

Article

Explicit Approximation Rates for Shallow Networks on Additive Besov Classes

Zhenliang Chen ¹ and Kai-Cheng Wang ^{2,*} 

¹ Office of the CPC Committee & President's Office, Sanming University, Sanming 365004, China; czl@fj-smu.edu.cn

² Department of Mathematics and Applied Mathematics, School of Information Engineering, Sanming University, No. 25, Jing Dong Road, Sanming 365004, China

* Correspondence: gtotony98@gmail.com

Abstract

We study the approximation of functions by localized shallow neural networks on bounded Lipschitz domains. For a non-polynomial activation, we establish a direct, Jackson-type estimate on the Besov scale $B_{p,q}^s(\Omega)$, the approximation error decaying as $N^{-s/d}$ in the width N with an explicit constant tracked through the activation and the domain. In one spatial variable we prove the corresponding Bernstein-type converse on the nonlinear free-knot Besov scale $B_{\tau,\tau}^\alpha$. For coordinate-subset additive classes we refine the direct estimate to the rate N^{-s/d_0} , in which the effective dimension d_0 governs the exponent while the ambient dimension enters only through the constant. A deterministic construction corroborates the predicted decay rates, and an effective-dimension reading of educational tabular data is consistent with the additive setting.

Keywords: shallow neural networks; Besov spaces; direct and inverse estimates; effective dimension; nonlinear approximation

MSC: 41A25; 41A27; 41A46; 41A63; 46E35; 68T07

1. Introduction

The approximation of functions by shallow neural networks lies at the meeting point of constructive approximation theory and the analysis of learning architectures. For a non-polynomial activation σ , the single-hidden-layer class of N -term ridge combinations is dense in L^p on bounded domains [1,2], and the quantitative question is the rate at which the best approximation error decreases as the width N grows. When the target carries smoothness measured on the Besov scale $B_{p,q}^s$, the governing rate on a domain of dimension d is $N^{-s/d}$, and the exponent degrades as d grows. This dependence of the attainable rate on the ambient dimension is the analytic form of the curse of dimensionality, and it is the structural obstruction that the present paper addresses.

A substantial body of work establishes direct, Jackson-type rates for shallow and deep networks. The variation-space framework of [3–5] yields L^2 rates for ReLU^k activations, the weighted variation spaces of [6] supply the domain-intrinsic formulation on bounded domains, the ridge-spline representer theory of [7] connects single-hidden-layer networks to free-knot spline families, and the Besov and Sobolev rates of [8] place these results within the classical smoothness scales. Here, the variation space is the space of functions represented as $f(x) = \int \sigma(w \cdot x + b) d\mu(w, b)$, normed by the infimum of the total variation



Academic Editor: Jüri Majak

Received: 6 June 2026

Revised: 7 July 2026

Accepted: 9 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

$|\mu|$ over such representations—equivalently, the gauge of the closed symmetric convex hull of the dictionary $\{\sigma(w \cdot x + b)\}$. Across this literature the direct theory is well developed. The converse, Bernstein-type theory, which would characterize the approximation classes of the network families, remains comparatively underdeveloped in dimension $d \geq 2$, and the absence of a usable multivariate inverse inequality is recorded as an open problem in the survey of [9].

A complementary line of work confronts the curse of dimensionality by restricting the structure of the target. When a function depends, up to superposition, on a small number of coordinates at a time, the attainable rate is governed by an effective dimension d_0 that can be considerably smaller than d . In the statistical estimation literature this phenomenon is the minimax theory of composition and additive models [10,11], and approximation-theoretic counterparts under low intrinsic dimension and tractability hypotheses have been developed in [12,13]. The additive structure is the conceptually delimited setting in which the dependence on the ambient dimension separates from the dependence on the effective dimension, and it is the object on which the refinement of this paper is built.

Tabular data arising in applied settings frequently exhibit precisely this structure. In educational prediction the response of interest, such as a final examination grade, is often governed by a small set of dominant factors rather than by high-order interactions among all recorded attributes [14–16]. The methodological premium in this domain is on transparent and feature-parsimonious models whose structure can be inspected, in contrast with the massive parameterization of an opaque predictor [17,18]. A construction whose approximation rate is governed by an effective dimension, and whose units are explicit rather than the endpoint of an opaque optimization, is aligned with this premium, and it is in this sense that the analytic theory developed below speaks to the applied setting.

This paper develops explicit approximation rates for a localized shallow network class on additive Besov classes, together with the converse theory available in one dimension, and validates the resulting rates by a deterministic construction. The first result, Theorem 5, is a direct estimate. For the localized class on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$, every $f \in B_{p,q}^s(\Omega)$ admits the bound $E_N^b(f)_{L^p(\Omega)} \leq C N^{-s/d} |f|_{B_{p,q}^s(\Omega)}$ with a constant whose dependence on the activation and on the Lipschitz character of the domain is tracked explicitly. The second result, Corollary 1, is the converse in dimension $d = 1$. An ℓ^τ summability condition on the dyadic approximation errors places f on the nonlinear free-knot Besov scale $B_{\tau,\tau}^\alpha$, where $\frac{1}{\tau} = \alpha + \frac{1}{p}$. The third result, Theorem 6, is the effective-dimension refinement. On the coordinate-subset additive class $S_{d_0}^s(\Omega)$ the rate improves to N^{-s/d_0} , with the ambient dimension entering the estimate only through the constant, by way of the multiplicity $\binom{d}{d_0}$ and a fixed embedding factor. A constructive validation in Section 4 exhibits the two regimes side by side and corroborates the effective-dimension reading on educational tabular data.

The Locus of Novelty: Localized Direct Rates, the Univariate Converse, and the Additive Refinement

The direct estimate of Theorem 5 is a localized and explicit form of the $N^{-s/d}$ rate. Its point of difference from the variation-space results cited above is twofold. The estimate is carried on the Besov seminorm itself, delivered through the homogeneous K -functional rather than through a passage to a complete norm, so that no spurious lower-order term enters the bound. The constant is then tracked deterministically through the activation parameters and the Lipschitz character of the domain, by way of an explicit quasi-interpolant built from the activation. The result is therefore positioned not as a first demonstration of the rate but as the explicit, constant-tracking instance on which the refinement of Theorem 6 can be erected as a black box. This attention to explicit, trackable error constants places the present estimates within the wider constructive-approximation program in which quantita-

tive error bounds are made explicit for concrete operator families, as in the recent fractional Schurer-type construction of [19].

The converse of Corollary 1 is delivered in dimension $d = 1$ by reducing the univariate network class to free-knot piecewise polynomials, on which a Bernstein-type inequality on the nonlinear scale is available [20], and by combining this inequality with a dyadic near-best construction through real interpolation on the diagonal scale. The restriction to one spatial variable is not incidental. The multivariate converse rests on an inverse inequality that is presently unavailable (Remark 7).

The effective-dimension refinement of Theorem 6 is the principal contribution. The effective-dimension phenomenon itself is established in the statistical estimation literature as a minimax risk rate under composition or additive hypotheses [10,11]. The present statement is of a different type. It is a constructive approximation estimate, in which the rate N^{-s/d_0} is produced by an explicit allocation of the width budget across the $\binom{d}{d_0}$ coordinate faces, each face inheriting the direct estimate of Theorem 5 in dimension d_0 , and the dependence on the ambient dimension is confined to the constant. The contrast is between a statistical minimax rate, which is a statement about estimation from samples and is given up to unspecified constants, and a deterministic approximation rate, which is a statement about the construction and carries a constant explicit in $\binom{d}{d_0}$ and in the embedding factor. The single-subset case $d_0 = d$ returns the ambient rate of Theorem 5, so the refinement is a genuine interpolation between the two dimensional regimes rather than a separate construction.

The remainder of the paper is organized as follows. Section 2 fixes the notation and recalls the background material, namely the modulus of smoothness on bounded Lipschitz domains, the Besov scale, the standing hypotheses on the activation function, the localized shallow class, and the coordinate-subset additive class, together with the results from the classical theory that are invoked without restatement. Section 3 establishes the approximation-theoretic results, namely, the polynomial reproduction and the σ -realized B-spline basis, the local Jackson estimate, the direct estimate of Theorem 5, the univariate converse of Corollary 1, the width-allocation lemma, and the effective-dimension refinement of Theorem 6. Section 4 reports the constructive validation of the predicted rates and the effective-dimension reading on educational tabular data.

2. Preliminaries

We begin by recalling the definitions and background material concerning shallow-neural-network approximation in the Besov scale, the modulus of smoothness on bounded Lipschitz domains, and the coordinate-subset decomposition that underlies the additive class introduced in Section 3. Familiarity with the classical theory of Besov and Triebel–Lizorkin spaces, the theory of nonlinear N -term approximation, and the approximation theory of shallow neural networks under non-polynomial activation is presupposed. Standard monographs for the classical theory of function spaces include [21,22]; for the constructive approximation framework and the modulus of smoothness, we refer to [23]; for the foundational density and approximation capabilities of shallow networks under non-polynomial activation, we refer to [1,2]; for recent advances concerning the tractability and dimension-independent bounds of such networks, see [13].

Throughout this paper we work in the Euclidean space $\mathbb{R}^d$ with dimension $d \geq 1$. The relation $A \lesssim B$ indicates that $A \leq CB$ for a positive constant C whose dependence is either irrelevant in the context or stated explicitly; $A \sim B$ means $A \lesssim B$ and $B \lesssim A$; the symbol $A \asymp B$ is used interchangeably with $A \sim B$. When the explicit value or the parameter dependence of a constant enters the assertion, the constant is denoted $C_1, C_2, \dots$

and tracked throughout the argument. The expression $\mathbb{N}_0$ denotes the set of non-negative integers and $\mathbb{N}$ the set of positive integers.

The Fourier transform of f and its inverse are normalized as

$$\widehat{f}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-2\pi i x \cdot \xi} dx, \quad f(x) = \int_{\mathbb{R}^d} \widehat{f}(\xi) e^{2\pi i x \cdot \xi} d\xi. \quad (1)$$

We denote by $\mathcal{S}(\mathbb{R}^d)$ the Schwartz space and by $\mathcal{S}'(\mathbb{R}^d)$ the space of tempered distributions. For $1 \leq p \leq \infty$ we denote by p' the conjugate exponent defined by $1/p + 1/p' = 1$, with the convention $1' = \infty$ and $\infty' = 1$.

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For a measurable function g on Ω and an exponent $0 < p \leq \infty$, the $L^p(\Omega)$ (quasi-)norm is defined in the standard manner. For an integer $r \geq 1$ and a vector $h \in \mathbb{R}^d$, the r th order forward difference operator, taken along the direction h , is

$$\Delta_h^r f(x) := \sum_{k=0}^r (-1)^{r-k} \binom{r}{k} f(x + kh), \quad (2)$$

which coincides with the r -fold iterate $(\Delta_h^1)^r f$. The modulus of smoothness of order r of $f \in L^p(\Omega)$ is

$$\omega_r(f, t)_p := \sup_{|h| \leq t} \|\Delta_h^r f\|_{L^p(\Omega_{rh})}, \quad t > 0, \quad (3)$$

where $\Omega_{rh} := \{x \in \Omega : x + kh \in \Omega \text{ for } k = 0, 1, \dots, r\}$ is the subset of Ω on which the r th difference is well defined. In one dimension, (2) and (3) reduce to the difference operator and the modulus of [23] (Ch. 2, § 7, Eqs. (7.1) and (7.2), p. 44). For bounded Lipschitz domains $\Omega \subset \mathbb{R}^d$, the modulus (3) agrees with the multivariate construction of [22] (§ 2.1.1); the identification is recorded in Remark 1.

Remark 1 (Identity with the adapted-difference convention). For a fixed $h \in \mathbb{R}^d$, let $\Delta_{h,\Omega}^r f$ denote the adapted difference of [22] (§ 2.1.1, Eq. (2.18), p. 32), defined to equal $\Delta_h^r f(x)$ when $x + kh \in \Omega$ for all $k = 0, \dots, r$ and to equal zero otherwise. The set on which $\Delta_{h,\Omega}^r f = \Delta_h^r f$ is exactly Ω_{rh} , and on the complement $\Omega \setminus \Omega_{rh}$, the adapted difference vanishes, contributing nothing to the integral $\int_{\Omega} |\cdot|^p$. Hence, for every h ,

$$\|\Delta_{h,\Omega}^r f\|_{L^p(\Omega)} = \|\Delta_h^r f\|_{L^p(\Omega_{rh})}. \quad (4)$$

Taking the supremum over $|h| \leq t$ and integrating in t , the domain-restriction convention (3) and the zero-extension convention of [22] (§ 2.1.1, Eqs. (2.18) and (2.19), p. 32) define the same seminorm (9) below.

We fix once and for all the activation function assumption used in this paper. Throughout, $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is a measurable non-polynomial function. The set of conditions (A_1) , (A_2) , (A_3) below is declared here once and is assumed in every definition, theorem, lemma, and corollary that follows; it is not restated at each occurrence.

Assumption 1 (Standing hypothesis on the activation function). The activation function $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfies:

(A_1) $\sigma \in L_{\text{loc}}^\infty(\mathbb{R})$. Here, $L_{\text{loc}}^\infty(\mathbb{R})$ denotes the measurable functions essentially bounded on every compact subset of $\mathbb{R}$; condition (A_1) is the minimal activation regularity ensuring that each ridge unit $\sigma(w \cdot x + b)$ is a well-defined element of $L^p(\Omega)$ on the bounded domain, and it is not a smoothness hypothesis on σ .

(A_2) There exist constants $K_\sigma > 0$ and $m_\sigma \in \mathbb{N}_0$ such that

$$|\sigma(t)| \leq K_\sigma(1 + |t|)^{m_\sigma}, \quad t \in \mathbb{R}. \quad (5)$$

(A₃) There exists an integer $r_\sigma \geq 1$ such that the distributional derivative $\sigma^{(r_\sigma)}$ belongs to $L^1_{\text{loc}}(\mathbb{R})$. The smoothness index s_σ permitted by σ is fixed in the direct theorem of Section 3. Under the additional condition (A₄) (Assumption 2) that theorem works with $s_\sigma := r_\sigma + 1$ on the open range $0 < s < s_\sigma$; this local convention supersedes the placeholder $s_\sigma := r_\sigma$ used in earlier drafts.

For the localized direct estimate of Section 3, one further structural condition on σ is invoked; it is stated separately because it is not implied by (A₁)–(A₃) and is used only where the localized rate is asserted.

