APPROXIMATION AND GENERALIZATION THEORY OF NEURAL OPERATOR TRANSFORMERS FOR FRÉCHET-HÖLDER OPERATOR CLASS

Anonymous authors
Paper under double-blind review

ABSTRACT

While Transformer-based networks have exhibited remarkable performance in operator learning, the theoretical analysis remains underdeveloped. In this paper, we establish a comprehensive approximation and generalization theory of neural operator Transformers for learning operators in the Fréchet-Hölder class. This class extends the Lipschitz continuous and k times Fréchet differentiable classes commonly considered in operator learning. For target operators in the considered Fréchet-Hölder class, we derive uniform approximation bounds on Sobolev balls of arbitrary radius, explicitly specifying network size required for a prescribed accuracy. Under a Sobolev sub-exponential assumption on the input distribution, we establish non-asymptotic L^2 generalization bounds for the neural operator Transformer estimator. By proving a matching lower bound, we demonstrate that neural operator Transformers achieve the minimax optimality over the Fréchet-Hölder class. We further extend our analysis to practical settings where input and output functions are observed only on random grids, providing explicit error bounds in the number of training functions and the number of per-function observations. Numerical experiments on PDE solution operators further verify and support our theoretical analysis.

1 INTRODUCTION

Many modern scientific computing tasks require evaluating maps between function spaces repeatedly, with examples including PDE solution operators that maps coefficients, sources, or initial states to solution trajectories, and forward models inside an inverse problem or a design loop. While conventional numerical methods require computationally expensive simulations for every evaluation, operator learning resolves this bottleneck by training neural surrogates on pairs of input-output functions (Chen & Chen, 1995; Li et al., 2021; Lu et al., 2021; Kovachki et al., 2023). Transformers and other attention-based architectures have emerged as state-of-the-art paradigms for operator learning (Cao, 2021; Kissas et al., 2022; Hao et al., 2023; Li et al., 2023; Alkin et al., 2024; Herde et al., 2024; Wu et al., 2024). These neual operators typically converts a continuous function into a structured sequence of tokens, such as point values, patches, or basis coefficients, and use the attention mechanisms to capture complex interactions over function spaces.

Despite the success achieved by Transformer-based neural operators, the theoretical foundations falls behind. In the literature of neural operators, approximation and generalization guarantees are available for DeepONet, the Fourier neural operator, PCA-Net, and related encoder-decoder architectures (Kovachki et al., 2021; Lanthaler et al., 2022; Lanthaler, 2023; Liu et al., 2024; Cheng et al., 2026; Liu et al., 2026). For Transformer neural operators, the universal approximation theorem has been established (Calvello et al., 2025). Quantitative approximation and estimation theories for Transformers are developed for finite-dimensional input and infinite sequential input with shift-equivariant regression functions (Edelman et al., 2022; Gurevych et al., 2022; Takakura & Suzuki, 2023; Havrilla & Liao, 2024). For Transformers operating on functions, three questions remain open:

Q1. How large a Transformer network must be to approximate a regular (e.g., smooth) nonlinear operator to a prescribed accuracy?

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.

Q2. How fast the least-squares estimator converges in the number n of training functions?

Q3. Whether any other estimator could converge faster?

Such questions are delicate because natural classes of operators are very large. For Lipschitz or Fréchet differentiable operators on infinite-dimensional spaces, the numbers of parameters and of samples needed for accuracy ϵ grow exponentially in a power of $1/\epsilon$ (Kovachki et al., 2024a; Lanthaler & Stuart, 2026). A useful theory therefore must identify how the regularity of the operator and the regularity of the input space enter the achievable rate, and whether a given architecture attains it.

We study neural operator Transformers acting on coefficient sequences. By fixing orthonormal bases of the input and output Hilbert spaces, we convert an operator $\Psi_0 : \mathcal{X} \to \mathcal{Y}$ into a map between sequences in $\ell^2(\mathbb{Z})$. A Transformer with a fixed positional encoding and sliding-window softmax attention reads one token per input coordinate and returns the output coordinates (Section 2.1). The target operator is assumed to lie in the Fréchet-Hölder class $\mathcal{H}^\beta$, which means it has bounded Fréchet derivatives up to order $\lfloor\beta\rfloor$, and the derivative of highest order is Hölder continuous. This class includes the bounded Lipschitz operators ($\beta = 1$) and the k times Fréchet differentiable operators with bounded derivatives. We allows the input distribution to have unbounded support, only assuming it has a sub-exponential tail in terms of a Sobolev-type norm of order α, which measures how fast the input sequences decay in the index. Our contributions are summarized as follows.

1. Approximation analysis. For every accuracy ϵ and radius R, we construct a Transformer with two attention blocks that approximates Ψ_0 with error uniformly bounded by ϵ on the Sobolev ball Ω_R^α (Theorem 1), with explicit bounds on its depth, width, heads, window size, weight count, and weight-norm bound. Up to a logarithmic factor, the logarithm of the network size is of order $(R\epsilon^{-1/\beta})^{1/\alpha}$, which is determined by the number of coordinates needed to achieved the prescribed sequence clipping error. In the construction, the self-attention layers aggregate these input coordinates into a single token and disperse the single-token outcomes to obtain the output coordinates, while the feedforward layers approximates the Taylor expansion locally.

2. Generalization analysis. We combine the approximation theorem with a truncation technique for Sobolev sub-exponential inputs, and a covering number bound for the Transformer class. This gives a non-asymptotic bound on the $L^2(\nu)$ risk of the least-squares Transformer estimator of order $(\log n)^{-2\alpha\beta}$, up to a $\log\log n$ term (Section 4). We also consider the setup where each training function is observed only at n_u random input grids and n_v random output grids. We estimate the coefficients by Monte Carlo quadrature and project the input coefficients onto a Sobolev ellipsoid. The estimator then retains the same convergence rate, up to an additional term $n_u^{-2\alpha(\beta\wedge 1)/(2\alpha+1)}$ caused by input discretization (Theorem 4). The number of output observations n_v only affects constant prefactors.

3. Minimax lower bound. No estimator converges faster than $(\log n)^{-2\alpha\beta}$ uniformly over the operator class and all input distributions with the same tail bound (Theorem 3). Hence, the neural operator Transformer is minimax optimal up to an iterated-logarithmic factor. The logarithmic rate reflects the complexity of learning operators on infinite-dimensional domains, which is consistent with the lower bounds of Kovachki et al. (2024a) and Lanthaler & Stuart (2026). The exponent $2\alpha\beta$ demonstrates that the difficulty of the problem is determined by the product of the operator smoothness and the input regularity.

4. Empirical verification. We implement the estimator for learning PDE solution operators to verify our theory, including a Kuramoto-Sivashinsky operator and an elliptic operator (Section 5). The generalization gap decreases consistently as the number of training functions grows, and the error caused by observing the inputs at random grids decays at the rate predicted by the theory. Additional experiments on Burgers and reaction-diffusion trajectory operators illustrate the effect of output truncation (Appendix F.5).

1.1 Related work

Theory of neural operators. Universal approximation of nonlinear operators traces back to Chen & Chen (1995). Error bounds have been derived for DeepONet (Lanthaler et al., 2022; Schwab et al., 2026; Liu et al., 2026), the Fourier neural operator (Kovachki et al., 2021) and PCA-Net (Bhattacharya et al., 2021; Lanthaler, 2023); see Kovachki et al. (2024b) for a survey. Generalization analyses of encoder-decoder neural operators include Chen et al. (2023) and Liu et al. (2024),

and Cheng et al. (2026). Closest to our operator class, Cheng et al. (2026) considers estimating k times Fréchet differentiable operators using prespecified encoder-decoder networks, assuming sub-Gaussian input measures on the whole space. In comparison, we study Transformers operating on the full coordinate sequence, and we complement the upper bound with a matching minimax lower bound. For linear operators and holomorphic operators, algebraic rates have been achieved (de Hoop et al., 2023; Jin et al., 2023; Adcock et al., 2024). For Lipschitz and Fréchet differentiable operators, Kovachki et al. (2024a) and Lanthaler & Stuart (2026) prove lower bounds on parametric complexity and on n-widths. Our lower bound is instead a statistical minimax bound for regression with additive noise, and it identifies the exponent $2\alpha\beta$ of the logarithmic rate.

Theory of Transformers. Transformers are known to be universal approximators of sequence-to-sequence functions (Yun et al., 2020; Kajitsuka & Sato, 2024). Their capacity and sample complexity have been studied through sparse variable creation (Edelman et al., 2022), for classification (Gurevych et al., 2022), and on data of low intrinsic dimension (Havrilla & Liao, 2024). Takakura & Suzuki (2023) shows that Transformers acting on infinite sequences avoid the curse of dimensionality for shift-equivariant anisotropic smooth function class. Calvello et al. (2025) formulates attention directly as an operator between function spaces, and proves universal approximation theorem for their Transformer neural operators. Our construction instead uses sliding-window attention on basis coefficients and exploits Fréchet-Hölder smoothness, providing explicit network size scaling and statistical rates.

Transformers for operator learning. Practical Transformer-based neural operators differ in how functions are tokenized and how attention is made efficient. Examples include Galerkin-type attention (Cao, 2021), cross-attention decoders at arbitrary query points (Hao et al., 2023; Li et al., 2023), attention in learned latent spaces (Alkin et al., 2024; Wu et al., 2024), and pretrained multiscale models (Herde et al., 2024). Benchmarks such as those of Li et al. (2021), Takamoto et al. (2022) and Raonić et al. (2023) are standard testbeds.

2 Problem setup

Let $\mathcal{X}$ and $\mathcal{Y}$ be two separable Hilbert spaces with inner products $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$, respectively. We consider the task of learning an unknown operator $\Psi_0 : \mathcal{X} \to \mathcal{Y}$ using the Transformer architecture. For this purpose, we represent $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ using coordinate sequences in $\mathbb{R}^\infty$. Then, the operator learning task is reduced to estimating an infinite-dimensional sequence-to-sequence function, which can be approached by a Transformer $\varphi : \mathbb{R}^\infty \to \mathbb{R}^\infty$.

