APPROXIMATION THEORY FOR TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We establish the first comprehensive approximation theory for positional transformers that parallels classical neural network approximation theory. (i) We prove universal approximation uniformly on compact sets; in L^p_μ; uniformly on compact sets together with derivatives up to order k; in weighted Sobolev norms. (ii) We establish L^p, Sobolev, C^k, and Barron approximation rates. (iii) The L^p rate is optimal under continuous parameter selection. All these results follow from an exact realization theorem: we show that a positional transformer acting on n tokens realizes suitable families of neural networks. This allows us to transfer classical neural network approximation theory, preserving approximation-rate exponents.

1 INTRODUCTION

1.1 MAIN RESULTS

Approximation theory for transformers remains fundamentally incomplete: existing results address particular architectures, target classes, or error criteria, while many foundational questions are still open. This contrasts with the extensive approximation theory available for classical neural networks (e.g. Hornik, 1991; Pinkus, 1999). In this manuscript, we consider positional encoder transformers acting on n tokens and prove: the four universal approximation theorems of Hornik; L^p, Sobolev, C^k, and Barron rates; optimality under continuous parameter selection. *To the best of our knowledge, this is the first comprehensive approximation theory for a transformer class that parallels the classical theory for neural networks*. We refer to Section 1.2 for a literature review and a comparison with our work.

Exact realization and transfer principle. All our results rest on an exact realization theorem (Theorem 3.2): a positional transformer acting on n tokens in $\mathbb{R}^d$, with enough attention heads and model width, realizes suitable families of shallow neural networks exactly as soon as its number Lq of feed-forward hidden units (L is the depth of the transformer and q the hidden width of the feed-forward network in each block) is at least their hidden width.[1] Classical density results and approximation rates thus transfer to sequence-to-sequence maps.

Universal Approximation. We first transfer the four Universal Approximation Theorems (UATs) of Hornik (1991) within our transformer class: compact-uniform and arbitrary-finite-measure L^p density, compact-open C^k density, and weighted Sobolev spaces (Section 4). Our UATs also apply to encoder–transformer–decoder architectures in operator learning (Remark 4.5).

Approximation rates. The quantitative consequences include L^p rates ($(Lq)^{-s/(nd)}$), Sobolev rates ($(Lq)^{-(s-k)/(nd)}$), C^k rates for Hölder targets ($(Lq)^{-(r+\alpha-k)/(nd)}$), and spectral Barron rates ($(Lq)^{-1/2}$) (Section 5). Here s is the Sobolev smoothness order; k is the highest derivative order measured in the approximation error; $r+\alpha$ is the Hölder smoothness, with integer $r \geq 0$ and $0 < \alpha \leq 1$. Since the number of parameters is proportional to Lq, these are rates in the number of parameters, and they have the same exponents as the source shallow-network bounds.

Optimality under continuous selection. Finally, for every smooth nonpolynomial activation, the transfer mechanism yields an affine, L^p-continuous parameter selector attaining the rate $P^{-s/(nd)}$

[1] In particular, a single encoder block suffices.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Section 6). The classical continuous-information lower bound makes this exponent optimal among all methods using P continuously selected real parameters.

Transformers with deep feed-forward sublayers. We also consider transformers whose feed-forward sublayers have several hidden layers: the exact realization theorem extends to them, and approximation results for deep networks transfer in the same way (Appendix C).

1.2 Relation to the state of the art

Exact realization and transfer. Exact realizations and transfer principles were established in Petersen & Voigtlaender (2020) for fully connected and convolutional networks. Gu et al. (2026) provide a conversion to transformers through exact hardmax realizations and compact-uniform softmax approximations of maxout and ReLU networks. Exact MLP-to-transformer conversions on a single token, without positional encoding, were established in Kratsios et al. (2026, Proposition A.1) (see also Kratsios & Furuya, 2025, and Hong et al., 2026, Proposition C.3 for ReLU activation). Our Theorem 3.2 is proved for ordered multi-token inputs and output-position labels using positional residual encoder blocks with depth L, feed-forward hidden width q, and number H of heads.

Universal Approximation. *Uniform approximation on compact sets.* Jiao et al. (2026a); Lai & Sun (2026) prove uniform approximation of Hölder targets on $[0,1]^{d\times n}$ by positional transformers with ReLU feed-forward networks.[2] To the best of our knowledge, these results leave open compact-uniform universality for positional residual encoders with general continuous nonpolynomial feed-forward activations. Our Theorem 4.1 fills this gap by transferring the shallow-network compact-open universal approximation theorems (Pinkus, 1999, Theorem 3.1) and (Hornik, 1991, Theorem 2).

Integral approximation in L^p_μ. For fixed-length positional transformers, Yun et al. (2020a) establish universality using fixed-width, variable-depth networks with ReLU feed-forward layers (see also Yang et al., 2023); Kajitsuka & Sato (2024) use one softmax attention layer between two ReLU feed-forward networks. Cheng et al. (2025) give a broader criterion requiring an affine-invariant feed-forward family containing a non-affine Lipschitz map and token-distinguishing attention, with unrestricted depth. These results concern continuous targets on compact domains in Lebesgue L^p.[3] Jiao et al. (2026a) also give L^p approximation bounds on compact domains for Hölder and $W^{1,p}$ targets. To the best of our knowledge, existing transformer universality results leave open universality under arbitrary finite Borel input measures and merely measurable activation functions. Our Theorem 4.2 establishes this result for positional transformers by transferring Hornik (1991, Theorem 1).

Derivative approximation. The preceding function-value approximation results leave open whether transformers can simultaneously approximate sufficiently smooth maps and their input derivatives. To the best of our knowledge, Theorems 4.3 and 4.4 provide the first UATs for positional transformers that simultaneously approximate arbitrary ordered sequence functions and their input derivatives in both the compact-open C^k topology and weighted derivative seminorms for arbitrary finite input measures.

To the best of our knowledge, Theorems 4.1 to 4.4 give the first complete analogue UATs to the four classical theorems of Hornik (1991) for transformers.

Rates and optimality. *Function-value rates.* For fixed-length positional transformers, the input dimension is nd. Writing P for the parameter budget, Lai & Sun (2026) gives an L^p error bound $O(P^{-(s+\lambda)/(nd)})$ for $C^{s,\lambda}$ targets, using deep ReLU feed-forward networks. Jiao et al. (2026a) gives

[2] Further compact-uniform universal approximation results are obtained via positional softmax-attention approximations of maxout and ReLU networks (Gu et al., 2026; Hu et al., 2026c), whose constructions add an auxiliary token. Other compact-uniform results include softmax-attention universality with additional sequence-wise linear maps (Hu et al., 2026b, Theorem D.1) and Hölder approximation with an accuracy-dependent number of latent tokens (Shi & Liao, 2026).

[3] Related results cover sparse positional transformers (Yun et al., 2020b; Zaheer et al., 2020), attention-only models (Hu et al., 2026b, Theorem G.1), set and equivariant models (Lee et al., 2019; Alberti et al., 2023; Selby et al., 2022), modified architectures (Luo et al., 2022; Ramapuram et al., 2025; Liu et al., 2025; Jiao et al., 2026b), and prompting, in-context, operator, or finite-vocabulary settings (Furuya et al., 2025; 2026b;a; Wang et al., 2023; Petrov et al., 2024; Hu et al., 2025; 2026a; Hsu & Lai, 2026; Ma et al., 2025).

L^p error bounds $O(P^{-\alpha/(nd)})$ for α-Hölder targets and $O(P^{-1/(nd)})$ for $W^{1,p}$ targets. For α-Hölder targets, He et al. (2026, Theorem 4.1) gives an L^2 error bound $O(P^{-\alpha/(nd)})$, using bounded-width ReLU blocks and one active softmax attention layer.[4] However, these cited results leave open the L^p rate $P^{-s/(nd)}$ for $W^{s,p}$ targets for positional transformers with general smooth nonpolynomial activations. Our Theorem 5.3 with $k = 0$ establishes this rate.

Derivative and Barron rates. To the best of our knowledge, rates for simultaneous approximation of sufficiently smooth target maps and their derivatives via transformers remain open. Our results in Section 5 provide the first Sobolev, C^k, and Barron derivative-approximation rates for transformers by transferring classical neural networks rates.

Optimality. Despite existing upper bounds, questions about optimality of transformer approximation rates remain, to the best of our knowledge, open.[5] Our Theorem 6.1 resolves one such question: it attains the function-value rate $P^{-s/(nd)}$ with an affine, L^p-continuous parameter selector, matching the classical continuous-selection lower bound (DeVore et al., 1989; Pinkus, 1999).

In view of the above comparisons, this manuscript establishes, to the best of our knowledge, the first comprehensive approximation theory for a transformer class that parallels the classical theory for neural networks.

Outline. Section 2 defines the transformer class; Section 3 states the exact realization theorem; Section 4 develops universal approximation theorems; Section 5 develops quantitative approximation rates; Section 6 establishes optimality under continuous selection; Appendix A introduces notations (token spaces and approximation norms); Appendix B collects the proofs, grouped by main-text section; Appendix C treats transformers with deep feed-forward sublayers.

2 Transformer architecture

Transformers were introduced in Vaswani et al. (2017). We use sequence-to-sequence positional transformers with unmasked softmax attention and, as in Kim et al. (2023); Kajitsuka & Sato (2025); Jiao et al. (2026a) (see also Yun et al., 2020a), without layer normalization.

Definition 2.1 (Positional transformer)**.** Fix the sequence length n, the token dimension d, an activation function $\sigma : \mathbb{R} \to \mathbb{R}$, and positive integers w (model width, i.e. the embedding dimension), H (number of attention heads), t (head dimension), q (feed-forward hidden width), and L (depth). Tokens are represented as columns.

(i) Embedding layer. Choose $E_{\mathrm{in}} \in \mathbb{R}^{w\times d}$ and positional vectors $p_1, \dots, p_n \in \mathbb{R}^w$ and let $P_{\mathrm{pos}} \coloneqq (p_1, \dots, p_n) \in \mathbb{R}^{w\times n}$. For an input $X \in \mathbb{R}^{d\times n}$, define

$$Z(X) = (z_1(X), \dots, z_n(X)) \coloneqq E_{\mathrm{in}}X + P_{\mathrm{pos}} \in \mathbb{R}^{w\times n}.$$

(ii) Encoder block. Choose query, key, and value projections, an output projection, and feed-forward parameters $W_Q^h, W_K^h, W_V^h \in \mathbb{R}^{t\times w}$ $(1 \le h \le H)$, $W_O \in \mathbb{R}^{w\times Ht}$, $B_1 \in \mathbb{R}^{q\times w}$, $B_2 \in \mathbb{R}^{w\times q}$, $b_1 \in \mathbb{R}^q$, and $b_2 \in \mathbb{R}^w$. For $Z \in \mathbb{R}^{w\times n}$, write $z_i \coloneqq Z_{:,i}$. Define the attention matrix $\Pi^{(h)}(Z) \in \mathbb{R}^{n\times n}$,

[4] Other rates concern pairwise coupling of temporal features (Jiang & Li, 2024), generalized D-retrieval tasks (Yu et al., 2026), long but sparse memories (Wang & E, 2024), column-symmetric polynomials (Takeshita & Imaizumi, 2025), permutation-equivariant Hölder functions (Siegel et al., 2026), infinite-dimensional inputs (Takakura & Suzuki, 2023), or low-dimensional manifolds with uniform rates depending on intrinsic dimension (Havrilla & Liao, 2024); other constructions use ReLU and floor activations and bias matrices with different columns (Jiao et al., 2026b).

[5] Lower bounds for transformer approximation exist but do not match the known upper bounds; see He et al. (2026); Yu et al. (2026).

the output of head h, the multi-head attention output, and its residual update, respectively, by

$$\Pi_{ij}^{(h)}(Z) \coloneqq \big[\mathrm{softmax}\big((\langle W_Q^h z_i, W_K^h z_\ell\rangle)_{\ell=1}^n\big)\big]_j = \frac{\exp\langle W_Q^h z_i, W_K^h z_j\rangle}{\sum_{\ell=1}^n \exp\langle W_Q^h z_i, W_K^h z_\ell\rangle}, \quad 1 \le i, j \le n,$$

$$\mathrm{head}_h(Z) \coloneqq W_V^h Z \left(\Pi^{(h)}(Z)\right)^\top \in \mathbb{R}^{t\times n}, \qquad [\mathrm{head}_h(Z)]_{:,i} = \sum_{j=1}^n \Pi_{ij}^{(h)}(Z) W_V^h z_j,$$

$$\mathrm{MHA}(Z) \coloneqq W_O \begin{pmatrix} \mathrm{head}_1(Z) \\ \vdots \\ \mathrm{head}_H(Z) \end{pmatrix} \in \mathbb{R}^{w\times n},$$

$$U(Z) \coloneqq Z + \mathrm{MHA}(Z) \in \mathbb{R}^{w\times n}, \qquad [U(Z)]_{:,i} = z_i + [\mathrm{MHA}(Z)]_{:,i}.$$

For $Y \in \mathbb{R}^{w\times n}$, define the feed-forward map $\mathrm{FF}(Y) \in \mathbb{R}^{w\times n}$ by

$$\mathrm{FF}(Y) \coloneqq B_2\sigma^\bullet(B_1 Y + b_1\mathbf{1}_n^\top) + b_2\mathbf{1}_n^\top, \quad [\mathrm{FF}(Y)]_{:,i} = B_2\sigma^\bullet(B_1 Y_{:,i} + b_1) + b_2, \quad 1 \le i \le n.$$

where $\sigma^\bullet$ applies σ entrywise (Appendix A), and $\mathbf{1}_n \in \mathbb{R}^n$ is the column vector of n ones. The encoder block $\mathcal{B}_\sigma : \mathbb{R}^{w\times n} \to \mathbb{R}^{w\times n}$ is

$$\mathcal{B}_\sigma(Z) \coloneqq U(Z) + \mathrm{FF}(U(Z)).$$

(iii) Readout. Choose $R_d \in \mathbb{R}^{d\times w}$, shared across all output tokens; the readout is $Y \mapsto R_d Y \in \mathbb{R}^{d\times n}$ for $Y \in \mathbb{R}^{w\times n}$.

(iv) Positional transformer. Let $\mathcal{B}_{(1)}, \ldots, \mathcal{B}_{(L)}$ be L blocks as in (ii), with the common dimensions w, H, t, q and independent parameters (indexed by the block, $W_{O,(u)}, B_{1,(u)}, \ldots$, when the block matters), and put

$$T(X) \coloneqq R_d\big(\mathcal{B}_{(L)} \circ \cdots \circ \mathcal{B}_{(1)}\big)\big(Z(X)\big) \in \mathbb{R}^{d\times n}.$$

The class of transformers $\mathcal{T}_{n,d}^{\mathrm{PE}}(\sigma)$ consists of all maps $T : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ so obtained, taking the union over all positive integers w, L, H, t, q and all real parameters of the stated shapes.

An *architecture* is a tuple $\mathfrak{a} = (w, L, H, t, q)$ of positive integers: the model width, depth, number of heads, head dimension, and feed-forward width in Definition 2.1. It leaves all weights and biases, including the positional vectors, unspecified; their number is[6]

$$P_{\mathfrak{a}} = w(2d + n) + L(4Htw + 2qw + q + w). \tag{1}$$

We write $T_{\mathfrak{a}}(\vartheta)$ for the transformer of architecture $\mathfrak{a}$ with parameters $\vartheta \in \mathbb{R}^{P_{\mathfrak{a}}}$, and $\mathcal{T}_{\mathfrak{a}}(\sigma) \coloneqq \{T_{\mathfrak{a}}(\vartheta) : \vartheta \in \mathbb{R}^{P_{\mathfrak{a}}}\}$, so that $\mathcal{T}_{n,d}^{\mathrm{PE}}(\sigma) = \bigcup_{\mathfrak{a}} \mathcal{T}_{\mathfrak{a}}(\sigma)$.

3 Exact realization

In this section, we state an exact realization theorem for transformers (Theorem 3.2), the mechanism on which our approximation theory is built. The reader interested only in the approximation results can go directly to Sections 4 to 6, reading only Definitions 3.1 and 3.6.

3.1 The exact realization theorem

Given a target $f : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$, $f(X) = (f_1(X), \ldots, f_n(X))$, we seek a transformer whose ith output token approximates $f_i(X) \in \mathbb{R}^d$ for all i. Classical approximation theory provides shallow-network approximants to the vector-valued map f. To transfer approximation to our transformer architecture, we construct approximants in a form compatible with its shared tokenwise feed-forward layer and readout: a family sharing input and output weights, with hidden biases depending affinely on position (see also Remark 3.4). The arguments in Appendices B.2 and B.3 show that this restriction

[6] The contributions are: input embedding, wd; positional vectors, nw; readout, dw; and, per block, query, key, and value matrices, $3Htw$; attention output matrix, Htw; feed-forward matrices, $2qw$; feed-forward biases, $q + w$.

preserves universality and the stated rate exponents. We then realize these families exactly, placing $\Phi_i(\operatorname{vec} X)$ at token i (Theorem 3.2); the positional vectors supply the position dependence internally.