Assumption 2 (Exact localization generator). In addition to Assumption 1, the activation σ admits an exact spline generator:

(A₄) There exist an integer $v_0 \geq 1$, a width $W \in \mathbb{N}$ with $W \geq r_\sigma + 1$, and a function $\Lambda \in \mathcal{N}_{v_0}(\sigma; 1)$ with $\text{supp } \Lambda \subseteq [0, W]$, $0 \leq \Lambda \leq 1$, $\sum_{k \in \mathbb{Z}} \Lambda(t - k) = 1$, and whose integer shifts reproduce every univariate polynomial of degree at most r_σ ;

(A'₄) The integer translates $\{\Lambda(\cdot - k)\}_{k \in \mathbb{Z}}$ are locally linearly independent and form an L^p -stable basis of their span for every $1 \leq p \leq \infty$; equivalently, Λ is the cardinal B-spline of order W .

The conditions (A₄) and (A'₄) hold for ReLU^k (with $v_0 = k + 2$, $W = k + 1$, Λ the cardinal B-spline of order $k + 1$) and for compactly supported spline-type activations; for activations whose finite σ -combinations cannot have compact support they are a genuine restriction. In particular, for the sigmoid and the hyperbolic tangent, whose finite σ -combinations are not compactly supported, the localized class does not embed naturally; (A₄) is retained because the exact polynomial reproduction it supplies is the mechanism underlying the explicit constants of Lemmas 1 and 2.

A single point of normalization is fixed for the shallow network parametrization used in this paper.

Definition 1 (Shallow network class). For $N \geq 1$, the shallow network class associated with the activation σ is

$$\mathcal{N}_N(\sigma) := \left\{ x \mapsto \sum_{i=1}^N a_i \sigma(w_i \cdot x + b_i) : a_i, b_i \in \mathbb{R}, w_i \in \mathbb{R}^d \right\}. \quad (6)$$

No norm constraint on the parameters (a_i, w_i, b_i) is imposed in (6); the variation-type constraints that enter the direct estimates appear in the hypotheses of the theorems of Section 3, and the parameter constraints of the trained models appear in the experimental protocol of Section 4.

Definition 2 (Localized shallow class). For $N \geq 1$ and input dimension $m \geq 1$, the localized shallow class is

$$\mathcal{N}_N^b(\sigma; m) := \left\{ x \mapsto \sum_{j=1}^M g_j(x) \prod_{l=1}^m \psi_{j,l}(x_l) : g_j \in \mathcal{N}_{v'}(\sigma; m), \psi_{j,l} \in \mathcal{N}_{v''}(\sigma; 1), M(v' + mv'') \leq N \right\}, \quad (7)$$

its size measured by the total ridge-unit count. The single-hidden-layer class (6) embeds via $M = 1$, $\psi_{1,l} \equiv 1$. The associated best-approximation quantity is

$$E_N^b(f)_{L^p(\Omega)} := \inf_{g \in \mathcal{N}_N^b(\sigma; m)} \|f - g\|_{L^p(\Omega)} \leq E_N(f)_p. \quad (8)$$

For functions defined on subsets of $\mathbb{R}^d$, the parametrization (6) serves as the structural template used throughout this paper. Results adapted from [1] (§ 2, p. 149) and [3] (Eq. (3), p. 314) correspond to finite-dimensional affine parametrizations that differ from (6) only by an elementary re-labeling of the affine terms; in particular, the threshold form $\sigma(w_i \cdot x - \theta_i)$ of [1] (§ 2) is recovered through $\theta_i = -b_i$. The kernel-based G-network framework of [13] (§ 1, (1.1)), formulated over compact metric input and parameter spaces, is invoked at the level of its approximation result in Section 3, where the compactness of the parameter domain is made explicit; it is not treated here as an affine reparametrization of (6).

Remark 2 (Relation to the standard shallow class). The class $\mathcal{N}_N^b(\sigma; m)$ contains $\mathcal{N}_N(\sigma)$ through $M = 1$, $\psi_{j,l} \equiv 1$, whence $E_N^b(f) \leq E_N(f)$; the localization enriches the standard class by univariate ridge products, at the bounded unit-budget cost κ of Lemma 3, which is independent of N and equals one when $d_0 = d$.

We now turn to the paper-specific function classes and associated quantities.

Definition 3 (Besov space on a bounded Lipschitz domain). Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For $1 \leq p \leq \infty$, $0 < q \leq \infty$, $s > 0$, and an integer $r > s$, the Besov seminorm is

$$|f|_{B_{p,q}^s(\Omega)} := \begin{cases} \left(\int_0^1 (t^{-s} \omega_r(f, t)_p)^q \frac{dt}{t} \right)^{1/q}, & 0 < q < \infty, \\ \sup_{0 < t \leq 1} t^{-s} \omega_r(f, t)_p, & q = \infty, \end{cases} \quad (9)$$

and the Besov norm is

$$\|f\|_{B_{p,q}^s(\Omega)} := \|f\|_{L^p(\Omega)} + |f|_{B_{p,q}^s(\Omega)}. \quad (10)$$

The Besov space $B_{p,q}^s(\Omega)$ consists of those $f \in L^p(\Omega)$ for which $\|f\|_{B_{p,q}^s(\Omega)} < \infty$.

The seminorm (9) is independent of the choice of $r > s$, up to a constant depending only on r, s, p, q , and Ω . In one dimension, this is [23] (Ch. 2, Thm 10.1, p. 55); for bounded Lipschitz domains it follows from the characterization via adapted differences in [22] (§ 2.1.1, 2.18 and 2.19) together with Remark 1. The statement is taken as axiomatic-level here and restated in Theorem 3.

Remark 3 (Function spaces and the localized class). Three space-level objects and one function class enter the estimates in distinct roles. The Besov space $B_{p,q}^s(\Omega)$ (Definition 3) is the smoothness class on which the direct rate is stated; the Bessel-potential space $W^{s,p}(\Omega) = F_{p,2}^s(\Omega)$ enters only as the bridge to the Sobolev-scale literature through Theorem 2; the homogeneous Sobolev seminorm $\dot{W}_p^r$ is the penalty in the K-functional of Theorem 4. The variation space of [3,6] is the representer framework against which the localized estimate is contrasted in Section 1 and is not a hypothesis here. The localized class $\mathcal{N}_N^b(\sigma; m)$ (Definition 2) contains the standard class $\mathcal{N}_N(\sigma)$ through $M = 1$, $\psi_{j,l} \equiv 1$, whence $E_N^b(f) \leq E_N(f)$. The direct estimate (Theorem 5) maps Besov smoothness to a decay rate of E_N^b ; the converse (Corollary 1) returns Besov membership from an ℓ^τ -summability of E_N^b .

Definition 4 (Best approximation error). For $f \in L^p(\Omega)$ and $N \geq 1$, the best approximation error of f by $\mathcal{N}_N(\sigma)$ in $L^p(\Omega)$ is

$$E_N(f)_p := \inf_{g \in \mathcal{N}_N(\sigma)} \|f - g\|_{L^p(\Omega)}. \quad (11)$$

This approximation-error quantity is formulated in accordance with the classical constructive approximation literature (see, e.g., [23] (Ch. 2, § 4, Eq. (4.1), p. 28)); for its structurally equiv-

alent instantiations concerning shallow neural network classes, see, e.g., [1] (Eq. (6.1), p. 168) and [9] (p. 376).

For the additive class introduced in Section 3 the following notation is used. Throughout the additive setting we assume that Ω is a bounded product domain, $\Omega = \prod_{i=1}^d \Omega_i$ with each $\Omega_i \subset \mathbb{R}$ a bounded open interval. Fix an integer $1 \leq d_0 \leq d$. For a subset $T \subset \{1, \dots, d\}$ with $|T| = d_0$ we denote by Ω_T the projection of Ω onto the coordinates indexed by T , and for $x \in \Omega$ we denote by $x_T = (x_i)_{i \in T}$ the corresponding sub-vector; under the product assumption, $\Omega_T = \prod_{i \in T} \Omega_i$ is a bounded Lipschitz domain in $\mathbb{R}^{d_0}$. The number of such subsets is $\binom{d}{d_0}$.

Definition 5 (Coordinate-subset additive Besov class). Fix an integer $1 \leq d_0 \leq d$ and parameters $s > 0$, $1 \leq p \leq \infty$, $0 < q \leq \infty$. The coordinate-subset additive Besov class is

$$\mathcal{S}_{d_0}^s(\Omega) := \left\{ f : \Omega \rightarrow \mathbb{R} \mid f(x) = \sum_{|T|=d_0} g_T(x_T), g_T \in B_{p,q}^s(\Omega_T) \right\}, \quad (12)$$

the sum running over coordinate subsets $T \subset \{1, \dots, d\}$ with $|T| = d_0$. The associated seminorm is

$$|f|_{\mathcal{S}_{d_0}^s(\Omega)} := \inf \left\{ \sum_{|T|=d_0} |g_T|_{B_{p,q}^s(\Omega_T)} : f = \sum_{|T|=d_0} g_T, g_T \in B_{p,q}^s(\Omega_T) \right\}. \quad (13)$$

The infimum is taken over all admissible decompositions of f ; since such a decomposition is not unique, the infimum renders $|\cdot|_{\mathcal{S}_{d_0}^s(\Omega)}$ a well-defined seminorm on $\mathcal{S}_{d_0}^s(\Omega)$, and it is this quantity that Step 3 of the proof of Theorem 6 returns. The structural assumption (12) explicitly constrains the target function to a superposition of low-dimensional interactions, providing a formally circumscribed framework to mitigate the curse of dimensionality. Targets carrying genuine high-order interactions among all coordinates lie outside $\mathcal{S}_{d_0}^s(\Omega)$; the additive class is offered as a formally circumscribed first-order description of low-effective-dimension structure, not as an exact model of every interaction pattern. For analogous theoretical frameworks that exploit low intrinsic dimensionality and compositional structures in neural network approximation, see, e.g., [12] (§ 4) and [13] (§ 1).

Four results from the classical theory close this section and are used in Section 3 without restatement of their proofs.

Theorem 1 (Universal density for non-polynomial activation). Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ satisfy Assumption 1. Then, $\bigcup_{N \geq 1} \mathcal{N}_N(\sigma)$ is dense in $L^p(\Omega)$ for every $1 \leq p < \infty$, and dense in $C(\Omega)$ when σ is continuous.

This is the shallow network density theorem under non-polynomial activation; it is cited from [1] (Thm 3.1, p. 153) and [2] (Thm 6.13, p. 74) and serves as the existence layer of the constructive estimate of Section 3.

Theorem 2 (Sobolev–Besov embedding on bounded Lipschitz domains). Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary, let $1 < p < \infty$ and $s > 0$, and let $W^{s,p}(\Omega)$ denote the Bessel-potential space, which coincides with $F_{p,2}^s(\Omega)$. Then, the continuous embeddings

$$B_{p,\min(p,2)}^s(\Omega) \hookrightarrow W^{s,p}(\Omega) \hookrightarrow B_{p,\max(p,2)}^s(\Omega) \quad (14)$$

hold with embedding constants depending only on d, s, p , and Ω . For $p = 2$, the two indices coincide, and $W^{s,2}(\Omega) = B_{2,2}^s(\Omega)$.

This is cited from [21] (§ 1.5.1, Thm (iii), p. 41, and § 2.3.2, Eq. (22), p. 96) for the relations on $\mathbb{R}^d$; their transfer to bounded Lipschitz domains uses the universal bounded linear extension operator of [24] (§ 4, Thm 4.1, p. 254), also recorded in [22] (§ 4.2, Cor. 4.12, p. 107). It provides the bridge used at every comparison point between the Besov estimates of this paper and the Sobolev estimates of [25] (Thm A), [26] (Thm 2.2), and [27] (Cor 1 & 2).

Theorem 3 (Modulus of smoothness equivalence). For $1 \leq p \leq \infty$, $0 < q \leq \infty$, $s > 0$, and integers $r_1, r_2 > s$,

$$\left(\int_0^1 (t^{-s} \omega_{r_1}(f, t)_p)^q \frac{dt}{t} \right)^{1/q} \asymp \left(\int_0^1 (t^{-s} \omega_{r_2}(f, t)_p)^q \frac{dt}{t} \right)^{1/q}, \quad (15)$$

with implicit constants depending only on r_1, r_2, s, p, q , and Ω .

This equivalence is established in [23] (Ch. 2, Thm 10.1, p. 55) in one dimension and in [22] (§ 2.1.1, Eqs. (2.18) and (2.19), p. 32) for bounded Lipschitz domains; by Remark 1 it applies to the seminorm (9) in the stated form.

Theorem 4 (K-functional/modulus equivalence). Let $1 \leq p \leq \infty$, let $r \in \mathbb{N}$, and let $\Omega \subset \mathbb{R}^m$ be either $\mathbb{R}^m$ itself or a bounded Lipschitz domain. For the homogeneous Sobolev seminorm

$|g|_{\dot{W}_p^r(\Omega)} := \left(\sum_{|\alpha|=r} \|\partial^\alpha g\|_{L^p(\Omega)}^p \right)^{1/p}$ (with the customary modification for $p = \infty$), define the homogeneous K-functional

$$K(f, t^r; L^p(\Omega), \dot{W}_p^r(\Omega)) := \inf_g \left\{ \|f - g\|_{L^p(\Omega)} + t^r |g|_{\dot{W}_p^r(\Omega)} \right\}, \quad t > 0, \quad (16)$$

for every $f \in L^p(\Omega)$, the infimum running over g with $|g|_{\dot{W}_p^r(\Omega)} < \infty$. Then, there exist constants $M_1, M_2 > 0$, depending only on r, p, m , and the Lipschitz character of Ω , such that

$$M_1 \omega_r(f, t)_{L^p(\Omega)} \leq K(f, t^r; L^p(\Omega), \dot{W}_p^r(\Omega)) \leq M_2 \omega_r(f, t)_{L^p(\Omega)}, \quad 0 < t \leq 1. \quad (17)$$

The penalty term in (16) is the homogeneous seminorm $|g|_{\dot{W}_p^r}$, so the equivalence (17) transfers a Besov seminorm rather than a complete-norm quantity; this is the form used in the localized direct estimate of Section 3. In one dimension, (17) is [23] (Ch. 6, § 2, Thm 2.4, Eq. (2.11), p. 177), with the homogeneous K-functional (16) introduced in [23] (Ch. 6, § 1, Eq. (1.11), p. 172); for $\mathbb{R}^m$ and for bounded Lipschitz domains the multivariate isotropic form, with the seminorm penalty stated as in (16), is [28] (Thm 1, Eq. (1.7), p. 120).

