Hastings. An area law for one-dimensional quantum systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(08):P08024–P08024, aug 2007.

[27] S. Holtz, T. Rohwedder, and R. Schneider. On manifolds of tensors of fixed TT-rank. *Numerische Mathematik*, 120(4):701–731, sep 2011.

[28] A. Hoorfar and M. Hassani. Inequalities on the lambert w functionand hyperpower function. *J. Inequalities in Pure and Applied Math.*, 9(2), 2008.

[29] V. Kazeev and C. Schwab. Quantized tensor-structured finite elements for second-order elliptic PDEs in two dimensions. *Numerische Mathematik*, 138(1):133–190, jul 2017.

[30] Y. Levine, O. Sharir, N. Cohen, and A. Shashua. Quantum entanglement in deep learning architectures. *Phys. Rev. Lett.*, 122:065301, Feb 2019.

[31] Y. Levine, D. Yakira, N. Cohen, and A. Shashua. Deep learning and quantum entanglement: Fundamental connections with implications to network design. In *International Conference on Learning Representations*, 2018.

[32] S. Mallat. *A Wavelet Tour of Signal Processing*. Elsevier LTD, Oxford, 2009.

[33] M. Marcellin and D. Taubman. *JPEG2000 Image Compression Fundamentals, Standards and Practice*. Springer US, 2013.

[34] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4):115–133, dec 1943.

[35] B. Michel and A. Nouy. Learning with tree tensor networks: complexity estimates and model selection, 2020.

[36] A. Nouy. *Low-Rank Methods for High-Dimensional Approximation and Model Order Reduction*, chapter 4. SIAM, Philadelphia, PA, 2017.

[37] A. Nouy. Higher-order principal component analysis for the approximation of tensors in tree-based low-rank formats. *Numerische Mathematik*, 141(3):743–789, Mar 2019.

[38] R. Orús. A practical introduction to tensor networks: Matrix product states and projected entangled pair states. *Annals of Physics*, 349:117–158, oct 2014.

[39] R. Orús. Tensor networks for complex quantum systems. *Nature Reviews Physics*, 1(9):538–550, aug 2019.

[40] I. Oseledets and E. Tyrtyshnikov. Breaking the curse of dimensionality, or how to use svd in many dimensions. *SIAM Journal on Scientific Computing*, 31(5):3744–3759, 2009.

[41] I. V. Oseledets. Approximation of matrices with logarithmic number of parameters. *Doklady Mathematics*, 80(2):653–654, oct 2009.

[42] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690. IEEE, 2011.

[43] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, jan 2015.

[44] M. Schwarz, O. Buerschaper, and J. Eisert. Approximating local observables on projected entangled pair states. *Physical Review A*, 95(6), jun 2017.

Appendix A. Proofs for Section 2

Proof of Lemma 2.6. We have

$$t_{b,\bar d}(i_1,\dots,i_{\bar d},y) = \sum_{k=1}^{\bar d} i_k b^{-k} + b^{-\bar d} y = \sum_{k=1}^{d} i_k b^{-k} + \sum_{k=1}^{\bar d - d} i_{k+d} b^{-k-d} + b^{-\bar d} y = \sum_{k=1}^{d} i_k b^{-k} + b^{-d} z,$$

with $z = \sum_{k=1}^{\bar d-d} i_{k+d} b^{-k} + b^{-(\bar d - d)} y = t_{b,\bar d - d}(i_{d+1},\dots,i_{\bar d},y)$, which proves the first statement. Then consider an elementary tensor $\boldsymbol{v} = v_1 \otimes \dots \otimes v_d \otimes g \in \mathbf{V}_{b,d}$, with $v_k \in \mathbb{R}^{I_b}$ and $g \in \mathbb{R}^{[0,1)}$. We have

$$\begin{aligned} T_{b,\bar d} \circ T_{b,d}^{-1} \boldsymbol{v}(i_1,\dots,i_{\bar d},y) &= \boldsymbol{v}(t_{b,d}^{-1} \circ t_{b,\bar d}(i_1,\dots,i_{\bar d},y)) \\ &= \boldsymbol{v}(i_1,\dots,i_d, t_{b,\bar d-d}(i_{d+1},\dots,i_{\bar d},y)) \\ &= v_1(i_1)\dots v_d(i_d) g(t_{b,\bar d - d}(i_{d+1},\dots,i_{\bar d},y)) \\ &= v_1(i_1)\dots v_d(i_d) T_{b,\bar d-d} g(i_{d+1},\dots,i_{\bar d},y) \\ &= (v_1 \otimes \dots \otimes v_d \otimes (T_{b,\bar d - d} g))(i_{d+1},\dots,i_{\bar d},y), \end{aligned}$$

which proves the second property. The last property simply follows from $T_{b,d} \circ T_{b,\bar d}^{-1} = (T_{b,\bar d} \circ T_{b,d}^{-1})^{-1} = (id_{\{1,\dots,d\}} \otimes T_{b,\bar d - d})^{-1} = id_{\{1,\dots,d\}} \otimes T_{b,\bar d-d}^{-1}$. ☐

Proof of Lemma 2.9. $\boldsymbol{f}$ is identified with a tensor in $\mathbf{V}_\beta \otimes \mathbf{V}_{\beta^c}$ with $\mathbf{V}_\beta \in (\mathbb{R}^{I_b})^{\otimes \# \beta}$ and $\mathbf{V}_{\beta^c} = \mathbf{V}_{b,d-\#\beta}$. We have $U^{\min}_{\beta^c}(\boldsymbol{f}) = \{(\boldsymbol{\varphi}_\beta \otimes id_{\beta^c})\boldsymbol{f} : \boldsymbol{\varphi}_\beta \in (\mathbf{V}_\beta)'\}$ with $(\mathbf{V}_\beta)'$ the algebraic dual of $\mathbf{V}_\beta$ (see [15, Corollary 2.19]). Then for any basis $\{\boldsymbol{\varphi}_\beta^{j_\beta} : j_\beta \in I_b^{\#\beta}\}$ of $(\mathbf{V}_\beta)'$, we have $U^{\min}_{\beta^c}(\boldsymbol{f}) = \operatorname{span}\{(\boldsymbol{\varphi}_\beta^{j_\beta} \otimes id_{\beta^c})\boldsymbol{f} : j_\beta \in I_b^{\#\beta}\}$. We conclude by introducing the particular basis $\boldsymbol{\varphi}_\beta^{j_\beta} = \delta_{j_\beta}$, with $\delta_{j_\beta} = \otimes_{\nu \in \beta} \delta_{j_\nu} \in \mathbb{R}^{I_b^{\#\beta}}$, and by noting that $(\delta_{j_\beta} \otimes id_{\beta^c})\boldsymbol{f} = \boldsymbol{f}(j_\beta, \cdot) \in \mathbf{V}_{\beta^c}$. ☐