Specifically, let $\{\phi_i : i \in \mathbb{Z}\}$ and $\{\psi_i : i \in \mathbb{Z}\}$ be orthonormal bases of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Then, any $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ admit the basis expansions

$$x = \sum_{i\in\mathbb{Z}} a_i\phi_i, \qquad a_i = \langle x, \phi_i\rangle_{\mathcal{X}}, \quad \sum_{i\in\mathbb{Z}} |a_i|^2 < \infty,$$
$$y = \sum_{i\in\mathbb{Z}} b_i\psi_i, \qquad b_i = \langle y, \psi_i\rangle_{\mathcal{Y}}, \quad \sum_{i\in\mathbb{Z}} |b_i|^2 < \infty.$$

By Parseval's identity, the coordinate maps $J_{\mathcal{X}} : x \mapsto (a_i)_{i\in\mathbb{Z}}$ and $J_{\mathcal{Y}} : y \mapsto (b_i)_{i\in\mathbb{Z}}$ are isometric isomorphisms from $\mathcal{X}$ and $\mathcal{Y}$ onto $\ell^2(\mathbb{Z})$. Learning the target operator Ψ_0 is thus equivalent to learning its coordinate-space representation $\tilde{\Psi}_0 = J_{\mathcal{Y}} \circ \Psi_0 \circ J_{\mathcal{X}}^{-1} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$. We approximate the surrogate operator $\tilde{\Psi}_0$ using a Transformer $\varphi : \mathbb{R}^\infty \to \ell^2(\mathbb{Z})$ with range restricted to $\ell^2(\mathbb{Z})$. By mapping coordinate representations back to the input and output spaces, $J_{\mathcal{Y}}^{-1} \circ \varphi \circ J_{\mathcal{X}} : \mathcal{X} \to \mathcal{Y}$ serves as an approximator of the target operator Ψ_0.

Remark 1. The coordinate map $J_{\mathcal{X}}$ converts functional inputs in $\mathcal{X}$ into sequential representations suitable for processing by Transformers, while $J_{\mathcal{Y}}^{-1}$ reconstructs outputs in $\mathcal{Y}$ from sequential predictions. Though conceptually reminiscent of the encoder-decoder pair in neural operator architectures such as PCA-Net, $J_{\mathcal{X}}$ and $J_{\mathcal{Y}}^{-1}$ are canonical isometric isomorphisms induced by the orthonormal bases, preserving all information without reconstruction error. In comparison, the PCA-Net encoder performs dimension reduction by projecting functional inputs into a finite-dimensional subspace, prior to network processing. Under our problem setup, the Transformer network φ operates directly on the infinite-dimensional sequence space $\ell^2(\mathbb{Z})$.

To measure the distance between any estimate $\Psi : \mathcal{X} \to \mathcal{Y}$ and target operator Ψ_0, we employ the $L^2(\nu)$ norm for some Borel probability measure ν on $\mathcal{X}$:

$$\|\Psi - \Psi_0\|^2_{L^2(\nu)} \coloneqq \mathbb{E}_{x\sim\nu}\|\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}} = \int_{\mathcal{X}}\|\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}}\nu(dx),$$

where $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ are norms induced by the inner products $\langle\cdot,\cdot\rangle_{\mathcal{X}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$, respectively. With $\Psi = J_{\mathcal{Y}}^{-1} \circ \varphi \circ J_{\mathcal{X}}$, the isometry of $J_{\mathcal{Y}}$ yields

$$\|\Psi - \Psi_0\|^2_{L^2(\nu)} = \mathbb{E}_{\mathbf{a}\sim\tilde{\nu}}\big\|\varphi(\mathbf{a}) - \tilde{\Psi}_0(\mathbf{a})\big\|_2^2,$$

where $\mathbf{a} \in \ell^2(\mathbb{Z})$, and $\tilde{\nu} = (J_{\mathcal{X}})_{\#}\nu$ is the pushforward measure on $\ell^2(\mathbb{Z})$. This identity allows theoretical error bounds to be derived in the infinite-dimensional sequence space, and the empirical network training to be implemented using the ℓ^2 loss.

In this paper, we study the estimation of a class of regular operators, with regularity characterized by Fréchet differentiability and Hölder continuity. The definition of Fréchet derivative can be found in Appendix A. Specifically, we consider target operators in the Fréchet-Hölder class $\mathcal{H}^\beta(\mathcal{X},\mathcal{Y};R)$ defined below. That is, all Fréchet derivatives of Ψ_0 up to order $\lfloor\beta\rfloor$ exist, and the $\lfloor\beta\rfloor$-th order Fréchet derivative is $(\beta - \lfloor\beta\rfloor)$-Hölder continuous.

Definition 1 (Fréchet-Hölder operator class)**.** *Let $\beta = s + r > 0$ with $s = \lfloor\beta\rfloor \in \mathbb{N}$ and $r \in (0,1]$, where $\lfloor\beta\rfloor$ denotes the largest integer strictly smaller than β. For a finite $B_0 > 0$, the Fréchet-Hölder operator class $\mathcal{H}^\beta(\mathcal{X},\mathcal{Y};B_0)$ is the set of all operators $\Psi : \mathcal{X} \to \mathcal{Y}$ satisfying*

$$\max_{0\le k\le s}\sup_{x\in\mathcal{X}}\|\mathcal{D}^k\Psi(x)\|_{B^k(\mathcal{X},\mathcal{Y})} \le B_0,\ \sup_{x_1\neq x_2}\frac{\|\mathcal{D}^s\Psi(x_2) - \mathcal{D}^s\Psi(x_1)\|_{B^s(\mathcal{X},\mathcal{Y})}}{\|x_2 - x_1\|^r_{\mathcal{X}}} \le B_0,$$

where $\mathcal{D}^k\Psi : \mathcal{X} \to B^k(\mathcal{X},\mathcal{Y})$ is the k-th order Fréchet derivative of Ψ, and $B^k(\mathcal{X},\mathcal{Y})$ is the space of all bounded k-linear operators from $\mathcal{X}^k$ to $\mathcal{Y}$ ($k \ge 1$). Moreover, $\mathcal{D}^0\Psi \coloneqq \Psi$ and $B^0(\mathcal{X},\mathcal{Y}) \coloneqq \mathcal{Y}$.

2.1 Transformer architecture

The Transformer architecture we consider consists of three main components: embedding layer, self-attention layer, and position-wise feedforward neural network (FNN) layer.

Embedding layer. Let D be the embedding dimension. The embedding layer $\mathrm{Enc}_P : \mathbb{R}^\infty \to \mathbb{R}^{D\times\infty}$ is defined as

$$\mathrm{Enc}_P(\mathbf{a}) = \eta\mathbf{a} + P, \qquad \mathbf{a} = (a_i)_{i\in\mathbb{Z}} \in \mathbb{R}^\infty,$$

where $\eta \in \mathbb{R}^D$ is a learnable embedding vector and $P \in \mathbb{R}^{D\times\infty}$ is a fixed positional encoding. Since the self-attention layers and position-wise FNN layers are permutation-equivariant, a positional encoding is introduced to provide positional information and break this equivariance. In this paper, we consider fixed P, since the input is an infinite-dimensional sequence.

Self-attention layer. Let H be the number of attention heads, and let U be the sliding window size. A self-attention layer $g : \mathbb{R}^{D\times\infty} \to \mathbb{R}^{D\times\infty}$ with parameters $Q^{(h)}, K^{(h)} \in \mathbb{R}^{D'\times D}, V^{(h)} \in \mathbb{R}^{D\times D} (D' \le D, h = 1,\dots,H)$ is defined by

$$[g(Z)]_i = z_i + \textstyle\sum_{h=1}^H V^{(h)}Z[i-U:i+U]\cdot\mathrm{Softmax}\big((K^{(h)}Z[i-U:i+U])^\top Q^{(h)}z_i\big),$$

where $Z = (z_i)_{i\in\mathbb{Z}} \in \mathbb{R}^{D\times\infty}$, $i \in \mathbb{Z}$, and $\mathrm{Softmax} : \mathbb{R}^{2U+1} \to \mathbb{R}^{2U+1}$ is defined in Appendix A. Without loss of generality, suppose that $D' \le D$, since $K^{(h)}$ and $Q^{(h)}$ are not of full rank if $D' > D$. We define the class of self-attention layers g with embedding dimension D, window size U, number of heads H, and weight-norm bound B by

$$\mathcal{G}(D,U,H,B) = \Big\{g : \max_{1\le h\le H}\big\{\big\|Q^{(h)}\big\|_\infty, \big\|K^{(h)}\big\|_\infty, \big\|V^{(h)}\big\|_\infty\big\} \le B\Big\}.$$

FNN layer. A feedforward neural network f with depth L and width W is defined as

$$f(x) = (A_L\boldsymbol{\sigma}(\cdot) + b_L)\circ\cdots\circ(A_1\boldsymbol{\sigma}(\cdot) + b_1)(A_0x + b_0), \qquad x \in \mathbb{R}^{d_0},$$

where $A_i \in \mathbb{R}^{d_{i+1}\times d_i}$, $b_i \in \mathbb{R}^{d_{i+1}}$ $(0 \le i \le L)$, $\max_{1\le i\le L}\{d_i\} = W$, and $\boldsymbol{\sigma}$ is the element-wise version of the ReLU activation function $\sigma(t) = \max\{t, 0\}$. For $X \in \mathbb{R}^{d_0\times\infty}$, f is applied position-wise. That is, $f(X) \coloneqq (f(x_i))_{i\in\mathbb{Z}} \in \mathbb{R}^{d_{L+1}\times\infty}$. We define the class of FNNs with depth L, width W, sparsity S, and weight-norm bound B as

$$\mathcal{F}(L,W,S,B) = \Big\{f : \textstyle\sum_{i=0}^L\big(\|A_i\|_0 + \|b_i\|_0\big) \le S,\ \max_{0\le i\le L}\big\{\|A_i\|_\infty \vee \|b_i\|_\infty\big\} \le B\Big\}.$$

The class of Transformers is defined by

$$\mathcal{T}(M,D,U,H,L,W,S,B) = \big\{ f_M \circ g_M \circ \cdots \circ f_1 \circ g_1 \circ \mathrm{Enc}_P : \mathbb{R}^\infty \to \mathbb{R}^\infty \big| \\ \|\eta\|_\infty \le B,\ g_l \in \mathcal{G}(D,U,H,B),\ f_m \in \mathcal{F}(L,W,S,B) \big\}.$$

To ensure the outputs of the Transformers always lie in $\ell^2(\mathbb{Z})$, we apply clipping to both the scale and index of the output. That is, we set all coordinates with index $|i| > N$ to zero, and clip the nonzero coordinates such that they are bounded by Γ. Then, the class of clipped Transformers is defined as

$$\mathcal{T}_{N,\Gamma}(M,D,U,H,L,W,S,B) = \big\{ \mathrm{Clip}_{N,\Gamma} \circ \varphi : \varphi \in \mathcal{T}(M,D,U,H,L,W,S,B) \big\}.$$

3 Approximation error analysis

In this section, we investigate the approximation ability of Transformers for operators in the Fréchet-Hölder space. One of the key differences between our operator approximation theory and conventional neural network approximation theory is that the operators are defined on the infinite-dimensional Hilbert space $\mathcal{X}$, rather than a compact domain. To tackle the complications associated with non-compactness, we leverage the truncated approximation technique. That is, we focus on approximating the target operator on a compact domain, and control the error due to truncation using the tail probability of $x \sim \nu$.