We close this section by recording the notational conventions governing the transfer of constants from cited results to this paper. All cited results are transferred to the conventions of (1), (3), and (6). Three conversion items are recorded here once and applied at every citation point:

- (C₁) Citations adopting the Fourier convention $\hat{f}(\xi) = \int f(x) e^{-ix \cdot \xi} dx$ contribute a factor of $(2\pi)^d$ in the relevant Plancherel-type or integral-representation constants.
- (C₂) Citations stating shallow-network results under a parametrization of the form $\sum_{i=1}^N a_i \sigma(w_i^\top x - \theta_i)$ or under a coefficient norm constraint $\sum_{i=1}^N |a_i| \leq M$ are transferred to (6) through the re-indexing $\theta_i = -b_i$, with the constraint level M recorded at the point of citation.

- (C₃) Citations stating estimates in $W^{s,p}(\Omega) = F_{p,2}^s(\Omega)$, $1 < p < \infty$, are transferred to the Besov scale via Theorem 2: an upper bound in $W^{s,p}$ yields an upper bound in $B_{p,\max(p,2)}^s(\Omega)$, and a lower bound transfers from $B_{p,\min(p,2)}^s(\Omega)$. The direction of each transfer is recorded at the point of citation.

3. The Constructive Machinery: Direct Estimates and the Effective-Dimension Refinement

This section establishes the approximation-theoretic results of the paper. The σ -realized B-spline basis of Lemma 1 is the structural ingredient from which the local Jackson estimate of Lemma 2 is assembled. Its ambient instance is the direct estimate of Theorem 5, and its univariate converse is Corollary 1. The deterministic width allocation of Lemma 3 then distributes a single budget across the coordinate subsets and yields the effective-dimension refinement of Theorem 6.

Lemma 1 (Exact σ -realized B-spline basis). *Let σ satisfy Assumptions 1 and 2, $m \geq 1$. Let $h > 0$ and $\{Q_j\}_{j \in \mathbb{Z}^m}$ be the side- h cubes from translating $[0, h]^m$ along $h\mathbb{Z}^m$. For $j \in \mathbb{Z}^m$, put $\Phi_j(x) := \prod_{l=1}^m \psi_{j,l}(x_l)$ with $\psi_{j,l}(t) := \Lambda(t/h - j_l) \in \mathcal{N}_{v_0}(\sigma; 1)$. Then,*

- (i) $0 \leq \Phi_j \leq 1$ and Φ_j is a product of m univariate single-hidden-layer networks, with mv_0 units in total, independent of h and of the index set;
- (ii) $\text{supp } \Phi_j \subseteq Q_j^* := \prod_l [hj_l, h(j_l + W)]$, cover multiplicity $\leq W^m$;
- (iii) $\sum_{j \in \mathbb{Z}^m} \Phi_j \equiv 1$ on $\mathbb{R}^m$;
- (iv) For every polynomial P of total degree $\leq r_\sigma$, there are scalars $c_j(P)$, linear and local in P , with $\sum_j c_j(P) \Phi_j \equiv P$;
- (v) The family $\{\Phi_j\}$ admits local de Boor–Fix dual functionals $\{\lambda_j\}$, λ_j supported in Q_j^* , with $\lambda_j(\Phi_{j'}) = \delta_{jj'}$ and $|\lambda_j(F)| \leq C_1 h^{-m/p} \|F\|_{L^p(Q_j^*)}$, $C_1 = C_1(\sigma, m, W)$.

Proof of Lemma 1. Let $\Lambda = \sum_{i=1}^{v_0} a_i \sigma(\lambda_i \cdot + \beta_i)$ be the (A_4) generator.

- (i) The reparametrization $\psi_{j,l}(t) = \sum_i a_i \sigma((\lambda_i/h)t + (\beta_i - \lambda_i j_l))$ shows $\psi_{j,l} \in \mathcal{N}_{v_0}(\sigma; 1)$; $0 \leq \Lambda \leq 1$ gives $0 \leq \Phi_j \leq 1$, a product of m factors, mv_0 units.
- (ii) $\Phi_j(x) \neq 0$ forces $x_l \in [hj_l, h(j_l + W)]$ for all l ; each x_l lies in at most W such intervals, so multiplicity $\leq W^m$.
- (iii) As only W^m summands are nonzero at fixed x ,

$$\sum_j \Phi_j(x) = \prod_l \left(\sum_k \Lambda(x_l/h - k) \right) = \prod_l 1 = 1.$$

- (iv) The Marsden identity for Λ [23] (Ch. 5, § 3, Eqs. (3.6) and (3.8), pp. 141–142) gives local linear functionals μ_n with $\sum_n \mu_n(p) \Lambda(\cdot - n) = p$ for univariate p of degree $\leq r_\sigma$; tensoring over the m coordinates and rescaling by h produce local linear $c_j(P)$ with $\sum_j c_j(P) \Phi_j = P$ for P of total degree $\leq r_\sigma$.
- (v) By (A_4') , the translates of Λ are locally linearly independent and L^p -stable, so the de Boor–Fix functionals of [23] (Ch. 5, § 3, Eq. (3.9), pp. 142–143) exist, are biorthogonal, and carry the L^p -stability bound of [23] (Ch. 5, § 4, Eqs. (4.11)–(4.12), p. 146);

the tensorization and the rescaling $t \mapsto t/h$ give the displayed h -scaling with $C_1 = C_1(\sigma, m, W)$.

□

Lemma 2 (Local Jackson estimate on a bounded Lipschitz domain). *Let σ satisfy Assumptions 1 and 2 with parameters $K_\sigma, m_\sigma, r_\sigma$ and generator data (v_0, W) . Let $m \geq 1$, $\Omega \subset \mathbb{R}^m$ a bounded Lipschitz domain, $1 \leq p \leq \infty$, $0 < q \leq \infty$, $0 < s < r_\sigma + 1$. Then, there are $C_J = C_J(\sigma, s, p, q, m, \Omega) > 0$ and an integer $N_0 = N_0(\sigma, s, p, q, m, \Omega) \geq 1$ such that for every $f \in B_{p,q}^s(\Omega)$ and every integer $N \geq N_0$,*

$$E_N^b(f)_{L^p(\Omega)} \leq C_J N^{-s/m} |f|_{B_{p,q}^s(\Omega)}. \quad (18)$$

C_J depends on σ through $K_\sigma, m_\sigma, r_\sigma, v_0, W$ and on Ω through its Lipschitz character via the extension constant C_W ; it is independent of N and f . The estimate holds for every $m \geq 1$.

Remark 4 (Domain dependence). *The Ω -dependence of C_J is carried by $C_W = C_W(\Omega, s, p, q, m)$, depending on Ω only through its Lipschitz character.*

Remark 5 (Smoothness range). *The open range $0 < s < r_\sigma + 1$ reflects the local degree r_σ : the quasi-interpolant reproduces $\mathbb{P}_{r_\sigma}$, so the local estimate delivers $h^{r_\sigma+1}$ on the seminorm, and the modulus monotonicity transfers it to h^s for $0 < s < r_\sigma + 1$. We set $s_\sigma := r_\sigma + 1$; the open endpoint is retained.*

Remark 6 (Relation to single-hidden-layer Sobolev rates). *For $\sigma = \text{ReLU}^k$ ($r_\sigma = k$) the estimate of Lemma 2 reads*

$$E_N^b(f)_{L^p(\Omega)} \leq C_J N^{-s/m} |f|_{B_{p,q}^s(\Omega)}, \quad 0 < s < k + 1,$$

with C_J explicit through $(v_0, W) = (k + 2, k + 1)$; the error is measured in $L^p(\Omega)$ and the dimension m enters only through the exponent s/m . A single-hidden-layer counterpart is recorded in [25] (Thm D) (after Siegel and Xu [4]): on $\Omega = [0, 1]^m$, for a target $f \in B^s(\Omega)$ with $s \geq \frac{1}{2}$ and a differentiation order $\mu \geq 0$ satisfying $\mu \leq s - \frac{1}{2}$ and $\mu < k + \frac{1}{2}$, a width- N single-hidden-layer class realizes

$$\inf_{f_N} \|f - f_N\|_{H^\mu(\Omega)} \lesssim \|f\|_{B^s(\Omega)} N^{-t} (\log N)^q, \quad t = \frac{1}{2} + \min\left(\frac{2(s-\mu)-1}{2(m+1)}, k - \mu + \frac{1}{2}\right),$$

the error being measured in the Sobolev norm $H^\mu(\Omega) = W^{\mu,2}(\Omega)$; here, μ denotes the order of differentiation and remains distinct from the spatial dimension m , while the logarithmic exponent $q \geq 0$ vanishes for s below the threshold $(m + 1)(k - \mu + \frac{1}{2}) + \mu + \frac{1}{2}$ and is positive beyond it. The two estimates are of the same algebraic type—decay polynomial in N , up to logarithmic factors, with an exponent increasing in the target smoothness and decreasing in the dimension—yet they do not share a common exponent: the present exponent s/m governs an L^p error, whereas t carries the smoothness deficit $s - \mu$ induced by the stronger H^μ norm, a dimension denominator $2(m + 1)$, and the logarithmic correction $(\log N)^q$. The construction of [25] (Thm D) proceeds through the algebraic expansion of the network output into monomials; the present construction proceeds through the localized class $\mathcal{N}_N^b(\sigma; m)$ and the Besov scale. The Sobolev and Besov scales are placed in a common frame by Theorem 2 (convention (C_3)); no identity of exponents and no identity of constants is asserted.

Proof of Lemma 2. The argument tracks every constant. Constants $C_1, C_2, \dots$ have their dependence displayed at the point of definition.

Since Ω is a bounded Lipschitz domain, there is a bounded linear extension operator, universal in the order of smoothness, $\mathcal{E} : B_{p,q}^s(\Omega) \rightarrow B_{p,q}^s(\mathbb{R}^m)$ with operator norm depending on Ω only through its Lipschitz character [24] (§ 4, Thm 4.1, p. 254); see also [22] (§ 4.2, Cor. 4.12, p. 107). Put $F := \mathcal{E}f$. We use the extension only at the seminorm level through the modulus: for the modulus of smoothness of order $r_\sigma + 1$,

$$\omega_{r_\sigma+1}(F, t)_{L^p(\mathbb{R}^m)} \leq C_W \omega_{r_\sigma+1}(f, t)_{L^p(\Omega)}, \quad 0 < t \leq 1, \quad (19)$$

with $C_W = C_W(\Omega, s, p, q, m)$ depending on Ω only through its Lipschitz character. Fix a closed cube $Q^* \supset \bar{\Omega}$ of side $R_0 := 1 + \text{diam}(\Omega)$. For $M \in \mathbb{N}$ set $h := R_0/M$ ($h \asymp M^{-1}$) and let $\mathcal{J} \subset \mathbb{Z}^m$ index the side- h cubes meeting $\bar{\Omega}$ together with a collar of width W layers; then every $x \in Q^*$ has all its W^m nonzero contributions $\Phi_j(x)$ indexed within $\mathcal{J}$, so $\sum_{j \in \mathcal{J}} \Phi_j \equiv 1$ on Q^* (Lemma 1 (iii)), and $\#\mathcal{J} \leq C_0 M^m$, $C_0 = C_0(m, W, R_0)$. Write $\tilde{Q}$ for the Wh -neighborhood of Q^* .

Let $\{\Phi_j\}_{j \in \mathcal{J}}$ be the basis of Lemma 1 for side h and $\{\lambda_j\}$ its de Boor–Fix duals (part (v)). Define $Q_h F := \sum_{j \in \mathcal{J}} \lambda_j(F) \Phi_j$. By Lemma 1 (iv)–(v) and biorthogonality, Q_h reproduces $\mathbb{P}_{r_\sigma}$: $Q_h P = P$ for $P \in \mathbb{P}_{r_\sigma}$. The L^p -stability of $\{\lambda_j\}$, $0 \leq \Phi_j \leq 1$, and multiplicity W^m give, on each cell,

$$\|Q_h F\|_{L^p(Q_j)} \leq C_2 \|F\|_{L^p(Q_j^*)}, \quad C_2 = C_2(\sigma, m, W),$$

hence $I - Q_h$ is bounded on L^p with constant $C_5 = C_5(\sigma, m, W)$ after summation with multiplicity W^m over $\tilde{Q}$. Since Q_h reproduces $\mathbb{P}_{r_\sigma}$, for every $P \in \mathbb{P}_{r_\sigma}$,

$$\|F - Q_h F\|_{L^p(Q_j)} = \|(I - Q_h)(F - P)\|_{L^p(Q_j)} \leq (1 + C_2) \|F - P\|_{L^p(Q_j^*)}; \quad (20)$$

taking the infimum over P and applying the Bramble–Hilbert lemma on Q_j^* (star-shaped of diameter $\asymp h$) [29] (Lemma 4.3.8, p. 102),

$$\inf_{P \in \mathbb{P}_{r_\sigma}} \|F - P\|_{L^p(Q_j^*)} \leq C_3 h^{r_\sigma+1} |F|_{W_p^{r_\sigma+1}(Q_j^*)}, \quad C_3 = C_3(m, r_\sigma, p). \quad (21)$$

We now deliver the Besov seminorm through the homogeneous K -functional, avoiding any complete-norm passage. For any decomposition $F = F_0 + F_1$ with $F_0 \in L^p$, $F_1 \in \dot{W}_p^{r_\sigma+1}$, the L^p -boundedness of $I - Q_h$ and the seminorm estimate (20) and (21) give, after raising to the p th power (with the customary modification for $p = \infty$), summing over $\mathcal{J}$ and using the multiplicity W^m once,

$$\|(I - Q_h)F\|_{L^p(Q^*)} \leq C_5 \|F_0\|_{L^p(\tilde{Q})} + C_4 h^{r_\sigma+1} |F_1|_{W_p^{r_\sigma+1}(\tilde{Q})}, \quad C_4 := (1 + C_2) C_3 W^{m/p}.$$

Taking the infimum over the decomposition identifies the right-hand side with the homogeneous K -functional $K(h^{r_\sigma+1}, F; L^p, \dot{W}_p^{r_\sigma+1})(\tilde{Q})$, and by the equivalence of the K -functional with the modulus of smoothness (Theorem 4),

$$\|(I - Q_h)F\|_{L^p(Q^*)} \leq C_6 \omega_{r_\sigma+1}(F, h)_{L^p(\tilde{Q})}, \quad C_6 = C_6(\sigma, m, r_\sigma, p).$$