Proof of Lemma 2.11. For any set $\beta \subset \{1,\dots,d+1\}$ and any partition $\beta = \gamma \cup \alpha$, the minimal subspaces from Definition 2.8 satisfy the hierarchy property (see [24, Corollary 6.18]) $U^{\min}_\beta(\boldsymbol{f}) \subset U^{\min}_\gamma(\boldsymbol{f}) \otimes U^{\min}_\alpha(\boldsymbol{f})$, from which we deduce that $r_\beta(\boldsymbol{f}) \le r_\gamma(\boldsymbol{f}) r_\alpha(\boldsymbol{f})$. Then for $1 \le \nu \le d-1$, by considering $\gamma = \{1,\dots,\nu\}$ and $\alpha = \{\nu+1\}$, we obtain $r_{\nu+1}(\boldsymbol{f}) \le r_\nu(\boldsymbol{f}) r_{\{\nu+1\}}(\boldsymbol{f})$, where $r_{\{\nu+1\}}(\boldsymbol{f}) = \dim U^{\min}_{\{\nu+1\}}(\boldsymbol{f}) \le b$, which yields the first inequality. By considering $\gamma = \{\nu+1\}$ and $\alpha = \{\nu+2,\dots,d+1\}$, we obtain $r_\nu(\boldsymbol{f}) = r_{\{\nu+1,\dots,d+1\}}(\boldsymbol{f}) \le r_{\{\nu+1\}}(\boldsymbol{f}) r_{\{\nu+2,\dots,d+1\}}(\boldsymbol{f}) = r_{\{\nu+1\}}(\boldsymbol{f}) r_{\nu+1}(\boldsymbol{f}) \le b r_{\nu+1}(\boldsymbol{f})$, that is the second inequality. ☐

Proof of Lemma 2.12. We have from Lemma 2.9 that

$$T_{b,d-\nu}^{-1}(U^{\min}_{\{\nu+1,\dots,d+1\}}(\boldsymbol{f}^d)) = \operatorname{span}\{T_{b,d-\nu}^{-1}(\boldsymbol{f}^d(j_1,\dots,j_\nu,\cdot)) : (j_1,\dots,j_\nu) \in I_b^\nu\},$$

where $\boldsymbol{f}^d(j_1,\dots,j_\nu,\cdot) \in \mathbf{V}_{b,d-\nu}$ is a partial evaluation of $\boldsymbol{f}^d$ along the first ν dimensions. We note that

$$\begin{aligned} T_{b,d-\nu}^{-1}(\boldsymbol{f}^d(j_1,\dots,j_\nu,\cdot)) &= ((id_{\{1,\dots,\nu\}} \otimes T_{b,d-\nu}^{-1})\boldsymbol{f}^d)(j_1,\dots,j_\nu,\cdot) \\ &= (T_{b,\nu} \circ T_{b,d}^{-1} \boldsymbol{f}^d)(j_1,\dots,j_\nu,\cdot) = \boldsymbol{f}^\nu(j_1,\dots,j_\nu,\cdot), \end{aligned}$$

where the second equality results from Lemma 2.6. The result then follows from Lemma 2.9 again. ☐

Proof of Proposition 2.14. Subsets of the form $J \times A$, with A a Borel set of $[0,1)$ and $J = \times_{k=1}^d J_k$ with $J_k \subset I_b$, $1 \le k \le d$, form a generating system of the Borel σ-algebra of $I_b^d \times [0,1)$. The image of such a set $J \times A$ through $t_{b,d}$ is $\bigcup_{j \in J} A_j$, where $A_{j_1,\dots,j_d} = b^{-d}(j + A)$ with $j = \sum_{k=1}^d j_k b^{d-k}$. Then

$$\lambda(t_{b,d}(J \times A)) = \lambda(\bigcup_{j\in J} A_j) = \#J b^{-d} \lambda(A) = \#J_1 \dots \#J_d b^{-d} \lambda(A) = \mu_b(J_1)\dots\mu_b(J_d)\lambda(A) = \mu_{b,d}(J \times A).$$

Then, we conclude on $T_{b,d}$ by noting that it is a linear bijection (Proposition 2.3) which preserves measurability. ☐

Lemma A.1. *Let S be a closed subspace of L^p, $1 \le p \le \infty$. The norm $\|\cdot\|_p$ is a reasonable crossnorm on $(\ell^p(I_b))^{\otimes d} \otimes S$.*

Proof. Let $v_k \in \ell^p(I_b)$, $1 \le k \le d$, and $g \in S$. For $p < \infty$, we have

$$\|v_1 \otimes \dots \otimes v_d \otimes g\|_p^p = \sum_{i_1 \in I_b} \dots \sum_{i_d \in I_b} |v_1(i_1)|^p \dots |v_d(i_d)|^p b^{-d} \int_0^1 |g(y)|^p dy = \|v_1\|_{\ell^p}^p \dots \|v_d\|_{\ell^p}^p \|g\|_p^p,$$

and for $p = \infty$,

$$\|v_1 \otimes \ldots \otimes v_d \otimes g\|_\infty = \max_{i_1 \in I_b} |v_1(i_1)| \ldots \max_{i_d \in I_b} |v_d(i_d)| \operatorname{ess\,sup}_y |g(y)| = \|v_1\|_{\ell^\infty} \ldots \|v_d\|_{\ell^\infty} \|g\|_\infty,$$

which proves that $\|\cdot\|_p$ is a crossnorm. Then, consider the dual norm $\|\varphi\|_p^* = \sup_{\|\boldsymbol{f}\| \le 1} |\varphi(\boldsymbol{f})|$ over the algebraic tensor space $(\ell^p(I_b)^*)^{\otimes d} \otimes S^*$, where V^* stands for the continuous dual of a space V. For $(v, \psi) \in \ell^p(I_b) \times \ell^p(I_b)^*$, we consider the duality pairing $\psi(v) = b^{-1} \sum_{k=0}^{b-1} \psi_k v_k$, such that $\ell^p(I_b)^* = \ell^q(I_b)$ with $1/p + 1/q = 1$. Consider $\phi \in S^*$ and $\varphi_\nu \in \ell^p(I_b)^*$, $1 \le \nu \le d$. To prove that $\|\cdot\|_p$ is a reasonable crossnorm, we have to prove that