For a scalar activation $\sigma : \mathbb{R} \to \mathbb{R}$, define the class of shallow network families with shared weights (recall that $\sigma^\bullet$ applies σ coordinatewise; see Appendix A)

$$\mathcal{S}_{n,d}(\sigma) := \bigcup_{m \geq 1} \{\mathbf{\Phi} = (\Phi_i)_{i=1}^n : \Phi_i(\xi) = C\sigma^\bullet(A\xi + Di + b) + c\}, \tag{2}$$

where $A \in \mathbb{R}^{m \times nd}$, $D, b \in \mathbb{R}^m$, $C \in \mathbb{R}^{d \times m}$, and $c \in \mathbb{R}^d$ are shared across positions; m is the common hidden width. For $\mathbf{\Phi} \in \mathcal{S}_{n,d}(\sigma)$, write its matrix-valued evaluation as

$$\mathbf{\Phi}(\xi) := (\Phi_1(\xi) \quad \cdots \quad \Phi_n(\xi)) \in \mathbb{R}^{d \times n}.$$

Equivalently, use the ordinary network $\Phi(\xi, \eta) = C\sigma^\bullet(A\xi + D\eta + b) + c$ on $\mathbb{R}^{nd} \times \mathbb{R}$ and set $\Phi_i(\xi) = \Phi(\xi, i)$.

We briefly explain how the exact realization of Theorem 3.2 works; Appendix B.1 gives a more detailed explanation of this idea and the full proof. We partition the hidden units of the shared network family into groups, each evaluated in the feed-forward sublayer of its own block. Their contributions are added to an output block through the residual connections. Fixed positional attention gives an invertible mixing of the input tokens, from which we obtain either the ordered input or the input-dependent terms in the current group's preactivations. The number and dimension of the attention heads determine how many blocks the attention computation requires, while the working width and feed-forward hidden width determine how many hidden units can be evaluated at each stage. The following definition records the resulting effective depth and width.

Definition 3.1 (Effective resources)**.** Let g be the dimension of a positional encoding,[7] and let $L, q, H, t \geq 1$ and $w \geq 2d + g + 1$ be the depth, feed-forward hidden width, number of heads, head dimension, and model width of an architecture. Its *required head count* $H_* := n\lceil d/t \rceil$ is the number of head evaluations used by the construction in the proof of Theorem 3.2 to gather the n tokens; its *stage length* $G := \lceil H_*/H \rceil$ is the number of blocks over which they are spread; its *working width* is $s_w := w - 2d - g$; its *effective width* is $q_* := q$ if $s_w \geq nd$ and $q_* := \min\{q, s_w\}$ otherwise; and its *effective depth* is $L_* := L - G + 1$ if $s_w \geq nd$ and $L_* := \lfloor L/G \rfloor$ otherwise.

The proof of the following result is given in Appendix B.1.

Theorem 3.2 (Exact realization)**.** *Fix $n \geq 2$, $d \geq 1$, and an arbitrary scalar function $\sigma : \mathbb{R} \to \mathbb{R}$. Let $\mathbf{\Phi} \in \mathcal{S}_{n,d}(\sigma)$ be represented with common hidden width m, and let L, q, H, t, w be as in Definition 3.1. If $L_* q_* \geq m$, there exists $T_{\mathbf{\Phi}} \in \mathcal{T}_{\mathfrak{a}}(\sigma)$, $\mathfrak{a} = (w, L, H, t, q)$, such that*

$$T_{\mathbf{\Phi}}(X) = (\Phi_1(\operatorname{vec} X) \quad \cdots \quad \Phi_n(\operatorname{vec} X)) \qquad \text{for every } X \in \mathbb{R}^{d \times n}. \tag{3}$$

The input embedding, positional vectors, query, key, and value projections, and readout can be chosen to depend only on n, d, q, H, t, w and the positional encoding. For fixed m, L, q, H, t, w, the attention output projections and the first feed-forward matrices and biases depend affinely on the matrices A, D, b representing the family in (2)*, and the second feed-forward matrices and biases depend linearly on C, c.*

Remark 3.3 (The condition $L_* q_* \geq m$)**.** The condition holds for sufficiently large L, with the other architecture dimensions fixed. When $s_w \geq nd$ and $L \geq G$, it becomes $(L - G + 1)q \geq m$; increasing either L or q then suffices, so depth and feed-forward width can be traded while preserving exact realization.

Remark 3.4 (Tokenwise output representation)**.** The label i identifies token i: the construction stores $\Phi_i(\operatorname{vec} X)$ in the last d coordinates of token i, so the same readout R_d extracts the output at every position.

Remark 3.5 (Scope of fixed attention)**.** The construction in the proof of Theorem 3.2 uses attention weights that are independent of the input content and the target function; only the attention output projections may depend on the target, and only when $s_w < nd$ (for the architectures of Definition 3.6, when $q < nd$). This fixed attention suffices to transfer the universal approximation theorems and upper approximation rates for shallow neural networks presented in Sections 4 and 5. Potential efficiency gains from input-dependent attention for structured targets lie beyond this construction.

[7] See Definition B.1; $g = n$ for the one-hot encoding, $g = 2$ for the Vandermonde encoding, and $g = 3$ for the Gaussian encoding. We call the last two *compact*, since their dimension does not depend on n.

3.2 ARCHITECTURES $\mathcal{A}$ AND PARAMETER COUNTS

In Sections 4 to 6 we use the architectures in which the construction of Theorem 3.2 applies ($L_* \geq 1$), uses all feed-forward units ($q_* = q$), and has no heads or head coordinates beyond those it needs. For these architectures the number of parameters is proportional to Lq (Remark 3.7), so that rates in Lq are rates in the number of parameters. This leads us to the set $\mathcal{A}$ below.

Definition 3.6 (The architectures $\mathcal{A}$)**.** Fix $g \in \{1, \dots, n\}$, the dimension of the positional encoding used in Theorem 3.2 (Definition B.1). Let $\mathcal{A}$ be the set of architectures $\mathfrak{a} = (w, L, H, t, q)$ with $t \leq d$, $H \leq H_* = n\lceil d/t \rceil$ (Definition 3.1), $L \geq G = \lceil H_*/H \rceil$ (that is, $L_* \geq 1$), and $w = 2d + g + \min\{q, nd\}$ (so that $q_* = q$).

For $\mathfrak{a} \in \mathcal{A}$, Theorem 3.2 realizes in $\mathcal{T}_{\mathfrak{a}}(\sigma)$ every family of hidden width $m \leq L_* q$.

Remark 3.7 (Parameter counts)**.** Let $\mathfrak{a} \in \mathcal{A}$; its number $P_{\mathfrak{a}}$ of parameters is given by (1) with $w = 2d + g + \min\{q, nd\}$. The construction in the proof of Theorem 3.2 is sparse: its number of nonzero entries is at most[8]

$$P_{\mathrm{nz}}(L, q) \coloneqq Lq\,(nd + d + g + 2) + L\,(nd + 2gH_*) + n^2 d + gn + 3d. \tag{4}$$

Since $2d + g + 1 \leq w \leq 2d + n + nd$ and $Ht \leq n\lceil d/t \rceil t \leq 2nd$,

$$2(2d + g + 1)\,Lq \leq P_{\mathfrak{a}} \leq C_{n,d}\,Lq, \qquad C_{n,d} \coloneqq (2d + n + nd)(2d + n + 8nd + 3) + 1, \tag{5}$$

so an approximation rate in Lq has the same exponent in the dense and nonzero budgets, uniformly in the split of Lq between L and q and in H.

4 UNIVERSAL APPROXIMATION THEOREMS

We can now transfer Theorems 1–4 of Hornik (1991). The proofs of the following results are given in Appendix B.2. By Theorem 3.2, the approximants in each theorem below can be taken in $\mathcal{T}_{\mathfrak{a}}(\sigma)$ for every $\mathfrak{a} \in \mathcal{A}$ with Lq large enough. In particular, H and q may be fixed arbitrarily (hence also the model width $2d + g + \min\{q, nd\}$) and only the depth L grows, or $L = 1$, $H = H_*$ and only q grows; with a single head, a fixed q, and a compact encoding, all block dimensions are independent of n.

Theorem 4.1 (Compact-uniform universal approximation)**.** *Suppose σ is continuous and nonpolynomial. For every compact $K \subset \mathbb{R}^{d\times n}$, the restrictions of $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$ to K are dense in $C^0(K; \mathbb{R}^{d\times n})$ in the uniform norm. Equivalently, for every $f \in C^0(K; \mathbb{R}^{d\times n})$ and $\varepsilon > 0$ there is $T \in \mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$ such that $\|f - T\|_{C^0_K} < \varepsilon$.*

Theorem 4.2 (Universal approximation in L^p_μ)**.** *Let $1 \leq p < \infty$ and let μ be any finite Borel measure on $\mathbb{R}^{d\times n}$. If $\sigma : \mathbb{R} \to \mathbb{R}$ is Borel measurable, bounded, and not almost everywhere equal to a constant (with respect to Lebesgue measure), then $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma) \cap L^p_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ is dense in $L^p_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$. Thus, for every $f \in L^p_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ and every $\varepsilon > 0$, there is a positional transformer $T \in \mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$ such that $\|f - T\|_{L^p_\mu} < \varepsilon$.*

Theorem 4.3 (Compact-uniform C^k universal approximation)**.** *Let $k \in \mathbb{N}_0$ and suppose $\sigma \in C^k(\mathbb{R}; \mathbb{R})$ is bounded and nonconstant. Then:*

- *(a) $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$ is dense in $C^k(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ for the compact-open C^k topology. Explicitly, for every $f \in C^k(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$, compact K, and $\varepsilon > 0$, there exists $T \in \mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$ such that $\|f - T\|_{C^k_K} < \varepsilon$.*
- *(b) If $1 \leq p < \infty$ and μ is a finite Borel measure with compact support, then the transformer class is dense in $C^{k,p}_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ with respect to* (8)*. Explicitly, for every $f \in C^{k,p}_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ and $\varepsilon > 0$, there exists $T \in \mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma) \cap C^{k,p}_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ such that $\|f - T\|_{C^{k,p}_\mu} < \varepsilon$.*

[8] Embedding, readout, positional vectors and the final bias contribute at most $gn + 3d$ entries. For $H = H_*$, each block contributes at most $nd + 2gH_*$ query, key and value entries, and $q(nd + d + g + 2)$ feed-forward and attention-output entries, plus at most $n^2 d$ entries for the initial gathering when $q \geq nd$. The same total bound holds when the heads are distributed over G blocks.

Theorem 4.4 (Universal approximation in weighted Sobolev norms)**.** *Let $k \in \mathbb{N}_0$, $1 \le p < \infty$, and let μ be any finite Borel measure on $\mathbb{R}^{d\times n}$. Suppose $\sigma \in C^k(\mathbb{R};\mathbb{R})$ is nonconstant and $\sigma^{(j)} \in L^\infty(\mathbb{R};\mathbb{R})$, $j = 0, 1, \dots, k$. Then $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma) \cap C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ is dense in $C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ with respect to the weighted derivative seminorm* (8)*. Explicitly, for every $f \in C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ and $\varepsilon > 0$, there exists $T \in \mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma) \cap C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ such that $\|f - T\|_{C^{k,p}_\mu} < \varepsilon$.*

Remark 4.5 (Application to operator learning)**.** Our UATs for transformers can be applied to encoder–transformer–decoder architectures (Shih et al., 2025) in operator learning thanks to Castro (2023); Cohen et al. (2026); de Feo (2026); Lanthaler et al. (2022); Saini (2026); Ismailov (2026). For instance, Theorems 4.1, 4.3 and 4.4 verify Assumptions 3.1 and 3.2 of de Feo (2026). Consequently, the operator UATs in de Feo (2026) apply to encoder–transformer–decoder architectures, yielding approximation of nonlinear operators and their derivatives up to order k.

5 Approximation rates

In this section, we transfer approximation rates. Throughout, $\mathfrak{a} = (w, L, H, t, q) \in \mathcal{A}$ (Definition 3.6). Let $B_R \coloneqq \{X \in \mathbb{R}^{d\times n} : \|X\|_{\mathrm{F}} \le R\}$ for fixed $R > 0$. The constant $C > 0$ may depend on the fixed dimensions n, d, radius R, activation σ, and regularity and integrability indices s, k, r, α, p, ℓ, but not on f and $\mathfrak{a}$; additional dependencies are indicated by subscripts.

Quantitative transfer. By (5), the number of parameters of $\mathfrak{a} \in \mathcal{A}$ is proportional to Lq, with constants depending only on n, d. Hence every rate below, stated in Lq, has the same exponent in the number of parameters.

5.1 L^p, Sobolev, and C^k rates

The Sobolev and Hölder norms are defined in Appendix A. The proofs of the following results are given in Appendix B.3.

Theorem 5.1 (L^p rate)**.** *Let $\ell \in \mathbb{N}_0$ and let s be an integer with $1 \le s \le \ell + (nd+1)/2$. Assume either*

(i) $\sigma(u) = u_+^\ell$, $\ell \ge 1$, and $2 \le p \le \infty$;

(ii) $p = 2$ and σ is continuous or piecewise continuous, with $\lim_{u\to-\infty} \sigma(u)/u^\ell = 0$ and $\lim_{u\to+\infty} \sigma(u)/u^\ell = 1$.

There exists $C > 0$ such that, for every $f \in W^{s,p}(B_R;\mathbb{R}^{d\times n})$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying

$$\|f - T\|_{L^p} \le C(Lq)^{-s/(nd)} \|f\|_{W^{s,p}}.$$

Theorem 5.2 (L^∞ rate)**.** *Let $\sigma(u) = u_+^\ell$, where $\ell \ge 1$ is an integer, and put $s = \ell + (nd+1)/2$. There exists $C > 0$ such that, for every $f \in W^{s,2}(B_R;\mathbb{R}^{d\times n})$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying*

$$\|f - T\|_{L^\infty} \le C(Lq)^{-s/(nd)} \|f\|_{W^{s,2}} = C(Lq)^{-\frac{1}{2}-\frac{2\ell+1}{2nd}} \|f\|_{W^{s,2}}.$$

Theorem 5.3 (Sobolev rate)**.** *Let $\sigma \in C^\infty(\mathbb{R};\mathbb{R})$ be nonpolynomial, let $s, k \in \mathbb{N}_0$ satisfy $k < s$, and let $1 < p < \infty$. There exists $C > 0$ such that, for every $f \in W^{s,p}(B_R;\mathbb{R}^{d\times n})$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying*

$$\|f - T\|_{W^{k,p}} \le C(Lq)^{-(s-k)/(nd)} \|f\|_{W^{s,p}}.$$

Theorem 5.4 (C^k rate)**.** *Let $\sigma \in C^\infty(\mathbb{R};\mathbb{R})$ be nonpolynomial, let $r \in \mathbb{N}_0$ and $0 < \alpha \le 1$, and let k be an integer with $0 \le k \le r$. There exists $C > 0$ such that, for every $f \in C^{r,\alpha}(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying*

$$\|f - T\|_{C^k_{B_R}} \le C(Lq)^{-(r+\alpha-k)/(nd)} \|f\|_{C^{r,\alpha}}.$$

Theorem 5.4 yields the following corollary by integrating the uniform derivative bound on B_R.

Corollary 5.5 (Weighted Sobolev rate). *Let $\sigma \in C^\infty(\mathbb{R};\mathbb{R})$ be nonpolynomial, let $r \in \mathbb{N}_0$ and $0 < \alpha \le 1$, and let k be an integer with $0 \le k \le r$. Let μ be a finite Borel measure on $\mathbb{R}^{d\times n}$ supported in B_R. There exists $C > 0$ such that, for every $f \in C^{r,\alpha}(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ such that, for every $1 \le p < \infty$,*[9]

$$\|f - T\|_{C^{k,p}_\mu} \le C\mu(B_R)^{1/p}(Lq)^{-(r+\alpha-k)/(nd)}\,\|f\|_{C^{r,\alpha}}\,,$$
$$\|f - T\|_{L^p_\mu} \le C\mu(B_R)^{1/p}(Lq)^{-(r+\alpha)/(nd)}\,\|f\|_{C^{r,\alpha}} \qquad (k=0).$$

5.2 Barron rates

We also transfer Barron's rates. The proofs of the following results are given in Appendix B.3.