For $0 < s < r_\sigma + 1$, the seminorm (9), the monotonicity of $t \mapsto t^{-s} \omega_{r_\sigma+1}(F, t)_p$, and the embedding $B_{p,q}^s \hookrightarrow B_{p,\infty}^s$ give $\omega_{r_\sigma+1}(F, h)_{L^p} \leq C_7 h^s |F|_{B_{p,q}^s(\tilde{Q})}$ for every $0 < q \leq \infty$, the right-hand side carrying only the Besov seminorm because $I - Q_h$ annihilates $\mathbb{P}_{r_\sigma}$. With the modulus extension bound (19),

$$\|F - Q_h F\|_{L^p(\Omega)} \leq C_8 h^s |f|_{B_{p,q}^s(\Omega)}, \quad C_8 := C_6 C_7 C_W. \quad (22)$$

Each $\lambda_j(F)$ is a scalar and $\Phi_j = \prod_l \psi_{j,l}$ is a product of m univariate single-hidden-layer networks (Lemma 1 (i)); writing the scalar as a degree-zero member of $\mathcal{N}_1(\sigma; m)$, each term $\lambda_j(F)\Phi_j$ has the form $g_j \prod_l \psi_{j,l}$ with $g_j \in \mathcal{N}_1(\sigma; m)$. Hence, $f_N := Q_h F \in \mathcal{N}_N^b(\sigma; m)$ with ridge-unit count $\sum_{j \in \mathcal{J}} (1 + m\nu_0) \leq C_0 M^m (1 + m\nu_0)$. Put $\nu := 1 + m\nu_0$ and choose $M := \left\lfloor (N/(C_0\nu))^{1/m} \right\rfloor$, $N_0 := C_0\nu 2^m$, so that $C_0 M^m \nu \leq N$ for $N \geq N_0$ and $M \geq \frac{1}{2} [N/(C_0\nu)]^{1/m}$, whence $h^s = R_0^s M^{-s} \leq R_0^s 2^s (C_0\nu)^{s/m} N^{-s/m}$. With (22),

$$E_N^b(f)_{L^p(\Omega)} \leq \|F - f_N\|_{L^p(\Omega)} \leq C_J N^{-s/m} |f|_{B_{p,q}^s(\Omega)}, \quad C_J := C_8 R_0^s 2^s (C_0\nu)^{s/m}.$$

Since $C_8 = C_6 C_7 C_W$ by (19), the domain enters C_J only through the extension constant $C_W = C_W(\Omega, s, p, q, m)$, which depends on Ω through its Lipschitz character alone (Remark 4). The number $\nu = 1 + m\nu_0$ is fixed by σ and m ; no tolerance shrinks with N and no unit count grows with N , so C_J is a constant. The reproduction of $\mathbb{P}_{r_\sigma}$ being exact under $(A_4)-(A'_4)$ is what keeps ν independent of N . $\square$

The ambient instance $m = d$ of the local estimate just established is the direct theorem of the paper, which we record next together with the converse it admits in one spatial variable.

Theorem 5 (Direct Jackson-type estimate for localized shallow networks on $B_{p,q}^s(\Omega)$). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. Let σ satisfy Assumption 1 $(A_1)-(A_3)$ together with the localization hypothesis Assumption 2 (A_4) , with associated parameters $K_\sigma, m_\sigma, r_\sigma$ and generator data (ν_0, W) , and set $s_\sigma := r_\sigma + 1$. Let $1 \leq p \leq \infty$, $0 < q \leq \infty$, and $0 < s < s_\sigma$. Then, there exist a constant $C = C(\sigma, s, p, q, d, \Omega) > 0$ and an integer $N_0 = N_0(\sigma, s, p, q, d, \Omega) \geq 1$ such that for every $f \in B_{p,q}^s(\Omega)$ and every integer $N \geq N_0$,*

$$E_N^b(f)_{L^p(\Omega)} \leq C N^{-s/d} |f|_{B_{p,q}^s(\Omega)}, \quad (23)$$

where $E_N^b(f)_{L^p(\Omega)} := \inf_{g \in \mathcal{N}_N^b(\sigma; d)} \|f - g\|_{L^p(\Omega)}$ is the localized best-approximation quantity of Definition 2, and $|\cdot|_{B_{p,q}^s(\Omega)}$ is the homogeneous Besov seminorm of Definition 3. The constant C depends on σ through $K_\sigma, m_\sigma, r_\sigma, \nu_0, W$ and on Ω through its Lipschitz character; it is independent of N and of f .

Proof of Theorem 5. The estimate is the ambient instance $m = d$ of Lemma 2, whose mechanism is the following chain.

- (i) A quasi-interpolant Q_h built from the activation (Lemma 1) reproduces $\mathbb{P}_{r_\sigma}$ on a scale- h cover of Ω ;
- (ii) The Bramble–Hilbert lemma gives the local bound $\|f - Q_h f\|_{L^p} \lesssim h^{r_\sigma+1} |f|$ on each cell;
- (iii) The homogeneous K -functional and its equivalence with the modulus (Theorem 4), followed by the inclusion $B_{p,q}^s \hookrightarrow B_{p,\infty}^s$, convert this into $h^s |f|_{B_{p,q}^s}$; and
- (iv) The choice $h \asymp N^{-1/d}$ within the unit budget of Definition 2 yields the rate $N^{-s/d}$. The constant is assembled from the generator data (ν_0, W) and the Lipschitz character of Ω . The detailed argument follows.

It identifies (23) as the ambient instance $m = d$ of the local Jackson estimate of Lemma 2 and then accounts for the provenance of the constant through the admissible mechanisms recorded in Section 2. Constants C_J, C_q, C have their dependence displayed at the point of definition. Throughout, the modulus of smoothness is the isotropic r th order modulus

$\omega_r(\cdot, t)_{L^p}$ of (3); this matches the geometry of the modulus– K -functional equivalence of Theorem 4, so no conversion between isotropic and coordinatewise moduli is required at any step.

We initialize the argument by explicitly fixing the admissible difference order. Since $0 < s < s_\sigma = r_\sigma + 1$, the integer

$$r := \lfloor s \rfloor + 1 \quad (24)$$

satisfies $s < r \leq r_\sigma + 1$. This is the order admitted by Lemma 2, whose quasi-interpolant reproduces $\mathbb{P}_{r_\sigma}$ and hence delivers the seminorm decay for every s in the open range $0 < s < r_\sigma + 1$ (see Remark 5); the value (24) is the order in the modulus characterization of $|\cdot|_{B_{p,q}^s}$ used below.

We proceed with a principled reduction to the ambient lemma for $m = d$. Fix $f \in B_{p,q}^s(\Omega)$ and an integer $N \geq N_0$, with N_0 as furnished by Lemma 2 in dimension $m = d$. The hypotheses of the present theorem are the hypotheses of that lemma with $m = d$: the activation satisfies Assumptions 1 and 2 with the same parameters $K_\sigma, m_\sigma, r_\sigma, (\nu_0, W)$; the domain $\Omega \subset \mathbb{R}^d$ is bounded with Lipschitz boundary; the indices satisfy $1 \leq p \leq \infty$, $0 < q \leq \infty$, and $0 < s < r_\sigma + 1$. Consequently Lemma 2 applies verbatim with $m = d$ and yields a constant $C_J = C_J(\sigma, s, p, q, d, \Omega) > 0$, independent of N and f , for which

$$E_N^b(f)_{L^p(\Omega)} \leq C_J N^{-s/d} |f|_{B_{p,q}^s(\Omega)}, \quad N \geq N_0. \quad (25)$$

This is (23) with $C = C_J$. The remaining steps record the membership underlying E_N^b and the provenance of C_J in the admissible mechanisms, so that no power of N or factor in the constant is left without attribution.

We subsequently establish membership and isolate the corresponding localized quantity. The estimate (25) is realized by an explicit element of the localized class. By Lemma 1, under Assumption 2 the integer shifts of the generator $\Lambda \in \mathcal{N}_{\nu_0}(\sigma; 1)$ furnish, after tensoring over the d coordinates and rescaling by the side $h \asymp N^{-1/d}$, a quasi-interpolant

$$Q_h f = \sum_{j \in \mathcal{J}} \lambda_j(f) \prod_{l=1}^d \psi_{j,l}(\cdot) \in \mathcal{N}_N^b(\sigma; d),$$

which reproduces $\mathbb{P}_{r_\sigma}$ and has cardinality $\#\mathcal{J} \leq C_0 h^{-d} \asymp N$, in accordance with the size convention $M(\nu' + d\nu'') \leq N$ of Definition 2; here, ν', ν'' are fixed by (ν_0, W) and are independent of N , so the budget $\#\mathcal{J} \asymp N$ is consistent with that convention. Since the single-hidden-layer class (6) embeds in $\mathcal{N}_N^b(\sigma; d)$ through $M = 1$, $\psi_{1,l} \equiv 1$, Definition 2 records $E_N^b(f)_{L^p(\Omega)} \leq E_N(f)_{L^p(\Omega)}$; the direct bound (25) controls the localized quantity itself, which is the assertion of the theorem.

We here explicitly extract the constant using the modulus and the homogeneous K -functional. The decay in (25) originates in the local Jackson mechanism of Lemma 2: the quasi-interpolant reproduces $\mathbb{P}_{r_\sigma}$, so on a Whitney-type cover of Ω at scale h it returns the local seminorm decay h^r , and the monotonicity of the modulus of smoothness transfers h^r to h^s for $0 < s < r_\sigma + 1$. In the notation of (3), this is the bound

$$E_N^b(f)_{L^p(\Omega)} \leq C_1 \omega_r(f, N^{-1/d})_{L^p(\Omega)}, \quad C_1 = C_1(\sigma, r, p, d), \quad (26)$$

with C_1 carrying the activation regularity through K_σ and r_σ by the distributional-derivative regularization mechanism of Section 2, the existence of an admissible network being furnished by the non-polynomial density mechanism (Theorem 1). To pass from (26) to the Besov seminorm we use the homogeneous-smoothness bridge. By Theorem 4 the modulus is equivalent to the homogeneous K -functional (the homogeneous K -functional and its equivalence with the modulus are those of Theorem 4, (16) and (17)),

$$\omega_r(f, t)_{L^p(\Omega)} \asymp K(f, t^r; L^p(\Omega), \dot{W}_p^r(\Omega)), \quad (27)$$

the penalty in the K -functional being the homogeneous seminorm $|\cdot|_{\dot{W}_p^r}$ and not the complete norm. Equivalence (27) is the multivariate isotropic statement recorded for $\Omega \subset \mathbb{R}^d$ in Theorem 4, whose multivariate provenance rests on [28] (Thm 1, Eq. (1.7), p. 120); there, the equivalence is stated for a Lipschitz-graph domain, with the homogeneous seminorm and K -functional introduced at [28] (Eqs. (1.1), (1.3), p. 119). The Besov seminorm of Definition 3 admits the corresponding modulus characterization

$$|f|_{B_{p,q}^s(\Omega)} \asymp \left\| t^{-s} \omega_r(f, t)_{L^p(\Omega)} \right\|_{L^q((0,1), dt/t)}, \quad (28)$$

with the usual supremum reading when $q = \infty$. For every $q \leq \infty$ the inclusion $B_{p,q}^s(\Omega) \hookrightarrow B_{p,\infty}^s(\Omega)$ gives, with an embedding constant $C_q = C_q(s, p, q)$,

$$\omega_r(f, t)_{L^p(\Omega)} \leq t^s |f|_{B_{p,\infty}^s(\Omega)} \leq C_q t^s |f|_{B_{p,q}^s(\Omega)}, \quad 0 < t \leq 1. \quad (29)$$

Here, C_q is the constant of the elementary inclusion $\ell^q((0,1), dt/t) \hookrightarrow \ell^\infty$ underlying $B_{p,q}^s(\Omega) \hookrightarrow B_{p,\infty}^s(\Omega)$; it depends only on s and q and may be taken as $C_q = 2^s (\ln 2)^{-1/q}$ for $0 < q < \infty$ and $C_q = 1$ for $q = \infty$. Evaluating (29) at $t = N^{-1/d} \leq 1$ (which holds for $N \geq N_0 \geq 1$) and substituting into (26) yields

$$E_N^b(f)_{L^p(\Omega)} \leq C_1 C_q N^{-s/d} |f|_{B_{p,q}^s(\Omega)}. \quad (30)$$

Setting $C := C_1 C_q$ recovers (23) and identifies its provenance: the exponent $-s/d$ is the decay carried by (27)–(29); the σ -dependence of C enters through via $K_\sigma, r_\sigma, v_0, W$; the Ω -dependence enters through the Lipschitz character of the domain, by the extension constant of the local estimate (Remark 4). No coordinate-subset mechanism enters, in agreement with the ambient setting $m = d$, for which the multiplicity factor $\binom{d}{d} = 1$ is absent from the constant.

Finally, the endpoint estimate is resolved via the associated canonical reduction. The hypothesis $0 < s < s_\sigma = r_\sigma + 1$ is displayed explicitly and the endpoint $s = s_\sigma$ is excluded, consistent with the open range of Lemma 2 (Remark 5): the order (24) requires $r \leq r_\sigma + 1$, hence $s < r_\sigma + 1$. As $s \rightarrow 0^+$ the factor $N^{-s/d} \rightarrow 1$ and (23) becomes vacuous, so the statement asserts no nontrivial decay at $s = 0$. For $p = 2$ and integer s the identification $B_{2,2}^s(\Omega) = W^{s,2}(\Omega)$ of Theorem 2 (convention (C_3)) aligns (23) with the classical L^2 régime; the same transfer between the Sobolev and Besov measures relates the present rate to the single-hidden-layer Sobolev estimates of [26] (Thm 2.2) and [27] (Cor. 1 & 2), which share the form $N^{-s/d}$ on the Sobolev scale; no identity of constants is asserted across these references. This completes the proof. $\square$

Corollary 1 (Inverse estimate, $d = 1$). *Let $d = 1$, $\Omega = (0,1)$, let $\sigma = \text{ReLU}^k$ satisfy Assumption 1 with $r_\sigma = k$ and $s_\sigma := r_\sigma + 1 = k + 1$, and let $1 \leq p < \infty$. Fix α with $0 < \alpha < s_\sigma$ and set $\tau := (\alpha + 1/p)^{-1}$. Suppose $f \in L^p(\Omega)$ obeys, for some $A > 0$, the approximation-space bound*

$$|f|_{\mathcal{A}} := \left[\sum_{j \geq 0} (2^{j\alpha} E_{2^j}^b(f)_{L^p(\Omega)})^\tau \right]^{1/\tau} \leq A. \quad (31)$$

Then, $f \in B_{\tau,\tau}^\alpha(\Omega)$ and

$$\|f\|_{B_{\tau,\tau}^\alpha(\Omega)} \leq C' (\|f\|_{L^p(\Omega)} + A), \quad (32)$$

with $C' = C'(k, \alpha, p) > 0$ explicit and independent of f .