Since $\mathcal{X}$ is infinite-dimensional, the Riesz's lemma suggests that closed balls of $\mathcal{X}$, $\{x \in \mathcal{X} : \|x\|_{\mathcal{X}} \le R\}$, are non-compact. Therefore, we instead consider the Sobolev balls

$$\Omega_R^\alpha = \{x \in \mathcal{X} : \|x\|_{\mathcal{X}^\alpha} \le R\}, \qquad \text{with } \|x\|_{\mathcal{X}^\alpha}^2 = \sum_{i\in\mathbb{Z}} (1+|i|)^{2\alpha} |\langle x, \phi_i\rangle_{\mathcal{X}}|^2, \quad \alpha > 0.$$

For $\gamma > 0$, $\|\cdot\|_{\mathcal{Y}^\gamma}$ is defined similarly. By Rellich-Kondrachov theorem, the Sobolev balls Ω_R^α are compact in $\mathcal{X}$. We thus decompose the global $L^2(\nu)$ approximation error into the approximation error on the truncated domain Ω_R^α, and the error due to domain truncation (Lemma 1 in Appendix B.1).

In the following theorem, we establish an L^∞ approximation error bound over the truncated domain, for any truncation level $R \ge 1$.

Theorem 1 (Main approximation result)**.** *Suppose that $\Psi_0 \in \mathcal{H}^\beta(\mathcal{X},\mathcal{Y};B_0)$ for some $\beta > 0$ and $B_0 \ge 1$, and $\sup_{x\in\mathcal{X}} \|\Psi_0(x)\|_{\mathcal{Y}^\gamma} \le B_{0,\gamma}$ for some $B_{0,\gamma} > 0$. Let Ω_R^α be the α-Sobolev ball with radius $R \ge 1$ in $\mathcal{X}$. Then, for any $\epsilon \in (0,1)$, there exists a Transformer $\bar\varphi \in \mathcal{T}_{N,\Gamma}(M,D,U,H,L,W,S,B)$ such that*

$$\sup_{x\in\Omega_R^\alpha} \|J_{\mathcal{Y}}^{-1} \circ \bar\varphi \circ J_{\mathcal{X}}(x) - \Psi_0(x)\|_{\mathcal{Y}} \le \epsilon,$$

where

$$M = 2, \qquad \max\{D,U,H,N\} \lesssim \max\left\{ \left(\frac{R}{\epsilon^{1/(\beta\wedge 1)}}\right)^{1/\alpha}, \epsilon^{-1/\gamma} \right\}, \qquad \Gamma \ge B_{0,\gamma},$$

$$L \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \left(\log\frac{R}{\epsilon}\right)^2, \qquad \log(W\vee S) \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon}, \qquad \log B \lesssim \log\frac{R}{\epsilon},$$

and we omit constants depending only on α, β, γ, B_0, and $B_{0,\gamma}$.

The proof of Theorem 1 can be found in Appendix B.3. In Theorem 1, we explicitly bound the network architecture scaling required for achieving a prespecified accuracy ϵ, which is important for subsequent generalization error analysis. The total number of nonzero parameters in our constructed Transformer approximator $\bar\varphi$ is $M(S + 3HD^2) \lesssim S$. That said, the FNN layers contain most of the nonzero parameters, which aligns well with many Transformer architectures used in practice Vaswani et al. (2017); Devlin et al. (2019). This is attributed to our construction, where self-attention layers focus on feature selection and aggregation, while FNN offers strong operator approximation capability. Specifically, given the infinite-dimensional input sequence, we first use one self-attention layer to select and aggregate a finite number of important features, whose dimension is determined

by the accuracy ϵ. Then, we use an FNN layer constructed based on Taylor expansion to approximate the target operator using the extracted features. Lastly, another self-attention layer aggregates the FNN output to obtain the coordinate prediction. Consequently, our approximation theory provides theoretical support that Transformer-based neural operators can adaptively select finite informative features from infinite-dimensional spaces, without using explicit encoders like PCA-Net.

Furthermore, by choosing $R \lesssim \log(1/\epsilon)$, we can express the approximation error in terms of number of network parameters:

$$\epsilon \lesssim \left(\frac{\log S}{(\log\log S)^{1+1/\alpha}} \right)^{-2\alpha\beta},$$

where we have used $\log\log S \gtrsim \log(1/\epsilon)$, which is implied by the proofs of Lemma 2 and Theorem 1. This rate is similar to the $(\log S/\log\log S)^{-\eta}$ rate in existing works (Cheng et al. (2026) and Liu et al. (2026), to name a few), where the exponent η is determined by corresponding class of target operators and setup of input space. Our rate has an additional $1/\alpha$ exponent on $\log\log S$, which arises from the choice of truncation level R and will not occur if R were considered a constant.

The truncation error can be well-control when ν has a fast-decaying tail probability. In this paper, we assume that ν is a Sobolev sub-exponential measure defined below.

Definition 2 (Sobolev sub-exponential measure)**.** *Let $(\mathcal{X}, \langle\cdot,\cdot\rangle_{\mathcal{X}})$ be a separable Hilbert space with orthonormal basis $\{\phi_i\}_{i\in\mathbb{Z}}$. For $\alpha > 0$, a Borel probability measure ν on $\mathcal{X}$ is α-Sobolev sub-exponential if $x \sim \nu$ has a finite $\|\cdot\|_{\mathcal{X}^\alpha}$-induced Orlicz ψ_1-norm, that is,*

$$\|x\|_{\psi_1,\mathcal{X}^\alpha} := \inf\left\{ \lambda > 0 : \mathbb{E}_{x\sim\nu}\left[\exp\left(\frac{\|x\|_{\mathcal{X}^\alpha}}{\lambda} \right) \right] \le 2 \right\} < \infty.$$

For $\kappa > 0$, let $\mathcal{P}_\alpha(\kappa)$ be the set of all α-Sobolev sub-exponential measures with $\|\cdot\|_{\psi_1,\mathcal{X}^\alpha}$ norm bounded by κ. That is, $\mathcal{P}_\alpha(\kappa) = \{\nu : \|x\|_{\psi_1,\mathcal{X}^\alpha} \le \kappa, x \sim \nu\}$.

In Definition 2, the finiteness of the Orlicz norm requires the probability measure ν to be supported on the Sobolev subspace $\mathcal{X}^\alpha := \{x \in \mathcal{X} : \|x\|_{\mathcal{X}^\alpha} < \infty\}$. The Sobolev subspace is dense in $\mathcal{X}$ in the sense that for any $x \in \mathcal{X}$ and $\epsilon > 0$, there exists $x_\epsilon \in \mathcal{X}^\alpha$ such that $\|x_\epsilon - x\|_{\mathcal{X}} < \epsilon$ (Adams & Fournier, 2003).

Theorem 1 establishes a uniform approximation error bound over Ω_R^α, and directly implies the $L^2(\nu;\Omega_R^\alpha)$ error bound. Under the Sobolev sub-exponential assumption on ν, we can bound the truncation error, and obtain the overall $L^2(\nu)$ approximation bound (Corollary 1 in Appendix B.4).

4 Generalization error analysis

4.1 Generalization error upper and lower bounds

We observe n training pairs $(x_1, y_1), \ldots, (x_n, y_n)$ generated by

$$y_i = \Psi_0(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $x_1, \ldots, x_n$ are i.i.d. with law ν and, independently of them, $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with a centered sub-exponential law $\nu_\varepsilon \in \mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$ on $\mathcal{Y}$ (Definition 5).

We estimate Ψ_0 by least squares over operators $J_{\mathcal{Y}}^{-1} \circ \varphi \circ J_{\mathcal{X}}$, where φ is a clipped Transformer applied to clipped input coefficients. Because ν may have unbounded support, the input clip serves to obtain covers of the class that are uniform over all inputs (Lemma 6). The class is

$$\mathcal{T}_n := \{\varphi \circ \Pi^{\rho_n} : \varphi \in \mathcal{T}_{N,\Gamma}(2, D, U, H, L, W, S, B)\},$$

where Π^ρ clips every coordinate to $[-\rho, \rho]$ and $\Gamma = \max\{1, B_{0,\gamma}\}$. For $n \ge n_1$, Definition 7 specifies the radius ρ_n, the output cutoff N, the remaining architecture parameters and the positional encoding. The threshold n_1 depends only on $(\alpha, \beta, \gamma, \kappa, B_0, B_{0,\gamma})$, and these choices depend only on n and the same parameters; in particular, the estimator requires no knowledge of Ψ_0, ν or ν_ε, not even of the noise level κ_ε. The input clip, whose radius ρ_n is of order $\log\log n$, leaves every network unchanged on the Sobolev sequence ball $\mathcal{B}_{\rho_n}^{2,\alpha} = J_{\mathcal{X}}(\Omega_{\rho_n}^\alpha)$, because all coordinates of its elements lie in $[-\rho_n, \rho_n]$; hence it preserves the approximation guarantee of Theorem 1 on $\Omega_{\rho_n}^\alpha$. The outputs

of every $\varphi \in \mathcal{T}_n$ vanish at all indices $|k| > N$, so on $\mathcal{T}_n$ the empirical risk $n^{-1}\sum_i \|\varphi(J_{\mathcal{X}}x_i) - J_{\mathcal{Y}}y_i\|_2^2$ differs by a finite term independent of φ from

$$R_n(\varphi) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \big\|\{\varphi(J_{\mathcal{X}}x_i) - J_{\mathcal{Y}}y_i\}[-N:N]\big\|_2^2;$$

hence the two criteria have the same minimizers, and R_n involves only finitely many coordinates of the responses. The estimator is

$$\hat{\Psi}_n = J_{\mathcal{Y}}^{-1} \circ \hat{\varphi}_n \circ J_{\mathcal{X}}, \qquad \hat{\varphi}_n \in \operatorname*{arg\,min}_{\varphi \in \mathcal{T}_n} R_n(\varphi).$$

Similar to Takakura & Suzuki (2023), we assume that a minimizer exists and that $\hat{\varphi}_n(\mathbf{a})$ is a measurable function of $\mathbf{a} \in \ell^2(\mathbb{Z})$ and the data $(J_{\mathcal{X}}x_i, (J_{\mathcal{Y}}y_i)[-N:N])_{i\le n}$.