$$\|\varphi_1 \otimes \ldots \otimes \varphi_d \otimes \phi\|_p^* \le \|\varphi_1\|_{\ell^q} \ldots \|\varphi_d\|_{\ell^q} \|\phi\|_p^*,$$

with $\|\phi\|_p^* = \sup_{f \in S, \|f\|_p \le 1} \phi(f)$. Let $\varphi = \varphi_1 \otimes \ldots \otimes \varphi_d \in (\ell^p(I_b)^*)^{\otimes d} = \ell^q(I_b^d)$. For $j \in I_b^d$, we let $\delta_j = \delta_{j_1} \otimes \ldots \otimes \delta_{j_d} \in \ell^p(I_b^d)$. Any $\boldsymbol{f} \in V_{b,d,S}$ admits a representation $\boldsymbol{f} = \sum_{j \in I_b^d} \delta_j \otimes g_j$ where $g_j = \boldsymbol{f}(j_1, \ldots, j_d, \cdot) \in L^p$, and

$$|(\varphi_1 \otimes \ldots \otimes \varphi_d \otimes \phi)(\boldsymbol{f})| = |\varphi(\sum_{j \in I_b^d} \delta_j \phi(g_j)| = |\varphi(\mathbf{v})|$$

where $\mathbf{v} \in \ell^p(I_b^d)$ is a tensor with entries $\mathbf{v}(j) = \phi(g_j)$. Also,

$$|\varphi(\mathbf{v})| \le \|\varphi\|_{\ell^p}^* \|\mathbf{v}\|_{\ell^p} \le \|\varphi\|_{\ell^q} \|\phi\|_p^* \|\mathbf{w}\|_{\ell^p},$$

where $\mathbf{w} \in \ell^p(I_b^d)$ is a tensor with entries $\mathbf{w}(j) = \|g_j\|_p = \|\boldsymbol{f}(j_1, \ldots, j_d, \cdot)\|_p$. From Theorem 2.15, we have $\|\mathbf{w}\|_{\ell^\infty} = \max_{j \in I_b^d} \|\boldsymbol{f}(j_1, \ldots, j_d, \cdot)\|_\infty = \|\boldsymbol{f}\|_\infty$, and for $p < \infty$

$$\|\mathbf{w}\|_{\ell^p}^p = b^{-d} \sum_{j \in I_b^d} |\mathbf{w}(j)|^p = b^{-d} \sum_{j \in I_b^d} \|\boldsymbol{f}(j_1, \ldots, j_d, \cdot)\|_p^p = \|\boldsymbol{f}\|_p^p.$$

Therefore, $|(\varphi_1 \otimes \ldots \otimes \varphi_d \otimes \phi)(\boldsymbol{f})| \le \|\varphi\|_{\ell^q} \|\phi\|_p^* \|\boldsymbol{f}\|_p$. We conclude by noting that $\|\cdot\|_{\ell^q}$ is a crossnorm on $\ell^q(I_b^d) = \ell^q(I_b)^{\otimes d}$, so that $\|\varphi\|_{\ell^q(I_b^d)} = \|\varphi_1\|_{\ell^q} \ldots \|\varphi_d\|_{\ell^q}$. □

Proof of Lemma 2.22. By definition, the result is true for $d = 1$. The result is then proved by induction. Assume that for all $f \in S$, $f(b^{-d}(\cdot + k)) \in S$ for all $k \in \{0, \ldots, b^d - 1\}$. Then for $f \in S$, consider the function $f(b^{-d-1}(\cdot + k))$ with $k \in \{0, \ldots, b^{d+1} - 1\}$. We can write $k = bk'' + k'$ for some $k' \in \{0, \ldots, b-1\}$ and $k'' \in \{0, \ldots, b^d - 1\}$. Then for any $x \in [0, 1)$,

$$f(b^{-d-1}(x + k)) = f(b^{-d}(b^{-1}(x + k') + k'')) = g(b^{-1}(x + k'))$$

for some $g \in S$, and $g(b^{-1}(x + k')) = h(x)$ for some $h \in S$. Therefore $f(b^{-d-1}(\cdot + k)) = h(\cdot) \in S$, which ends the proof. □

Proof of Proposition 2.23. For $f \in S$, we have $(T_{b,1} f)(i_1, \cdot) = f(b^{-1}(\cdot + i_1))$. Then from Lemma 2.22, we have $(T_{b,1} f)(i_1, \cdot) \in S$, which implies $f \in V_{b,1,S}$. Now assume $f \in V_{b,d,S}$ for $d \in \mathbb{N}$, i.e. $T_{b,d} f = \boldsymbol{f}^d \in \mathbf{V}_{b,d,S}$. Then $\boldsymbol{f}^d(i_1, \ldots, i_d, \cdot) \in S$ and from Lemma 2.22, $\boldsymbol{f}^d(i_1, \ldots, i_d, b^{-1}(i_{d+1} + \cdot)) \in S$. Then using Lemma 2.6, we have that $\boldsymbol{f}^d(i_1, \ldots, i_d, b^{-1}(i_{d+1} + \cdot)) = f \circ t_{b,d}(i_1, \ldots, i_d, t_{b,1}(i_{d+1}, \cdot)) = f \circ t_{b,d+1}(i_1, \ldots, i_{d+1}, \cdot)$, which implies that $(T_{b,d+1} f)(i_1, \ldots, i_{d+1}, \cdot) \in S$, and therefore $f \in V_{b,d+1,S}$. □

Proof of Proposition 2.24. Since $0 \in V_{b,d,S}$ for any d, we have $0 \in V_{b,S}$. For $f_1, f_2 \in V_{b,S}$, there exists $d_1, d_2 \in \mathbb{N}$ such that $f_1 \in V_{b,d_1,S}$ and $f_2 \in V_{b,d_2,S}$. Letting $d = \max\{d_1, d_2\}$, we have from Proposition 2.23 that $f_1, f_2 \in V_{b,d,S}$, and therefore $cf_1 + f_2 \in V_{b,d,S} \subset V_{b,S}$ for all $c \in \mathbb{R}$, which ends the proof. □