Proof of Corollary 1. This is the standard converse: the Besov smoothness of f is bounded by the ℓ^τ sum of the dyadic best-approximation errors, obtained by the standard Bernstein route: reduction to free-knot piecewise polynomials, the Bernstein-type inequality, dyadic summation, and real interpolation [23] (Ch. 12). The argument pairs the direct mechanism underlying Theorem 5 with a Bernstein-type inequality for the univariate network class, the two being combined through the discrete characterization of a real-interpolation space and the diagonal interpolation identity of the nonlinear Besov scale. Every constant is tracked; the construction uses a 2-near-best element at each scale, so attainment of the infimum in E_N^b is not required.

As the first step, we circumscribe the univariate structure of the underlying class. For $x \in (0, 1)$, each summand $a_i (w_i x + b_i)_+^k$ is a polynomial of degree k on $\{w_i x + b_i > 0\}$ and vanishes on the complementary half-line, with a single breakpoint at $x = -b_i/w_i$. Hence, every $g \in \mathcal{N}_N(\sigma)$ restricted to $(0, 1)$ is a piecewise polynomial of degree at most k with at most N interior breakpoints, that is, with at most $N + 1 \leq 2N$ polynomial pieces, and is of class C^{k-1} . By Definition 2 the localized class $\mathcal{N}_N^b(\sigma; 1)$ embeds in the same family of piecewise polynomials, the number of pieces of an element of width N being at most $c_0 N$ for a constant c_0 fixed by the generator data (ν_0, W) and independent of N .

In the second stage, the intermediate estimates are structurally bounded by invoking a cited univariate Bernstein inequality. Fix once and for all the index

$$\beta := \frac{1}{2}(\alpha + k + 1), \quad \text{so that} \quad \alpha < \beta < k + 1, \quad (33)$$

and set $\rho := (\beta + 1/p)^{-1}$. For piecewise polynomials of degree at most k with n pieces on $(0, 1)$, with free coefficients, the Bernstein-type inequality

$$|g|_{B_{\rho, \rho}^\beta(\Omega)} \leq C_B n^\beta \|g\|_{L^p(\Omega)}, \quad 0 < \beta < k + 1, \quad \rho = (\beta + 1/p)^{-1}, \quad (34)$$

holds with $C_B = C_B(k, \beta, p)$; this is recorded in [20] (Thm 5.14) (the admissible range there being $p \in (0, \infty)$, which is the reason the present statement is confined to $1 \leq p < \infty$). By Step 1, an element of the localized class of width n_* has at most $c_0 n_*$ pieces, so (34) applies to it with $n = c_0 n_*$. The power of the width thus produced is the distributional-derivative regularization mechanism of Section 2.

Advancing to the third step, we explicitly construct the dyadic increments and enforce their seminorm control. For each $j \geq 0$ choose $g_{2^j} \in \mathcal{N}_{2^j}^b(\sigma; 1)$ with

$$\|f - g_{2^j}\|_{L^p(\Omega)} \leq 2 E_{2^j}^b(f)_{L^p(\Omega)}. \quad (35)$$

The best-approximation numbers are monotone, $E_{2^{j+1}}^b \leq E_{2^j}^b$, so by the triangle inequality with (35),

$$\begin{aligned} \|g_{2^{j+1}} - g_{2^j}\|_{L^p(\Omega)} &\leq \|f - g_{2^{j+1}}\|_{L^p(\Omega)} + \|f - g_{2^j}\|_{L^p(\Omega)} \\ &\leq 2(E_{2^{j+1}}^b + E_{2^j}^b) \leq 4 E_{2^j}^b(f)_{L^p(\Omega)}. \end{aligned} \quad (36)$$

The increment $g_{2^{j+1}} - g_{2^j}$ lies in $\mathcal{N}_{3 \cdot 2^j}^b(\sigma; 1)$ (the class is additively closed with the width adding), hence has at most $3c_0 2^j$ pieces; applying (34) with $n = 3c_0 2^j$ and inserting (36),

$$|g_{2^{j+1}} - g_{2^j}|_{B_{\rho, \rho}^\beta(\Omega)} \leq C_B (3c_0 2^j)^\beta \cdot 4 E_{2^j}^b(f)_{L^p(\Omega)} = C_2 2^{j\beta} E_{2^j}^b(f)_{L^p(\Omega)}, \quad C_2 := 4 C_B (3c_0)^\beta. \quad (37)$$

Write $a_j := 2^{j\alpha} E_{2^j}^b(f)_{L^p(\Omega)}$, so that $\|(a_j)\|_{\ell^\tau} \leq A$ by (31) and $2^{j\beta} E_{2^j}^b = 2^{j(\beta-\alpha)} a_j$. Summing the increments and isolating the base element $g_1 = g_{2^0}$,

$$|g_{2^n}|_{B_{\rho, \rho}^\beta(\Omega)} \leq \sum_{j=0}^{n-1} |g_{2^{j+1}} - g_{2^j}|_{B_{\rho, \rho}^\beta(\Omega)} + |g_1|_{B_{\rho, \rho}^\beta(\Omega)} \leq C_2 \sum_{j=0}^{n-1} 2^{j(\beta-\alpha)} a_j + |g_1|_{B_{\rho, \rho}^\beta(\Omega)}. \quad (38)$$

Let $K(\cdot, t) := K(f, t; L^p(\Omega), B_{\rho, \rho}^\beta(\Omega))$ be the real-interpolation K -functional of the couple. With the dyadic scale $t_n := 2^{-n\beta}$ and the splitting $f = (f - g_{2^n}) + g_{2^n}$ for $n \geq 0$,

$$K(f, 2^{-n\beta}) \leq \|f - g_{2^n}\|_{L^p(\Omega)} + 2^{-n\beta} \|g_{2^n}\|_{B_{\rho, \rho}^\beta(\Omega)}, \quad (39)$$

while for $n < 0$ the trivial splitting gives $K(f, 2^{-n\beta}) \leq K(f, 1) \leq \|f\|_{L^p(\Omega)}$. The discrete equivalent norm of the space $(L^p, B_{\rho, \rho}^\beta)_{\theta, \tau}$ with $\theta := \alpha/\beta$ uses the weight $t_n^{-\theta} = 2^{n\beta\theta} = 2^{n\alpha}$ and the index τ ; we bound its two contributions in ℓ^τ separately.

For the first term of (39), (35) gives $2^{n\alpha} \|f - g_{2^n}\|_{L^p} \leq 2a_n$, hence

$$\left\| (2^{n\alpha} \|f - g_{2^n}\|_{L^p})_{n \geq 0} \right\|_{\ell^\tau} \leq 2 \|(a_n)\|_{\ell^\tau} \leq 2A. \quad (40)$$

For the second term, (38) and $2^{n\alpha} 2^{-n\beta} 2^{j(\beta-\alpha)} = 2^{-(\beta-\alpha)(n-j)}$ yield

$$2^{n\alpha} 2^{-n\beta} \|g_{2^n}\|_{B_{\rho, \rho}^\beta} \leq C_2 \sum_{j=0}^{n-1} 2^{-(\beta-\alpha)(n-j)} a_j + 2^{-(\beta-\alpha)n} \|g_1\|_{B_{\rho, \rho}^\beta}. \quad (41)$$

The first summand is the convolution $(k * a)_n$ with the kernel $k_m := 2^{-(\beta-\alpha)m}$, $m \geq 1$, which is summable since $\beta > \alpha$. For $\tau \geq 1$, Young's inequality gives $\|k * a\|_{\ell^\tau} \leq \|k\|_{\ell^1} \|a\|_{\ell^\tau} = (2^{\beta-\alpha} - 1)^{-1} A$, while for $0 < \tau < 1$, the τ -subadditivity of $|\cdot|^\tau$ gives $\|k * a\|_{\ell^\tau} \leq \|k\|_{\ell^\tau} \|a\|_{\ell^\tau} = (2^{\tau(\beta-\alpha)} - 1)^{-1/\tau} A$; in either case

$$\|k * a\|_{\ell^\tau} \leq \kappa A, \quad \kappa := \|k\|_{\ell^{\min(\tau, 1)}} = \begin{cases} (2^{\beta-\alpha} - 1)^{-1}, & \tau \geq 1, \\ (2^{\tau(\beta-\alpha)} - 1)^{-1/\tau}, & 0 < \tau < 1. \end{cases} \quad (42)$$

The base contribution in (41) is geometric: since $\alpha < \beta$,

$$\left\| (2^{-(\beta-\alpha)n})_{n \geq 0} \right\|_{\ell^\tau} = (1 - 2^{-\tau(\beta-\alpha)})^{-1/\tau}, \quad (43)$$

$$\|g_1\|_{B_{\rho, \rho}^\beta} \leq C_B c_0^\beta \|g_1\|_{L^p} \leq C_B c_0^\beta (\|f\|_{L^p} + 2A), \quad (44)$$

the last step using $\|g_1\|_{L^p} \leq \|f\|_{L^p} + \|f - g_1\|_{L^p} \leq \|f\|_{L^p} + 2E_1^\beta(f)$ and $E_1^\beta(f) = a_0 \leq A$. Finally, the $n < 0$ tail contributes $\|f\|_{L^p} \left(\sum_{n \leq 0} 2^{n\alpha\tau} \right)^{1/\tau} = \|f\|_{L^p} (1 - 2^{-\alpha\tau})^{-1/\tau}$. Collecting (40)–(44) through the ℓ^τ quasi-triangle inequality,

$$\left(\sum_{n \in \mathbb{Z}} [2^{n\alpha} K(f, 2^{-n\beta})]^\tau \right)^{1/\tau} \leq C_5 (\|f\|_{L^p(\Omega)} + A), \quad (45)$$

with $C_5 = C_5(k, \alpha, p)$ assembled explicitly from $C_2, \kappa, c_0, \beta, \tau$. By the discrete characterization of the real-interpolation space, (45) states that $f \in (L^p(\Omega), B_{\rho, \rho}^\beta(\Omega))_{\alpha/\beta, \tau}$ with interpolation norm at most $C_5(\|f\|_{L^p} + A)$.

The diagonal real-interpolation identity for the nonlinear (free-knot) Besov scale, recorded in [20] (Eq. (5.3)) (therein attributed to [23] (Ch. 12, Cor. 8.5)), reads

$$(L^p(\Omega), B_{\rho, \rho}^\beta(\Omega))_{\theta, q} = B_{q, q}^{\theta\beta}(\Omega), \quad q = (\theta\beta + 1/p)^{-1}, \quad 0 < \theta < 1. \quad (46)$$

With $\theta = \alpha/\beta$, one has $\theta\beta = \alpha$ and $q = (\alpha + 1/p)^{-1} = \tau$, so the right-hand side is $B_{\tau, \tau}^\alpha$ and the choice of $\beta \in (\alpha, k + 1)$ does not enter the resulting scale. Combining with (45),

$$\|f\|_{B_{\tau,\tau}^\alpha(\Omega)} \leq C'(\|f\|_{L^p(\Omega)} + A), \quad C' = C'(k, \alpha, p),$$

which is (32). The existence of the near-best elements used at each scale is furnished by the non-polynomial density mechanism (Theorem 1). This completes the proof. $\square$

Remark 7 (Direct and inverse theory). *Theorem 5 and Corollary 1 act in opposite directions on the same approximation process: the direct estimate sends Besov smoothness to a decay rate of best approximation, while the inverse estimate returns an ℓ^τ approximation-space condition as membership on a Besov scale. The two statements are recorded on the two scales natural to each direction: the linear scale $B_{p,q}^s$ for Theorem 5 and the nonlinear (free-knot) diagonal scale $B_{\tau,\tau}^\alpha$, $\frac{1}{\tau} = \alpha + \frac{1}{p}$, for Corollary 1. Since $\alpha > 0$ gives $\tau < p$, the two scales carry different integrability and are kept distinct. The inverse estimate is presented under an ℓ^τ summability hypothesis rather than a single power rate: a bare rate $E_N^p(f) \leq A N^{-\alpha}$ locates f only in the interpolation space $(L^p, B_{p,\rho}^\beta)_{\alpha/\beta, \infty}$, which on this change-of-integrability scale is not identified with a Besov space; the diagonal identity (46) returns a Besov membership precisely on the diagonal fine index $q = \tau$.*

The converse is presented in a single spatial variable, where the elements of $\mathcal{N}_N(\sigma)$ are univariate free-knot piecewise polynomials and the Bernstein-type inequality (34) is available in explicit form [20] (Thm 5.14). A Bernstein-type inequality for the multivariate shallow-network class with unconstrained parameters, which is the ingredient an inverse estimate in dimension $d \geq 2$ would rest upon, is presently an open problem in the approximation theory of neural networks [9] (Sec. 7.1.1, p. 391). For activations other than ReLU^k the reduction of Step 1 to piecewise polynomials no longer applies, so the Bernstein-type inequality does not transfer directly; an extension would rest on a Bernstein-type inequality for the corresponding nonlinear family within the framework of [20]. The present development therefore advances the direct theory in its multivariate generality (Theorem 5) and supplies the converse in the setting where it admits an explicit constant; the additive refinement of the following section operates on the direct side, where the multivariate estimates are available.

The refinement to additive classes rests on a deterministic distribution of the width budget across the coordinate subsets. We isolate this allocation as the following lemma, from which the effective-dimension estimate of Theorem 6 follows.

Lemma 3 (Width allocation and superposition over coordinate subsets). *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary, let σ satisfy Assumption 1 (A_1)–(A_3) together with the localization hypothesis Assumption 2 (A_4), whose generator data contribute the per-term costs v', v'' of the size convention in Definition 2, and let $1 \leq p \leq \infty$, $0 < q \leq \infty$, $0 < s < s_\sigma$. Fix an integer $1 \leq d_0 \leq d$ and write $\mathcal{T} = \{T \subset \{1, \dots, d\} : |T| = d_0\}$, so that $|\mathcal{T}| = \binom{d}{d_0}$. For $T \in \mathcal{T}$ let $T^c = \{1, \dots, d\} \setminus T$, and for $y \in \Omega_T$ let $\Omega_T^y := \{z \in \mathbb{R}^{T^c} : x \in \Omega, x_T = y, x_{T^c} = z\}$ denote the fibre of Ω over y . Put*

$$\kappa := \frac{v' + d v''}{v' + d_0 v''} \in [1, \infty), \quad \Lambda := \begin{cases} \max_{T \in \mathcal{T}} \left(\sup_{y \in \Omega_T} |\Omega_T^y|_{d-d_0} \right)^{1/p}, & 1 \leq p < \infty, \\ 1, & p = \infty, \end{cases}$$

where $|\cdot|_{d-d_0}$ is a $(d - d_0)$ -dimensional Lebesgue measure; both quantities are finite (since Ω is bounded) and independent of N . The per-term costs v', v'' are fixed by the generator data (v_0, W) of Assumption 2 and are independent of N ; hence κ depends on the activation only through (v_0, W) .