Theorem 2 (Upper bound of generalization error)**.** *Let $\alpha, \beta, \gamma, \kappa, \kappa_\varepsilon, B_{0,\gamma} > 0$ and $B_0 \ge 1$, and let $\mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$ be the set of $\Psi \in \mathcal{H}^{\beta}(\mathcal{X}, \mathcal{Y}; B_0)$ with $\sup_{x\in\mathcal{X}} \|\Psi(x)\|_{\mathcal{Y}^\gamma} \le B_{0,\gamma}$. There is a constant C, depending only on $(\alpha, \beta, \gamma, \kappa, \kappa_\varepsilon, B_0, B_{0,\gamma})$, such that for all $n \ge n_1$,*

$$\sup_{\Psi_0, \nu, \nu_\varepsilon} \mathbb{E}\|\hat{\Psi}_n - \Psi_0\|_{L^2(\nu)}^2 \le C\left(\frac{\log n}{(\log\log n)^{1+1/\alpha}}\right)^{-2\alpha\beta},$$

where the supremum is over $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, $\nu \in \mathcal{P}_\alpha(\kappa)$ and $\nu_\varepsilon \in \mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$, and $\mathbb{E}$ is taken over the sample $\{(x_i, y_i)\}_{i=1}^n$ from model (1)*.*

Theorem 2 establishes a logarithmic rate whose exponent is determined by the product of the input decay α and the operator smoothness β; the output regularity γ, the tail parameter κ and the noise distribution do not affect the rate, and the noise need only be sub-exponential. This rate reflects a trade-off specific to infinite-dimensional inputs. To attain accuracy ϵ on a Sobolev ball whose complement has negligible ν-probability, the construction of Theorem 1 resolves of order $\epsilon^{-1/(\alpha\beta)}$ input coefficients, up to a logarithmic factor, and its size, and hence the bound on the metric entropy of the class, is exponential in this number. Balancing the resulting stochastic error against the squared approximation error then requires this number to be of order $\log n$, up to a power of $\log\log n$, which gives the stated rate. The proof is given in Appendix C.5. Theorem 3 below shows that no estimator can improve on the exponent $2\alpha\beta$, so the logarithmic rate is a feature of the problem rather than of the Transformer.

Theorem 3 (Minimax lower bound)**.** *Let $\alpha, \beta, \gamma, \kappa, \kappa_\varepsilon, B_0, B_{0,\gamma} > 0$. There are constants $c_0 > 0$ and n_0, depending only on $(\alpha, \beta, \kappa, \kappa_\varepsilon, B_0, B_{0,\gamma})$, such that for all $n \ge n_0$,*

$$\inf_{\hat{\Psi}} \sup_{\Psi_0, \nu, \nu_\varepsilon} \mathbb{E}\|\hat{\Psi} - \Psi_0\|_{L^2(\nu)}^2 \ge c_0(\log n)^{-2\alpha\beta},$$

where the supremum is over $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, $\nu \in \mathcal{P}_\alpha(\kappa)$ and $\nu_\varepsilon \in \mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$, $\mathbb{E}$ is taken over the sample $\{(x_i, y_i)\}_{i=1}^n$ from model (1)*, and the infimum is over all estimators, that is, all maps that assign to the sample an operator $\hat{\Psi} : \mathcal{X} \to \mathcal{Y}$ such that $\hat{\Psi}(x)$ is jointly Borel measurable in x and the sample.*

Together with Theorem 2, Theorem 3 shows that, for $B_0 \ge 1$, the least-squares Transformer estimator is minimax optimal up to the factor $(\log\log n)^{2\beta(\alpha+1)}$. The lower bound stems from the size of the Sobolev ball: at scale r, it contains a number of well-separated points that is exponential in $r^{-1/\alpha}$, and the operators obtained by placing bumps of height of order r^β with arbitrary signs at these points cannot be distinguished from n samples once $r^{-1/\alpha}$ is of order $\log n$, which leaves a squared error of order $r^{2\beta} \asymp (\log n)^{-2\alpha\beta}$. Since the construction uses Gaussian noise and input laws supported on finitely many points of a Sobolev ball, the logarithmic rate is caused neither by the tails of the noise nor by the unbounded support of the inputs. The proof is given in Appendix D.

4.2 Setting of discretized observations on random grids

In many applications, the input and output functions are observed only at finitely many points. Let $\mathcal{X} = L^2(\Omega_{\mathcal{X}}, \mu_u)$ and $\mathcal{Y} = L^2(\Omega_{\mathcal{Y}}, \mu_v)$ for probability measures μ_u and μ_v, and assume that $\sup_i \|\phi_i\|_\infty \le C_\phi$ and $\sup_i \|\psi_i\|_\infty \le C_\psi$, as for trigonometric bases and the uniform measure on

the torus. If $\alpha, \gamma > 1/2$, then for every x with $\|x\|_{\mathcal{X}^\alpha} < \infty$ the basis expansions of x and of $\Psi_0(x)$, whose $\mathcal{Y}^\gamma$-norm is at most $B_{0,\gamma}$, converge absolutely and uniformly, and their sums define the point values used below. For each $i \le n$, we observe the pairs $(u_{i,j}, x_i(u_{i,j}))$, $j \le n_u$, and $(v_{i,j}, z_{i,j})$, $j \le n_v$, where $z_{i,j} = \Psi_0(x_i)(v_{i,j}) + \xi_{i,j}$. All $x_i \sim \nu$, $u_{i,j} \sim \mu_u$ and $v_{i,j} \sim \mu_v$ are independent, and the $\xi_{i,j}$ are i.i.d. with a law in $\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$ on $\mathbb{R}$, independent of all other variables.

The observations are encoded by Monte Carlo quadrature with the input cutoff $K := \lceil c_K n_u^{1/(2\alpha+1)} \rceil$ for a fixed $c_K > 0$,

$$\hat{a}_{i,k} = \frac{1}{n_u} \sum_{j=1}^{n_u} x_i(u_{i,j}) \phi_k(u_{i,j}) \ \ (|k| \le K), \qquad \hat{b}_{i,k} = \frac{1}{n_v} \sum_{j=1}^{n_v} z_{i,j} \psi_k(v_{i,j}) \ \ (|k| \le N), \tag{2}$$

with $\hat{a}_{i,k} = 0$ for $|k| > K$ and $\hat{b}_{i,k} = 0$ for $|k| > N$, where N is the output cutoff of $\mathcal{T}_n$. The encoded inputs are then projected onto $\mathcal{B}^{2,\alpha}_{\rho_n}$, on which $\mathcal{T}_n$ retains the approximation guarantee of Theorem 1: $\tilde{\mathbf{a}}_i = P_{\rho_n}(\hat{\mathbf{a}}_i)$, where P_ρ is the metric projection of $\ell^2(\mathbb{Z})$ onto $\mathcal{B}^{2,\alpha}_\rho$. The projection is non-expansive and fixes this ball, so it does not increase the encoding error of inputs in $\Omega^\alpha_{\rho_n}$, and it is computed by a one-dimensional root search (Appendix C.6). The network $\hat{\varphi}_n$ minimizes the empirical risk $n^{-1} \sum_i \|\varphi(\tilde{\mathbf{a}}_i) - \hat{\mathbf{b}}_i\|_2^2$ over $\mathcal{T}_n$. As in Section 4.1, we assume that a minimizer exists and that $\hat{\varphi}_n(\mathbf{c})$ is a measurable function of $\mathbf{c} \in \ell^2(\mathbb{Z})$ and $(\tilde{\mathbf{a}}_i, \hat{\mathbf{b}}_i)_{i \le n}$. A new input x is observed at fresh grids $U = (u_1, \dots, u_{n_u})$, encoded into $\tilde{\mathbf{a}}(x, U)$ in the same way, and mapped to $\hat{\Psi}_n(x, U) := J_{\mathcal{Y}}^{-1} \hat{\varphi}_n(\tilde{\mathbf{a}}(x, U))$. The risk is $\mathcal{E}(\hat{\Psi}_n) := \mathbb{E}\|\hat{\Psi}_n(x, U) - \Psi_0(x)\|^2_{\mathcal{Y}}$, where the expectation is over the sample and an independent pair $(x, U) \sim \nu \otimes \mu_u^{\otimes n_u}$.

Theorem 4 (Discretized observations). *Let $\alpha, \gamma > 1/2$, $\beta, \kappa, \kappa_\varepsilon, B_{0,\gamma} > 0$ and $B_0 \ge 1$, and let $\mathcal{T}_n$, ρ_n and n_1 be as in Section 4.1. There exist a constant C_u and, for each $n_v \ge 1$, a constant C such that for all $n \ge n_1$ and $n_u \ge 1$,*

$$\sup \mathcal{E}(\hat{\Psi}_n) \le C \left(\frac{\log n}{(\log \log n)^{1+1/\alpha}} \right)^{-2\alpha\beta} + C_u \, n_u^{-2\alpha(\beta \wedge 1)/(2\alpha+1)},$$

where the supremum is over $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, $\nu \in \mathcal{P}_\alpha(\kappa)$ and the common law of the $\xi_{i,j}$ in $\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$. The constant C_u depends only on $(\alpha, \beta, \kappa, B_0, c_K, C_\phi)$; the constant C depends only on n_v and $(\alpha, \beta, \gamma, \kappa, \kappa_\varepsilon, B_0, B_{0,\gamma}, C_\psi)$, and it can be chosen nonincreasing in n_v and nondecreasing in κ_ε.

In the bound of Theorem 4, the two discretizations enter differently. Sampling the outputs leaves each encoded response unbiased given the input and adds centered sub-exponential noise, so n_v enters only the constant C, as the noise level does in Theorem 2. Sampling the inputs instead perturbs the argument of Ψ_0, both in training and at test time, which yields a second term that does not decrease with n. By the Hölder continuity of Ψ_0, this term is controlled by the $(\beta \wedge 1)$-th power of the mean squared encoding error, which is at most of order $n_u^{-2\alpha/(2\alpha+1)}$ for the chosen K. Because the first term decays only logarithmically in n, the rate of Theorem 2 is retained whenever $n_u \gtrsim (\log n)^{\beta(2\alpha+1)/(\beta \wedge 1)}$, a number of observations per input function that is polylogarithmic in n. The proof is given in Appendix C.6.

5 Numerical experiments

We verify two implications of the theory: the generalization gap decreases as the number n of training functions grows (Theorem 2), and discretizing the inputs adds an error that decreases with the number n_u of input observations but not with n, whereas discretizing the outputs affects only the constant (Theorem 4). Implementation details and further results are given in Appendix F.

Setup. The Kuramoto–Sivashinsky (KS) operator is the endpoint map $\Psi_0 : u(\cdot, 0) \mapsto u(\cdot, 4)$ of $u_t + uu_x + u_{xx} + u_{xxxx} = 0$ on the torus $[0, 64)$, with inputs in the chaotic regime. The elliptic operator is $\Psi_0 : f \mapsto u$ on the unit torus, where $u - 0.03^2 u'' + \frac{1}{2} \tanh u = \tanh(Sf)$ with $S = (I - 0.02^2 \partial_{xx})^{-1}$, and f is a centered Gaussian field with independent Fourier coefficients of standard deviation $0.3/\{1 + (k/6)^2\}$ at frequency k, so that its law belongs to $\mathcal{P}_\alpha(\kappa)$ for every $\alpha < 3/2$. Functions are represented by their real Fourier coefficients, and the Transformer embeds

the coefficient a_i either as a learned token or, as in $\mathcal{T}_{N,\Gamma}$, as $\eta a_i + P_i$ with a shared vector η and a sinusoidal encoding P_i. It is trained by AdamW on the empirical squared L^2 loss with a budget that does not depend on n, so its risk includes optimization error. We report the test risk $\mathbb{E}\|\hat{\Psi}_n(x) - \Psi_0(x)\|_{\mathcal{Y}}^2$ on 2048 test functions, the corresponding relative error, and the generalization gap, the test risk minus the training loss, which reflects the stochastic error.