Proof of Theorem 2.25. The set of simple functions over $[0, 1)$ is dense in $L^p([0, 1))$ for $1 \le p < \infty$ (see, e.g., [9, Lemma 4.2.1]). Then, it remains to prove that $V_{b,S}$ is dense in the set of simple functions over $[0, 1)$. Consider a simple function $f = \sum_{i=0}^{n-1} a_i \mathbb{1}_{[x_i, x_{i+1})} \ne 0$, with $0 = x_0 < x_1 < \ldots < x_n = 1$, and $\|f\|_p^p = \sum_{i=0}^{n-1} |a_i|^p (x_{i+1} - x_i)$. Let $x_i^d = b^{-d} \lfloor b^d x_i \rfloor$, $0 \le i \le n$, and consider the function $f_d = \sum_{i=0}^{n-1} a_i \mathbb{1}_{[x_i^d, x_{i+1}^d)}$ which is such that $f_d \in V_{b,d,S}$. Then, noting that $x_0^d = x_0 = 0$ and $x_n^d = x_n = 1$, it holds

$$f - f_d = \sum_{i=0}^{n-1} a_i (\mathbb{1}_{[x_i, x_{i+1})} - \mathbb{1}_{[x_i^d, x_{i+1}^d)}) = \sum_{i=0}^{n-1} a_i (\mathbb{1}_{[x_{i+1}^d, x_{i+1})} - \mathbb{1}_{[x_i^d, x_i)}) = \sum_{i=0}^{n-2} (a_i - a_{i+1}) \mathbb{1}_{[x_{i+1}^d, x_{i+1})}.$$

Then, noting that $0 \leq x_i - x_i^d \leq b^{-d}$ for all $0 < i < n$, we have

$$\|f - f_d\|_p^p = \sum_{i=0}^{n-2} |a_i - a_{i+1}|^p (x_{i+1} - x_{i+1}^d) \leq 2^p b^{-d} \sum_{i=0}^{n-1} |a_i|^p = 2^p b^{-d} \|f\|_p^p \big(\min_{0 \leq i \leq n-1} (x_{i+1} - x_i) \big)^{-1},$$

so that $\|f - f_d\|_p \to 0$ as $d \to \infty$, which ends the proof. $\square$

Proof of Lemma 2.27. (i) If $f \in S$, from Proposition 2.23, we have $f \in V_{b,\nu,S}$ for any ν, so that $r_{\nu,\nu}(f) \leq \dim S$. Then, using Corollary 2.13, we have $r_{\nu,d}(f) = r_{\nu,\nu}(f) \leq \dim S$. The other bound $r_{\nu,d}(f) \leq b^\nu$ results from Lemma 2.26.
(ii) The fact that $f \in V_{b,\bar{d},S}$ follows from Proposition 2.23. Then, from Corollary 2.13, we have that $r_{\nu,\bar{d}}(f) = r_{\nu,\nu}(f)$ for all $1 \leq \nu \leq \bar{d}$. For $1 \leq \nu \leq d$, Corollary 2.13 also implies $r_{\nu,\nu}(f) = r_{\nu,d}(f)$ and we obtain the desired inequality from Lemma 2.26. For $\nu > d$, we note that $r_{\nu,\nu}(f) = \dim U^{\min}_{\{\nu+1\}}(T_{b,\nu} f)$. From Proposition 2.23, we know that $T_{b,\nu} f \in V_{b,\nu,S}$ for $\nu \geq d$, so that $U^{\min}_{\{\nu+1\}}(T_{b,\nu} f) \subset S$ and $r_{\nu,\nu}(f) \leq \dim(S)$. The other bound $r_{\nu,\bar{d}}(f) \leq b^\nu$ results from Lemma 2.26. $\square$

Proof of Lemma 2.29. Let $\{\phi_l\}_{1 \leq l \leq \dim S}$ be a basis of S, such that for $g \in L^p$, $\mathcal{I}_S(g) = \sum_{l=1}^{\dim S} \phi_l \sigma_l(g)$, with σ_l a linear map from L^p to $\mathbb{R}$. For $f \in L^p$, and $x \in [b^{-d} j, b^{-d}(j+1))$,

$$\mathcal{I}_{b,d,S} f(x) = \sum_{l=1}^{\dim S} \phi_l(b^d x - j) \sigma_l(f(b^{-d}(j + \cdot)).$$

We have $f(b^{-d}(j + \cdot) = \boldsymbol{f}(j_1, \ldots, j_d, \cdot)$, with $j = \sum_{k=1}^d b^{d-k} j_k$ and $\boldsymbol{f} = T_{b,d} f$, so that

$$T_{b,d}(\mathcal{I}_{b,d,S} f)(j_1, \ldots, j_d, y) = \sum_{l=1}^{\dim S} \phi_l(y) \sigma_l(\boldsymbol{f}(j_1, \ldots, j_d, \cdot)).$$

For $\boldsymbol{f} = \varphi_1 \otimes \ldots \varphi_d \otimes g$, using the linearity of σ_l, we then have

$$T_{b,d}(\mathcal{I}_{b,d,S}(T_{b,d}^{-1} \boldsymbol{f}))(j_1, \ldots, j_d, y) = \varphi_1(j_1) \ldots \varphi_d(j_d) (\sum_{l=1}^{\dim S} \phi_l(y) \sigma_l(g)) = \varphi_1(j_1) \ldots \varphi_d(j_d) \mathcal{I}_S(g)(y),$$

which proves (2.14). $\square$

Appendix B. Proofs for Section 3

Proof of Proposition 3.5. (ii). Let $\varphi_A,\ \varphi_B \in \Phi_n$ with $\varphi_A \in V_{b,d_A,S}$, $\varphi_B \in V_{b,d_B,S}$ and w.l.o.g. $d_B \geq d_A$. Set $r_A := r_{\max}(\varphi_A)$ and $r_B := r_{\max}(\varphi_B)$. Then,

$$\begin{aligned} \operatorname{compl}(\varphi_A + \varphi_B) &\leq b d_A (\max(r_A, \dim S) + r_B)^2 + (\max(r_A, \dim S) + r_B) \dim S \\ &\leq 2 b d_A r_B^2 + 4 b d_A r_A^2 + 4 b d_A (\dim S)^2 + r_A \dim S + r_B \dim S + (\dim S)^2 \\ &\leq [4 + 4(\dim S)^2 + \dim S] n + 4n^2 \leq [8 + 4(\dim S)^2 + \dim S] n^2. \end{aligned}$$

(i). Let W_0 denote the principal Branch of the Lambert W function. Take $n \in \mathbb{N}$ large enough such that

$$d_A := \left\lfloor \frac{1}{\ln(b)} W_0 \left[n \frac{\ln(b)}{2 \max\{b, \dim S\}} \right] \right\rfloor \geq 2, \quad d_A := \left\lfloor \frac{n - \dim S}{b} \right\rfloor \geq 2$$

Pick a full-rank function $\varphi_A \in V_{b,d_A,S}$ such that $r_A^2 := r_{\max}^2(\varphi_A) = b^{2\left\lfloor \frac{d_A}{2} \right\rfloor}$. Then,

$$\operatorname{compl}(\varphi_A) \leq b d_A b^{d_A} + b^{\frac{d_A}{2}} \dim S \leq 2 \max\{b, \dim S\} d_A b^{d_A} \leq n,$$

by the choice of d_A and the properties of the Lambert W function.