Suppose that for each $T \in \mathcal{T}$ there is $g_T \in B_{p,q}^s(\Omega_T)$ and that the direct estimate of Theorem 5, applied on Ω_T in dimension d_0 , furnishes a constant $C_1 = C_1(\sigma, s, p, q, d_0, \Omega_T) > 0$ and an integer $N_1 \geq 1$ with

$$E_n^b(g_T)_{L^p(\Omega_T)} \leq C_1 n^{-s/d_0} |g_T|_{B_{p,q}^s(\Omega_T)}, \quad n \geq N_1. \quad (47)$$

Set

$$C^\sharp := \Lambda C_1 (2\kappa)^{s/d_0} \binom{d}{d_0}^{s/d_0}, \quad N_0 := 2\kappa \binom{d}{d_0} N_1 \left(\geq N_1 \binom{d}{d_0} \right).$$

Then, for every integer $N \geq N_0$ there is an allocation $\{N_T\}_{T \in \mathcal{T}}$ of nonnegative integers with $\sum_{T \in \mathcal{T}} N_T \leq N$ such that

$$E_N^b \left(\sum_{T \in \mathcal{T}} g_T \right)_{L^p(\Omega)} \leq C^\sharp N^{-s/d_0} \sum_{T \in \mathcal{T}} |g_T|_{B_{p,q}^s(\Omega_T)}. \quad (48)$$

The exponent $-s/d_0$ is governed by the subset dimension d_0 ; the ambient dimension d enters (48) only through the constant $C^\sharp$, via the multiplicity $\binom{d}{d_0}$ and the embedding factor κ .

Proof of Lemma 3. Throughout, each g_T is read both as an element of $B_{p,q}^s(\Omega_T)$ and as the function $x \mapsto g_T(x_T)$ on Ω , which depends on x only through the sub-vector x_T . We argue in four steps: an explicit embedding of a d_0 -variable localized approximant into the ambient localized class, the transfer of the $L^p(\Omega_T)$ error to $L^p(\Omega)$, the deterministic width allocation, and the superposition with attribution of the exponent.

As zeroth step, we explicitly embed a d_0 -active component into the ambient function class. Fix $T \in \mathcal{T}$ and let $\phi \in \mathcal{N}_n^b(\sigma; d_0)$ be any competitor for $E_n^b(g_T)_{L^p(\Omega_T)}$, so that, by Definition 2,

$$\phi(y) = \sum_{j=1}^M h_j(y) \prod_{l \in T} \psi_{j,l}(y_l), \quad h_j \in \mathcal{N}_{\nu'}(\sigma; d_0), \quad \psi_{j,l} \in \mathcal{N}_{\nu''}(\sigma; 1), \quad M(\nu' + d_0\nu'') \leq n,$$

the univariate factors being indexed by the coordinates of T . Define $\tilde{\phi}(x) := \phi(x_T)$ on Ω as follows. Each ridge unit of $h_j(y) = \sum_{i=1}^{\nu'} a_i \sigma(w_i \cdot y + b_i)$, $w_i \in \mathbb{R}^{d_0}$, is zero-padded to $\tilde{w}_i \in \mathbb{R}^d$ by setting $(\tilde{w}_i)_l = (w_i)_l$ for $l \in T$ and $(\tilde{w}_i)_l = 0$ for $l \in T^c$; then, $\tilde{w}_i \cdot x = w_i \cdot x_T$, so the resulting $\tilde{h}_j \in \mathcal{N}_{\nu'}(\sigma; d)$ has the same number ν' of ridge units and satisfies $\tilde{h}_j(x) = h_j(x_T)$. Completing the tensor structure to all d coordinates by the admissible trivial factors $\tilde{\psi}_{j,l} := \psi_{j,l}$ for $l \in T$ and $\tilde{\psi}_{j,l} \equiv 1$ for $l \in T^c$ (the constant unit being admissible by the embedding convention of Definition 2), we obtain

$$\tilde{\phi}(x) = \sum_{j=1}^M \tilde{h}_j(x) \prod_{l=1}^d \tilde{\psi}_{j,l}(x_l) \in \mathcal{N}_m^b(\sigma; d), \quad m := M(\nu' + d\nu'').$$

The size inflates from $M(\nu' + d_0\nu'') \leq n$ to

$$m = M(\nu' + d\nu'') = M(\nu' + d_0\nu'') \cdot \frac{\nu' + d\nu''}{\nu' + d_0\nu''} \leq \kappa n, \quad (49)$$

the factor κ being the deterministic cost of carrying a d_0 -active component in the d -coordinate localized class; it is independent of N and reduces to $\kappa = 1$ when $d_0 = d$.

Moving into the primary analysis (step 1), we establish a transparent per-component approximation that facilitates the subsequent error transfer. Apply (47) on Ω_T in dimension d_0 : for $n \geq N_1$, choose a competitor $\phi_T \in \mathcal{N}_n^b(\sigma; d_0)$ with

$$\|g_T - \phi_T\|_{L^p(\Omega_T)} \leq 2C_1 n^{-s/d_0} |g_T|_{B_{p,q}^s(\Omega_T)} \quad (50)$$

(any factor exceeding the infimum suffices; the displayed slack is absorbed below). Let $\tilde{\phi}_T(x) = \phi_T(x_T) \in \mathcal{N}_{m_T}^{\phi}(\sigma; d)$ be its embedding from zeroth step, with $m_T \leq \kappa n$. Since $g_T(x) - \tilde{\phi}_T(x) = (g_T - \phi_T)(x_T)$ depends on x only through x_T , Tonelli's theorem applied to the nonnegative integrand $|(g_T - \phi_T)(x_T)|^p$ gives, for $1 \leq p < \infty$,

$$\|g_T - \tilde{\phi}_T\|_{L^p(\Omega)}^p = \int_{\Omega_T} |(g_T - \phi_T)(y)|^p |\Omega_T^y|_{d-d_0} dy \leq \Lambda^p \|g_T - \phi_T\|_{L^p(\Omega_T)}^p,$$

so that $\|g_T - \tilde{\phi}_T\|_{L^p(\Omega)} \leq \Lambda \|g_T - \phi_T\|_{L^p(\Omega_T)}$; for $p = \infty$, the same bound holds with $\Lambda = 1$, since the essential supremum over Ω of a function of x_T equals that over Ω_T . Combining with (50),

$$\|g_T - \tilde{\phi}_T\|_{L^p(\Omega)} \leq 2 \Lambda C_1 n^{-s/d_0} |g_T|_{B_{p,q}^s(\Omega_T)}, \quad n \geq N_1. \quad (51)$$

Progressing to the second phase, we employ a width allocation, favoring mathematical parsimony over arbitrary parameter scaling. For $N \geq N_0$, allocate equally,

$$n := \left\lfloor \frac{N}{\kappa \binom{d}{d_0}} \right\rfloor, \quad N_T := m_T \leq \kappa n \quad (T \in \mathcal{T}).$$

The total budget is admissible: using $\lfloor t \rfloor \leq t$ and (49),

$$\sum_{T \in \mathcal{T}} N_T \leq \binom{d}{d_0} \kappa n = \kappa \binom{d}{d_0} \left\lfloor \frac{N}{\kappa \binom{d}{d_0}} \right\rfloor \leq \kappa \binom{d}{d_0} \cdot \frac{N}{\kappa \binom{d}{d_0}} = N.$$

The threshold is met: for $N \geq N_0 = 2\kappa \binom{d}{d_0} N_1$, one has $N / (\kappa \binom{d}{d_0}) \geq 2N_1 \geq 1$; hence, by the elementary bound $\lfloor t \rfloor \geq t/2$ for $t \geq 1$,

$$n \geq \frac{1}{2} \cdot \frac{N}{\kappa \binom{d}{d_0}} \geq N_1, \quad (52)$$

so (47) is applicable for this n on every Ω_T .

In this third and concluding step, the superposition is systematically evaluated so that the corresponding exponent is attributed. By the zeroth step, the budgets are additive over $\mathcal{T}$. For each $T \in \mathcal{T}$, the embedded competitor $\tilde{\phi}_T$ of step 1, written out as a member of $\mathcal{N}_{m_T}^{\phi}(\sigma; d)$, is

$$\tilde{\phi}_T(x) = \sum_{j=1}^{M_T} \tilde{h}_{T,j}(x) \prod_{l=1}^d \tilde{\psi}_{T,j,l}(x_l), \quad M_T(v' + dv'') = m_T,$$

its blocks and factors being those produced by the zeroth-step embedding from the data of the d_0 -variable competitor $\phi_T(y) = \sum_{j=1}^{M_T} h_{T,j}(y) \prod_{l \in T} \psi_{T,j,l}(y_l)$, where $h_{T,j}(y) = \sum_{i=1}^{v'} a_{T,j,i} \sigma(w_{T,j,i} \cdot y + b_{T,j,i})$ with $w_{T,j,i} \in \mathbb{R}^{d_0}$. Explicitly, the zero-padded ridge block and the trivially completed tensor factors are

$$\tilde{h}_{T,j}(x) = \sum_{i=1}^{v'} a_{T,j,i} \sigma(\tilde{w}_{T,j,i} \cdot x + b_{T,j,i}) \in \mathcal{N}_{v'}(\sigma; d), \quad (\tilde{w}_{T,j,i})_l = \begin{cases} (w_{T,j,i})_l, & l \in T, \\ 0, & l \in T^c, \end{cases}$$

$$\tilde{\psi}_{T,j,l} = \begin{cases} \psi_{T,j,l}, & l \in T, \\ 1, & l \in T^c, \end{cases} \quad \psi_{T,j,l} \in \mathcal{N}_{v''}(\sigma; 1) \quad (l \in T).$$

The factors indexed by $l \in T^c$ equal 1 in value, so that $\tilde{\phi}_T(x) = \phi_T(x_T)$ as recorded in step 1; under the size convention of Definition 2, however, each of the d univariate factors is counted at cost v'' , whence every term costs $v' + dv''$ and the M_T terms contribute the per-component budget $M_T(v' + dv'') = m_T$. The budgets are thus additive across $\mathcal{T}$, and

$$\Phi := \sum_{T \in \mathcal{T}} \tilde{\phi}_T \in \mathcal{N}_{\sum_T m_T}^b(\sigma; d) \subseteq \mathcal{N}_N^b(\sigma; d),$$

because $\sum_T m_T = \left(\sum_T M_T \right) (v' + dv'') \leq N$ by the second phase, the common per-term cost $v' + dv''$ factoring out of the sum. Hence, Φ is an admissible competitor for

$$E_N^b \left(\sum_T g_T \right)_{L^p(\Omega)}, \text{ and by the triangle inequality in } L^p(\Omega) \text{ together with (51),}$$

$$E_N^b \left(\sum_T g_T \right)_{L^p(\Omega)} \leq \left\| \sum_T g_T - \Phi \right\|_{L^p(\Omega)} \leq \sum_{T \in \mathcal{T}} \|g_T - \tilde{\phi}_T\|_{L^p(\Omega)} \leq 2\Lambda C_1 n^{-s/d_0} \sum_{T \in \mathcal{T}} |g_T|_{B_{p,q}^s(\Omega_T)}.$$

It remains to convert the per-component budget n into the total budget N . The lower bound (52) gives

$$n^{-s/d_0} \leq \left(\frac{N}{2\kappa \binom{d}{d_0}} \right)^{-s/d_0} = (2\kappa)^{s/d_0} \left(\frac{d}{d_0} \right)^{s/d_0} N^{-s/d_0},$$

and substituting,

$$E_N^b \left(\sum_T g_T \right)_{L^p(\Omega)} \leq \underbrace{\Lambda C_1 (2\kappa)^{s/d_0} \left(\frac{d}{d_0} \right)^{s/d_0}}_{=C^\sharp} N^{-s/d_0} \sum_{T \in \mathcal{T}} |g_T|_{B_{p,q}^s(\Omega_T)},$$

where the absolute factor 2 from (50) has been absorbed into $C^\sharp$.

The dependence on the ambient dimension d is confined to the constant: the allocation (52) is linear in N , the multiplicity $\binom{d}{d_0}$ and the embedding factor κ entering only as N -independent divisors, so substitution preserves the per-component exponent $-s/d_0$ supplied by (47) in dimension d_0 and produces $\left(\frac{d}{d_0} \right)^{s/d_0}$ and $(2\kappa)^{s/d_0}$ as constant prefactors. No step contributes a power of N other than the per-component rate together with the linear allocation; consequently, the exponent $-s/d_0$ is governed by the subset dimension d_0 , while d acts only through $C^\sharp$. When $d_0 = d$ one has $\binom{d}{d} = 1$, $\kappa = 1$, $\Lambda = 1$, and the allocation $n = N$ involves no rounding, so $C^\sharp$ reduces to C_1 and (48) recovers Theorem 5. $\square$

Theorem 6 (Direct estimate for the coordinate-subset additive class). *Let $\Omega = \prod_{i=1}^d \Omega_i \subset \mathbb{R}^d$ be a bounded product domain with each $\Omega_i \subset \mathbb{R}$ a bounded open interval. Let σ satisfy Assumption 1 (A_1)–(A_3) together with Assumption 2 (A_4), let $1 \leq p \leq \infty$, $0 < q \leq \infty$, $0 < s < s_\sigma$, and fix an integer $1 \leq d_0 \leq d$. Then, there exist a constant $C'' = C''(\sigma, s, p, q, d, d_0) > 0$ and an integer $N_0'' \geq 1$ such that, for every $f \in \mathcal{S}_{d_0}^s(\Omega)$ and every integer $N \geq N_0''$,*

$$E_N^b(f)_{L^p(\Omega)} \leq C'' N^{-s/d_0} |f|_{\mathcal{S}_{d_0}^s(\Omega)}. \quad (53)$$

The exponent $-s/d_0$ is governed by the effective dimension d_0 ; the ambient dimension d enters only through C'' , via $\binom{d}{d_0}$ and the factor κ recorded below. For $d_0 = d$, the estimate (53) reduces to the rate $N^{-s/d}$ of Theorem 5.