Effect of the sample size. On the KS operator (Figure 1(a)), the generalization gap of the Transformer with the shared embedding decreases at every doubling of n in each of five runs, while its relative error falls from 13.3% at $n = 128$ to 3.9% at $n = 8192$. DeepONet (Lu et al., 2021), trained with the same protocol as a reference, attains a risk within a factor of about two of that of the Transformer, and its gap also decreases at every doubling of n. On the elliptic operator, the relative error of the Transformer with learned tokens falls from 30.9% at $n = 256$ to 3.7% at $n = 16{,}384$ (black curves in Figure 1(c) and (d)), and its gap decreases monotonically (Appendix F.2).

Effect of discretization. Discretized observations estimate the coefficients by the Monte Carlo averages (2) from n_u input or n_v output values at i.i.d. uniform grids; we report the elliptic operator, and the KS operator behaves similarly (Appendix F.3). The input encoding error decays approximately as $n_u^{-0.73}$ (Figure 1(b)), close to the rate $n_u^{-0.74}$ of Lemma 17 for $\alpha = 1.4$. With discretized inputs, the test risk decreases towards the error incurred by applying Ψ_0 to the encoded input rather than towards zero (Figure 1(c)); at $n = 16{,}384$, it exceeds the sum of this error and the risk under exact observations by 3–14%, consistent with the additive form of the bound in Theorem 4. With discretized outputs, the test risk keeps decreasing over the whole range of n (Figure 1(d)), consistent with n_v affecting only the constant.

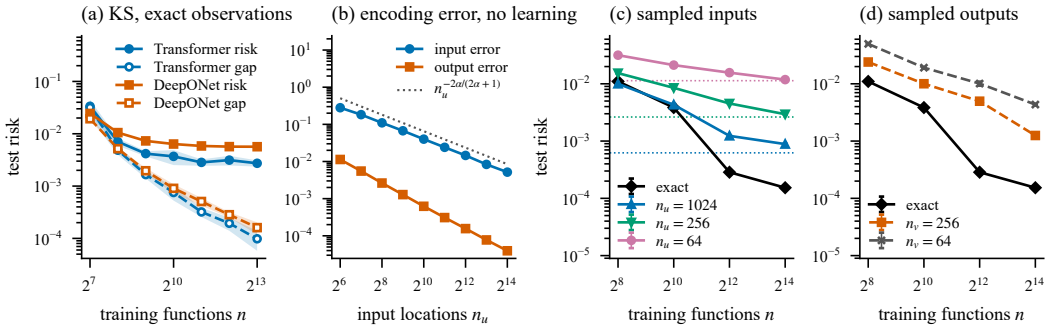


Figure 1: (a) KS operator, exact observations: Transformer and DeepONet (five runs). (b)–(d) Elliptic operator. (b) Input error from encoding and the resulting output error of Ψ_0; dotted: the rate of Lemma 17 for $\alpha = 1.4$. (c), (d) Transformer with inputs (c) or outputs (d) at random locations (three runs); dotted in (c): output errors from (b). Bands and error bars: ± 1 standard deviation.

6 Conclusions

We developed an approximation and generalization theory for learning Fréchet-Hölder operators with Transformers operating on the basis coefficients of functions. Under our smoothness and tail conditions, the least-squares Transformer estimator converges at a rate logarithmic in the sample size n that is minimax optimal up to an iterated-logarithmic factor. This logarithmic rate is intrinsic to the estimation problem rather than to the architecture, and is retained under the discretized observation setup when the number of input grids grows sufficiently fast relative to $\log n$. Numerical experiments on a Kuramoto-Sivashinsky operator and an elliptic operator support the analysis: the generalization gap decreases consistently as the sample size grows, and the error of encoding discretized inputs decays at the rate predicted by the theory.

Our results describe the worst case over the operator class and concern the exact least-squares minimizer; they do not account for the optimization error of gradient-based training, and they do not imply that Transformers outperform other architectures on a specific operator. Identifying structures of the operator or the input distribution that permit faster rates is a direction of future works.

AI USE STATEMENT

We used generative AI tools for the following tasks with required disclosure: refining and polishing the language and proofs to improve readability and structure, and checking mathematical derivations. Additionally, we used generative AI tools for checking and formatting references. We have reviewed all AI-assisted work. The authors checked AI-assisted reference list and searched related works. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is primarily theoretical and studies the approximation and generalization properties of neural operator Transformers between function spaces. The experiments use synthetic data generated by numerical solvers of partial differential equations and, for the Burgers equation, publicly available input functions (Li et al., 2021); they involve no human subjects, personal data, or sensitive information. We are not aware of potential harms specific to this work.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results are given in the appendix, and the experimental settings, including data generation, network architectures, and training procedures, are described in Appendix F.

REFERENCES

Ralph Abraham, Jerrold E. Marsden, and Tudor Ratiu. *Manifolds, Tensor Analysis, and Applications*. Springer, New York, NY, 1988.

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*. Elsevier, 2nd edition, 2003.

Ben Adcock, Nick Dexter, and Sebastian Moraga. Optimal deep learning of holomorphic operators between Banach spaces. In *Advances in Neural Information Processing Systems*, volume 37, pp. 27725–27789, 2024.

Benedikt Alkin, Andreas Fürst, Simon Schmid, Lukas Gruber, Markus Holzleitner, and Johannes Brandstetter. Universal physics Transformers: A framework for efficiently scaling neural operators. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25152–25194, 2024.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric PDEs. *The SMAI Journal of Computational Mathematics*, 7:121–157, 2021.

Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2241–2256. PMLR, 2022.

Edoardo Calvello, Nikola B. Kovachki, Matthew E. Levine, and Andrew M. Stuart. Continuum attention for neural operators. *Journal of Machine Learning Research*, 26(300):1–52, 2025.

Shuhao Cao. Choose a Transformer: Fourier or Galerkin. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24924–24940, 2021.

Ke Chen, Chunmei Wang, and Haizhao Yang. Deep operator learning lessens the curse of dimensionality for PDEs. *Transactions on Machine Learning Research*, 2023.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

Kun Cheng, Jun Fan, Linhao Song, and Ding-Xuan Zhou. Learning Fréchet differentiable operators via prespecified neural operators. *Applied and Computational Harmonic Analysis*, 84:101878, 2026.

S. M. Cox and P. C. Matthews. Exponential time differencing for stiff systems. *Journal of Computational Physics*, 176(2):430–455, 2002.

Maarten V. de Hoop, Nikola B. Kovachki, Nicholas H. Nelsen, and Andrew M. Stuart. Convergence rates for learning linear operators from noisy data. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):480–513, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional Transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

Benjamin L. Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5793–5831. PMLR, 2022.

L. E. Fraenkel. Formulae for high derivatives of composite functions. *Mathematical Proceedings of the Cambridge Philosophical Society*, 83(2):159–165, 1978.

Iryna Gurevych, Michael Kohler, and Gözde Gül Şahin. On the rate of convergence of a classifier based on a Transformer encoder. *IEEE Transactions on Information Theory*, 68(12):8139–8155, 2022.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator Transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 2023.

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.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624, 2024.

Jikai Jin, Yiping Lu, Jose Blanchet, and Lexing Ying. Minimax optimal kernel operator learning via multilevel training. In *International Conference on Learning Representations*, 2023.

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.

Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005.

Georgios Kissas, Jacob H. Seidman, Leonardo Ferreira Guilhoto, Victor M. Preciado, George J. Pappas, and Paris Perdikaris. Learning operators with coupled attention. *Journal of Machine Learning Research*, 23(215):1–63, 2022.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *Journal of Machine Learning Research*, 22(290):1–76, 2021.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Nikola B. Kovachki, Samuel Lanthaler, and Hrushikesh Mhaskar. Data complexity estimates for operator learning. arXiv:2405.15992, 2024a.

Nikola B. Kovachki, Samuel Lanthaler, and Andrew M. Stuart. Operator learning: Algorithms and analysis. In *Numerical Analysis Meets Machine Learning*, volume 25 of *Handbook of Numerical Analysis*, pp. 419–467. Elsevier, 2024b.

Andreas Kriegl and Peter W. Michor. *The Convenient Setting of Global Analysis*, volume 53. American Mathematical Society, 1997.

Samuel Lanthaler. Operator learning with PCA-Net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023.

Samuel Lanthaler and Andrew M. Stuart. The parametric complexity of operator learning. *IMA Journal of Numerical Analysis*, 46(2):647–712, 2026.

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):1–141, 2022.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023.

Zongyi Li, Nikola B. Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Phillip Lippe, Bas Veeling, Paris Perdikaris, Richard Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023.

Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *Journal of Machine Learning Research*, 25 (24):1–67, 2024.

Hao Liu, Zecheng Zhang, Wenjing Liao, and Hayden Schaeffer. Neural scaling laws of deep ReLU and deep operator network: A theoretical study. arXiv:2410.00357, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77187–77200, 2023.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Christoph Schwab, Andreas Stein, and Jakob Zech. Deep operator network approximation rates for Lipschitz operators. *Analysis and Applications*, 24(1):199–239, 2026.

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.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1596–1611, 2022.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, New York, NY, 2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast Transformer solver for PDEs on general geometries. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53681–53705. PMLR, 2024.

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*, 2020.

A SUPPORTING DEFINITIONS

This section includes the notations used throughout this paper, and the definitions of Fréchet derivative and sub-exponential noise distribution.