Pick any $\varphi_B \in V_{b,d_B,S}$ with $r_B := r_{\max}(\varphi_B) = 1$ and $d_B = d_A$, so that $\operatorname{compl}(\varphi_B) = b d_B + \dim S \leq n$. Then, $\varphi_A,\ \varphi_B \in \Phi_n$. On the other hand, $r_A \geq r_B$ and from [28] we can estimate the Lambert W function from below as

$$W_0 \left[n \frac{\ln(b)}{2 \max\{b, \dim S\}} \right] \geq \ln \left[n \frac{\ln(b)}{2 \max\{b, \dim S\}} \right] - \ln \ln \left[n \frac{\ln(b)}{2 \max\{b, \dim S\}} \right]$$

Then

$$
\begin{aligned}
\operatorname{compl}(\varphi_A+\varphi_B) &\geq b d_A r_A^2 + r_A \dim S \\
&\geq b\left(\frac{n-\dim S}{b}-1\right)\left(b^{\frac{1}{\ln(b)}W_0\left[n\frac{\ln(b)}{2\max\{b,\dim S\}}\right]-1}\right) \\
&= \left(\frac{n-\dim S}{b}-1\right)\left(n\frac{\ln(b)}{2\max\{b,\dim S\}}\right)\left[\ln\left(n\frac{\ln(b)}{2\max\{b,\dim S\}}\right)\right]^{-1},
\end{aligned}
$$

The leading term in the latter expression is

$$
\frac{\ln(b)}{2b\max\{b,\dim S\}}n^2\left[\ln\left(n\frac{\ln(b)}{2\max\{b,\dim S\}}\right)\right]^{-1}.
$$

This cannot be bounded by cn for any $c>0$ and thus (i) follows. ☐

Proof of Lemma 3.6. Let $\varphi_A,\varphi_B\in\Phi_n^{\mathcal{N}}$ with $d_A:=d(\varphi_A)$, $d_B:=d(\varphi_B)$, $\boldsymbol{r}^A:=\boldsymbol{r}^A(\varphi_A)$, $\boldsymbol{r}^B:=\boldsymbol{r}^B(\varphi_B)$ and w.l.o.g. $d_A\leq d_B$. Then using Lemma 2.27,

$$
\begin{aligned}
\operatorname{compl}_{\mathcal{N}}(\varphi_A+\varphi_B) &\leq \sum_{\nu=1}^{d_B}(r_\nu^A+r_\nu^B) \leq \sum_{\nu=1}^{d_A} r_\nu^A + (d_B-d_A)\dim S + \sum_{\nu=1}^{d_B} r_\nu^B \\
&\leq \operatorname{compl}_{\mathcal{N}}(\varphi_A)+\operatorname{compl}_{\mathcal{N}}(\varphi_B)(1+\dim S) \leq (2+\dim S)n.
\end{aligned}
$$

☐

Proof of Lemma 3.8. Let $\varphi_A,\varphi_B\in\Phi_n^{\mathcal{C}}$ with $d_A:=d(\varphi_A)$, $d_B:=d(\varphi_B)$, $\boldsymbol{r}^A:=\boldsymbol{r}^A(\varphi_A)$, $\boldsymbol{r}^B:=\boldsymbol{r}^B(\varphi_B)$ and w.l.o.g. $d_A\leq d_B$. Then

$$
\begin{aligned}
&\operatorname{compl}_{\mathcal{C}}(\varphi_A+\varphi_B) \leq b(r_1^A+r_1^B)+\sum_{k=2}^{d_B} b(r_{k-1}^A+r_{k-1}^B)(r_k^A+r_k^B)+(r_{d_B}^A+r_{d_B}^B)\dim S \\
&=\underbrace{br_1^A+\sum_{k=2}^{d_A} br_{k-1}^A r_k^A + r_{d_A}^A\dim S}_{N_1}+\underbrace{br_1^B+\sum_{k=2}^{d_B} br_{k-1}^B r_k^B + br_{d_B}^B\dim S}_{N_2}+\underbrace{\sum_{k=1}^{d_A} br_{k-1}^A r_k^B + br_{k-1}^B r_k^A}_{N_3} \\
&\quad+\underbrace{\sum_{k=d_A+1}^{d_B} br_{k-1}^A r_k^A}_{N_4}+\underbrace{(r_{d_B}^A-r_{d_A}^A)\dim S}_{N_5}+\underbrace{\sum_{k=d_A+1}^{d_B} br_{k-1}^A r_k^B + br_{k-1}^B r_k^A}_{N_6}.
\end{aligned}
$$

Since $\varphi_A,\varphi_B\in\Phi_n^{\mathcal{C}}$, we have $N_1=\operatorname{compl}_{\mathcal{C}}(\varphi_A)\leq n$ and $N_2=\operatorname{compl}_{\mathcal{C}}(\varphi_B)\leq n$. Then, using Lemma 2.11, we have

$$
\begin{aligned}
N_3 \leq& b\Big(\sum_{k=2}^{d_A}(r_{k-1}^A)^2\Big)^{1/2}\Big(\sum_{k=2}^{d_A}(r_k^B)^2\Big)^{1/2}+b\Big(\sum_{k=2}^{d_A}(r_{k-1}^B)^2\Big)^{1/2}\Big(\sum_{k=2}^{d_A}(r_k^A)^2\Big)^{1/2} \\
\leq& b\Big(\sum_{k=2}^{d_A} br_{k-1}^A r_k^A\Big)^{1/2}\Big(\sum_{k=2}^{d_A} br_{k-1}^B r_k^B)^2\Big)^{1/2}+b\Big(\sum_{k=2}^{d_A} br_{k-1}^B r_k^B\Big)^{1/2}\Big(\sum_{k=2}^{d_A} br_{k-1}^A r_k^A\Big)^{1/2} \\
\leq& 2b\operatorname{compl}_{\mathcal{C}}(\varphi_A)^{1/2}\operatorname{compl}_{\mathcal{C}}(\varphi_B)^{1/2}\leq 2bn.
\end{aligned}
$$