Theorem 5.6 (Barron L^2_μ rate). *Let μ be a Borel probability measure on B_R, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and Borel measurable, with limits 0 and 1 at $-\infty$ and $+\infty$, respectively. Suppose $f : B_R \to \mathbb{R}^{d\times n}$ has scalar entries $f_{a,i}$ satisfying*

$$f_{a,i}(x) - f_{a,i}(0) = \int_{\mathbb{R}^{d\times n}} (e^{\mathrm{i}\langle\omega,x\rangle_{\mathrm{F}}} - 1)\,dF_{a,i}(\omega), \qquad C_F := \sum_{a=1}^{d}\sum_{i=1}^{n}\int \|\omega\|_{\mathrm{F}}\,d|F_{a,i}|(\omega) < \infty.$$

Here the $F_{a,i}$ are complex measures and the entries $f_{a,i}$ are real. There is a constant $C_{n,d,R} > 0$ such that for every $\mathfrak{a} \in \mathcal{A}$ there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ with[10]

$$\|f - T\|_{L^2_\mu} \le C_{n,d,R}C_F(Lq)^{-1/2}.$$

Theorem 5.7 (Barron–Sobolev rate). *Let $k \in \mathbb{N}_0$ and $\sigma \in W^{k,\infty}_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$. Suppose that, for some $J \ge 1$ and real coefficients a_j, b_j, c_j, the function $\nu(u) = \sum_{j=1}^{J} a_j\sigma(b_ju + c_j)$ is nonzero and satisfies $|\partial^\ell\nu(u)| \le C_0(1+|u|)^{-\zeta}$, $0 \le \ell \le k$, a.e. $u \in \mathbb{R}$, for some $C_0 > 0$ and $\zeta > 1$, with derivatives understood weakly. There exists $C_{\nu,C_0,\zeta,J} > 0$ such that, for every f with $\|f\|_{\mathcal{B}^{k+1}} < \infty$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying*

$$\|f - T\|_{W^{k,2}} \le C_{\nu,C_0,\zeta,J}(Lq)^{-1/2}\,\|f\|_{\mathcal{B}^{k+1}}\,.$$

Theorem 5.8 (Sharper Barron–Sobolev rate). *Let $k, \sigma, \nu, \zeta, C_0, J$ be as in Theorem 5.7, and suppose that $\nu \in W^{k+1,\infty}_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$ and $|\partial^{k+1}\nu(u)| \le C_0(1+|u|)^{-\zeta}$ for a.e. $u \in \mathbb{R}$. Let $\varepsilon > 0$ and put $\tau = \min\{\zeta - 1, \varepsilon\}$. There exists $C_{\nu,C_0,\zeta,J,\varepsilon} > 0$ such that, for every f with $\|f\|_{\mathcal{B}^{k+1+\varepsilon}} < \infty$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying*

$$\|f - T\|_{W^{k,2}} \le C_{\nu,C_0,\zeta,J,\varepsilon}\,(Lq)^{-\frac{1}{2}-\frac{\tau}{(2+\tau)(nd+1)}}\,\|f\|_{\mathcal{B}^{k+1+\varepsilon}}\,.$$

For cosine activation, higher spectral regularity gives a further family of transferred rates.

Theorem 5.9 (Spectral rate with cosine). *Let $\sigma(u) = \cos u$, let $k \in \mathbb{N}_0$, and let $s \ge k$. There exists $C > 0$ such that, for every real matrix-valued f with $\|f\|_{\mathcal{B}^s} < \infty$ and every $\mathfrak{a} \in \mathcal{A}$, there exists $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ satisfying*

$$\|f - T\|_{W^{k,2}} \le C(Lq)^{-\frac{1}{2}-\frac{s-k}{nd}}\,\|f\|_{\mathcal{B}^s}\,.$$

Further quantitative transfers. The preceding results are representative applications of the exact transfer principle. Other examples include the improved approximation rates on variation spaces of Siegel & Xu (2024), and their higher-order spectral Barron estimates for ReLU powers (Siegel & Xu, 2022, Theorems 3–4). More generally, a shallow-network upper rate transfers to the sequence-to-sequence class whenever the relevant target norm admits the localization and translated extension used above.

[9] Here C is also independent of μ, p.

[10] The offsets $f_i(0)$ are included in the positional vectors and do not enter the error bound.

6 Optimal Rates under Continuous Selection

The transferred upper rates in Section 5 assert the existence of an approximant for each target. To establish optimality, we require an additional property: the approximating parameters must be chosen continuously as the target varies in L^p. We transfer a classical affine selector for L^p error, then combine it with the classical, architecture-free continuous-information lower bound. The targets and reconstructions take values in $\mathbb{R}^{d\times n}$. Throughout this section, $B_R = \{X \in \mathbb{R}^{d\times n} : \|X\|_{\mathrm{F}} \le R\}$ is the closed ball of fixed radius $R > 0$; $s \ge 1$ is an integer and $1 \le p < \infty$. Positive constants c, C and budget thresholds may depend on n, d, s, p, R, σ, but never on P. In this section, we express rates in the total parameter budget P; for $\mathfrak{a} \in \mathcal{A}$, $P_{\mathfrak{a}} \le C_{n,d}Lq$ uniformly in $\mathfrak{a}$ at fixed input dimensions, by (5).

Recall from Section 2 the number $P_{\mathfrak{a}}$ of parameters of an architecture $\mathfrak{a}$ and its realization map $\vartheta \mapsto T_{\mathfrak{a}}(\vartheta)$. For positive integer budgets P, set

$$\mathcal{E}^{s,p}_{P,\mathrm{cont}}(\sigma) = \inf_{\mathfrak{a}:\, P_{\mathfrak{a}}\le P} \inf_{\substack{\mathsf{S}:\mathcal{U}_{s,p}\to\mathbb{R}^{P_{\mathfrak{a}}} \\ \mathsf{S}\text{ continuous in } L^p}} \sup_{f\in\mathcal{U}_{s,p}} \|f - T_{\mathfrak{a}}(\mathsf{S}(f))\|_{L^p}\,, \tag{6}$$

where $\mathcal{U}_{s,p} = \{f \in W^{s,p}(B_R;\mathbb{R}^{d\times n}) : \|f\|_{W^{s,p}} \le 1\}$ is the sequence-valued Sobolev unit ball, with all nd scalar output entries included in its norm. Thus, for a given budget P, one architecture $\mathfrak{a}$ with $P_{\mathfrak{a}} \le P$ is used for all targets, and its weights and biases are chosen by a map S that depends continuously on the target; the outer infimum optimizes over the architecture and the inner one over the selector. The two function spaces play different roles: $W^{s,p}$ determines the class of targets, whereas L^p measures the error and gives the topology on $\mathcal{U}_{s,p}$ (the relative $L^p(B_R;\mathbb{R}^{d\times n})$ topology) in which S must be continuous, the parameter space $\mathbb{R}^{P_{\mathfrak{a}}}$ carrying the Euclidean topology.[11]

The main result combines the upper and lower bounds stated in Appendices B.4.1 and B.4.2. The proof is given in Appendix B.4.

Theorem 6.1 (Optimal function-value rate under continuous selection)**.** *Let $\sigma \in C^\infty(\mathbb{R};\mathbb{R})$ be nonpolynomial, let $s \ge 1$ be an integer, and let $1 \le p < \infty$. There exist $c, C > 0$ and $P_0 \in \mathbb{N}$ such that*

$$cP^{-s/(nd)} \le \mathcal{E}^{s,p}_{P,\mathrm{cont}}(\sigma) \le CP^{-s/(nd)}, \qquad P \ge P_0,$$

and the lower bound holds for every $P \ge 1$. The upper bound is attained, with an affine L^p-continuous selector, by every $\mathfrak{a} \in \mathcal{A}$ with $P_{\mathfrak{a}} \le P \le 2C_{n,d}Lq$ (Proposition B.8). The lower bound holds for every method with at most P continuously selected real parameters (Proposition B.10).

In particular, a single block of growing feed-forward width ($H = H_*$) and blocks of fixed feed-forward width and growing depth (any $H \le H_*$) attain the optimal exponent among all continuously selected P-parameter methods, including transformer variants beyond Definition 2.1.

AI use statement

The author proposed the initial mathematical approach and used ChatGPT (OpenAI) and Claude (Anthropic) to assist in reviewing related literature, developing mathematical claims and proofs, and drafting and polishing the manuscript. Throughout the project, the author reviewed all AI-assisted material, identified gaps in the arguments, and directed revisions and extensions. The author takes full responsibility for the content of the manuscript.

References

Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning*, volume 221 of *Proceedings of Machine Learning Research*, pp. 72–86. PMLR, 2023. URL https://proceedings.mlr.press/v221/alberti23a.html.

[11] Since this topology is coarser than that of $W^{s,p}$, continuity is a strong requirement; it rules out, for instance, parameters that encode the target in the digits of a few real numbers. This is the setting of Pinkus (1999, Section 6) (see also DeVore et al., 1989).

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

Javier Castro. The Kolmogorov infinite dimensional equation in a Hilbert space via deep learning methods. *Journal of Mathematical Analysis and Applications*, 527(2):127413, 2023. doi: 10.1016/j.jmaa.2023.127413. URL https://doi.org/10.1016/j.jmaa.2023.127413.

Jingpu Cheng, Ting Lin, Zuowei Shen, and Qianxiao Li. A unified framework for establishing the universal approximation of transformer-type architectures. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129131–129163, 2025. doi: 10.52202/085713-4300. URL https://arxiv.org/abs/2506.23551.

Samuel N. Cohen, Filippo de Feo, Jackson Hebner, and Justin Sirignano. Deep Hilbert–Galerkin methods for infinite-dimensional PDEs and optimal control. *arXiv preprint arXiv:2603.19463*, 2026. doi: 10.48550/arXiv.2603.19463. URL https://arxiv.org/abs/2603.19463.

Filippo de Feo. Universal approximation of nonlinear operators and their derivatives, 2026. URL https://arxiv.org/abs/2605.15285v3. arXiv:2605.15285v3.

Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.

Ronald A. DeVore, Ralph Howard, and Charles A. Micchelli. Optimal nonlinear approximation. *Manuscripta Mathematica*, 63(4):469–478, 1989. URL https://eudml.org/doc/155392.

Takashi Furuya, Maarten V. de Hoop, and Gabriel Peyré. Transformers are universal in-context learners. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.01367.

Takashi Furuya, David Mis, Ivan Dokmanić, Maarten V. de Hoop, and Matti Lassas. Function graph transformers universally approximate operators between function spaces, 2026a. URL https://arxiv.org/abs/2605.17968. arXiv:2605.17968.

Takashi Furuya, Davide Murari, and Carola-Bibiane Schönlieb. Approximation theory for Lipschitz continuous transformers. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b. URL https://arxiv.org/abs/2602.15503. arXiv:2602.15503.

Steffen Goebbels. On sharpness of error bounds for multivariate neural network approximation. *Ricerche di Matematica*, 71:633–653, 2022. doi: 10.1007/s11587-020-00549-x. URL https://doi.org/10.1007/s11587-020-00549-x. Published online in 2020.

Linyan Gu, Lihua Yang, and Feng Zhou. On the expressive power of transformers for maxout networks and continuous piecewise linear functions, 2026. URL https://arxiv.org/abs/2603.03084. arXiv:2603.03084.

Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Analysis and Applications*, 18(5):803–859, 2020.

Alex Havrilla and Wenjing Liao. Understanding scaling laws with statistical and approximation theory for transformer neural networks on intrinsically low-dimensional data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 42162–42210, 2024. doi: 10.52202/079017-1334. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4a68bb0e075a3215d4ff37382ea598dc-Abstract-Conference.html.

Xin He, Yuling Jiao, Xiliang Lu, and Jerry Zhijian Yang. Approximation error upper and lower bounds for Hölder class with transformers. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2605.07463v2. arXiv:2605.07463v2.

Ruiyang Hong, Hrad Ghoukasian, and Anastasis Kratsios. A closed-form formula for consistent Lipschitz regression on metric spaces with sparse neural network realizations, 2026. URL https://arxiv.org/abs/2609.03129. arXiv:2609.03129.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991. doi: 10.1016/0893-6080(91)90009-T. URL https://doi.org/10.1016/0893-6080(91)90009-T.

Alexander Hsu and Rongjie Lai. Training-free universal approximation by prompting random transformers, 2026. URL https://arxiv.org/abs/2608.09558. arXiv:2608.09558.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.16525.

Jerry Yao-Chieh Hu, Hong-Yu Chen, Po-Chiao Lin, Maojiang Su, and Han Liu. In-context universal approximation, compositional generalization, and algorithm emulation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a. URL https://openreview.net/forum?id=ZK0U30zmDY.

Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026b. URL https://openreview.net/forum?id=GoW78qJ3qM. arXiv:2504.15956.

Jerry Yao-Chieh Hu, Mingcheng Lu, Yi-Chen Lee, and Han Liu. Transformer approximations from ReLUs, 2026c. URL https://arxiv.org/abs/2604.24878. arXiv:2604.24878.

Vugar E. Ismailov. Universal approximation theorem for neural networks with inputs from a topological vector space. *Information Processing Letters*, 193:106623, 2026. doi: 10.1016/j.ipl.2026.106623. URL https://doi.org/10.1016/j.ipl.2026.106623.

Yoshifusa Ito. Extension of approximation capability of three layered neural networks to derivatives. In *IEEE International Conference on Neural Networks*, volume 1, pp. 377–381, 1993.

Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68926–68955, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7f64034009f4a5fa417a57e1a987c5cd-Abstract-Conference.html.

Yuling Jiao, Yanming Lai, Defeng Sun, Yang Wang, and Bokai Yan. Approximation bounds for transformer networks with application to regression. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026a. doi: 10.48550/arXiv.2504.12175. URL https://arxiv.org/abs/2504.12175. arXiv:2504.12175.

Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Transformers can overcome the curse of dimensionality: A theoretical study from an approximation perspective. *Journal of Machine Learning Research*, 27(50):1–34, 2026b. URL https://jmlr.org/papers/v27/25-1214.html.

Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.14023.

Tokio Kajitsuka and Issei Sato. On the optimal memorization capacity of transformers. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.17677. arXiv:2409.17677.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. PMLR, 2020.

Junghwan Kim, Michelle YoungJin Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8JCg5xJCTPR.

Anastasis Kratsios and Takashi Furuya. Is in-context universality enough? MLPs are also universal in-context, 2025. URL https://arxiv.org/abs/2502.03327. arXiv:2502.03327.

Anastasis Kratsios, A. Martina Neuman, and Philipp Petersen. Adaptivity under realizability constraints: Comparing in-context and agentic learning, 2026. URL https://arxiv.org/abs/2605.04995. arXiv:2605.04995.

Yanming Lai and Defeng Sun. Standard transformers achieve the minimax rate in nonparametric regression with $C^{s,\lambda}$ targets, 2026. URL https://arxiv.org/abs/2602.20555v2. arXiv:2602.20555v2.

Samuel Lanthaler, Siddhartha Mishra, and George E. Karniadakis. Error estimates for DeepONets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1):tnac001, 2022. doi: 10.1093/imatrm/tnac001.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Huiyuan Li and Yuan Xu. Spectral approximation on the unit ball. *SIAM Journal on Numerical Analysis*, 52(6):2647–2675, 2014. doi: 10.1137/130940591.

Hude Liu, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. Attention mechanism, max-affine partition, and universal approximation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 120443–120511, 2025. doi: 10.52202/085713-4020. URL https://arxiv.org/abs/2504.19901.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Shengjie Luo, Shanda Li, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. Your transformer may not be as powerful as you expect. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4301–4315, 2022. URL https://arxiv.org/abs/2205.13401.

Qian Ma, Ruoxiang Xu, and Yongqiang Cai. Vocabulary in-context learning in transformers: Benefits of positional encoding. In *Advances in Neural Information Processing Systems*, volume 38, pp. 135250–135283, 2025. doi: 10.52202/085713-4513. URL https://arxiv.org/abs/2511.06376.

Tong Mao, Jonathan W. Siegel, and Jinchao Xu. Approximation rates for shallow ReLU^k neural networks on Sobolev spaces via the Radon transform. *SIAM Journal on Mathematical Analysis*, 58(2):1171–1186, 2026. doi: 10.1137/24M1686693.

Hrushikesh N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Computation*, 8(1):164–177, 1996. doi: 10.1162/neco.1996.8.1.164.

Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148 (4):1567–1581, 2020. doi: 10.1090/proc/14789. URL https://arxiv.org/abs/1809.00973.

Aleksandar Petrov, Philip Torr, and Adel Bibi. Prompting a pretrained transformer can be a universal approximator. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 40523–40550, 2024. URL https://proceedings.mlr.press/v235/petrov24a.html.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999. doi: 10.1017/S0962492900002919. URL https://pinkus.net.technion.ac.il/files/2021/02/acta.pdf.

Jason Ramapuram, Federico Danieli, Eeshan Dhekane, Floris Weers, Dan Busbridge, Pierre Ablin, Tatiana Likhomanenko, Jagrit Digani, Zijin Gu, Amitis Shidani, and Russ Webb. Theory, analysis, and best practices for sigmoid self-attention. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.04431.

Sachin Saini. A universal approximation theorem for neural networks with outputs in locally convex spaces, 2026. URL https://arxiv.org/abs/2603.07242. arXiv:2603.07242.

Kira A. Selby, Ahmad Rashid, Ivan Kobyzev, Mehdi Rezagholizadeh, and Pascal Poupart. Learning functions on multiple sets using multi-set transformers. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1760–1770, 2022. URL https://proceedings.mlr.press/v180/selby22a.html.

Zhongjie Shi and Wenjing Liao. Learning theory of transformers: Local-to-global approximation via softmax partition of unity, 2026. URL https://arxiv.org/abs/2605.08811. arXiv:2605.08811.

Benjamin Shih, Ahmad Peyvan, Zhongqiang Zhang, and George Em Karniadakis. Transformers as neural operators for solutions of differential equations with finite regularity. *Computer Methods in Applied Mechanics and Engineering*, 434:117560, 2025. doi: 10.1016/j.cma.2024.117560. URL https://doi.org/10.1016/j.cma.2024.117560.

Jonathan W. Siegel. Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces. *Journal of Machine Learning Research*, 24(357):1–52, 2023.

Jonathan W. Siegel and Jinchao Xu. Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321, 2020. doi: 10.1016/j.neunet.2020.05.019. URL https://arxiv.org/abs/1904.02311.