Proof of Theorem 6. The argument proceeds in four steps: discharge of the domain regularity required on each coordinate-subset face, the per-component direct estimate, the deterministic width allocation across subsets, and passage to the additive seminorm by an infimum over admissible decompositions. The degenerate case $d_0 = d$ is recovered at the end. The mechanism that produces the improved exponent is the coordinate-subset decomposition; every other ingredient is carried, as a black box, by the two results established earlier in the paper.

To establish a transparent baseline, the zeroth step formally discharges the requisite domain regularity across the coordinate-subset faces. Fix $T \subset \{1, \dots, d\}$ with $|T| = d_0$. Since $\Omega = \prod_{i=1}^d \Omega_i$ is a product of bounded open intervals, the projection onto the coordinates indexed by T is the product of the corresponding factors,

$$\Omega_T = \prod_{i \in T} \Omega_i \subset \mathbb{R}^{d_0}, \quad (54)$$

which is a bounded box and is therefore a bounded Lipschitz domain in $\mathbb{R}^{d_0}$. Identity (54) furnishes the domain regularity under which both Theorem 5 and Lemma 3 are applicable on each Ω_T . The product structure further renders the fibre $\Omega_T^y = \prod_{i \in T^c} \Omega_i$ independent of $y \in \Omega_T$ for each T , so the supremum over $y \in \Omega_T$ in the geometric factor of Lemma 3 reduces to the corresponding product, and the lemma aggregates these across subsets by the maximum,

$$\Lambda = \max_{|T|=d_0} \left(\prod_{i \in T^c} |\Omega_i| \right)^{1/p} \quad (1 \leq p < \infty), \quad \Lambda = 1 \quad (p = \infty), \quad (55)$$

finite and independent of N .

Transitioning to the first principal stage, the per-component direct estimate is established via the earlier results. Let $f = \sum_{|T|=d_0} g_T(x_T)$ be any admissible decomposition, with $g_T \in B_{p,q}^s(\Omega_T)$ as in Definition 5. For each T , Theorem 5, applied on the domain Ω_T in dimension d_0 (legitimate by zeroth step), supplies a constant $C_1 = C_1(\sigma, s, p, q, d_0, \Omega_T) > 0$ and an integer $N_1 \geq 1$ with

$$E_n^b(g_T)_{L^p(\Omega_T)} \leq C_1 n^{-s/d_0} |g_T|_{B_{p,q}^s(\Omega_T)}, \quad n \geq N_1. \quad (56)$$

Estimate (56) is precisely the per-component hypothesis required by Lemma 3: a single-face direct bound at the rate n^{-s/d_0} against the face seminorm.

Advancing to the second step, we explicitly execute the deterministic width allocation. With (56) in force on every face, apply Lemma 3. Writing

$$\kappa := \frac{v' + d v''}{v' + d_0 v''} \in [1, \infty), \quad (57)$$

($\kappa \geq 1$ because $d_0 \leq d$ and $v', v'' \geq 0$), and setting

$$C^\sharp := \Lambda C_1 (2\kappa)^{s/d_0} \left(\frac{d}{d_0} \right)^{s/d_0}, \quad N_0 := 2\kappa \left(\frac{d}{d_0} \right) N_1, \quad (58)$$

the lemma yields, for every integer $N \geq N_0$, an allocation $\{N_T\}_{|T|=d_0}$ of nonnegative integers with $\sum_{|T|=d_0} N_T \leq N$ such that

$$E_N^b \left(\sum_{|T|=d_0} g_T \right)_{L^p(\Omega)} \leq C^\sharp N^{-s/d_0} \sum_{|T|=d_0} |g_T|_{B_{p,q}^s(\Omega_T)}. \quad (59)$$

Because $f = \sum_{|T|=d_0} g_T$ for the chosen decomposition, the left side of (59) equals $E_N^b(f)_{L^p(\Omega)}$.

Penultimately in the third phase, the coordinate-subset decomposition transparently facilitates the passage to the additive seminorm, isolating the structural origin of the improved exponent. The bound (59) holds for the fixed but arbitrary admissible decomposition $f = \sum_{|T|=d_0} g_T$ of stage 1. The prefactor $C^\sharp N^{-s/d_0}$ is a nonnegative constant independent of that decomposition, hence it may be drawn outside an infimum taken over the right side. Passing to the infimum over all admissible decompositions $f = \sum_{|T|=d_0} g_T$ gives

$$E_N^b(f)_{L^p(\Omega)} \leq C^\sharp N^{-s/d_0} \inf \left\{ \sum_{|T|=d_0} |g_T|_{B_{p,q}^s(\Omega_T)} : f = \sum_{|T|=d_0} g_T, g_T \in B_{p,q}^s(\Omega_T) \right\}. \quad (60)$$

By the seminorm of Definition 5, the infimum on the right of (60) is exactly $|f|_{\mathcal{S}_{d_0}^s(\Omega)}$. Setting $C'' := C^\sharp$ and $N_0'' := \lceil N_0 \rceil$ (an integer) establishes (53). The infimum acts on the seminorm factor alone and leaves both the constant $C^\sharp$ and the power N^{-s/d_0} unchanged.

As the fourth and final step, we directly recover the estimates for the degenerate case $d_0 = d$. For $d_0 = d$ the index set reduces to the single subset $T = \{1, \dots, d\}$, so $\binom{d}{d} = 1$ and, by (57), $\kappa = 1$; the complementary set T^c is empty, whence $\Lambda = 1$ by (55) and $\Omega_T = \Omega$ by (54). With a single coordinate subset no allocation across faces is invoked, and Definition 5 gives $\mathcal{S}_d^s(\Omega) = B_{p,q}^s(\Omega)$ with $|f|_{\mathcal{S}_d^s(\Omega)} = |f|_{B_{p,q}^s(\Omega)}$. Applying Theorem 5 directly on Ω in dimension d therefore yields

$$E_N^b(f)_{L^p(\Omega)} \leq C_1 N^{-s/d} |f|_{B_{p,q}^s(\Omega)}, \quad N \geq N_1,$$

which is the estimate of Theorem 5. The ambient-dimension rate is thereby recovered as the single-subset specialization of (53). $\square$

4. Constructive Validation: Effective-Dimension Decay Rates

The theoretical results of the preceding sections are the principal contribution of this paper; the present section is an illustrative validation of their constructive content. Theorem 5 predicts the ambient-dimension decay $E_N^b(f)_{L^2(\Omega)} \sim N^{-s/d}$, while Theorem 6 predicts the effective-dimension decay $E_N^b(f)_{L^2(\Omega)} \sim N^{-s/d_0}$ on the additive class $\mathcal{S}_{d_0}^s(\Omega)$. We exhibit a deterministic construction whose measured decay exponents are on par with these two predictions, and a held-out diagnostic on educational tabular data whose effective dimension is consistent with the modeling object of Theorem 6.

4.1. Schema of the Localized Shallow Network and Parameter Configuration

The localized shallow construction used in this section is summarized in Figure 1. The input $x \in \Omega = \prod_{i=1}^d \Omega_i \subset \mathbb{R}^3$ is processed by the σ -realized localized system of Definition 2: each univariate factor $\psi_{j,l}$ is the linear B-spline (hat function), the finite-difference realization of the ReLU activation at $k = 1$, and the ridge units are assembled into the localized basis

of Lemma 1; the output is a linear read-out of that basis, obtained by a deterministic least-squares projection. Two coordinate-structured branches are compared at ambient dimension $d = 3$: the additive construction ($d_0 = 1$), which approximates the target coordinatewise at width $N = d n_1$, and the full construction ($d_0 = d$), which employs a tensor grid at width $N = n_1^d$. Throughout, N denotes the dimension of the localized system (the basis count), the constructive surrogate for the network width, and not a count of trained units.

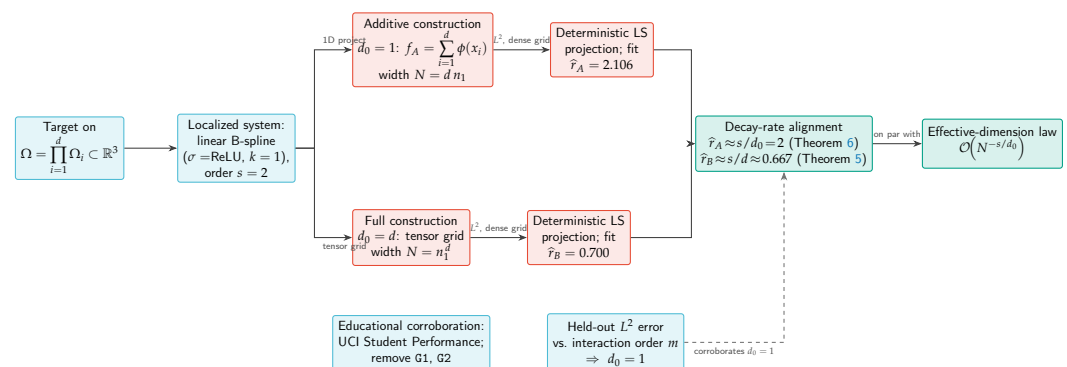


Figure 1. Schema of the localized shallow construction. The localized system is the linear B-spline, the finite-difference realization of the ReLU activation at $k = 1$, of approximation order $s = 2$ on the product domain $\Omega \subset \mathbb{R}^3$. The construction proceeds along two coordinate-structured branches: the additive construction ($d_0 = 1$) approximates the target coordinatewise at width $N = d n_1$, while the full construction ($d_0 = d$) employs a tensor grid at width $N = n_1^d$; in both, N denotes the dimension of the localized system rather than a count of trained units, and the approximant is obtained by a deterministic least-squares projection on a dense evaluation grid. The fitted decay exponents $\hat{r}_A = 2.106$ and $\hat{r}_B = 0.700$ are on par with the effective-dimension rate $s/d_0 = 2$ of Theorem 6 and the ambient rate $s/d \approx 0.667$ of Theorem 5, respectively. The educational corroboration (dashed link) estimates the effective dimension of the response by the held-out L^2 error as a function of the admitted interaction order m , returning $d_0 = 1$.

We recorded the parameter configuration of the construction. Because the approximant was obtained by a deterministic least-squares projection, with no iterative training and no restart variance, the construction carried no trained hyperparameters in the sense of an optimized model; the settings in Table 1 were fixed inputs of the localized system. The ambient dimension was $d = 3$ and the system order was $s = 2$, an axiomatic-level property of the linear B-spline system held fixed and not inferred from the measured decay. The error was measured in $L^2(\Omega)$ ($p = 2$) on a dense evaluation grid, and reproducibility was fixed by the deterministic seed 20260601. The additive branch used the width grid $N \in \{6, 12, 24, 48, 96, 192\}$ ($N = d n_1$) and the full branch used $N \in \{8, 27, 64, 125, 216, 343\}$ ($N = n_1^d$); the corresponding fitted exponents and their comparison with the predicted rates are reported in Table 1 and Figure 1.

4.2. The Constructive Scheme

The validation was carried out by an explicit, coordinatewise construction rather than by a trained network, so that the measured exponent reflected the constructive upper bound itself and not the contingent endpoint of an optimization. The localized system was the linear B-spline (hat function), which is the finite-difference realization of the ReLU activation at $k = 1$ admitted by Assumptions 1 and 2; its approximation order was s_σ with the associated value $r = 2$, so the present experiment probed the range up to second order. Two constructions were compared at ambient dimension $d = 3$. The additive construction ($d_0 = 1$) approximated the target coordinatewise, $f_A(x) = \sum_{i=1}^d \phi(x_i)$, and consumed width

$N = d \cdot n_1$; this is the transparent, constructive analogue of the per-component step of Theorem 6. The full construction ($d_0 = d$) approximated a genuinely d -variate target on a full tensor grid and consumed width $N = n_1^d$. Both used a deterministic least-squares projection; in the notation of Theorem 6, they instantiated the per-component estimate (Step 1), the deterministic width allocation (Step 2), and the degenerate single-subset case (Step 4), with the black-box interface to Theorem 5 and Lemma 3 left intact. We emphasize that the width N here denotes the dimension of the localized system, i.e., the basis count, which is the constructive surrogate for the network width of Definition 2.

Table 1. Parameter configuration of the constructive validation. The settings are fixed inputs of the deterministic localized system; the approximant in each branch is a deterministic least-squares projection on a dense evaluation grid, with no iterative training and no restart variance. The system order $s = 2$ is held fixed and is not inferred from the measured decay.

Parameter/Component	Configuration Value
Ambient dimension	$d = 3$
Domain	$\Omega = \prod_{i=1}^d \Omega_i$, bounded product of intervals
Localized system	Linear B-spline (hat function), ReLU $k = 1$ realization
System order s	$s = 2$ (axiomatic-level; held fixed)
Error metric	$L^2(\Omega)$ ($p = 2$), dense evaluation grid
Projection	Deterministic least-squares (no iterative training)
Random seed	20260601 (deterministic reproducibility)
Additive branch ($d_0 = 1$)	Width $N = d n_1$, $N \in \{6, 12, 24, 48, 96, 192\}$
Full branch ($d_0 = d$)	Width $N = n_1^d$, $N \in \{8, 27, 64, 125, 216, 343\}$

4.3. Protocol and the Choice of $p = 2$

All errors were computed at ambient dimension $d = 3$ under a single deterministic least-squares projection (seed 20260601), evaluated on a dense grid, with no iterative training and no restart variance; the widths N are provided in Table 2. The design matrix was assembled by evaluating the localized B-spline system on a dense evaluation grid over Ω ; the coefficient vector was the solution of the associated normal equations, computed once under seed 20260601 with no iterative training and no restart variance. The additive construction ($d_0 = 1$) used the coordinatewise system $\{\phi(x_i)\}_{i=1}^d$, and the full construction ($d_0 = d$) used the tensor system on the full grid; in both cases the reported error is the L^2 discrepancy on the same dense grid. We instantiated the $L^p(\Omega)$ error of Theorems 5 and 6 at $p = 2$. Beyond its coincidence with the root-mean-square error, this value renders the empirical functional a transparent, inner-product-induced quantity, keeping the numerical diagnostic on the same analytic footing as the seminorms $|\cdot|_{B_{p,q}^s}$ that the theory constrains, rather than treating it as a black-box objective.