Let $\mathbb{Z}$, $\mathbb{N}$, and $\mathbb{Z}_+$ denote the set of integers, non-negative integers, and positive integers, respectively. For $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We write $X \lesssim Y$ if $X \leq CY$ for a constant C that does not depend on n, ϵ, or R, and $X \asymp Y$ if $X \lesssim Y \lesssim X$. Let $\mathbb{R}^{d\times\infty}$ be the set of all sequences $Z = (z_i)_{i\in\mathbb{Z}}$ with $z_i \in \mathbb{R}^d$, and $\mathbb{R}^\infty = \mathbb{R}^{1\times\infty}$. For $l \leq r$, let $Z[l:r] = (z_l, z_{l+1} \ldots, z_r) \in \mathbb{R}^{d\times(r-l+1)}$. For a sequence $\mathbf{a} \in \mathbb{R}^\infty$, we use $\|\mathbf{a}\|_2$ to denote its ℓ^2 norm, and $\ell^2(\mathbb{Z})$ is the space of square-summable sequences. For a matrix W, $\|W\|_\infty$ is the largest absolute value of its entries, and $\|W\|_0$ is the number of nonzero entries. Let e_i be the i-th standard basis vector. The push-forward measure $g_\#\nu$ of a measurable set A is defined as $g_\#\nu(A) = \nu(g^{-1}(A))$. Let $\mathbb{1}_\Omega$ be the indicator function of a set Ω, and $\sigma(t) = \max\{t, 0\}$ be the ReLU function. The Softmax function is defined as

$$\mathrm{Softmax}(v) = \left[\frac{e^{v_1}}{\sum_{j=1}^k e^{v_j}}, \ldots, \frac{e^{v_k}}{\sum_{j=1}^k e^{v_j}}\right], \qquad v \in \mathbb{R}^k.$$

For two Banach spaces $\mathcal{X}$ and $\mathcal{Y}$, let $B(\mathcal{X}, \mathcal{Y})$ be the Banach space of all bounded linear operators L from $\mathcal{X}$ to $\mathcal{Y}$ with norm

$$\|L\|_{B(\mathcal{X},\mathcal{Y})} = \sup_{x\in\mathcal{X}, x\neq 0} \frac{\|Lx\|_\mathcal{Y}}{\|x\|_\mathcal{X}}.$$

We further define the space $B^k(\mathcal{X}, \mathcal{Y})$ with $k \in \mathbb{Z}_+$ by $B^1(\mathcal{X}, \mathcal{Y}) = B(\mathcal{X}, \mathcal{Y})$ and recursively

$$B^k(\mathcal{X}, \mathcal{Y}) = B(\mathcal{X}, B^{k-1}(\mathcal{X}, \mathcal{Y})), \quad k \geq 2.$$

For $k \geq 2$, $B^k(\mathcal{X}, \mathcal{Y})$ is isometrically isomorphic to the space of all bounded k-linear operators from $\mathcal{X}^k$ to $\mathcal{Y}$, $B(\mathcal{X}^k, \mathcal{Y})$. The norm of $B^k(\mathcal{X}, \mathcal{Y})$ is defined as

$$\|L\|_{B^k(\mathcal{X},\mathcal{Y})} = \sup_{x_i\in\mathcal{X}, x_i\neq 0, 1\leq i\leq k} \frac{\|L(x_1)(x_2)\cdots(x_k)\|_\mathcal{Y}}{\|x_1\|_\mathcal{X}\|x_2\|_\mathcal{X}\cdots\|x_k\|_\mathcal{X}}.$$

For ease of notation, we denote $L(x_1)(x_2)\cdots(x_k) = L[x_1, x_2, \ldots, x_k]$ and $L(x)(x)\cdots(x) = L[x^k]$ for $L \in B^k(\mathcal{X}, \mathcal{Y})$.

Definition 3 (Fréchet derivative)**.** *Let $\Omega \subseteq \mathcal{X}$ be an open set. An operator $\Psi : \Omega \to \mathcal{Y}$ is Fréchet differentiable at $x_0 \in \Omega$ if there exists $A \in B(\mathcal{X}, \mathcal{Y})$ such that*

$$\lim_{\|h\|_\mathcal{X}\to 0} \frac{\|\Psi(x_0 + h) - \Psi(x_0) - A(h)\|_\mathcal{Y}}{\|h\|_\mathcal{X}} = 0.$$

When such A exists, we call it the Fréchet derivative of Ψ at x_0, and denote $A = \mathcal{D}\Psi(x_0)$.

Definition 4 (Higher-order Fréchet derivative)**.** *For $k \geq 2$, an operator $\Psi : \mathcal{X} \to \mathcal{Y}$ is k times Fréchet differentiable at $x_0 \in \mathcal{X}$ if it is $(k-1)$ times Fréchet differentiable in a neighborhood of x_0, and the $(k-1)$-th order derivative $\mathcal{D}^{k-1}\Psi : \mathcal{X} \to B^{k-1}(\mathcal{X}, \mathcal{Y})$ is Fréchet differentiable at x_0. In such cases, we denote*

$$\mathcal{D}^k\Psi(x_0) = \mathcal{D}(\mathcal{D}^{k-1}\Psi)(x_0).$$

Definition 5 (Sub-exponential noise distribution)**.** *Let $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be a Banach space. A Borel probability measure ν_ε on $\mathcal{Y}$ is sub-exponential if $\varepsilon \sim \nu_\varepsilon$ has a finite Orlicz ψ_1-norm, that is,*

$$\|\varepsilon\|_{\psi_1} \coloneqq \inf\left\{\lambda > 0 : \mathbb{E}_{\varepsilon\sim\nu_\varepsilon}\left[\exp\left(\frac{\|\varepsilon\|_{\mathcal{Y}}}{\lambda}\right)\right] \le 2\right\} < \infty.$$

For $\kappa_\varepsilon > 0$, let $\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$ be the set of all sub-exponential measures with mean zero and Orlicz ψ_1-norm bounded by κ_ε. That is,

$$\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon) = \left\{\nu_\varepsilon : \mathbb{E}[\varepsilon] = 0, \|\varepsilon\|_{\psi_1} \le \kappa_\varepsilon, \varepsilon \sim \nu_\varepsilon\right\}.$$

B Approximation error analysis

In this section, we analyze the approximation error and particularly prove Theorem 1. We first decompose the approximation error using Lemma 1. Then, we establish an important lemma on the operator approximation error of FNN (Lemma 2), based on which we show Theorem 1. Lastly, under the Sobolev sub-exponential assumption, we derive the overall approximation error bound (Corollary 1).

Before proceeding, we introduce some notations used in the section. Let

$$\mathcal{B}_R^{2,\alpha} = \left\{\mathbf{a} \in \ell^2(\mathbb{Z}) : \sum_{i\in\mathbb{Z}} (1+|i|)^{2\alpha}|a_i|^2 \le R^2\right\}$$

be the Sobolev sequence ball with radius R. Then, $\mathcal{B}_R^{2,\alpha}$ is compact in $\ell^2(\mathbb{Z})$, and $J_{\mathcal{X}}$ maps Ω_R^α to $\mathcal{B}_R^{2,\alpha}$. Moreover, for any subset $\Omega \subseteq \mathcal{X}$, we define

$$\|\Psi - \Psi_0\|^2_{L^2(\nu;\Omega)} \coloneqq \mathbb{E}_{x\sim\nu}\left[\|\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}} \mathbb{1}_\Omega(x)\right] = \int_\Omega \|\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}} \nu(dx).$$

B.1 Truncated approximation

Lemma 1 (Truncated approximation)**.** *Let $\mathcal{T}_n$ be a class of Transformers $\varphi : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$. For any $\bar\varphi \in \mathcal{T}_n$ and $R > 0$, the $L^2(\nu)$ approximation error for Ψ_0 using $\mathcal{T}_n$ satisfies*

$$\inf_{\varphi\in\mathcal{T}_n} \|J_{\mathcal{Y}}^{-1} \circ \varphi \circ J_{\mathcal{X}} - \Psi_0\|^2_{L^2(\nu)} \le \|\bar\Psi - \Psi_0\|^2_{L^2(\nu;\Omega_R^\alpha)} + \|\bar\Psi - \Psi_0\|^2_{L^2(\nu;\mathcal{X}\setminus\Omega_R^\alpha)},$$

where $\bar\Psi \coloneqq J_{\mathcal{Y}}^{-1} \circ \bar\varphi \circ J_{\mathcal{X}}$.

Proof. Since $\bar\varphi \in \mathcal{T}_n$, we have

$$\begin{aligned}
&\inf_{\varphi\in\mathcal{T}_n} \|J_{\mathcal{Y}}^{-1} \circ \varphi \circ J_{\mathcal{X}} - \Psi_0\|^2_{L^2(\nu)} \le \|J_{\mathcal{Y}}^{-1} \circ \bar\varphi \circ J_{\mathcal{X}} - \Psi_0\|^2_{L^2(\nu)} = \mathbb{E}_{x\sim\nu}\|\bar\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}} \\
&= \mathbb{E}_{x\sim\nu}\left[\|\bar\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}} \mathbb{1}_{\Omega_R^\alpha}(x)\right] + \mathbb{E}_{x\sim\nu}\left[\|\bar\Psi(x) - \Psi_0(x)\|^2_{\mathcal{Y}} \mathbb{1}_{\mathcal{X}\setminus\Omega_R^\alpha}(x)\right] \\
&= \|\bar\Psi - \Psi_0\|^2_{L^2(\nu;\Omega_R^\alpha)} + \|\bar\Psi - \Psi_0\|^2_{L^2(\nu;\mathcal{X}\setminus\Omega_R^\alpha)}.
\end{aligned}$$

This concludes the proof of Lemma 1. □

B.2 Key operator approximation result of FNN

Lemma 2. *Suppose that $\alpha > 0$, $R \ge 1$, and $\Phi_0 \in \mathcal{H}^\beta(\ell^2(\mathbb{Z}), \ell^2(\mathbb{Z}); B_0)$ for some $\beta > 0$ and $B_0 \ge 1$. Then, for any $\epsilon \in (0,1)$ and $N \in \mathbb{Z}_+$, there exists a feedforward neural network $\phi : \mathbb{R}^{2K+1} \to \mathbb{R}^{2N+1}$ in $\mathcal{F}(L, W, S, B)$ such that*

$$\sup_{\mathbf{a}\in\mathcal{B}_R^{2,\alpha}} \left\|\phi(\mathbf{a}[-K:K]) - \Phi_0(\mathbf{a})[-N:N]\right\|_2 \le \epsilon,$$

where

$$K = \left\lceil \left\{ R\left(\frac{12B_0}{\epsilon}\right)^{1/(\beta\wedge 1)} \right\}^{1/\alpha} \right\rceil - 1,$$

$$L \lesssim \left\{ \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N \right\} \log\frac{R}{\epsilon}, \qquad \log W \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N,$$

$$\log B \lesssim \log\frac{R}{\epsilon}, \qquad \log S \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N,$$

and we omit constants depending only on α, β, and B_0.