Jonathan W. Siegel and Jinchao Xu. High-order approximation rates for shallow neural networks with cosine and ReLU^k activation functions. *Applied and Computational Harmonic Analysis*, 58:1–26, 2022. doi: 10.1016/j.acha.2021.12.005. URL https://arxiv.org/abs/2012.07205.

Jonathan W. Siegel and Jinchao Xu. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *Foundations of Computational Mathematics*, 24(2):481–537, 2024. doi: 10.1007/s10208-022-09595-3. URL https://arxiv.org/abs/2101.12365.

Jonathan W. Siegel, Snir Hordan, Hannah Lawrence, Ali Syed, and Nadav Dym. Quantitative approximation rates for group equivariant learning, 2026. URL https://arxiv.org/abs/2602.20370v2. arXiv:2602.20370v2.

Elias M. Stein. *Singular Integrals and Differentiability Properties of Functions*. Princeton University Press, 1970.

Shokichi Takakura and Taiji Suzuki. Approximation and estimation ability of transformers for sequence-to-sequence functions with infinite dimensional input. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 33416–33447. PMLR, 2023. URL https://proceedings.mlr.press/v202/takakura23a.html.

Naoki Takeshita and Masaaki Imaizumi. Approximation of permutation invariant polynomials by transformers: Efficient construction in column-size, 2025. URL https://arxiv.org/abs/2502.11467. arXiv:2502.11467.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Mingze Wang and Weinan E. Understanding the expressive power and mechanisms of transformer for sequence modeling. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25781–25856, 2024. doi: 10.52202/079017-0811. URL https://arxiv.org/abs/2402.00522.

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. Universality and limitations of prompt tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/eef6aecfe050b556c6a48d9c16b15558-Abstract-Conference.html. arXiv:2305.18787.

Haotong Yang, Fanxu Meng, Zhouchen Lin, and Muhan Zhang. Parrot mind: Towards explaining the complex task reasoning of pretrained large language models with template-content structure, 2023. URL https://arxiv.org/abs/2310.05452. arXiv:2310.05452.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Penghao Yu, Haotian Jiang, Zeyu Bao, Ruoxi Yu, and Qianxiao Li. The effect of attention head count on transformer approximation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.06662v2. arXiv:2510.06662v2.

Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2020a. doi: 10.48550/arXiv.1912.10077. URL https://arxiv.org/abs/1912.10077v2.

Chulhee Yun, Yin-Wen Chang, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J. Reddi, and Sanjiv Kumar. $O(n)$ connections are expressive enough: Universal approximability of sparse transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13783–13794, 2020b. URL https://arxiv.org/abs/2006.04862.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17283–17297, 2020. URL https://arxiv.org/abs/2007.14062.

A NOTATIONS

Dimensions and basic conventions. Fix integers $n \geq 2$ and $d \geq 1$. We use k for the differentiability order and N for a shallow network's input dimension. We write $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$ and use the zero-order conventions $\mathrm{D}^0 f := f$ and $\sigma^{(0)} := \sigma$. For $\sigma : \mathbb{R} \to \mathbb{R}$, $\sigma^\bullet$ denotes its entrywise application to vectors or matrices of any shape. The superscript $\top$ denotes transpose, and $\mathbf{1}_n \in \mathbb{R}^n$ is the column vector of n ones. For matrices with the same number of columns, $(M_1; \ldots; M_k)$ denotes vertical stacking; for column vectors, it denotes concatenation. Unless another norm is displayed, vectors carry the Euclidean norm and matrices the Frobenius norm. We write I_a for the $a \times a$ identity matrix, $0_{a\times b}$ for the $a \times b$ zero matrix, and $0_a := 0_{a\times 1}$ for the zero column vector.

Kronecker products. For matrices, $\otimes$ denotes the Kronecker product, the matrix representation of the tensor product of linear maps. If $M = (M_{ij}) \in \mathbb{R}^{a_1\times a_2}$ and $K \in \mathbb{R}^{b_1\times b_2}$, then $M \otimes K := (M_{ij}K)_{1\leq i\leq a_1,\, 1\leq j\leq a_2} \in \mathbb{R}^{a_1b_1\times a_2b_2}$: each scalar entry M_{ij} is replaced by the block $M_{ij}K$. For matrices of compatible dimensions, the multiplication rule is

$$(M_1 \otimes K_1)(M_2 \otimes K_2) = (M_1M_2) \otimes (K_1K_2).$$

Token spaces and vectorization. We represent vectors as columns. Inputs are matrices $X \in \mathbb{R}^{d\times n}$; the i-th token is the column $x_i := X_{:,i} \in \mathbb{R}^d$. Thus the matrix representation and its column-major vectorization are

$$X = (x_1 \quad \cdots \quad x_n) \in \mathbb{R}^{d\times n}, \qquad \mathrm{vec}(X) := \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^{nd}.$$

The map $\mathrm{vec} : \mathbb{R}^{d\times n} \to \mathbb{R}^{nd}$ is a linear isometric isomorphism between the Frobenius norm and the Euclidean norm. Therefore, when convenient, we identify $\mathbb{R}^{d\times n}$ with $\mathbb{R}^N$, $N = nd$, through vec.

Function spaces and measures. Function spaces are defined on the token space, with derivatives and Lebesgue measure taken with respect to the matrix entries. A sequence target is a map $f : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$; equivalently, it is a map on $\mathbb{R}^{nd}$ via $f \circ \mathrm{vec}^{-1}$. The scalar components f_ℓ enumerate the nd output entries in column-major order; auxiliary $\mathbb{R}^b$-valued functions use b components. We display domains and codomains in function-space notation, but omit them from norm subscripts when unambiguous. For a Borel measure μ on a measurable domain Ω, the space $L^p_\mu(\Omega; \mathbb{R}^{d\times n})$ consists of measurable maps modulo equality μ-almost everywhere with finite norm

$$\|f\|_{L^p_\mu} := \begin{cases} \left(\displaystyle\sum_{\ell=1}^{nd} \int_\Omega |f_\ell(x)|^p \, d\mu(x) \right)^{1/p}, & 1 \leq p < \infty, \\ \displaystyle\max_{1\leq \ell\leq nd} \operatorname*{ess\,sup}_{x\in\Omega} |f_\ell(x)|, & p = \infty. \end{cases}$$

The essential supremum is taken with respect to μ. These componentwise conventions agree with the Sobolev norms below at order zero. When μ is Lebesgue measure, we suppress its subscript and write $L^p(\Omega; \mathbb{R}^{d\times n})$ and $\|f\|_{L^p}$.

Derivatives and compact-open norms. We use $\partial^\alpha f$ for coordinate derivatives indexed by a multi-index α, and $\mathrm{D}^j f$ for the jth Fréchet derivative. For a compact $K \subset \mathbb{R}^{d\times n}$, let $C^0(K; \mathbb{R}^{d\times n})$ be the space of continuous maps on K, with

$$\|f\|_{C^0_K} := \max_{1\leq \ell\leq nd} \sup_{X\in K} |f_\ell(X)|.$$

This agrees with the following C^k seminorm when $k = 0$. If $f \in C^k(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$, define

$$\|f\|_{C^k_K} := \max_{1\leq \ell\leq nd} \max_{|\alpha|\leq k} \sup_{X\in K} \left|\partial^\alpha f_\ell(X)\right|. \tag{7}$$

This is understood as a compact-open seminorm on the space of globally defined C^k functions.

Hölder norms. For $r \in \mathbb{N}_0$ and $0 < \alpha \le 1$, let $C^{r,\alpha}(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ consist of matrix-valued C^r functions with finite norm

$$\|f\|_{C^{r,\alpha}} := \sum_{|\beta|\le r} \left\|\partial^\beta f\right\|_{L^\infty} + \sum_{|\beta|=r} \sup_{\substack{x,y\in\mathbb{R}^{d\times n}\\ x\neq y}} \frac{\left\|\partial^\beta f(x) - \partial^\beta f(y)\right\|_{\mathrm{F}}}{\|x-y\|_{\mathrm{F}}^{\ \alpha}}.$$

The endpoint $\alpha = 1$ means Lipschitz continuity of the order-r derivatives. The support is $\operatorname{supp} f := \overline{\{x \in \mathbb{R}^{d\times n} : f(x) \neq 0\}}$; f is compactly supported when this set is compact. For a compact $K \subset \mathbb{R}^{d\times n}$, we use $\|g\|_{C^k_K} := \max_{1\le \ell\le nd} \max_{|\beta|\le k} \sup_{x\in K} |\partial^\beta g_\ell(x)|$ for globally C^k matrix-valued functions, consistently with (7).

Weighted derivative seminorms. For a finite Borel measure μ on $\mathbb{R}^{d\times n}$ and $1 \le p < \infty$, put

$$\|f\|_{C^{k,p}_\mu} := \left(\sum_{\ell=1}^{nd} \sum_{|\alpha|\le k} \int_{\mathbb{R}^{d\times n}} |\partial^\alpha f_\ell(X)|^p \, d\mu(X)\right)^{1/p}, \tag{8}$$

$$C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n}) := \left\{ f \in C^k(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n}) : \|f\|_{C^{k,p}_\mu} < \infty \right\}.$$

As in Hornik (1991), this is a class of globally classically differentiable functions, not the completion defining a weak Sobolev space. For a general measure μ, (8) can be only a seminorm. A class $\mathcal{G} \subset C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ is *dense* in this space with respect to (8) if, for every $f \in C^{k,p}_\mu(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ and every $\varepsilon > 0$, there exists $g \in \mathcal{G}$ such that $\|f - g\|_{C^{k,p}_\mu} < \varepsilon$. For approximation in Sobolev spaces with Lebesgue measure, we use $W^{k,p}(\Omega;\mathbb{R}^{d\times n})$, defined below.

Fourier convention. Following Mao et al. (2026, Eq. (1.7)), for $g \in L^1(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ we set

$$\widehat{g}(\omega) := \int_{\mathbb{R}^{d\times n}} g(x) e^{\mathrm{i}\langle x,\omega\rangle_{\mathrm{F}}} \, dx.$$

Here i denotes the imaginary unit. We use its Plancherel extension on $L^2(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$ and its distributional extension when needed.

Sobolev norms. On an open domain $\Omega \subset \mathbb{R}^{d\times n}$, write $W^{s,p}(\Omega;\mathbb{R}^{d\times n})$ for the weak Sobolev space, with $s \in \mathbb{N}_0$ and $1 \le p \le \infty$. Its norm is

$$\|f\|_{W^{s,p}} := \begin{cases} \left(\displaystyle\sum_{\ell=1}^{nd} \sum_{|\alpha|\le s} \int_\Omega |\partial^\alpha f_\ell(x)|^p \, dx\right)^{1/p}, & 1 \le p < \infty, \\ \displaystyle\max_{1\le\ell\le nd} \max_{|\alpha|\le s} \operatorname*{ess\,sup}_{x\in\Omega} |\partial^\alpha f_\ell(x)|, & p = \infty, \end{cases}$$

where derivatives are weak and dx is Lebesgue measure. In one dimension, $\partial^\ell g$ denotes the weak derivative of order ℓ, with $\partial^0 g = g$. For the closed ball B_R, the notation $W^{s,p}(B_R;\mathbb{R}^{d\times n})$ means the Sobolev space on its interior.

For noninteger $s > 0$, the matrix-valued space $W^{s,2}(B_R;\mathbb{R}^{d\times n})$ is defined by restriction from the Fourier Sobolev space on $\mathbb{R}^{d\times n}$, with norm

$$\|f\|_{W^{s,2}} := \inf_{g|_{B_R}=f} \left(\int_{\mathbb{R}^{d\times n}} (1 + \|\omega\|_{\mathrm{F}})^{2s} \sum_{\ell=1}^{nd} |\widehat{g}_\ell(\omega)|^2 \, d\omega\right)^{1/2}.$$

Here $g \in L^2(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$. For integer s, this restriction norm is equivalent to the integer-order Sobolev norm defined above.

Remark A.1 (Sobolev and Hölder extensions)**.** For integer $s \ge 0$ and $1 \le p \le \infty$, the extension theorem of Stein (1970, Chapter VI, Theorem 5, p. 181) provides a linear operator E extending functions from B_R to $\mathbb{R}^{nd}$, with

$$\|Ef\|_{W^{s,p}(\mathbb{R}^{nd})} \le C \, \|f\|_{W^{s,p}(B_R)}.$$

The same operator is bounded for $s = 0$, hence in L^p; this simultaneous boundedness is used for continuous parameter selection. Hölder extensions with controlled $C^{r,\alpha}$ norms follow from Stein (1970, Chapter VI, Theorem 4, p. 177)[12]. For the noninteger $W^{s,2}$ restriction norm, the defining infimum provides extensions with whole-space norm arbitrarily close to the norm on B_R. Multiplication by a fixed $\theta \in C_c^\infty(\mathbb{R}^{nd})$ equal to 1 near B_R gives a compactly supported extension with controlled norm: for integer Sobolev and Hölder norms this follows from the product rule; for fractional $W^{s,2}$ it follows from the Fourier product formula and weighted convolution estimates. These constructions apply componentwise to vector- and matrix-valued maps.

Spectral Barron norms. For $s \geq 0$ and $\Omega \subset \mathbb{R}^{d\times n}$ (e.g. B_R or $\mathbb{R}^{d\times n}$), define the spectral extension norm

$$\|f\|_{\mathcal{B}^s} \coloneqq \inf_{g|_\Omega = f} \sum_{\ell=1}^{nd} \int_{\mathbb{R}^{d\times n}} (1 + \|\omega\|_{\mathrm{F}})^s \, |\widehat{g}_\ell(\omega)| \, d\omega.$$

The infimum runs over extensions with a function-valued distributional Fourier transform satisfying the displayed weighted integrability, represented by

$$g_\ell(x) = (2\pi)^{-nd} \int_{\mathbb{R}^{d\times n}} \widehat{g}_\ell(\omega) e^{-\mathrm{i}\langle x,\omega\rangle_{\mathrm{F}}} \, d\omega.$$

Remark A.2 (Spectral Barron extensions and cutoffs)**.** The restriction norm above is defined by minimizing over whole-space extensions. Thus, for every $\varepsilon > 0$, a map f with finite $\mathcal{B}^s(B_R)$ norm admits an extension h satisfying

$$\|h\|_{\mathcal{B}^s(\mathbb{R}^{nd})} \leq \|f\|_{\mathcal{B}^s(B_R)} + \varepsilon.$$

Let $\theta \in C_c^\infty(\mathbb{R}^{nd})$ equal 1 near B_R. Then θh still extends f and has compact support. The identities $\widehat{\theta h} = (2\pi)^{-nd}\widehat{\theta} * \widehat{h}$, $(1+|u+v|)^s \leq (1+|u|)^s(1+|v|)^s$ and the triangle inequality for convolution give

$$\|\theta h\|_{\mathcal{B}^s(\mathbb{R}^{nd})} \leq C_{\theta,s} \, \|h\|_{\mathcal{B}^s(\mathbb{R}^{nd})}, \qquad C_{\theta,s} \coloneqq (2\pi)^{-nd} \int_{\mathbb{R}^{nd}} (1+|u|)^s |\widehat{\theta}(u)| \, du < \infty.$$

The constant is finite because $\widehat{\theta}$ decays faster than every power. Translation multiplies the Fourier transform by a phase of modulus one, so it preserves this norm; finite sums of translated cutoff extensions are therefore bounded by the sum of their norms. This justifies the spectral norm control in the translated-domain construction.

B Proofs

B.1 Proofs for Section 3: Exact realization

In this subsection, δ_{hj} denotes the Kronecker delta, $h, j \in \{1, \ldots, n\}$.

B.1.1 Positional encodings and softmax mixing

The construction requires linear functions of the positional encoding that yield a constant (for the queries), the token label (for the feed-forward sublayers), and scores (for the keys) whose softmax weights form an invertible matrix. Definition B.1 states these requirements as (C1)–(C3). The model width of the realization is $2d + g + \min\{q, nd\}$ for $\mathfrak{a} \in \mathcal{A}$ and w in general, where g is the dimension of the encoding.

Definition B.1 (Positional encoding and mixing matrix)**.** Let g be an integer with $1 \leq g \leq n$. A *positional encoding of dimension g* is a family of vectors $\rho_1, \ldots, \rho_n \in \mathbb{R}^g$ such that

(C1) there is $a \in \mathbb{R}^g$ with $\langle a, \rho_i \rangle = 1$ for $i = 1, \ldots, n$;

[12] Stein's notation Lip_γ (see Stein, 1970, Chapter VI, p. 176), with $\gamma = r + \alpha$, $r \in \mathbb{N}_0$ and $0 < \alpha \leq 1$, denotes Hölder regularity of order $r + \alpha$: derivatives through order r are bounded and those of order r are α-Hölder continuous. On a closed set, the definition is formulated in terms of derivative data satisfying the corresponding Taylor remainder bounds.

(C2) there is $\lambda \in \mathbb{R}^g$ with $\langle \lambda, \rho_i \rangle = i$ for $i = 1, \dots, n$;

(C3) there are $k_1, \dots, k_n \in \mathbb{R}^g$ such that the *mixing matrix* $S \in \mathbb{R}^{n \times n}$ defined below is invertible:

$$S_{hj} \coloneqq \frac{\exp\langle k_h, \rho_j \rangle}{\sum_{l=1}^{n} \exp\langle k_h, \rho_l \rangle}, \qquad h, j = 1, \dots, n. \tag{9}$$

In the transformer of Definition 2.1, a positional encoding enters through the positional vectors, $p_i = (0_d; \rho_i; 0) \in \mathbb{R}^w$ (see the proof of Theorem 3.2).