If $d_A=d_B$, we have $N_4=N_5=N_6=0$. If $d_A<d_B$, using Lemma 2.27, we have

$$
\begin{aligned}
N_4 &\leq (\dim S)^2 b(d_B-d_A)\leq(\dim S)^2\operatorname{compl}_{\mathcal{C}}(\varphi_B)\leq n(\dim S)^2, \\
N_5 &\leq (\dim S)^2\leq(\dim S)\operatorname{compl}_{\mathcal{C}}(\varphi_A)\leq n\dim S, \\
N_6 &\leq (\dim S)\Big(\sum_{k=d_A+1}^{d_B} br_k^B+br_{k-1}^B\Big)\leq 2(\dim S)\operatorname{compl}_{\mathcal{C}}(\varphi_B)\leq 2n\dim S.
\end{aligned}
$$

Thus, putting all together

$$
\operatorname{compl}_{\mathcal{C}}(\varphi_A+\varphi_B)\leq[(\dim S)^2+3\dim S+2b+2]n,
$$

and (P4) is satisfied with $c:=(\dim S)^2+3\dim S+2b+2$. ☐

Proof of Lemma 3.11. We have the representation

$$T_{b,d}(\varphi)(i_1,\dots,i_d,y) = \sum_{k_1=1}^{r_1}\cdots\sum_{k_d=1}^{r_d}\sum_{q=1}^{\dim S} v_1^{k_1}(i_1)\cdots v_d^{k_{d-1},k_d}(i_d)v_{d+1}^{k_d,q}\varphi_q(y).$$

Then, from Lemma 2.6,

$$\begin{aligned}
&T_{b,\bar d}(\varphi)(i_1,\dots,i_{\bar d},y)\\
&= \sum_{k_1=1}^{r_1}\cdots\sum_{k_d=1}^{r_d}\sum_{q=1}^{\dim S} v_1^{k_1}(i_1)\cdots v_d^{k_{d-1},k_d}(i_d)v_{d+1}^{k_d,q}T_{b,\bar d-d}(\varphi_q)(i_{d+1},\dots,i_{\bar d},y)\\
&= \sum_{k_1=1}^{r_1}\cdots\sum_{k_d=1}^{r_d}\sum_{q=1}^{\dim S}\sum_{j_{d+1}=1}^{b} v_1^{k_1}(i_1)\cdots v_d^{k_{d-1},k_d}(i_d)\underbrace{v_{d+1}^{k_d,q}\delta_{j_{d+1}}(i_{d+1})}_{\bar v_{d+1}^{k_d,(q,j_{d+1})}(i_{d+1})}T_{b,\bar d-d}(\varphi_q)(j_{d+1},i_{d+2},\dots,i_{\bar d},y),
\end{aligned}$$

where $\bar v_{d+1}\in\mathbb{R}^{b\times r_d\times(b\dim S)}$. Since S is closed under b-adic dilation, we know from Lemma 2.27 that $r_\nu(\varphi_q)\le\dim S$ for all $\nu\in\mathbb{N}$. Let $l=\bar d-d$ and first assume $l\ge 2$. Then, $T_{b,\bar d-d}(\varphi_q)$ admits a representation

$$\begin{aligned}
&T_{b,\bar d-d}(\varphi_q)(j_{d+1},i_{d+2},\dots,i_{\bar d},y)\\
&= \sum_{\alpha_1=1}^{\dim S}\dots\sum_{\alpha_l=1}^{\dim S}\sum_{p=1}^{\dim S} w_1^{q,\alpha_1}(j_{d+1})w_2^{q,\alpha_1,\alpha_2}(i_{d+2})\dots w_l^{q,\alpha_{l-1},\alpha_l}(i_{\bar d})w_{l+1}^{q,\alpha_l,p}\varphi_p(y)\\
&= \sum_{\alpha_2=1}^{\dim S}\dots\sum_{\alpha_l=1}^{\dim S}\sum_{p=1}^{\dim S} w_{1,2}^{q,\alpha_2}(j_{d+1},i_{d+2})\dots w_l^{q,\alpha_{l-1},\alpha_l}(i_{\bar d})w_{l+1}^{q,\alpha_l,p}\varphi_p(y)
\end{aligned}$$

with $w_{1,2}^{q,\alpha_2}(j_{d+1},i_{d+2})=\sum_{\alpha_1=1}^{\dim S}w_1^{q,\alpha_1}(j_{d+1})w_2^{q,\alpha_1,\alpha_2}(i_{d+2})$. Then,

$$\begin{aligned}
&T_{b,\bar d-d}(\varphi_q)(j_{d+1},i_{d+2},\dots,i_{\bar d},y)\\
&= \sum_{\alpha_2,q_2=1}^{\dim S}\dots\sum_{\alpha_l,q_l=1}^{\dim S}\sum_{p=1}^{\dim S} \underbrace{\delta_{q,q_2}w_{1,2}^{q,\alpha_2}(j_{d+1},i_{d+2})}_{\bar v_{d+2}^{(q,j_{d+1}),(q_2,\alpha_2)}(i_{d+2})}\dots\underbrace{\delta_{q_{l-1},q_l}w_l^{q_{l-1},\alpha_{l-1},\alpha_l}(i_{\bar d})}_{\bar v_{\bar d}^{(q_{l-1},\alpha_{l-1}),(q_l,\alpha_l)}(i_{\bar d})}\underbrace{w_{l+1}^{q_l,\alpha_l,p}}_{\bar v_{\bar d+1}^{(q_l,\alpha_l),p}}\varphi_p(y)
\end{aligned}$$

with $\bar v_{d+2}\in\mathbb{R}^{b\times(b\dim S)\times(\dim S)^2}$, $\bar v_\nu\in\mathbb{R}^{b\times(\dim S)^2\times(\dim S)^2}$ for $d+3\le\nu\le\bar d$, and $\bar v_{\bar d+1}\in\mathbb{R}^{(\dim S)^2\times\dim S}$. Then, we have $\varphi\in\mathcal{R}_{b,\bar d,S,\bar{\boldsymbol r}}(\bar{\mathbf v})$ with $\bar{\mathbf v}=(\bar v_1,\dots,v_d,\bar v_{d+1},\dots,\bar v_{\bar d+1})$, with $\bar v_\nu$ defined above for $\nu>d$, and $\bar r_\nu=r_\nu$ for $\nu\le d$, $\bar r_{d+1}=b\dim S$, and $\bar r_\nu=(\dim S)^2$ for $d+1<\nu\le\bar d$. From the definition of $\bar v_\nu$, we easily deduce that $\|\bar v_{d+1}\|_{\ell_0}=b\|v_{d+1}\|_{\ell_0}$, $\|\bar v_{d+2}\|_{\ell^0}\le b^2(\dim S)^2$, $\|\bar v_\nu\|_{\ell_0}\le b(\dim S)^3$ for $d+3\le\nu\le\bar d$, and $\|\bar v_{\bar d+1}\|_{\ell_0}\le(\dim S)^3$. Then, for $l=\bar d-d\ge 2$, we obtain

$$\mathrm{compl}_{\mathcal S}(\bar{\mathbf v})=\sum_{\nu=1}^{d+1}\|\bar v_\nu\|_{\ell_0}+\sum_{\nu=d+2}^{\bar d+1}\|\bar v_\nu\|_{\ell_0}\le b\,\mathrm{compl}_{\mathcal S}(\mathbf v)+b^2(\dim S)^2+b(\dim S)^3(\bar d-d-2)+(\dim S)^3.$$