Proof. We first consider the case of $\beta > 1$. Given any $\epsilon \in (0,1)$, define $K = \lceil (12B_0R/\epsilon)^{1/\alpha} \rceil - 1$ and

$$K_0 = \left\lceil \left\{ R\left(\frac{8B_0}{\epsilon}\right)^{1/\beta} \right\}^{1/\alpha} \right\rceil - 1.$$

Then, we have $1 \le K_0 \le K$. For any $\mathbf{a} = (a_i)_{i\in\mathbb{Z}} \in \mathcal{B}_R^{2,\alpha}$, define

$$\mathbf{a}_{\le K_0} = \big(a_i \mathbb{1}_{\{|i|\le K_0\}}\big)_{i\in\mathbb{Z}}, \qquad \mathbf{a}_{>K_0} = \big(a_i \mathbb{1}_{\{|i|>K_0\}}\big)_{i\in\mathbb{Z}},$$
$$\mathbf{a}_{\le K} = \big(a_i \mathbb{1}_{|i|\le K\}}\big)_{i\in\mathbb{Z}}, \qquad \mathbf{a}_{>K} = \big(a_i \mathbb{1}_{\{|i|>K\}}\big)_{i\in\mathbb{Z}},$$
$$\mathbf{a}_{K_0<\cdot\le K} = \big(a_i \mathbb{1}_{\{K_0<|i|\le K\}}\big)_{i\in\mathbb{Z}}.$$

By definition of $B_R^{2,\alpha}$,

$$\|\mathbf{a}_{>K}\|_2^2 = \sum_{|i|>K} |a_i|^2 \le (1+K)^{-2\alpha} \sum_{|i|>K} (1+|i|)^{2\alpha}|a_i|^2 \le R^2(1+K)^{-2\alpha}.$$

Since $K \ge (12B_0R/\epsilon)^{1/\alpha} - 1$, we have $\|\mathbf{a}_{>K}\|_2 \le \epsilon/(12B_0)$. Similarly, $\|\mathbf{a}_{>K_0}\|_2 \le \{\epsilon/(8B_0)\}^{1/\beta}$. Expanding $\Phi_0(\mathbf{a})$ around $\mathbf{a}_{\le K_0}$ yields

$$\begin{aligned}\Phi_0(\mathbf{a}) &= \Phi_0(\mathbf{a}_{\le K_0}) + \sum_{k=1}^{s} \frac{1}{k!} \mathcal{D}^k \Phi_0(\mathbf{a}_{\le K_0})\big[(\mathbf{a}_{>K_0})^k\big] + R_s(\mathbf{a}) \\ &:= \Phi_0(\mathbf{a}_{\le K_0}) + T_0(\mathbf{a}) + R_s(\mathbf{a})\end{aligned}$$

By Lemma 28,

$$\|R_s(\mathbf{a})\|_2 \le \frac{B_0}{s!}\|\mathbf{a}_{>K_0}\|_2^\beta \le \frac{\epsilon}{8}.$$

In the following, we construct subnetworks to approximate $\Phi_0(\mathbf{a}_{\le K_0})$ and $T_0(\mathbf{a})$, respectively, and combine them to obtain an approximation of $\Phi_0(\mathbf{a})$. We proceed in three main steps.

Step 1: Approximate $\Phi_0(\mathbf{a}_{\le K_0})$.

Firstly, we consider the approximation of $\Phi_0(\mathbf{a}_{\le K_0})$ for any $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$. Let $\Upsilon_{K_0} : \mathbb{R}^{2K_0+1} \hookrightarrow \ell^2(\mathbb{Z})$ be the canonical isometric embedding defined by

$$[\Upsilon_{K_0}(\mathbf{x})]_i = \begin{cases} x_i, & \text{if} \quad |i| \le K_0, \\ 0 & \text{if} \quad |i| > K_0, \end{cases}, \qquad \mathbf{x} = (x_{-K_0}, \dots, x_{K_0}) \in \mathbb{R}^{2K_0+1}.$$

Then, our target is to approximate $\Phi_{0,K_0} := \Phi_0 \circ \Upsilon_{K_0}$ on a bounded domain $\Omega_{\alpha,R,\delta} \subset \mathbb{R}^{2K_0+1}$ using a feedforward neural network. Since Υ_{K_0} is a bounded linear operator with $\|\Upsilon_{K_0}\|_{B(\mathbb{R}^{2K_0+1},\ell^2(\mathbb{Z}))} = 1$, Φ_{0,K_0} belongs to the Fréchet-Hölder class $\mathcal{H}^\beta(\mathbb{R}^{2K_0+1}, \ell^2(\mathbb{Z}); B_0)$, and can be approximated using Taylor expansion. By Lemma 28, for any $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^{2K_0+1}$, we have

$$\left\| \Phi_{0,K_0}(\mathbf{x}) - \sum_{k=0}^{s} \frac{1}{k!} \mathcal{D}^k \Phi_{0,K_0}(\mathbf{x}_0)\big[(\mathbf{x}-\mathbf{x}_0)^k\big] \right\|_2 \le \frac{B_0}{s!}\|\mathbf{x}-\mathbf{x}_0\|_2^\beta.$$

That is, the Taylor remainder can be well controlled when $\mathbf{x}$ lies within a fairly small neighborhood of $\mathbf{x}_0$. This motivates us to construct localized approximations of Φ_{0,K_0}, and combine them using a partition of unity. This proof step is further divided into six sub-steps.

Step 1.1: Construct partition of unity.
For any $\eta \in (0,1)$, define $\delta = \eta/\sqrt{2K_0+1}$ $J_i = \lceil (1+|i|)^{-\alpha} R/\delta \rceil$ $(-K_0 \le i \le K_0)$, and

$$\Omega_{\alpha,R,\delta} = \big\{\mathbf{x} = (x_{-K_0}, \dots, x_{K_0}) \in \mathbb{R}^{2K_0+1} : |x_i| \le J_i\delta,\ -K_0 \le i \le K_0\big\}.$$

Note that $J_i\delta \ge R/(1+|i|)^\alpha$, we have $\{\mathbf{a}[-K_0 : K_0] : \mathbf{a} \in \mathcal{B}_R^{2,\alpha}\} \subset \Omega_{\alpha,R,\delta}$. For any $j \in \mathbb{Z}$ and $x \in \mathbb{R}$, we define

$$\psi_j(x) = \frac{1}{\delta}\big\{\sigma(x-(j-1)\delta) + \sigma(x-(j+1)\delta) - 2\sigma(x-j\delta)\big\}.$$

Then, ψ_j is the hat function supported on $[(j-1)\delta, (j+1)\delta]$ with gradient $1/\delta$. For any $-K_0 \le i \le K_0$ and $x \in [-J_i\delta, J_i\delta]$, we have $\sum_{j=-J_i}^{J_i} \psi_j(x) = 1$. Moreover, ψ_j is an FNN with depth 1, width 3, and weight-norm bound $\max\{(1+|j|)\delta, 1/\delta\}$. Let $\Theta = \{(\theta_{-K_0}, \theta_{-K_0+1}, \dots, \theta_{K_0})^\top \in \mathbb{Z}^{2K_0+1} : |\theta_i| \le J_i, -K_0 \le i \le K_0\}$. For each $\theta \in \Theta$, we further define

$$\psi_\theta(\mathbf{x}) = \prod_{i=-K_0}^{K_0} \psi_{\theta_i}(x_i), \qquad \mathbf{x} = (x_{-K_0}, \dots, x_{K_0}) \in \mathbb{R}^{2K_0+1}.$$

Then, for any $\mathbf{x} \in \Omega_{\alpha,R,\delta}$,

$$\sum_{\theta\in\Theta} \psi_\theta(\mathbf{x}) = \sum_{\theta_{-K_0}=-J_{-K_0}}^{J_{-K_0}} \cdots \sum_{\theta_{K_0}=-J_{K_0}}^{J_{K_0}} \prod_{i=-K_0}^{K_0} \psi_{\theta_i}(x_i) = \prod_{i=-K_0}^{K_0} \left\{ \sum_{\theta_i=-J_i}^{J_i} \psi_{\theta_i}(x_i) \right\} = 1.$$

That is, $\{\psi_\theta : \theta \in \Theta\}$ is a partition of unity on $\Omega_{\alpha,R,\delta}$. Note that for any $j \in \mathbb{Z}$ and $x \in \mathbb{R}$, $\psi_j(x) \in [0,1]$. According to Lemma 26, for any $\theta \in \Theta$ and $\epsilon > 0$, there exist an FNN $\psi_{\theta,\times}$ with depth $L_1 = \lceil \log_2((2K_0+1)^2/\epsilon_1) + 5\rceil \lceil \log_2(2K_0+1)\rceil + 1$, width $6(2K_0+1)$ and weight-norm bound $\max\{(1+J_0)\delta, 1/\delta\}$ such that

$$\sup_{\mathbf{x}\in\Omega_{\alpha,R,\eta}} |\psi_{\theta,\times}(\mathbf{x}) - \psi_\theta(\mathbf{x})| \le \epsilon_1.$$

Moreover, $\psi_{\theta,\times}(\mathbf{x}) = 0$ whenever $\psi_\theta(\mathbf{x}) = 0$.

Step 1.2: Evaluate Taylor expansion coordinate-wise.
For each output index $j \in \mathbb{Z}$, define the canonical projection functional $\pi_j \in (\ell^2(\mathbb{Z}))^*$ by $\pi_j(\mathbf{y}) = y_j$, where $\mathbf{y} \in \ell^2(\mathbb{Z})$. Since π_j is a bounded linear functional with $\|\pi_j\|_{(\ell^2(\mathbb{Z}))^*} = 1$, the coordinate-output function $f_j := \pi_j \circ \Phi_{0,K_0} : \mathbb{R}^{2K_0+1} \to \mathbb{R}$ has the same Fréchet differentiability as Φ_{0,K_0}. Moreover, π_j is commutative with the Fréchet derivative operator:

$$\pi_j\big(\mathcal{D}^k\Phi_{0,K_0}(\mathbf{x}_0)\big[(\mathbf{x}-\mathbf{x}_0)^k\big]\big) = \mathcal{D}^k(\pi_j\circ\Phi_{0,K_0})(\mathbf{x}_0)\big[(\mathbf{x}-\mathbf{x}_0)^k\big] = \mathcal{D}^k f_j(\mathbf{x}_0)\big[(\mathbf{x}-\mathbf{x}_0)^k\big].$$

Evaluating the k-th order Fréchet derivative $\mathcal{D}^k f_j(\mathbf{x}_0)$ of the real-valued function f_j on k copies of the vector $\mathbf{x}-\mathbf{x}_0 \in \mathbb{R}^{2K_0+1}$ yields

$$\frac{1}{k!}\mathcal{D}^k f_j(\mathbf{x}_0)\big[(\mathbf{x}-\mathbf{x}_0)^k\big] = \sum_{\|\mathbf{m}\|_1=k} \frac{1}{\mathbf{m}!}\partial^{\mathbf{m}} f_j(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^{\mathbf{m}}.$$

where $\mathbf{m} = (m_{-K_0}, \dots, m_{K_0}) \in \mathbb{N}^{2K_0+1}$ is a multi-index, $\|\mathbf{m}\|_1 = \sum_{i=-K_0}^{K_0} m_i$, $\mathbf{m}! := \prod_{i=-K_0}^{K_0} m_i!$, $\mathbf{x}^{\mathbf{m}} := x_{-K_0}^{m_{-K_0}} \cdots x_{K_0}^{m_{K_0}}$ and $\partial^{\mathbf{m}} := \partial^{\|\mathbf{m}\|_1}/(\partial x_{-K_0}^{m_{-K_0}} \cdots \partial x_{K_0}^{m_{K_0}})$. That said, it suffices to approximate the Taylor polynomial

$$\sum_{\|\mathbf{m}\|_1\le s} \frac{1}{\mathbf{m}!}\partial^{\mathbf{m}} f_j(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0)^{\mathbf{m}}$$

for each j.