By construction, S has strictly positive entries and every row sums to 1. The condition $g \le n$ holds for all encodings considered below; it ensures that the model width $2d + g + \min\{q, nd\}$ of the architectures in $\mathcal{A}$ is at most $2d + n + nd$, a bound used in Remark 3.7.

Example B.2 (Vandermonde encoding)**.** Let $g = 2$ and $\rho_i \coloneqq (1, i)^\top$. Then (C1) holds with $a = e_1$ and (C2) with $\lambda = e_2$ (the standard basis of $\mathbb{R}^2$). With $k_h \coloneqq (0, h-1)^\top$ the scores are $\langle k_h, \rho_j \rangle = (h-1)j$, so that, writing $x_j \coloneqq e^j$,

$$S_{hj} = \frac{x_j^{\,h-1}}{\sum_{l=1}^{n} x_l^{\,h-1}}, \qquad \text{that is,} \qquad S = \Delta^{-1} V, \quad V_{hj} \coloneqq x_j^{\,h-1}, \quad \Delta \coloneqq \operatorname{diag}(V \mathbf{1}_n).$$

Here V is the Vandermonde matrix of the distinct nodes $e^1, \dots, e^n$, hence invertible, and Δ is diagonal with positive entries; thus (C3) holds.

The condition number of S grows rapidly with n, but the exact realization and approximation results proved here require only invertibility. The following two examples give well-conditioned alternatives.

Example B.3 (One-hot encoding)**.** Let $g = n$ and let $\rho_i \coloneqq \tau_i \in \mathbb{R}^n$ have entry 1 in position i and entry 0 elsewhere. Then (C1) holds with $a = \mathbf{1}_n$, since the coordinates of τ_i sum to 1; (C2) holds with $\lambda = (1, \dots, n)^\top$, since $\lambda^\top \tau_i = i$; and with $k_h \coloneqq \log(n+1)\, \tau_h$, the scores are $\langle k_h, \rho_j \rangle = \delta_{hj} \log(n+1)$. We claim that (C3) holds with the matrix

$$S \coloneqq \frac{1}{2} I_n + \frac{1}{2n} \mathbf{1}_n \mathbf{1}_n^\top \in \mathbb{R}^{n \times n}, \qquad S_{hj} = \frac{1}{2}\delta_{hj} + \frac{1}{2n}. \tag{10}$$

This S has strictly positive entries, every row sums to 1, and

$$S_{hj} = \frac{\exp(\delta_{hj} \log(n+1))}{\sum_{l=1}^{n} \exp(\delta_{hl} \log(n+1))}.$$

Moreover, S is invertible with

$$S^{-1} = 2 I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top.$$

Indeed, each entry is positive and, for every h, $\sum_{j=1}^{n} S_{hj} = \frac{1}{2} \sum_{j=1}^{n} \delta_{hj} + \frac{1}{2n} \sum_{j=1}^{n} 1 = 1$. For each h, precisely one score equals $\log(n+1)$ and the remaining $n-1$ scores equal 0. Thus the softmax denominator is

$$\begin{aligned} \sum_{l=1}^{n} \exp(\delta_{hl} \log(n+1)) &= \exp(\delta_{hh} \log(n+1)) + \sum_{\substack{l=1 \\ l \neq h}}^{n} \exp(\delta_{hl} \log(n+1)) \\ &= \exp(\log(n+1)) + (n-1) = (n+1) + (n-1) = 2n. \end{aligned}$$

The numerator is $n+1$ when $j = h$ and 1 otherwise. Consequently,

$$\frac{\exp(\delta_{hj} \log(n+1))}{\sum_{l=1}^{n} \exp(\delta_{hl} \log(n+1))} = \frac{1 + n\delta_{hj}}{2n} = \frac{1}{2}\delta_{hj} + \frac{1}{2n} = S_{hj}.$$

Put $J \coloneqq n^{-1} \mathbf{1}_n \mathbf{1}_n^\top$. Since $\mathbf{1}_n^\top \mathbf{1}_n = n$, we have $J^2 = \frac{1}{n^2} \mathbf{1}_n (\mathbf{1}_n^\top \mathbf{1}_n) \mathbf{1}_n^\top = J$. Thus $S = (I_n + J)/2$ and

$$S(2I_n - J) = \frac{1}{2}(I_n + J)(2I_n - J) = \frac{1}{2}(2I_n + J - J^2) = I_n,$$

which proves the inverse formula.

Head h thus averages its designated token with the mean of all tokens: since $S_{hj} = \frac{1}{2}\delta_{hj} + \frac{1}{2n}$, we have $\sum_{j=1}^{n} S_{hj} x_j = \frac{1}{2} x_h + \frac{1}{2}\big(\frac{1}{n} \sum_{j=1}^{n} x_j\big)$, independently of the receiving position.

Example B.4 (Gaussian encoding)**.** Let $n \geq 3$, $g = 3$, $\rho_i := (1, i, i^2)^\top$, and $\gamma := \log 16$. Then (C1) holds with $a = e_1$ and (C2) with $\lambda = e_2$. With $k_h := (0, 2\gamma h, -\gamma)^\top$ the scores are $\langle k_h, \rho_j \rangle = 2\gamma hj - \gamma j^2$, and multiplying numerator and denominator of (9) by $e^{-\gamma h^2}$ gives

$$S_{hj} = \frac{e^{-\gamma(h-j)^2}}{\sum_{l=1}^{n} e^{-\gamma(h-l)^2}}, \qquad \text{that is,} \qquad S = \Delta^{-1}\widetilde{S}, \quad \widetilde{S}_{hj} := 16^{-(h-j)^2}, \quad \Delta := \operatorname{diag}(\widetilde{S}\mathbf{1}_n).$$

The diagonal entries of $\widetilde{S}$ equal 1, while $\sum_{j \neq h} \widetilde{S}_{hj} \leq 2\sum_{k=1}^{\infty} 16^{-k^2} \leq 2\sum_{k=1}^{\infty} 16^{-k} = \frac{2}{15} < 1$. Thus $\widetilde{S}$ is strictly diagonally dominant, hence invertible by the Lévy–Desplanques theorem, and so is S; thus (C3) holds.

B.1.2 Proof of the exact realization theorem

Proof of Theorem 3.2. We prove the theorem first in the *base case* $t = d$, $H = H_* = n$, and $w = 2d + g + \min\{q, nd\}$, in which $G = 1$, $q_* = q$, $L_* = L$, and the hypothesis reads $Lq \geq m$; the general case is reduced to it at the end of the proof. We first explain the main idea. After padding the source network to hidden width Lq, we partition its hidden units into L groups of q units. Constant queries and positional keys produce the same collection of weighted token averages at every query position. Softmax weights with finite scores are strictly positive, so no head can select a single token: the one-hot weights $\Pi_{ij}^{(h)} = \delta_{hj}$, for which $[\mathrm{head}_h(Z(X))]_{:,i} = \sum_{j=1}^{n} \delta_{hj} x_j = x_h$, would need infinite scores. Instead, head h has the weights $\Pi_{ij}^{(h)} = S_{hj}$, the hth row of the mixing matrix S of (9); since S is invertible, the output projection reverses the mixing of the n heads exactly, through the factor $S^{-1} \otimes I_d$. These weighted averages determine the whole ordered input $\mathrm{vec}(X)$. In each token, we store a linear image of this input in a working block: the input itself when $q \geq nd$, and the input-dependent terms $A_{(u)}\,\mathrm{vec}(X)$ for group u otherwise. The feed-forward sublayer of block u evaluates q hidden units of the source network and adds their contribution to an output block that the residual connections accumulate across the L blocks. Summing the contributions of the L groups and adding the output bias c once gives $\Phi_i(\mathrm{vec}(X))$, which the readout extracts.

Chunks and factorization. Write the given family $\mathbf{\Phi} = (\Phi_i)_{i=1}^{n}$ as

$$\Phi_i(\xi) = C\sigma^{\bullet}(A\xi + Di + b) + c, \quad A \in \mathbb{R}^{m \times nd},\ D \in \mathbb{R}^{m},\ b \in \mathbb{R}^{m},\ C \in \mathbb{R}^{d \times m},\ c \in \mathbb{R}^{d}.$$

Let $L, q \geq 1$ with $Lq \geq m$. Padding A, D, and b with $Lq - m$ zero rows and C with $Lq - m$ zero columns does not change $\mathbf{\Phi}$, because each added hidden unit is multiplied by a zero column of C; hence we may assume $m = Lq$. Partition the hidden units into L consecutive groups of size q,

$$A = \begin{pmatrix} A_{(1)} \\ \vdots \\ A_{(L)} \end{pmatrix}, \quad D = \begin{pmatrix} D_{(1)} \\ \vdots \\ D_{(L)} \end{pmatrix}, \quad b = \begin{pmatrix} b_{(1)} \\ \vdots \\ b_{(L)} \end{pmatrix}, \quad C = \begin{pmatrix} C_{(1)} & \cdots & C_{(L)} \end{pmatrix},$$

with $A_{(u)} \in \mathbb{R}^{q \times nd}$, $D_{(u)}, b_{(u)} \in \mathbb{R}^{q}$, and $C_{(u)} \in \mathbb{R}^{d \times q}$. Then

$$\Phi_i(\xi) = \sum_{u=1}^{L} C_{(u)}\sigma^{\bullet}\big(A_{(u)}\xi + D_{(u)}i + b_{(u)}\big) + c \qquad \text{for every } \xi \in \mathbb{R}^{nd}. \tag{11}$$

Put $s := \min\{q, nd\}$ and factor each chunk through dimension s: for $u = 1, \ldots, L$ choose $M_{(u)} \in \mathbb{R}^{s \times nd}$ and $\Gamma_{(u)} \in \mathbb{R}^{q \times s}$ with $\Gamma_{(u)} M_{(u)} = A_{(u)}$, namely

$$M_{(u)} := I_{nd}, \quad \Gamma_{(u)} := A_{(u)} \quad \text{if } q \geq nd, \qquad\qquad M_{(u)} := A_{(u)}, \quad \Gamma_{(u)} := I_q \quad \text{if } q < nd,$$

and put $M_{(0)} := 0 \in \mathbb{R}^{s \times nd}$. After the attention update and residual addition in block u, the working block equals $M_{(u)}\,\mathrm{vec}(X)$. Applying $\Gamma_{(u)}$ gives $A_{(u)}\,\mathrm{vec}(X)$, since $\Gamma_{(u)} M_{(u)} = A_{(u)}$. We use s working coordinates to store the shorter of the two vectors: $\mathrm{vec}(X)$, with nd coordinates, when $q \geq nd$, and $A_{(u)}\,\mathrm{vec}(X)$, with q coordinates, when $q < nd$. Thus $s = \min\{q, nd\} \geq \operatorname{rank} A_{(u)}$ for every target. The convention $M_{(0)} = 0$ matches the initially zero working block.

Architecture. Fix a positional encoding $g, \rho_1, \ldots, \rho_n, a, \lambda, k_1, \ldots, k_n$ as in Definition B.1, with mixing matrix S from (9), and use L blocks with $w = 2d + g + s$, $H = n$, $t = d$, and feed-forward

hidden width q. The four coordinate blocks of a token, of sizes d, g, s, d, hold the input token, its positional encoding, the working block $M_{(u)}\,\mathrm{vec}(X)$, and the accumulated output: the first two are never modified, the third is written by the attention sublayers, and the fourth receives one summand of (11) from the feed-forward sublayer of every block. Larger dimensions are treated in the general case below.

Embedding. Choose

$$E_{\mathrm{in}} \coloneqq \begin{pmatrix} I_d \\ 0_{g\times d} \\ 0_{s\times d} \\ 0_{d\times d} \end{pmatrix}, \quad p_i \coloneqq \begin{pmatrix} 0_d \\ \rho_i \\ 0_s \\ 0_d \end{pmatrix}; \quad \textit{hence:} \quad z_i(X) \coloneqq [Z(X)]_{:,i} = E_{\mathrm{in}}x_i + p_i = \begin{pmatrix} x_i \\ \rho_i \\ 0_s \\ 0_d \end{pmatrix}.$$

The positional encoding ρ_i provides the constant $\langle a, \rho_i\rangle = 1$, the key scores $\langle k_h, \rho_i\rangle$, and the label $\langle \lambda, \rho_i\rangle = i$. The working and output blocks are initially zero.

Attention. Let $e_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$. In every block and for $h = 1, \ldots, n$, set (with $e_1 a^\top, e_1 k_h^\top \in \mathbb{R}^{d\times g}$)

$$\begin{aligned} W_Q^h &\coloneqq \begin{pmatrix} 0_{d\times d} & e_1 a^\top & 0_{d\times s} & 0_{d\times d} \end{pmatrix} \in \mathbb{R}^{d\times w}, \\ W_K^h &\coloneqq \begin{pmatrix} 0_{d\times d} & e_1 k_h^\top & 0_{d\times s} & 0_{d\times d} \end{pmatrix} \in \mathbb{R}^{d\times w}, \\ W_V^h &\coloneqq \begin{pmatrix} I_d & 0_{d\times g} & 0_{d\times s} & 0_{d\times d} \end{pmatrix} \in \mathbb{R}^{d\times w}. \end{aligned}$$

The query matrix reads the constant of (C1), the key matrix of head h reads the score $\langle k_h, \cdot\rangle$ of (C3), and the value matrix reads the original content. These matrices read only the first two coordinate blocks. For a fixed input X, let $\widetilde{Z} \in \mathbb{R}^{w\times n}$ be any state whose columns $\widetilde{z}_i$ have the same first two coordinate blocks as $z_i(X)$, that is, $\widetilde{z}_i = (x_i; \rho_i; *; *)$. We first compute attention for such a state $\widetilde{Z}$; in the induction below, we will take $\widetilde{Z} = Z_{(u-1)}(X)$. Then

$$W_Q^h \widetilde{z}_i = e_1\langle a, \rho_i\rangle = e_1, \qquad W_K^h \widetilde{z}_j = \langle k_h, \rho_j\rangle e_1, \qquad W_V^h \widetilde{z}_j = x_j,$$

and the attention score is $\langle W_Q^h \widetilde{z}_i, W_K^h \widetilde{z}_j\rangle = \langle k_h, \rho_j\rangle\langle e_1, e_1\rangle = \langle k_h, \rho_j\rangle$. Here h indexes heads, i the receiving token, and j the token being averaged. By (9), substitution into the softmax and head definitions gives

$$\Pi_{ij}^{(h)}(\widetilde{Z}) = \frac{\exp\langle k_h, \rho_j\rangle}{\sum_{l=1}^n \exp\langle k_h, \rho_l\rangle} = S_{hj}, \qquad [\mathrm{head}_h(\widetilde{Z})]_{:,i} = \sum_{j=1}^n S_{hj} x_j.$$

These weights are independent of the receiving token i, of the input values, and of the target. Every query token receives the same mixture from head h. Stacking the n head outputs gives (notice that $S \otimes I_d \in \mathbb{R}^{nd\times nd}$)

$$\begin{pmatrix} [\mathrm{head}_1(\widetilde{Z})]_{:,i} \\ \vdots \\ [\mathrm{head}_n(\widetilde{Z})]_{:,i} \end{pmatrix} = \begin{pmatrix} S_{11}I_d & \cdots & S_{1n}I_d \\ \vdots & \ddots & \vdots \\ S_{n1}I_d & \cdots & S_{nn}I_d \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (S \otimes I_d)\,\mathrm{vec}(X).$$

By (C3), S is invertible, and the Kronecker product multiplication rule gives $(S^{-1} \otimes I_d)(S \otimes I_d) = (S^{-1}S) \otimes (I_d I_d) = I_{nd}$. In block $u \in \{1, \ldots, L\}$ choose the output projection

$$W_{O,(u)} \coloneqq \begin{pmatrix} 0_{d\times nd} \\ 0_{g\times nd} \\ (M_{(u)} - M_{(u-1)})(S^{-1} \otimes I_d) \\ 0_{d\times nd} \end{pmatrix} \in \mathbb{R}^{w\times nd}.$$

The concatenation of the n head outputs has dimension $Ht = nd$. The third block row of $W_{O,(u)}$ reverses the mixing and supplies $(M_{(u)} - M_{(u-1)})\,\mathrm{vec}(X)$, which the residual connection adds to the previous working block. Its other block rows are zero. In particular, if $q \geq nd$, then $M_{(u)} = I_{nd}$ for every $u \geq 1$. The third block row of $W_{O,(1)}$ is $S^{-1} \otimes I_d$, which inverts the mixing and does not depend on the target, and $W_{O,(u)} = 0$ for $u \geq 2$. If $q < nd$, the third block row of $W_{O,(u)}$

is $(A_{(u)} - A_{(u-1)})(S^{-1} \otimes I_d)$, with $A_{(0)} = 0$, so the attention output projections depend on A. Consequently, for $\widetilde{Z}$ as above and every token i,

$$[\mathrm{MHA}(\widetilde{Z})]_{:,i} = \begin{pmatrix} 0_d \\ 0_g \\ (M_{(u)} - M_{(u-1)})(S^{-1} \otimes I_d)(S \otimes I_d)\,\mathrm{vec}(X) \\ 0_d \end{pmatrix} = \begin{pmatrix} 0_d \\ 0_g \\ (M_{(u)} - M_{(u-1)})\,\mathrm{vec}(X) \\ 0_d \end{pmatrix}. \tag{12}$$

If $q \geq nd$, the working block of every token equals $\mathrm{vec}(X)$ after the first block and remains unchanged, since $W_{O,(u)} = 0$ for $u \geq 2$. If $q < nd$, the working block after block u equals $A_{(u)}\,\mathrm{vec}(X)$.