4.4. Decay-Rate Alignment

The measured errors are collected in Table 2 and displayed in Figure 2. A least-squares fit of $\log \hat{E}_N$ against $\log N$ returned the decay exponent $\hat{r}_A = 2.106$ for the additive construction, on par with the prediction $s/d_0 = 2$ of Theorem 6, and $\hat{r}_B = 0.700$ for the full construction, on par with the rate $s/d \approx 0.667$ of Theorem 5 recovered at $d_0 = d$. The additive construction decayed systematically faster than the full construction across the entire range of widths; that separation is the constructive signature of the mechanism of Theorem 6, namely, that the decay exponent is governed by the effective dimension d_0 and not by the ambient dimension d , the latter entering only through the constant C'' via the

multiplicity $\binom{d}{d_0}$ and the factor κ . The full construction, the single-subset specialization $d_0 = d$, serves as the consistency checkpoint against the ambient rate of Theorem 5; the two reference slopes $-s/d_0$ and $-s/d$ in Figure 2 are anchored at the leading datum of each series, and the fitted slopes track them across the decade in N .

Table 2. Constructive approximation error $\hat{E}_N$ in the L^2 metric for the additive construction ($d_0 = 1$) and the full construction ($d_0 = d$), at ambient dimension $d = 3$, under the linear (ReLU, $k = 1$) localized system of approximation order $s = 2$. The width N denotes the dimension of the localized system. The fitted decay exponents are $\hat{r}_A = 2.106$, on par with the prediction $s/d_0 = 2$ of Theorem 6, and $\hat{r}_B = 0.700$, on par with the rate $s/d \approx 0.667$ of Theorem 5 recovered at $d_0 = d$.

Additive Construction ($d_0 = 1$)		Full Construction ($d_0 = d$)	
N	$\hat{E}_N$	N	$\hat{E}_N$
6	0.9815	8	0.3463
12	0.6330	27	0.2939
24	0.1107	64	0.1291
48	0.0192	125	0.0694
96	0.0043	216	0.0477
192	0.0010	343	0.0278

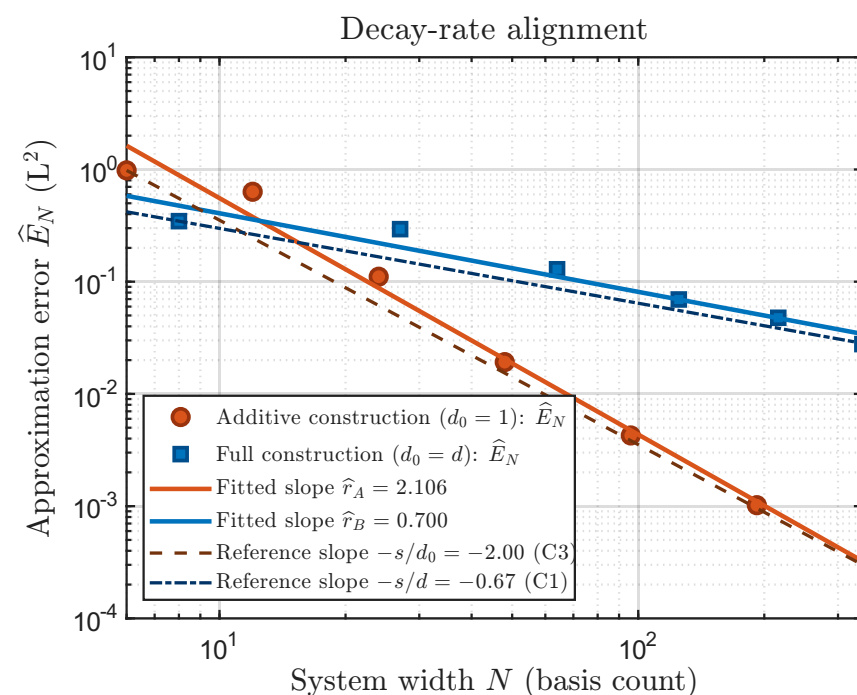


Figure 2. Decay-rate alignment of the constructive approximant. The additive construction ($d_0 = 1$) decays at the fitted exponent $\hat{r}_A = 2.106$, on par with $s/d_0 = 2$; the full construction ($d_0 = d$) decays at $\hat{r}_B = 0.700$, on par with $s/d \approx 0.667$. Reference slopes $-s/d_0$ and $-s/d$ are anchored at the leading datum of each series.

4.5. Educational Corroboration

To connect the additive class with applied tabular data, we estimated the effective dimension of an educational response. On the UCI Student Performance data, with the response taken to be the final grade and the two intermediate grades G1, G2 discarded so that the task was nontrivial, we recorded the held-out L^2 error as a function of the admitted interaction order m . The error was systematically governed by a single coordinate, the selected effective dimension being $d_0 = 1$ (Figure 3); this is consistent with the modeling

object of the additive class $\mathcal{S}_{d_0}^s(\Omega)$ of Definition 5. The relevance of a transparent, low-effective-dimension construction in this setting is in keeping with the place of explainable and feature-parsimonious models in educational prediction [14,17], where principled structure is preferred to the massive parameterization of a black-box predictor. The data are publicly available under the CC BY 4.0 license with DOI 10.24432/C5TG7T.

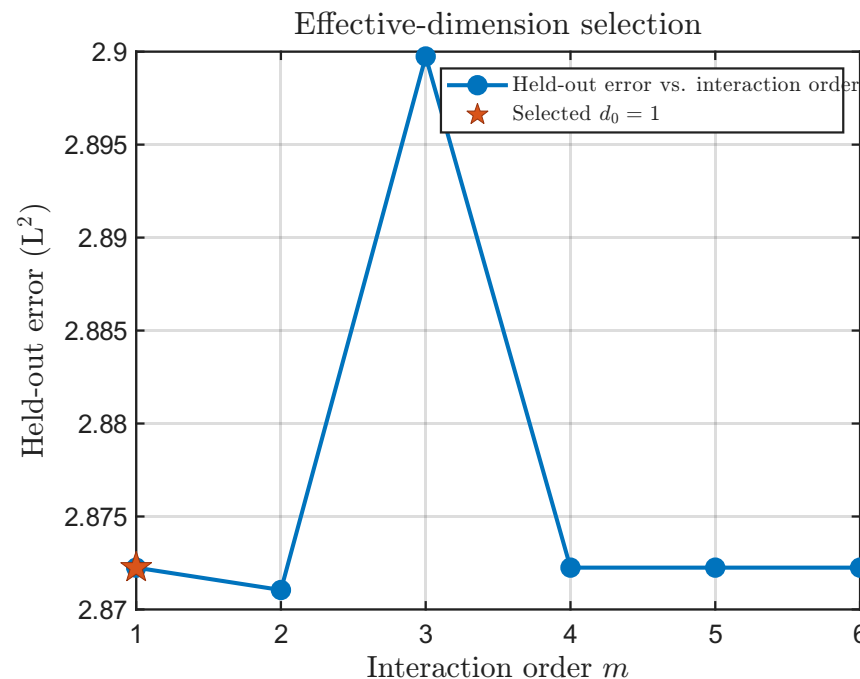


Figure 3. Effective-dimension selection on the educational response. The held-out L^2 error as a function of the admitted interaction order m is systematically governed by a single coordinate, the selected effective dimension being $d_0 = 1$.

Author Contributions: Conceptualization, Z.C. and K.-C.W.; methodology, K.-C.W.; formal analysis, K.-C.W.; investigation, K.-C.W.; resources, Z.C.; data curation, K.-C.W.; writing—original draft preparation, K.-C.W.; writing—review and editing, Z.C. and K.-C.W.; supervision, Z.C.; project administration, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The empirical foundation supporting the reported findings relies on publicly accessible records. The original data are systematically archived in the UCI Machine Learning Repository and can be explicitly accessed via <https://doi.org/10.24432/C5TG7T> under the designation of the Student Performance dataset.

Acknowledgments: The authors extend mutual gratitude for the constructive theoretical dialogue that shaped this work. Additionally, institutional support from the Department of Mathematics and Applied Mathematics, School of Information Engineering, Sanming University, and the IOT Application Engineering Research Center of Fujian Province Colleges and Universities is systematically recognized for providing a structurally stable research environment.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pinkus, A. Approximation Theory of the MLP Model in Neural Networks. *Acta Numer.* **1999**, *8*, 143–195.
2. Pinkus, A. *Ridge Functions*; Cambridge University Press: Cambridge, UK, 2015.
3. Siegel, J.W.; Xu, J. Approximation Rates for Neural Networks with General Activation Functions. *Neural Netw.* **2020**, *128*, 313–321. <https://doi.org/10.1016/j.neunet.2020.05.019>.
4. Siegel, J.W.; Xu, J. High-Order Approximation Rates for Shallow Neural Networks with Cosine and ReLU^k Activation Functions. *Appl. Comput. Harmon. Anal.* **2022**, *58*, 1–26. <https://doi.org/10.1016/j.acha.2021.12.005>.
5. Siegel, J.W.; Xu, J. Sharp Bounds on the Approximation Rates, Metric Entropy, and n-Widths of Shallow Neural Networks. *Found. Comput. Math.* **2024**, *24*, 481–537. <https://doi.org/10.1007/s10208-022-09595-3>.
6. DeVore, R.; Nowak, R.D.; Parhi, R.; Siegel, J.W. Weighted Variation Spaces and Approximation by Shallow ReLU Networks. *Appl. Comput. Harmon. Anal.* **2025**, *74*, 101713. <https://doi.org/10.1016/j.acha.2024.101713>.
7. Parhi, R.; Nowak, R.D. Banach Space Representer Theorems for Neural Networks and Ridge Splines. *J. Mach. Learn. Res.* **2021**, *22*, 1960–1999.
8. Siegel, J.W. Optimal Approximation Rates for Deep ReLU Neural Networks on Sobolev and Besov Spaces. *J. Mach. Learn. Res.* **2023**, *24*, 17197–17248.
9. DeVore, R.; Hanin, B.; Petrova, G. Neural Network Approximation. *Acta Numer.* **2021**, *30*, 327–444. <https://doi.org/10.1017/S0962492921000052>.
10. Schmidt-Hieber, J. Nonparametric Regression Using Deep Neural Networks with ReLU Activation Function. *Ann. Stat.* **2020**, *48*, 1875–1897. <https://doi.org/10.1214/19-AOS1875>.
11. Kohler, M.; Langer, S. On the Rate of Convergence of Fully Connected Deep Neural Network Regression Estimates. *Ann. Stat.* **2021**, *49*, 2231–2249. <https://doi.org/10.1214/20-AOS2034>.
12. Nagler, J.; Langer, S. Optimal Neural Network Approximation of Smooth Compositional Functions on Sets with Low Intrinsic Dimension. *arXiv* **2026**, arXiv:2602.03539. <http://arxiv.org/abs/2602.03539>.
13. Mhaskar, H.N.; Mao, T. Tractability of Approximation by General Shallow Networks. *Anal. Appl.* **2024**, *22*, 535–568. <https://doi.org/10.1142/S0219530524400013>.
14. Yağcı, M. Educational Data Mining: Prediction of Students' Academic Performance Using Machine Learning Algorithms. *Smart Learn. Environ.* **2022**, *9*, 11. <https://doi.org/10.1186/s40561-022-00192-z>.
15. Jang, Y.; Choi, S.; Jung, H.; Kim, H. Practical Early Prediction of Students' Performance Using Machine Learning and eXplainable AI. *Educ. Inf. Technol.* **2022**, *27*, 12855–12889. <https://doi.org/10.1007/s10639-022-11120-6>.
16. Deho, O.B.; Joksimovic, S.; Li, J.; Zhan, C.; Liu, J.; Liu, L. Should Learning Analytics Models Include Sensitive Attributes? Explaining the Why. *IEEE Trans. Learn. Technol.* **2023**, *16*, 560–572. <https://doi.org/10.1109/TLT.2022.3226474>.
17. Khosravi, H.; Buckingham Shum, S.; Chen, G.; Conati, C.; Tsai, Y.S.; Kay, J.; Knight, S.; Maldonado, R.M.; Sadiq, S.W.; Gašević, D. Explainable Artificial Intelligence in Education. *Comput. Educ. Artif. Intell.* **2022**, *3*, 100074.
18. Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; Leanpub: Victoria, BC, Canada, 2022.
19. Mishra, T.; Singh, K.K.; Ayman-Mursaleen, M.; Rao, N. Novel Fractional Schurer Type Operators with Explicit Error Bounds and Applications. *J. Inequalities Appl.* **2026**, *2026*, 70. <https://doi.org/10.1186/s13660-026-03481-0>.
20. Gribonval, R.; Kutyniok, G.; Nielsen, M.; Voigtlaender, F. Approximation Spaces of Deep Neural Networks. *Constr. Approx.* **2022**, *55*, 259–367. <https://doi.org/10.1007/s00365-021-09543-4>.
21. Triebel, H. *Theory of Function Spaces II*; Monographs in Mathematics 84; Birkhäuser: Basel, Switzerland, 1992.
22. Triebel, H. *Function Spaces and Wavelets on Domains*; European Mathematical Society: Helsinki, Finland, 2008.
23. DeVore, R.A.; Lorentz, G.G. Constructive Approximation. In *Grundlehren Der Mathematischen Wissenschaften*; Springer: Berlin/Heidelberg, Germany; New York, NY, USA, 1993; Volume 303.
24. Rychkov, V.S. On Restrictions and Extensions of the Besov and Triebel-Lizorkin Spaces with Respect to Lipschitz Domains. *J. Lond. Math. Soc.* **1999**, *60*, 237–257.
25. Lu, J.; Huang, X. Approximation Rates of Shallow Neural Networks: Barron Spaces, Activation Functions and Optimality Analysis. *arXiv* **2025**, arXiv:2510.18388. <http://arxiv.org/abs/2510.18388>.
26. Liu, X.; Mao, T.; Xu, J. Integral Representations of Sobolev Spaces via ReLU Activation Function and Optimal Error Estimates for Linearized Networks. *arXiv* **2025**, arXiv:2505.00351. <http://arxiv.org/abs/2505.00351>.
27. Mao, T.; Siegel, J.W.; Xu, J. Approximation Rates for Shallow ReLU^k Neural Networks on Sobolev Spaces via the Radon Transform. *SIAM J. Math. Anal.* **2024**, *58*. <https://doi.org/10.1137/24M1686693>.

28. Johnen, H.; Scherer, K. On the Equivalence of the K-functional and Moduli of Continuity and Some Applications. In *Proceedings of the Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach 25 April–1 May 1976*; Lecture Notes in Mathematics; Schempp, W., Zeller, K., Eds.; Springer: Berlin/Heidelberg, Germany, 1977; Volume 571, pp. 119–140.
29. Brenner, S.C.; Scott, L.R. *The Mathematical Theory of Finite Element Methods*, 3rd ed.; Texts in Applied Mathematics; Springer: New York, NY, USA, 2008; Volume 15.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.