For $l=1$, we have a representation

$$T_{b,\bar d-d}(\varphi_q)(j_{d+1},y)=\sum_{p=1}^{\dim S}\bar v_{d+2}^{(q,j_{d+1}),p}\varphi_p(y)$$

with $\bar v_{d+2}\in\mathbb{R}^{(b\dim S)\times\dim S}$ such that $\bar v_{d+2}^{(q,j_{d+1}),p}=\sum_{\alpha_1=1}^{\dim S}w_1^{q,\alpha_1}(j_{d+1})\varphi_2^{q,\alpha_1,p}$. Then, for $l=1$, $\varphi\in\mathcal{R}_{b,\bar d,S,\bar{\boldsymbol r}}(\bar{\mathbf v})$ with $\bar{\mathbf v}=(\bar v_1,\dots,v_d,\bar v_{d+1},\bar v_{\bar d+2})$, and

$$\mathrm{compl}_{\mathcal S}(\bar{\mathbf v})\le b\,\mathrm{compl}_{\mathcal S}(\mathbf v)+b(\dim S)^2=b\,\mathrm{compl}_{\mathcal S}(\mathbf v)+b(\dim S)^2(\bar d-d).$$

For any $l\ge 1$, we then deduce

$$\mathrm{compl}_{\mathcal S}(\bar{\mathbf v})\le b\,\mathrm{compl}_{\mathcal S}(\mathbf v)+b^2(\dim S)^3(\bar d-d).$$

□

Proof of Lemma 3.12. φ_A and φ_B admit representations

$$T_{b,d}(\varphi_C)(i_1,\ldots,i_d,y) = \sum_{k_1=1}^{r_1^C}\cdots\sum_{k_d=1}^{r_d^C}\sum_{q=1}^{\dim S} v_1^{C,k_1}(i_1)\cdots v_d^{C,k_{d-1},k_d}(i_d)v_{d+1}^{C,k_d,q}\varphi_q(y),$$

with $C=A$ or B. Then, $\varphi_A+\varphi_B$ admit the representation

$$T_{b,d}(\varphi_A+\varphi_B)(i_1,\ldots,i_d,y) = \sum_{k_1=1}^{r_1^A+r_1^B}\cdots\sum_{k_d=1}^{r_d^A+r_d^B}\sum_{q=1}^{\dim S} v_1^{k_1}(i_1)\cdots v_d^{k_{d-1},k_d}(i_d)v_{d+1}^{k_d,q}\varphi_q(y),$$

with $v_1^{k_1}=v_1^{A,k_1}$ if $1\le k_1\le r_1^A$ and $v_1^{k_1}=v_1^{B,k_1}$ if $r_1^A<k_1\le r_1^A+r_1^B$,

$$v_\nu^{k_{\nu-1},k_\nu} = \begin{cases} v^{A,k_{\nu-1},k_\nu} & \text{if } 1\le k_{\nu-1},k_\nu\le r_1^A \\ v^{B,k_{\nu-1},k_\nu} & \text{if } r_1^A<k_{\nu-1},k_\nu\le r_1^A+r_1^B \\ 0 & \text{elsewhere,}\end{cases}$$

and $v_{d+1}^{k_d,q}=v_{d+1}^{A,k_d,q}$ if $1\le k_d\le r_1^A$ and $v_{d+1}^{k_d,q}=v_{d+1}^{B,k_d,q}$ if $r_1^A<k_d\le r_1^A+r_1^B$. From the above, we deduce that $\|v^C\|_{\ell_0}\le\|v^A\|_{\ell_0}+\|v^B\|_{\ell_0}$, so that

$$\mathrm{compl}_{\mathcal{S}}(\mathbf{v}) = \sum_{\nu=1}^{d+1}\|v^C\|_{\ell_0} \le \mathrm{compl}_{\mathcal{S}}(\mathbf{v}) \le \mathrm{compl}_{\mathcal{S}}(\mathbf{v}_A)+\mathrm{compl}_{\mathcal{S}}(\mathbf{v}_B).$$

□

Proof of Lemma 3.14. The norm defined in Theorem 2.15 is a reasonable crossnorm (see Lemma A.1) and thus, in particular, not weaker than the injective norm on $V_{b,d,S}$. Thus, by [24, Lemma 8.6] $\mathcal{TT}_{\boldsymbol{r}}(V_{b,d,S})$ for $\boldsymbol{r}\in\mathbb{N}^d$ is a weakly closed subset of L^p. Moreover, the set Φ_n, with either $\Phi_n=\Phi_n^{\mathcal{N}}$ or $\Phi_n^{\mathcal{C}}$, is a finite union of the sets $\mathcal{TT}_{\boldsymbol{r}}(V_{b,d,S})$ for different $d\in\mathbb{N}$ and $\boldsymbol{r}\in\mathbb{N}^d$. Since finite unions of closed sets (in the weak topology) are closed, it follows that Φ_n is weakly closed in L^p, and a fortiori, Φ_n is also closed in the strong topology. Since L^p is reflexive for $1<p<\infty$ and Φ_n is weakly closed, Φ_n is proximinal in L^p (see [24, Theorem 4.28]).