Feed-forward sublayers and readout. In block u choose

$$B_{1,(u)} \coloneqq \begin{pmatrix} 0_{q\times d} & D_{(u)}\lambda^\top & \Gamma_{(u)} & 0_{q\times d} \end{pmatrix} \in \mathbb{R}^{q\times w}, \qquad b_{1,(u)} \coloneqq b_{(u)} \in \mathbb{R}^q,$$

$$B_{2,(u)} \coloneqq \begin{pmatrix} 0_{d\times q} \\ 0_{g\times q} \\ 0_{s\times q} \\ C_{(u)} \end{pmatrix} \in \mathbb{R}^{w\times q}, \qquad b_{2,(u)} \coloneqq \begin{pmatrix} 0_d \\ 0_g \\ 0_s \\ \mathbf{1}_{\{u=L\}}\, c \end{pmatrix} \in \mathbb{R}^w.$$

Thus $B_{1,(u)}\widetilde{z}_i$ depends only on the label $\langle \lambda, \rho_i \rangle = i$ and the working block of $\widetilde{z}_i$, $B_{2,(u)}$ maps into the output block, and c appears only in the last bias $b_{2,(L)}$. Finally, take

$$R_d \coloneqq \begin{pmatrix} 0_{d\times d} & 0_{d\times g} & 0_{d\times s} & I_d \end{pmatrix}.$$

Induction. Denote by $Z_{(u)}(X) \coloneqq \mathcal{B}_{(u)} \circ \cdots \circ \mathcal{B}_{(1)}(Z(X))$ the state after u blocks, with $Z_{(0)}(X) = Z(X)$. We show by induction on $u \in \{0, \ldots, L\}$ that, for every token i,

$$[Z_{(u)}(X)]_{:,i} = \begin{pmatrix} x_i \\ \rho_i \\ M_{(u)}\,\mathrm{vec}(X) \\ \sum_{u'=1}^{u} C_{(u')}\sigma^\bullet\big(A_{(u')}\,\mathrm{vec}(X) + D_{(u')}i + b_{(u')}\big) + \mathbf{1}_{\{u=L\}}\, c \end{pmatrix}. \tag{13}$$

For $u = 0$ the right-hand side is $z_i(X)$, since $M_{(0)} = 0$. Assume (13) for $u-1$ and consider block u. By the induction hypothesis, the first two coordinate blocks of $Z_{(u-1)}(X)$ are those of $Z(X)$, so (12) applies with $\widetilde{Z} = Z_{(u-1)}(X)$, and the first residual connection gives

$$[U(Z_{(u-1)}(X))]_{:,i} = \begin{pmatrix} x_i \\ \rho_i \\ M_{(u)}\,\mathrm{vec}(X) \\ \sum_{u'=1}^{u-1} C_{(u')}\sigma^\bullet\big(A_{(u')}\,\mathrm{vec}(X) + D_{(u')}i + b_{(u')}\big) \end{pmatrix}.$$

Since $\langle \lambda, \rho_i \rangle = i$ and $\Gamma_{(u)}M_{(u)} = A_{(u)}$, we have $B_{1,(u)}[U(Z_{(u-1)}(X))]_{:,i} + b_{1,(u)} = A_{(u)}\,\mathrm{vec}(X) + D_{(u)}i + b_{(u)}$, the preactivation of the u-th group of hidden units of Φ_i at $\xi = \mathrm{vec}\,X$. After applying the activation, the second matrix and bias place the group's contribution in the final d coordinates:

$$\big[\mathrm{FF}(U(Z_{(u-1)}(X)))\big]_{:,i} = \begin{pmatrix} 0_d \\ 0_g \\ 0_s \\ C_{(u)}\sigma^\bullet\big(A_{(u)}\,\mathrm{vec}(X) + D_{(u)}i + b_{(u)}\big) + \mathbf{1}_{\{u=L\}}\, c \end{pmatrix}.$$

The second residual connection,

$$[Z_{(u)}(X)]_{:,i} = [U(Z_{(u-1)}(X))]_{:,i} + \big[\mathrm{FF}(U(Z_{(u-1)}(X)))\big]_{:,i},$$

yields (13) for u, completing the induction.

Conclusion. For $u = L$, the fourth block of (13) equals $\Phi_i(\mathrm{vec}(X))$ by (11). Denote by $T_{\mathbf{\Phi}}$ the transformer defined by these parameters and the shared tokenwise readout. Then

$$T_{\mathbf{\Phi}}(X) = R_d Z_{(L)}(X), \qquad [T_{\mathbf{\Phi}}(X)]_{:,i} = \Phi_i(\mathrm{vec}(X)), \quad 1 \leq i \leq n.$$

The identity holds for every $X \in \mathbb{R}^{d\times n}$ and every scalar activation σ. Both residual connections are present in every block, each block has n heads of dimension d and feed-forward hidden width exactly

q, the model width is $2d + g + \min\{q, nd\}$, and the depth is L. The input embedding, positional vectors, query, key, and value matrices, and readout depend only on n, d, q and the positional encoding. At fixed m, L, q: if $q \geq nd$, the output projections are independent of the target and $B_{1,(u)}$ depends linearly on $(A_{(u)}, D_{(u)})$; if $q < nd$, $W_{O,(u)}$ depends linearly on A and $B_{1,(u)}$ depends affinely on $D_{(u)}$, its constant part $(0\ 0\ I_q\ 0)$ being independent of the target. In both cases $b_{1,(u)}$ depends linearly on b, $B_{2,(u)}$ on C, and $b_{2,(L)}$ on c. This proves the base case of Theorem 3.2.

General parameters. We now adapt the preceding construction to the prescribed architecture. Step (a) constructs L_* blocks, each evaluating a group of q_* hidden units, within the available model width. Step (b) implements each attention update using $H_* = n\lceil d/t \rceil$ heads of dimension t. Step (c) distributes these heads over $G = \lceil H_*/H \rceil$ consecutive blocks. This expansion is needed for every block when $s_w < nd$, and only for the first block when $s_w \geq nd$, because the full input is then stored for all subsequent blocks. Steps (d) and (e) complete the construction to the prescribed depth and dimensions.

(a) Model width and feed-forward width. The base case stores in the working block either $\mathrm{vec}(X)$ or the input-dependent terms $A_{(u)}\,\mathrm{vec}(X)$ for the current group, whichever is shorter; a given model width leaves only $s_w = w - 2d - g$ coordinates for it. If $s_w \geq nd$, the working block holds $\mathrm{vec}(X)$ and every block evaluates q hidden units; if $s_w < nd$, it holds at most s_w input-dependent terms, so every block evaluates only $q_* = \min\{q, s_w\}$ hidden units and the remaining $q - q_*$ units are set to zero. Precisely, we take

$$s \coloneqq \begin{cases} nd, & s_w \geq nd, \\ q_*, & s_w < nd, \end{cases} \qquad (M_{(u)}, \Gamma_{(u)}) \coloneqq \begin{cases} (I_{nd},\, A_{(u)}), & s_w \geq nd, \\ (A_{(u)},\, I_{q_*}), & s_w < nd, \end{cases}$$

where the hidden units of $\mathbf{\Phi}$, padded with zero units to hidden width $L_* q_*$ as in the base case (possible since $L_* q_* \geq m$ by hypothesis, so $L_* \geq 1$), are partitioned into L_* groups of q_* units, with $A_{(u)} \in \mathbb{R}^{q_* \times nd}$, $D_{(u)}, b_{(u)} \in \mathbb{R}^{q_*}$, and $C_{(u)} \in \mathbb{R}^{d \times q_*}$. All parameters are those of the base case, with L_* blocks in place of L, s in place of $\min\{q, nd\}$, q_* in place of q, and $(M_{(u)}, \Gamma_{(u)})$ as above; in particular, $\Gamma_{(u)} M_{(u)} = A_{(u)}$, and $b_{2,(u)}$ contains c only for $u = L_*$. If $s_w \geq nd$, then $q_* = q$, and the working block equals $\mathrm{vec}(X)$ after the first block even when $q < nd$. Since no later block changes the working block, only block 1 requires replacement by a stage in (c) when $H < H_*$. If $q_* < q$ (which requires $s_w < nd$), the feed-forward sublayers are padded with $q - q_*$ zero hidden units:

$$B_{1,(u)} = \begin{pmatrix} 0_{q_* \times d} & D_{(u)}\lambda^\top & \Gamma_{(u)} & 0_{q_* \times d} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad b_{1,(u)} = \begin{pmatrix} b_{(u)} \\ 0_{q - q_*} \end{pmatrix},$$

$$B_{2,(u)} = \begin{pmatrix} 0_{(d+g+s) \times q_*} & 0 \\ C_{(u)} & 0_{d \times (q - q_*)} \end{pmatrix},$$

while $b_{2,(u)}$ is unchanged. The padded hidden units contribute nothing because their columns in $B_{2,(u)}$ are zero. In both cases, L_* blocks with n heads of dimension d and model width $w_0 \coloneqq 2d + g + s \leq w$ realize every family of hidden width $L_* q_*$.

(b) Head dimension. A head of dimension $t < d$ cannot output the d-dimensional average $\sum_j S_{hj} x_j$ of the base case. We therefore split each of the n averages into $\lceil d/t \rceil$ slices of at most t coordinates, $H_* = n\lceil d/t \rceil$ slices in total, let each slice be computed by a separate head with the same attention weights, and regroup the slices in the output projection. Precisely, let $p \coloneqq \lceil d/t \rceil$ and $I_\ell \coloneqq \{(\ell - 1)t + 1, \ldots, \min\{\ell t, d\}\}$, $\ell = 1, \ldots, p$, so that $|I_\ell| = t$ for $\ell < p$ and $1 \leq |I_p| \leq t$. Head h of (a) is replaced by p heads $(h, 1), \ldots, (h, p)$ of dimension t with

$$W_Q^{(h,\ell)} \coloneqq \begin{pmatrix} 0_{t \times d} & e_1 a^\top & 0_{t \times s} & 0_{t \times d} \end{pmatrix}, \qquad W_K^{(h,\ell)} \coloneqq \begin{pmatrix} 0_{t \times d} & e_1 k_h^\top & 0_{t \times s} & 0_{t \times d} \end{pmatrix},$$

$$W_V^{(h,\ell)} \coloneqq \begin{pmatrix} E_\ell & 0_{t \times g} & 0_{t \times s} & 0_{t \times d} \end{pmatrix},$$

where now $e_1 \in \mathbb{R}^t$, and $E_\ell \in \mathbb{R}^{t \times d}$ has the rows $e_i^\top \in \mathbb{R}^{1 \times d}$, $i \in I_\ell$, followed by $t - |I_\ell|$ zero rows. The scores $\langle W_Q^{(h,\ell)} \widetilde{z}_i, W_K^{(h,\ell)} \widetilde{z}_j \rangle = \langle k_h, \rho_j \rangle$ are those of head h, so $\Pi_{ij}^{(h,\ell)}(\widetilde{Z}) = \Pi_{ij}^{(h)}(\widetilde{Z}) = S_{hj}$, and, since a head output is linear in its values, head (h, ℓ) outputs $\sum_{j=1}^n S_{hj} E_\ell x_j = E_\ell \sum_{j=1}^n S_{hj} x_j$, i.e. the coordinates I_ℓ of $\sum_{j=1}^n S_{hj} x_j$ followed by $t - |I_\ell|$ zeros. With $E \coloneqq (E_1; \ldots; E_p) \in \mathbb{R}^{pt \times d}$ we have $E^\top E = I_d$. Ordering the heads as $(1, 1), \ldots, (1, p), (2, 1), \ldots, (n, p)$, the concatenated

head outputs are $(I_n \otimes E)(S \otimes I_d)\,\mathrm{vec}(X)$,[13] and we take the output projections

$$\widehat{W}_{O,(u)} \coloneqq W_{O,(u)}(I_n \otimes E^\top) \in \mathbb{R}^{w_0 \times npt}, \quad \text{so that} \quad \widehat{W}_{O,(u)}(I_n \otimes E) = W_{O,(u)},$$

by the Kronecker multiplication rule and $E^\top E = I_d$; applied to $(S \otimes I_d)\,\mathrm{vec}(X)$, this gives the base-case attention output, now with $H_* = n\lceil d/t \rceil$ heads of dimension t. For $t > d$ we have $p = 1$, $I_1 = \{1, \dots, d\}$, and $E_1 = (I_d; 0_{(t-d)\times d})$, so the same formulas complete each head by $t - d$ zero coordinates, and $H_* = n$. For $t = d$, step (b) changes nothing ($E = I_d$).

(c) Number of heads. In the base case, $H = H_* = n$. After (b), each attention sublayer uses $H_* = n\lceil d/t \rceil$ heads of dimension t ($H_* = n$ if $t \geq d$), and a transformer with $H < H_*$ heads cannot compute all of them in one block. Since these head outputs depend only on coordinates that remain unchanged, they can be computed in any block. We distribute their contributions over $G = \lceil H_*/H \rceil$ consecutive blocks and sum them in the working block through residual additions. Within each stage, the feed-forward output is set to zero in the first $G - 1$ blocks. The original feed-forward sublayer is applied in the last block, after all head contributions have been accumulated. Precisely, number these heads as $\kappa = 1, \dots, H_*$ in the order of (b), with matrices $W_Q^\kappa, W_K^\kappa, W_V^\kappa$, and write $\widehat{W}_{O,(u)} = (W_{O,(u)}^{[1]} \ \cdots \ W_{O,(u)}^{[H_*]})$ with $W_{O,(u)}^{[\kappa]} \in \mathbb{R}^{w_0 \times t}$, so that block u adds $\sum_{\kappa=1}^{H_*} W_{O,(u)}^{[\kappa]}[\mathrm{head}_\kappa(\widetilde{Z})]_{:,i}$ to token i. When $H < H_*$, a block u can be replaced by a stage of G blocks $(u,1), \dots, (u,G)$. Block (u,j) computes the heads $\kappa \in J_j \coloneqq \{(j-1)H + 1, \dots, \min\{jH, H_*\}\}$, so that $|J_j| = H$ for $j < G$ and $1 \leq |J_G| \leq H$: its first $|J_j|$ heads have the matrices $W_Q^\kappa, W_K^\kappa, W_V^\kappa$, $\kappa \in J_j$, its other heads are zero, and

$$W_{O,(u,j)} \coloneqq \big((W_{O,(u)}^{[\kappa]})_{\kappa \in J_j} \ \ 0_{w_0 \times (H - |J_j|)t} \big), \qquad \big(B_2, b_2\big)_{(u,j)} \coloneqq \begin{cases} (0,0), & j < G, \\ (B_{2,(u)}, b_{2,(u)}), & j = G, \end{cases}$$

with B_1, b_1 of block (u,j) equal to those of block u. As noted, $[\mathrm{head}_\kappa(\widetilde{Z})]_{:,i}$ is the same in every block of the stage, and the blocks (u,j), $j < G$, have zero feed-forward output. Hence, before its last feed-forward sublayer, the stage adds

$$\sum_{j=1}^{G} \sum_{\kappa \in J_j} W_{O,(u)}^{[\kappa]}[\mathrm{head}_\kappa(\widetilde{Z})]_{:,i} = \sum_{\kappa=1}^{H_*} W_{O,(u)}^{[\kappa]}[\mathrm{head}_\kappa(\widetilde{Z})]_{:,i},$$

which is what block u adds, so the state (13), with the changes of (a), is reached after each stage. Only blocks whose attention sublayer writes to the working block need the H_* heads, and which blocks these are depends only on the architecture. If $s_w < nd$, every block writes the input-dependent terms $A_{(u)}\,\mathrm{vec}(X)$ for its group ($M_{(u)} = A_{(u)}$) and is replaced by a stage. If $s_w \geq nd$, only block 1 writes (it stores $\mathrm{vec}(X)$, $M_{(u)} = I_{nd}$), so $\widehat{W}_{O,(u)} = 0$ for $u \geq 2$; only block 1 is replaced, and blocks $u \geq 2$ remain single blocks, with H zero heads and their feed-forward sublayers unchanged. Queries, keys, and values are those of (b), and the output projections are made of columns of $\widehat{W}_{O,(u)}$, so the parameter dependence stated in the theorem is unchanged. If $H \geq H_*$, then $G = 1$ and (c) changes nothing; extra heads are handled in (e).

(d) Depth. By (c), the construction has GL_* blocks if $s_w < nd$ and $G - 1 + L_*$ blocks if $s_w \geq nd$, in the order of the blocks of (a). By Definition 3.1, $GL_* = G\lfloor L/G \rfloor \leq L$ if $s_w < nd$ and $G - 1 + L_* = L$ if $s_w \geq nd$, so the construction fits into the L given blocks. If $s_w < nd$, the remaining $L - GL_* < G$ blocks are appended with $W_O = 0$, $B_2 = 0$, and $b_2 = 0$; they act as the identity and leave the output block unchanged.

(e) Unused dimensions. The given architecture may have a larger model width than w_0 and more heads than H_*, which the construction does not use. They receive zero weights, so that they stay zero and do not affect the other coordinates. Precisely, if $w > w_0$, every token receives $w - w_0$ additional coordinates, which we place last, and we replace the matrices as follows:

$$W \mapsto (W \quad 0) \ \ \big(W \in \{W_Q^\kappa, W_K^\kappa, W_V^\kappa, B_1, R_d\}\big), \quad W \mapsto \begin{pmatrix} W \\ 0 \end{pmatrix} \ \ \big(W \in \{E_{\mathrm{in}}, p_i, W_O, B_2, b_2\}\big),$$

[13] For fixed h, the p heads $(h,1), \dots, (h,p)$ output $E_1 \sum_j S_{hj}x_j, \dots, E_p \sum_j S_{hj}x_j$, which together form $E \sum_j S_{hj}x_j$. The vector $(S \otimes I_d)\,\mathrm{vec}(X)$ consists of the n blocks $\sum_j S_{hj}x_j$, $h = 1, \dots, n$, and applying E to each block is multiplication by the block-diagonal matrix $I_n \otimes E$.

with $w - w_0$ zero columns, respectively rows. Heads beyond H_* have $W_Q = W_K = W_V = 0$, and the corresponding columns of W_O are zero. The additional token coordinates therefore stay zero in every block, and the additional heads contribute nothing to the output projection. The resulting transformer has L blocks, model width w, H heads of dimension t, and feed-forward hidden width q.

By (a)–(e), the output at token i is $\Phi_i(\operatorname{vec} X)$. This proves Theorem 3.2.

□

Corollary B.5 (Exact preservation of derivatives)**.** *Suppose $\sigma \in C^k(\mathbb{R};\mathbb{R})$ and $\mathbf{\Phi} \in \mathcal{S}_{n,d}(\sigma)$. Each transformer $T_{\mathbf{\Phi}}$ of Theorem 3.2 belongs to $C^k(\mathbb{R}^{d\times n};\mathbb{R}^{d\times n})$, satisfies* (3)*, and for $1 \le j \le k$ and directions $\Xi_1, \dots, \Xi_j \in \mathbb{R}^{d\times n}$ satisfies, for every $i = 1, \dots, n$,*

$$\mathrm{D}^j[T_{\mathbf{\Phi}}]_{:,i}(X)[\Xi_1, \dots, \Xi_j] = \mathrm{D}^j\Phi_i(\operatorname{vec} X)[\operatorname{vec}\Xi_1, \dots, \operatorname{vec}\Xi_j].$$

Proof of Corollary B.5. Equation (3) is an identity on the whole Euclidean space and vec is linear. Differentiate the identity j times. The positional vectors p_i are fixed parameters. □

Corollary B.6 (Density transfer principle)**.** *Let $\mathcal{F}$ be a class of maps $\mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$ equipped with a topology τ (e.g. induced by a metric or by seminorms). If $\{X \mapsto \mathbf{\Phi}(\operatorname{vec} X) : \mathbf{\Phi} \in \mathcal{S}_{n,d}(\sigma)\} \subseteq \mathcal{F}$ is dense in $\mathcal{F}$, then $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma) \cap \mathcal{F}$ is dense in $\mathcal{F}$ with respect to τ.*

Proof of Corollary B.6. By Theorem 3.2, the dense shallow-network subclass is contained in $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma) \cap \mathcal{F}$. □

B.2 Proofs for Section 4: Universal Approximation Theorems

Definition B.7 (Shallow networks)**.** For integers $N, d \ge 1$, $\mathcal{N}_{N,d}(\sigma)$ denotes the set of maps $z \mapsto C\sigma^\bullet(Bz + b) + c$ from $\mathbb{R}^N$ to $\mathbb{R}^d$ with $B \in \mathbb{R}^{m\times N}$, $b \in \mathbb{R}^m$, $C \in \mathbb{R}^{d\times m}$, and $c \in \mathbb{R}^d$ for some hidden width $m \ge 1$, and, for $M \ge 1$, $\mathcal{N}^M_{N,d}(\sigma)$ denotes its subset of networks of hidden width $m \le M$.

In the proofs below we proceed as follows.

Given a target[14] $f : \mathbb{R}^{d\times n} \to \mathbb{R}^{d\times n}$, $f(X) = (f_1(X), \dots, f_n(X))$, we first construct an auxiliary map $\widetilde{f}$ whose restriction to each positional slice recovers the corresponding component f_i:

$$\widetilde{f} : \mathbb{R}^{nd} \times \mathbb{R} \to \mathbb{R}^d, \quad \widetilde{f}(\xi, \eta) := \sum_{i=1}^{n} \chi_i(\eta) f_i(\operatorname{vec}^{-1}\xi), \qquad \widetilde{f}(\operatorname{vec} X, i) = f_i(X),$$

where the smooth cutoffs $\chi_i : \mathbb{R} \to \mathbb{R}$ have disjoint supports and equal 1 near i. For a finite Borel measure μ, put $\widetilde{\mu} := (\operatorname{vec}_{\#}\mu) \otimes \sum_{i=1}^n \delta_i$. For an approximant $\Phi \in \mathcal{N}_{nd+1,d}(\sigma)$, set $\Phi_i(\xi) := \Phi(\xi, i)$. The identification on positional slices preserves the componentwise approximation errors: for a compact set K, $1 \le p < \infty$, and $\alpha \in \mathbb{N}_0^{nd}$ (including $\alpha = 0$), whenever the derivatives exist,

$$\max_{1\le i\le n} \sup_{X\in K} \left\| \partial_X^\alpha [\Phi_i(\operatorname{vec} X) - f_i(X)] \right\| = \max_{1\le i\le n} \sup_{\xi \in \operatorname{vec} K} \left\| \partial_\xi^\alpha (\Phi - \widetilde{f})(\xi, i) \right\|,$$

$$\sum_{i=1}^{n} \int \left\| \partial_X^\alpha [\Phi_i(\operatorname{vec} X) - f_i(X)] \right\|^p d\mu(X) = \int \left\| \partial_\xi^\alpha (\Phi - \widetilde{f}) \right\|^p d\widetilde{\mu}.$$

Applying the corresponding classical approximation theorem from Hornik (1991); Pinkus (1999) (detailed below) to each output coordinate of $\widetilde{f}$ and concatenating the resulting hidden units gives a single vector-valued shallow network $\Phi(\xi, \eta) = C\sigma^\bullet(B(\xi, \eta) + b) + c$ with the displayed errors sufficiently small. Writing $B = (A, D)$, its restrictions to the positional slices satisfy

$$\Phi_i(\xi) := \Phi(\xi, i) = C\sigma^\bullet(A\xi + Di + b) + c.$$

Thus $\mathbf{\Phi} = (\Phi_i)_{i=1}^n$ belongs to the shared-family class $\mathcal{S}_{n,d}(\sigma)$ in (2) and is realized exactly by Theorem 3.2 as a transformer $T_{\mathbf{\Phi}}$. For the derivative statements, $\widetilde{f}$ is globally C^k whenever f is.

[14] For the continuous target in Theorem 4.1, first extend it continuously, componentwise, from the compact set to the whole input space.

Since the cutoffs are constant near each i, all derivatives involving η vanish on the support of $\widetilde{\mu}$; thus $\widetilde{f}$ also belongs to the required weighted derivative space whenever f does. The required content derivatives correspond to $(\alpha, 0)$, so classical control of all derivatives in (ξ, η) controls every required content derivative.

Proof of Theorem 4.1. Theorem 3.1 of Pinkus (1999) provides the approximation of $\widetilde{f}$ required in the positional-slice construction above. Exact realization of the resulting shared family via Theorem 3.2 yields the stated transformer density. □

Proof of Theorem 4.2. Theorem 1 of Hornik (1991) provides the approximation of $\widetilde{f}$ required in the positional-slice construction above. Exact realization of the resulting shared family via Theorem 3.2 yields the stated transformer density. Since σ is bounded, the shallow approximants are bounded; exact realization preserves this property, so the resulting transformers belong to $L^p_\mu(\mathbb{R}^{d\times n}; \mathbb{R}^{d\times n})$ for every finite measure μ. □

Proof of Theorem 4.3. Both conclusions of Theorem 3 of Hornik (1991), applied in dimension $nd+1$, provide the approximation of $\widetilde{f}$ required in the positional-slice construction above. Exact realization of the resulting shared family via Theorem 3.2 yields the stated transformer density. By Corollary B.5, realization changes neither the function nor any derivative through order k. □

Proof of Theorem 4.4. Theorem 4 of Hornik (1991), applied in dimension $nd+1$ with the measure $\widetilde{\mu}$, provides the approximation of $\widetilde{f}$ required in the positional-slice construction above. Exact realization of the resulting shared family via Theorem 3.2 yields the stated transformer density. The shallow approximants and their derivatives through order k are bounded, hence integrable to the power p against the finite measure. Exact realization and Corollary B.5 preserve every term in (8); in particular, every realized approximant belongs to the displayed intersection. □

B.3 Proofs for Section 5: Rates

We transfer classical shallow-network approximation rates to transformers. Unlike the auxiliary map $\widetilde{f}(\xi, \eta)$ in Appendix B.2, the construction below encodes position by translating the input within $\mathbb{R}^{nd}$, rather than adding an input coordinate (which might worsen exponents). The argument has four steps.

1. *Combine the target components into one function.*[15] Given $f = (f_1, \dots, f_n)$ on B_R, identify B_R with the ball in $\mathbb{R}^{nd}$. Choose $v = 6Re_1$ (where e_1 is the first standard basis vector of $\mathbb{R}^{nd}$), set $v_i = iv$, and put $R' = (6n+2)R$. The balls $v_i + B_{2R}$ are disjoint and lie in $B_{R'}$. Define

$$\overline{f} : \mathbb{R}^{nd} \to \mathbb{R}^d, \qquad \overline{f}(z) := \sum_{i=1}^{n} \theta(z - v_i) f_i(\mathrm{vec}^{-1}(z - v_i)), \qquad z \in \mathbb{R}^{nd},$$

   where each f_i is extended from B_R to all of $\mathbb{R}^{d\times n}$ with the norm control described in Remarks A.1 and A.2, and $\theta \in C^\infty_c(B_{2R}; \mathbb{R})$ equals 1 on a neighbourhood of B_R. Then, disjoint supports ensure

$$\overline{f}(\xi + iv) = f_i(\mathrm{vec}^{-1}\,\xi), \qquad \xi \in B_R.$$

2. *Approximate $\overline{f}$ by one shallow network.* Consider $\mathcal{N}^{L_* q}_{nd,d}(\sigma)$ (Definition B.7), whose hidden widths are those realized in $\mathcal{T}_{\mathfrak{a}}(\sigma)$ by Theorem 3.2. Apply the classical scalar rate estimates componentwise (detailed below for each theorem), allocating $\overline{m} := \lfloor L_* q/d \rfloor$ hidden units to each of the d output coordinates $\overline{f}_a$, $a = 1, \dots, d$. Stacking these scalar networks[16]

[15] This disjoint-translation construction does not directly cover all the density results of Appendix B.2: for a measure μ with unbounded support, translated copies of $\operatorname{supp}\mu$ need not be disjoint. The positional-slice construction used there avoids this obstruction.

[16] If the scalar networks are $\phi_a(z) = c_a^\top \sigma^\bullet(A_a z + b_a) + c_{0,a}$, $a = 1, \dots, d$, with $A_a \in \mathbb{R}^{\overline{m}\times nd}$ and $b_a, c_a \in \mathbb{R}^{\overline{m}}$, then $\phi(z) = C\sigma^\bullet(Az + b) + c$ with $A = (A_1; \dots; A_d)$, $b = (b_1; \dots; b_d)$, $C = \mathrm{diag}(c_1^\top, \dots, c_d^\top) \in \mathbb{R}^{d\times d\overline{m}}$, and $c = (c_{0,1}, \dots, c_{0,d})$, so that $\phi(z)_a = \phi_a(z)$.

gives an approximant ϕ of width $m = d\overline{m} \le L_* q$, to which Theorem 3.2 applies. Here the extension and cutoff bounds in Remarks A.1 and A.2 control the norm of $\overline{f}$ appearing in the classical rate by the corresponding norm of the original target f, with constants independent of f, L, q, and H. For example, under the assumptions of Theorem 5.3 and Theorem 5.4, respectively,

$$\left\|\overline{f}-\phi\right\|_{W^{k,p}(B_{R'})} \le C\overline{m}^{-(s-k)/(nd)} \left\|\overline{f}\right\|_{W^{s,p}(B_{R'})} \le C'(Lq)^{-(s-k)/(nd)} \left\|f\right\|_{W^{s,p}(B_R)},$$
$$\left\|\overline{f}-\phi\right\|_{C^k_{B_{R'}}} \le C\overline{m}^{-(r+\alpha-k)/(nd)} \left\|\overline{f}\right\|_{C^{r,\alpha}} \le C'(Lq)^{-(r+\alpha-k)/(nd)} \left\|f\right\|_{C^{r,\alpha}}.$$

The first inequalities are the classical rates with $\overline{m}$ units per component. For the second, the extension and cutoff bounds of Remarks A.1 and A.2 give $\|\overline{f}\| \le C_{\mathrm{ext}}\|f\|$, and $\mathfrak{a} \in \mathcal{A}$ implies[17] $\overline{m} \ge Lq/(4nd^2)$ whenever $L_* q \ge d$. Hence $C' = C\,C_{\mathrm{ext}}(4nd^2)^{\gamma}$ with $\gamma = (s-k)/(nd)$, respectively $(r+\alpha-k)/(nd)$; it depends only on n, d and the classical constant, not on $\mathfrak{a}$. The same conversion applies to all rates below. For small $\overline{m}$ (including $\overline{m} = 0$, i.e. $L_* q < d$; hence small Lq), we take $T = 0$ and, possibly, enlarge the constant.

3. *Realize the translated evaluations with a transformer.* Write $\phi(z) = C\sigma^{\bullet}(Az + b) + c$, $D := Av$ and
$$\Phi_i(\xi) := \phi(\xi + iv) = C\sigma^{\bullet}(A\xi + Di + b) + c.$$
This is precisely the shared-family form in (2). Since ϕ has at most $L_* q$ hidden units, Theorem 3.2 gives $T \in \mathcal{T}_{\mathfrak{a}}(\sigma)$ with
$$[T(X)]_{:,i} = \phi(\mathrm{vec}\, X + iv).$$
The depth is L, the feed-forward hidden width is q, the number of heads is H, the head dimension is t, and the model width is $2d + g + \min\{q, nd\}$.

4. *Transfer the approximation error.* On $\Omega = \bigcup_i (v_i + B_R)$, translation preserves the componentwise Lebesgue L^p, Sobolev, and uniform derivative errors. For the transformer T constructed in Step 3, $k \in \mathbb{N}_0$, and $1 \le p \le \infty$, whenever the functions involved belong to the indicated spaces, we have
$$\begin{aligned} \|f - T\|_{W^{k,p}(B_R)} &= \left\|\overline{f}-\phi\right\|_{W^{k,p}(\Omega)} \le \left\|\overline{f}-\phi\right\|_{W^{k,p}(B_{R'})}, \\ \|f - T\|_{C^k_{B_R}} &= \left\|\overline{f}-\phi\right\|_{C^k_{\Omega}} \le \left\|\overline{f}-\phi\right\|_{C^k_{B_{R'}}}. \end{aligned} \tag{14}$$
Combining (14) with the bounds of Step 2, we obtain the required bounds. For a finite measure μ, use the sum of its translates on Ω instead of Lebesgue measure, normalizing it when a probability measure is required. Thus approximation of $\overline{f}$ gives simultaneous approximation of all f_i, with all the rate exponents below governed by dimension nd.

Each proof below, except that of Theorem 5.6, which modifies Step 1, supplies the scalar bound of Step 2 for a fixed component $\overline{f}_a$, $a \in \{1, \dots, d\}$, on $B_{R'}$ with $\overline{m} = \lfloor L_* q/d \rfloor$ hidden units.

Proof of Theorem 5.1. Mao et al. (2026, Corollary 1.3) in (i), and Pinkus (1999, Corollary 6.11) in (ii), provide the scalar shallow-network approximation bounds required in Step 2. Applying Steps 1–4 above yields the stated transformer bound. □

Proof of Theorem 5.2. Mao et al. (2026, Corollary 1.2) provide the scalar shallow-network approximation bound required in Step 2. Applying Steps 1–4 above yields the stated transformer bound. □

Proof of Theorem 5.3. For sufficiently large $\overline{m}$, choose an integer κ such that $(\kappa+1)^{nd} \le \overline{m}$ and $\kappa \asymp \overline{m}^{1/(nd)}$. As in Li & Xu (2014, Theorem 4.2), there is a polynomial P_κ of total degree at most κ with
$$\left\|\overline{f}_a - P_\kappa\right\|_{W^{k,p}(B_{R'})} \le C\kappa^{-(s-k)} \left\|\overline{f}_a\right\|_{W^{s,p}(B_{R'})}.$$

[17] We have $L_* \ge L/(2nd)$. Indeed, $L \ge G \ge 1$ for $\mathfrak{a} \in \mathcal{A}$, and in both cases of Definition 3.1, $L_* \ge \lfloor L/G \rfloor$, since $(L-G+1)G - L = (L-G)(G-1) \ge 0$ gives $L - G + 1 \ge L/G$. As $\lfloor x \rfloor \ge x/2$ for $x \ge 1$, it follows that $L_* \ge L/(2G) \ge L/(2nd)$, because $G = \lceil H_*/H \rceil \le H_* = n\lceil d/t \rceil \le nd$. Hence, if $L_* q \ge d$, then $\overline{m} = \lfloor L_* q/d \rfloor \ge L_* q/(2d) \ge Lq/(4nd^2)$, using $\lfloor x \rfloor \ge x/2$ once more.