Now consider that S is finite-dimensional. There exists d such that $\Phi_n\subset V_{b,d,S}$ and $V_{b,d,S}$ is finite-dimensional. Since Φ_n is a closed subset of a finite-dimensional space $V_{b,d,S}$, it is proximinal in L^p for any $1\le p\le\infty$. □

Proof of Proposition 3.18. Consider a function $0\ne\varphi\in V_{b,S}$ and let $d=d(\varphi)$ and $\boldsymbol{r}=\boldsymbol{r}(\varphi)$. We have

$$\mathrm{compl}_{\mathcal{N}}(\varphi) = \sum_{\nu=1}^d r_\nu \le br_1+\sum_{r_\nu=2}^d br_{\nu-1}r_\nu + b\dim S = \mathrm{compl}_{\mathcal{C}}(\varphi),$$

which implies $\Phi_n^{\mathcal{C}}\subset\Phi_n^{\mathcal{N}}$. Also

$$\begin{aligned}\mathrm{compl}_{\mathcal{C}}(\varphi) &\le br_1+b(\sum_{\nu=1}^{d-1}r_\nu^2)^{1/2}(\sum_{\nu=2}^{d}r_\nu^2)^{1/2}+b\dim S \le br_1+b(\sum_{\nu=1}^{d-1}r_\nu)(\sum_{\nu=2}^{d}r_\nu)+b\dim S\\ &\le b(\sum_{\nu=1}^d r_\nu)^2+b\dim S = b\,\mathrm{compl}_{\mathcal{N}}(\varphi)^2+b\dim S,\end{aligned}$$

which yields $\Phi_n^{\mathcal{N}}\subset\Phi_{b\dim S+bn^2}^{\mathcal{C}}$. Also, we clearly have $\mathrm{compl}_{\mathcal{S}}(\varphi)\le\mathrm{compl}_{\mathcal{C}}(\varphi)$, which implies $\Phi_n^{\mathcal{C}}\subset\Phi_n^{\mathcal{S}}$. Now consider any tensor network $\mathbf{v}\in\mathcal{P}_{b,d,S,\boldsymbol{r}}$ such that $\varphi=\mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v})$, with $d(\varphi)\le d$ and $\boldsymbol{r}(\varphi)\le\boldsymbol{r}$. We have that $r_1(\varphi)\le\dim\{v_1^{k_1}(\cdot)\in\mathbb{R}^b : 1\le k_1\le r_1\}\le\|v_1\|_{\ell_0}$ and for $2\le\nu\le d$, $r_\nu(\varphi)\le\dim\{v_\nu^{\cdot,k_\nu}(\cdot)\in\mathbb{R}^{b\times r_{\nu-1}} : 1\le k_\nu\le r_\nu\}\le\|v_\nu\|_{\ell^0}$. Therefore

$$\mathrm{compl}_{\mathcal{N}}(\varphi) = \sum_{\nu=1}^d r_\nu(\varphi) \le \sum_{\nu=1}^d \|v_\nu\|_{\ell^0} \le \mathrm{compl}_{\mathcal{S}}(\mathbf{v}).$$

The inequality being true for any tensor network $\mathbf{v}$ such that $\varphi=\mathcal{R}_{b,d,S,\boldsymbol{r}}(\mathbf{v})$, we deduce $\mathrm{compl}_{\mathcal{N}}(\varphi)\le\mathrm{compl}_{\mathcal{S}}(\varphi)$, which yields $\Phi_n^{\mathcal{S}}\subset\Phi_n^{\mathcal{N}}$. □

Proof of Lemma 3.21. (i). Consider $\varphi_A, \varphi_B \in \Phi_n^{\mathcal{R}}$, and let $d_A = d(\varphi_A)$, $d_B = d(\varphi_B)$, $r_A = r(\varphi_A)$ and $r_B = r(\varphi_B)$. Assume w.l.o.g. that $d_A \leq d_B$. The function φ_A admits a representation

$$T_{b,d_A}\varphi_A(i_1, \ldots, i_{d_A}, y) = \sum_{k=1}^{r_A} w_1^{A,k}(i_1) \ldots w_d^{A,k}(i_d) w_{d+1}^{A,k}(y),$$

and

$$T_{b,d_B}\varphi_A(i_1, \ldots, i_{d_B}, y) = \sum_{k=1}^{r_A} w_1^{A,k}(i_1) \ldots w_d^{A,k}(i_d) T_{b,d_B}(w_{d+1}^{A,k})(y).$$

From the assumption on S, we have $T_{b,d_B}(w_{d+1}^{A,k})$ of rank 1, so that $r(T_{b,d_B}\varphi_A) \leq r_A$. We easily deduce that $r(\varphi_A+\varphi_B) \leq r_A+r_B$ and $\mathrm{compl}_{\mathcal{R}}(\varphi_A+\varphi_B) \leq bd_B(r_A+r_B)+(r_A+r_B)b\dim S \leq 2n+br_A(d_B-d_A) \leq 2n + n^2 \leq 3n^2$.
(ii). The proof idea is analogous to Proposition 3.5: we take a rank-one tensor $\varphi_B \in \Phi_n^{\mathcal{R}}$ such that $d_B \sim n$ and a full-rank tensor $\varphi_A \in \Phi_n^{\mathcal{R}}$ with $d_A < d_B$ such that $r_A \sim b^{d_A} \sim n$. Then, as in Proposition 3.5, $\mathrm{compl}_{\mathcal{R}}(\varphi_A + \varphi_B) \sim n^2$. □

Proof of Lemma 3.22. Let $\varphi \in \Phi_n^{\mathcal{R}}$, $d = d(\varphi)$, $r = r(\varphi)$. The function φ admits a representation

$$T_{b,d}\varphi(i_1, \ldots, i_d, y) = \sum_{k=1}^{r} \sum_{q=1}^{\dim S} w_1^k(i_1) \ldots w_d^k(i_d) w_{d+1}^{q,k} \varphi_q(y).$$

Letting $v_1 = w_1$, $v_{d+1} = w_{d+1}$ and $v_\nu \in \mathbb{R}^{b \times r \times r}$ such that $v_\nu^{k_{\nu-1},k_\nu} = \delta_{k_{\nu-1},k_\nu} w_\nu^{k_\nu}$ for $2 \leq \nu \leq d$, and letting $\boldsymbol{r} = (r, \ldots, r) \in \mathbb{N}^d$, we have

$$T_{b,d}\varphi(i_1, \ldots, i_d, y) = \sum_{k_1=1}^{r} \ldots \sum_{k_d=1}^{r} \sum_{q=1}^{\dim S} v_1^k(i_1) \ldots w_d^{k_{d-1},k_d}(i_d) w_{d+1}^{q,k_d} \varphi_q(y),$$

which proves that $T_{b,d}\varphi \in \Phi_{b,d,S,\boldsymbol{r}}$ with

$$\mathrm{compl}_{\mathcal{S}}(\varphi) = \sum_{\nu=1}^{d+1} \|v_\nu\|_{\ell_0} = \sum_{\nu=1}^{d+1} \|w_\nu\|_{\ell_0} \leq brd + r\dim S = \mathrm{compl}_{\mathcal{R}}(\varphi) \leq n,$$

that is $\varphi \in \Phi_n^{\mathcal{S}}$. □