As in Ito (1993, Proposition 4 and the subsequent shift argument), P_κ can be approximated arbitrarily closely in $C^k(B_{R'};\mathbb{R})$ by a shallow network Φ_κ of width at most $(\kappa+1)^{nd} \leq \overline{m}$ and such that

$$\|P_\kappa - \Phi_\kappa\|_{W^{k,p}(B_{R'})} \leq \kappa^{-(s-k)} \left\|\overline{f}_a\right\|_{W^{s,p}(B_{R'})}.$$

The triangle inequality gives the shallow-network bound required in Step 2; Steps 3–4 then yield the stated transformer bound. □

Proof of Theorem 5.4. As in Goebbels (2022, Lemmas 1–2), use a common polynomial to approximate all required derivatives; the lemmas apply since $\overline{f}_a$ has compact support in $B_{R'}$ (Step 1). For sufficiently large $\overline{m}$, choose an integer $\kappa \geq k$ such that $(\kappa+1)^{nd} \leq \overline{m}$ and $\kappa \asymp \overline{m}^{1/(nd)}$. Lemma 2 gives one polynomial of degree at most κ, and Lemma 1 approximates it and all derivatives through order k arbitrarily closely using $(\kappa+1)^{nd} \leq \overline{m}$ neurons. Since $\kappa \asymp \overline{m}^{1/(nd)}$ and $\omega_1(\partial^\beta \overline{f}_a, \delta)_\infty \leq \delta^\alpha \left\|\overline{f}_a\right\|_{C^{r,\alpha}}$ for $|\beta| = r$ (the modulus of continuity Goebbels, 2022, p. 638), the bound of Lemma 2 gives the C^k bound of Step 2 for the resulting network. Steps 3–4 then yield the transformer bound. □

Proof of Corollary 5.5. Take the transformer T of Theorem 5.4; since μ is supported in B_R, each of the ndN_k terms in (8), where $N_k := \#\{\beta \in \mathbb{N}_0^{nd} : |\beta| \leq k\}$, is at most $\mu(B_R)\,\|f - T\|^p_{C^k_{B_R}}$, and $(ndN_k)^{1/p} \leq ndN_k$. □

Proof of Theorem 5.6. We apply Steps 1–4 with the following modification: in Step 1, subtract $a_i := f_i(0)$ from each component f_i, so that $\overline{f}(\xi + iv) = f_i(\mathrm{vec}^{-1}\,\xi) - a_i, \xi \in B_R$. The estimate $\left\|\theta(e^{\mathrm{i}\omega\cdot(\cdot)} - 1)\right\|_{\mathcal{B}^1} \leq C_\theta|\omega|$ (obtained by differentiating the translated Fourier transform of θ for $|\omega| \leq 1$ and by the triangle inequality otherwise) controls the localized extension by C_F. In Step 2, apply Barron's theorem (Barron, 1993, Theorem 1) componentwise. In Step 3, additionally place a_i in the final d coordinates of p_i; these coordinates are ignored by the attention and feed-forward input matrices and pass through the residual connection. The resulting transformer $T_{\phi,a}$ retains the stated architecture. The offsets add at most nd nonzero positional entries, so its nonzero parameter count is at most $P_{\mathrm{nz}}(L,q)+nd$, while its dense count remains $P_{\mathfrak{a}}$. It satisfies $[T_{\phi,a}(X)]_{:,i} = \phi(\mathrm{vec}\,X + iv) + a_i$. The offsets cancel in the approximation error, so Step 4 yields the stated bound. For total hidden width $L_* q < d$, use the constant outputs a_i and enlarge the constant in the bound. □

Proof of Theorem 5.7. Siegel & Xu (2020, Corollary 1) provide the scalar shallow-network approximation bound required in Step 2, with the fixed width factor depending on J absorbed into the constant. Applying Steps 1–4 above yields the stated transformer bound. □

Proof of Theorem 5.8. Siegel & Xu (2020, Theorem 5) provide the scalar shallow-network approximation bound with activation ν. Expanding each ν-neuron into J σ-neurons, as in Siegel & Xu (2020, Corollary 1), gives the bound required in Step 2, with the fixed width factor absorbed into the constant. Applying Steps 1–4 above yields the stated transformer bound. □

Proof of Theorem 5.9. Siegel & Xu (2022, Theorem 1) provide the scalar shallow-network approximation bound required in Step 2. Applying Steps 1–4 above yields the stated transformer bound. □

B.4 Proofs for Section 6: Optimal Rates under Continuous Selection

B.4.1 Upper bound with continuous selection

Proposition B.8 (Function-value upper bound with continuous selection)**.** *Let $\sigma \in C^\infty(\mathbb{R};\mathbb{R})$ be nonpolynomial, let $s \geq 1$ be an integer, and let $1 \leq p < \infty$. There exists $C > 0$ such that, for every $\mathfrak{a} \in \mathcal{A}$, there is an affine map $\mathsf{S}_\mathfrak{a} : \mathcal{U}_{s,p} \to \mathbb{R}^{P_\mathfrak{a}}$, continuous in L^p, such that*

$$\sup_{f \in \mathcal{U}_{s,p}} \|f - T_\mathfrak{a}(\mathsf{S}_\mathfrak{a}(f))\|_{L^p} \leq C(Lq)^{-s/(nd)}.$$

Only the feed-forward output coefficients depend on f, linearly; all other parameters are fixed for each $\mathfrak{a}$. Consequently, with $C_{n,d}$ from (5)*, there is $C' > 0$ such that, for every integer budget P and every $\mathfrak{a} \in \mathcal{A}$ with $P_{\mathfrak{a}} \le P \le 2C_{n,d}Lq$,*

$$\sup_{f \in \mathcal{U}_{s,p}} \|f - T_{\mathfrak{a}}(\mathsf{S}_{\mathfrak{a}}(f))\|_{L^p} \le C' P^{-s/(nd)}.$$

Remark B.9. In Proposition B.8, note that:

- every $\mathfrak{a} \in \mathcal{A}$ with $P/(2C_{n,d}) \le Lq \le P/C_{n,d}$ satisfies $P_{\mathfrak{a}} \le P \le 2C_{n,d}Lq$. Indeed, $P_{\mathfrak{a}} \le C_{n,d}Lq \le P$ by (5), and $P \le 2C_{n,d}Lq$;
- such architectures exist for every $P \ge P_0 := 2C_{n,d}$, for instance a single block ($L = 1$, $H = H_*$). Indeed, for $P \ge 2C_{n,d}$ the interval $[P/(2C_{n,d}), P/C_{n,d}]$ has length at least 1 and hence contains an integer q, and with $L = 1$ and $H = H_*$ we have $G = 1 = L$, so this architecture lies in $\mathcal{A}$.

Proof of Proposition B.8. We run Steps 1–4 of Appendix B.3 with the following additions. In Step 2 we take, for each component, the approximant of Pinkus (1999, Theorem 6.8 and the following remark, pp. 176–177)[18] (see also Mhaskar, 1996), which attains the rate $\overline{m}^{-s/(nd)}$ with $\overline{m}$ hidden units. Its input weights and hidden biases are fixed and its output coefficients depend linearly on the target, continuously in L^p in the sense of Pinkus (1999, Theorem 6.6). Since the map $f \mapsto \overline{f}$ of Step 1 is linear and bounded in L^p (Remark A.1 with $s = 0$; the cutoff and the translations are L^p-bounded), the coefficients C, c of ϕ depend linearly and continuously on $f \in \mathcal{U}_{s,p}$. In Step 3, Theorem 3.2 with A, D, b fixed makes only the second feed-forward parameters of $T_{\mathfrak{a}}(\mathsf{S}_{\mathfrak{a}}(f))$ depend on f, linearly. Steps 2 and 4 then give the bound $C(Lq)^{-s/(nd)}$, and the hypothesis $P \le 2C_{n,d}Lq$ turns it into $C' P^{-s/(nd)}$. □

B.4.2 Classical continuous-information lower bound

As in Pinkus (1999, Theorem 6.5, pp. 173–175), we obtain the following lower bound.

Proposition B.10 (Classical continuous-information lower bound)**.** *Let $s \ge 1$ be an integer and let $1 \le p < \infty$. There exists $c > 0$ such that, for every integer $P \ge 1$, every L^p-continuous map $\mathsf{S} : \mathcal{U}_{s,p} \to \mathbb{R}^P$ and every map $\mathsf{R} : \mathbb{R}^P \to L^p(B_R; \mathbb{R}^{d\times n})$ satisfy*[19]

$$\sup_{f \in \mathcal{U}_{s,p}} \|f - \mathsf{R}(\mathsf{S}(f))\|_{L^p} \ge c P^{-s/(nd)}.$$

The scalar lower bound applies by embedding a scalar target in one output entry and projecting the reconstruction onto that entry; projection cannot increase the full sequence error. Here R is an arbitrary reconstruction map from parameter vectors to functions in $L^p(B_R; \mathbb{R}^{d\times n})$. For transformers, it is the realization map of a fixed architecture, as used in the proof of Theorem 6.1.

Proof of Theorem 6.1. For the upper bound, let $P \ge P_0$, take an architecture as in Remark B.9, and use its selector from Proposition B.8. For the lower bound, fix an architecture $\mathfrak{a}$ with $P_{\mathfrak{a}} \le P$ and a continuous selector in (6).[20] Pad the selector's parameter vector with zeros to obtain a continuous map into $\mathbb{R}^P$. Apply Proposition B.10 with $\mathsf{R}(\vartheta) = T_{\mathfrak{a}}(\vartheta_1, \dots, \vartheta_{P_{\mathfrak{a}}})$. The lower constant is independent of the selector and architecture, so taking both infima gives the result. □

[18] Recall also the notation at the beginning of Section 6 of Pinkus (1999, p. 168): Pinkus writes $B_p^m = \{f \in W_p^m(B^n) : \|f\|_{m,p} \le 1\}$ for the Sobolev unit ball. Here his B^n is the unit ball in ambient dimension n, corresponding to our dimension $N = nd$, and his smoothness index m corresponds to our s.

[19] No continuity assumption on R is needed.

[20] By (1), every architecture has $P_{\mathfrak{a}} = w(2d + n) + L(4Htw + 2qw + q + w) \ge 2d + n + 8$, since $w, L, H, t, q \ge 1$. If, for possibly small P, no architecture satisfies $P_{\mathfrak{a}} \le P$, then $\mathcal{E}^{s,p}_{P,\mathrm{cont}}(\sigma) = \inf \emptyset = +\infty$ and the lower bound holds trivially.

C TRANSFORMERS WITH DEEP FEED-FORWARD SUBLAYERS

The main text treats the encoder block of Definition 2.1, whose feed-forward sublayer has one hidden layer. This choice matches the shallow-network approximation results used in the preceding sections. We now extend Theorem 3.2 to feed-forward sublayers with several hidden layers and indicate how approximation results for deep networks can be transferred.

The transformer class. Fix $K \geq 1$ and replace the feed-forward map of Definition 2.1(ii) by a network with K hidden layers of common width q. For $Y \in \mathbb{R}^{w\times n}$, define

$$\mathrm{FF}(Y) \coloneqq B_{K+1}\,\sigma^{\bullet}\big(B_K\sigma^{\bullet}(\cdots B_2\,\sigma^{\bullet}(B_1Y + b_1\mathbf{1}_n^\top) + b_2\mathbf{1}_n^\top\cdots) + b_K\mathbf{1}_n^\top\big) + b_{K+1}\mathbf{1}_n^\top.$$

Here $B_1 \in \mathbb{R}^{q\times w}$, $B_j \in \mathbb{R}^{q\times q}$ for $2 \leq j \leq K$, $B_{K+1} \in \mathbb{R}^{w\times q}$, $b_1, \ldots, b_K \in \mathbb{R}^q$, and $b_{K+1} \in \mathbb{R}^w$; for $K = 1$ this is the standard sublayer. This map applies the same network with K hidden layers separately to each column of Y. Let $\mathcal{T}^{\mathrm{PE}}_{n,d,\mathrm{deep}}(\sigma) \supseteq \mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$ be the resulting class of transformers, with the union also taken over K, and let $\mathcal{S}_{n,d,\mathrm{deep}}(\sigma) \supseteq \mathcal{S}_{n,d}(\sigma)$ be the class of deep shared families

$$\Phi_i(\xi) = C\sigma^{\bullet}\big(A^{(K)}\sigma^{\bullet}(\cdots A^{(2)}\sigma^{\bullet}(A^{(1)}\xi + Di + b^{(1)}) + b^{(2)}\cdots) + b^{(K)}\big) + c, \quad i = 1, \ldots, n, \tag{15}$$

with $A^{(1)} \in \mathbb{R}^{q\times nd}$, $D, b^{(1)} \in \mathbb{R}^q$, $A^{(j)} \in \mathbb{R}^{q\times q}$ and $b^{(j)} \in \mathbb{R}^q$ for $2 \leq j \leq K$, $C \in \mathbb{R}^{d\times q}$, and $c \in \mathbb{R}^d$ shared across positions; for $K = 1$ these are the families of (2) with $m = q$.

Proposition C.1 (Exact realization with deep feed-forward sublayers)**.** *Fix $n \geq 2$, $d \geq 1$, and an arbitrary activation function $\sigma : \mathbb{R} \to \mathbb{R}$. Every $\mathbf{\Phi} \in \mathcal{S}_{n,d,\mathrm{deep}}(\sigma)$, represented as in* (15)*, is realized in the sense of* (3) *by a transformer in $\mathcal{T}^{\mathrm{PE}}_{n,d,\mathrm{deep}}(\sigma)$ with $L = 1$, n heads of dimension d, model width $2d + g + \min\{q, nd\}$, where g is the dimension of the positional encoding (Definition B.1), and a feed-forward sublayer with K hidden layers of width q. Its parameters depend on those of $\mathbf{\Phi}$ as in Theorem 3.2, and its hidden layers $2, \ldots, K$ are those of $\mathbf{\Phi}$. Corollaries B.5 and B.6 hold with $\mathcal{S}_{n,d,\mathrm{deep}}(\sigma)$ and $\mathcal{T}^{\mathrm{PE}}_{n,d,\mathrm{deep}}(\sigma)$ in place of $\mathcal{S}_{n,d}(\sigma)$ and $\mathcal{T}^{\mathrm{PE}}_{n,d}(\sigma)$.*

Proof of Proposition C.1. The proof is the same as the base case of the proof of Theorem 3.2 with $L = 1$ and with $(A^{(1)}, D, b^{(1)})$ in place of (A, D, b), so that $s = \min\{q, nd\}$, $\Gamma_{(1)} = A^{(1)}$ if $q \geq nd$, and $\Gamma_{(1)} = I_q$ if $q < nd$; only the feed-forward sublayer changes. We choose its parameters as

$$\begin{aligned}
&B_1 \coloneqq \begin{pmatrix} 0_{q\times d} & D\lambda^\top & \Gamma_{(1)} & 0_{q\times d}\end{pmatrix}, && b_1 \coloneqq b^{(1)},\\
&B_j \coloneqq A^{(j)}, && b_j \coloneqq b^{(j)}, \quad 2 \leq j \leq K,\\
&B_{K+1} \coloneqq \begin{pmatrix} 0_{d\times q}\\ 0_{g\times q}\\ 0_{s\times q}\\ C\end{pmatrix}, && b_{K+1} \coloneqq \begin{pmatrix} 0_d\\ 0_g\\ 0_s\\ c\end{pmatrix}.
\end{aligned}$$

At token i, the first hidden preactivation is $A^{(1)}\,\mathrm{vec}(X) + Di + b^{(1)}$, since $\Gamma_{(1)}M_{(1)} = A^{(1)}$ and $\lambda^\top\rho_i = i$. The subsequent hidden layers coincide with those of $\mathbf{\Phi}$. Thus the feed-forward output has $\Phi_i(\mathrm{vec}(X))$ in its final d coordinates and zero elsewhere. Since the output block is zero before this sublayer, the residual addition followed by the readout gives $\Phi_i(\mathrm{vec}(X))$. □

Transfer of deep-network results. By Proposition C.1, the positional-slice and translation arguments of Sections 4 and 5 extend, with a single block, to deep feed-forward networks under the activation and norm assumptions of the source results. The realization preserves the hidden depth and width of the source network, so the resulting density statements transfer to $\mathcal{T}^{\mathrm{PE}}_{n,d,\mathrm{deep}}(\sigma)$, and so do the approximation bounds, with the same exponents in the width and depth of the feed-forward sublayer or in its number of nonzero weights. Examples are the universal approximation theorem for deep narrow networks (Kidger & Lyons, 2020) and the rates of Yarotsky (2017); Gühring et al. (2020); Lu et al. (2021); Siegel (2023); De Ryck et al. (2021).