Step 1.3: Construct localized approximation.
Given j, for each $\theta \in \Theta$, we construct an FNN to approximate the Taylor polynomial of f_j expanded at $\theta\delta := [\theta_{-K_0}\delta, \theta_{-K_0+1}\delta, \dots, \theta_{K_0}\delta]^\top$, that is,

$$T_{j,\theta}(\mathbf{x}) = \sum_{\|\mathbf{m}\|_1\le s} \frac{1}{\mathbf{m}!}\partial^{\mathbf{m}} f_j(\theta\delta)(\mathbf{x}-\theta\delta)^{\mathbf{m}}.$$

According to Lemma 27, for any multi-index $\mathbf{m}$ with $1 \le \|\mathbf{m}\|_1 \le s$ and $\epsilon_2 \in (0,1)$, there exists an FNN $P_{\mathbf{m}}$ with depth $\lceil \log_2(s^2/\epsilon_2) + s + 5 \rceil \lceil \log_2 s \rceil + 1$, width $6s2^s$, and weight-norm bound 1 such that for all $\mathbf{u} \in [-1,1]^{2K_0+1}$,

$$|P_{\mathbf{m}}(\mathbf{u}) - \mathbf{u}^{\mathbf{m}}| \le \epsilon_2.$$

Define

$$\tilde{T}_{j,\theta}(\mathbf{x}) \coloneqq f_j(\theta\delta) + \sum_{1 \le \|\mathbf{m}\|_1 \le s} \frac{\partial^{\mathbf{m}} f_j(\theta\delta)}{\mathbf{m}!} P_{\mathbf{m}}(\mathbf{x} - \theta\delta).$$

Note that the number of terms in the summation can be bounded by

$$\sum_{1 \le \|\mathbf{m}\|_1 \le s} 1 = \sum_{k=1}^{s} \sum_{\|\mathbf{m}\|_1 = k} 1 \le \sum_{k=1}^{s} (2K_0+1)^s = s(2K_0+1)^s,$$

and the partial derivatives are bounded as

$$\begin{aligned} |\partial^{\mathbf{m}} f_j(\theta\delta)| &= \left| \mathcal{D}^{\|\mathbf{m}\|_1} f_j(\theta\delta)[e_{-K_0}^{m_{-K_0}}, \dots, e_{K_0}^{m_{K_0}}] \right| \le \left\| \mathcal{D}^{\|\mathbf{m}\|_1} f_j(\theta\delta) \right\|_{(\mathbb{R}^{2K_0+1})^*} \prod_{i=-K_0}^{K_0} \|e_i\|_2^{m_i} \\ &= \left\| \mathcal{D}^{\|\mathbf{m}\|_1} f_j(\theta\delta) \right\|_{(\mathbb{R}^{2K_0+1})^*} \le B_0, \end{aligned}$$

where $e_{-K_0}, \dots, e_{K_0} \in \mathbb{R}^{2K_0+1}$ are standard basis vectors. Then, $\tilde{T}_{j,\theta}$ can be implemented by an FNN with depth $L_2 = \lceil \log_2(s^2/\epsilon_2) + s + 5 \rceil \lceil \log_2 s \rceil + 1$, width $6s^2(4K_0+2)^s$, and weight-norm bound $\max\{J_0\delta, B_0\}$. For all $\mathbf{x}$ satisfying $\|\mathbf{x} - \theta\delta\|_\infty \le \delta < 1$, we have

$$|\tilde{T}_{j,\theta}(\mathbf{x}) - T_{j,\theta}(\mathbf{x})| \le \sum_{1 \le \|\mathbf{k}\|_1 \le s} \frac{|\partial^{\mathbf{m}} f_j(\theta\delta)|}{\mathbf{m}!} \cdot |P_{\mathbf{m}}(\mathbf{x} - \theta\delta) - (\mathbf{x} - \theta\delta)^{\mathbf{m}}| \le sB_0(2K_0+1)^s \epsilon_2.$$

Step 1.4: Obtain approximation on $\Omega_{\alpha,R,\delta}$.
In this step, we combine the localized approximation using the partition of unity. According to Lemma 27, for any $\epsilon_3 > 0$, there exists an FNN $\phi_\times$ with depth $L_3 = \lceil \log_2(1/\epsilon_3) + 9 \rceil$, width 48, and weight-norm bound 1 such that

$$\sup_{x,y \in [-1,1]} |\phi_\times(x,y) - xy| \le \epsilon_3.$$

Moreover, $\phi_\times(x,y) = 0$ if $xy = 0$. Note that for any $\theta \in \Theta$, we have $\psi_{\theta,\times}(\mathbf{x}) = \psi_\theta(\mathbf{x}) = 0$ whenever $\|\mathbf{x} - \theta\delta\|_\infty \ge \delta$. When $\|\mathbf{x} - \theta\delta\|_\infty \le \delta$, we have $|P_{\mathbf{m}}(\mathbf{x} - \theta\delta)| \le |(\mathbf{x} - \theta\delta)^{\mathbf{m}}| + \epsilon_2 \le 2$ and $|\tilde{T}_{j,\theta}(\mathbf{x})| \le 3sB_0(2K_0+1)^s \coloneqq M_T$. Furthermore, $|\psi_{\theta,\times}(\mathbf{x})| \le |\psi_\theta(\mathbf{x})| + \epsilon_1 \le 2$. Define

$$\phi_j(\mathbf{x}) = 2M_T \sum_{\theta \in \Theta} \phi_\times\left(\frac{\tilde{T}_{j,\theta}(\mathbf{x})}{M_T}, \frac{\psi_{\theta,\times}(\mathbf{x})}{2} \right),$$

then ϕ_j can be implemented by an FNN with with depth $\{(L_1 \vee L_2) + L_3\}$, width $9s^2(4K_0+2)^s|\Theta|$, and weight-norm bound $\max\{(1+J_0)\delta, 1/\delta, 2M_T\}$. Further define

$$\phi_j^\circ(\mathbf{x}) = \sum_{\theta \in \Theta} T_{j,\theta}(\mathbf{x}) \psi_\theta(\mathbf{x}).$$

Then, for any $\mathbf{x} \in \Omega_{\alpha,R,\eta}$,

$$\begin{aligned} &|\phi_j(\mathbf{x}) - \phi_j^\circ(\mathbf{x})| \\ &\le 2M_T \sum_{\theta \in \Theta} \left| \phi_\times\left(\frac{\tilde{T}_{j,\theta}(\mathbf{x})}{M_T}, \frac{\psi_{\theta,\times}(\mathbf{x})}{2} \right) - \frac{T_{j,\theta}(\mathbf{x})\psi_\theta(\mathbf{x})}{2M_T} \right| \\ &\le 2M_T \sum_{\theta \in \Theta} \left\{ \left| \phi_\times\left(\frac{\tilde{T}_{j,\theta}(\mathbf{x})}{M_T}, \frac{\psi_{\theta,\times}(\mathbf{x})}{2} \right) - \frac{\tilde{T}_{j,\theta}(\mathbf{x})\psi_{\theta,\times}(\mathbf{x})}{2M_T} \right| + \left| \frac{\tilde{T}_{j,\theta}(\mathbf{x})\psi_{\theta,\times}(\mathbf{x})}{2M_T} - \frac{T_{j,\theta}(\mathbf{x})\psi_\theta(\mathbf{x})}{2M_T} \right| \right\} \\ &\le 2M_T \sum_{\theta \in \Theta} \epsilon_3 + \sum_{\theta \in \Theta} \left\{ |\psi_\theta(\mathbf{x})||\tilde{T}_{j,\theta}(\mathbf{x}) - T_{j,\theta}(\mathbf{x})| + |\tilde{T}_{j,\theta}(\mathbf{x})||\psi_{\theta,\times}(\mathbf{x}) - \psi_\theta(\mathbf{x})| \right\} \\ &\le 2M_T|\Theta|\epsilon_3 + sB_0(2K_0+1)^s\epsilon_2 + |\Theta|M_T\epsilon_1 \\ &= sB_0(2K_0+1)^s\big(3|\Theta|\epsilon_1 + \epsilon_2 + 6|\Theta|\epsilon_3\big) \coloneqq \mathcal{E}_0. \end{aligned}$$

Step 1.5: Combine coordinate-wise approximation.
For any $N \in \mathbb{Z}_+$, we define

$$\phi^\dagger(\mathbf{x}) = [\phi_{-N}(\mathbf{x}), \phi_{-N+1}(\mathbf{x}), \ldots, \phi_N(\mathbf{x})],$$
$$\phi^\circ(\mathbf{x}) = [\phi^\circ_{-N}(\mathbf{x}), \phi^\circ_{-N+1}(\mathbf{x}), \ldots, \phi^\circ_N(\mathbf{x})].$$

Then, $\phi^\dagger$ can be implemented by an FNN with with depth $L = \{(L_1 \vee L_2) + L_3\}$, width $W = 9(2N+1)s^2(4K_0+2)^s|\Theta|$, and weight-norm bound $B = \max\{(1+J_0)\delta, 1/\delta, 2M_T\}$. For any $\mathbf{x} \in \Omega_{\alpha,R,\eta}$,

$$\|\phi^\dagger(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x})[-N:N]\|_2 \le \|\phi^\dagger(\mathbf{x}) - \phi^\circ(\mathbf{x})\|_2 + \|\phi^\circ(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x})[-N:N]\|_2.$$

Note that the first term can be bounded by $\sqrt{2N+1}\mathcal{E}_0$. We thus focus on bounding the second term $\|\phi^\circ(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x})[-N:N]\|_2$. Define

$$T_\theta(\mathbf{x}) = \sum_{k=0}^{s} \frac{1}{k!} \mathcal{D}^k \Phi_{0,K_0}(\theta\delta)\big[(\mathbf{x}-\theta\delta)^k\big],$$

then we have $\pi_j(T_\theta(\mathbf{x})) = T_{j,\theta}(\mathbf{x})$. In other words, $T_\theta(\mathbf{x}) = (T_{j,\theta}(\mathbf{x}))_{j\in\mathbb{Z}}$. According to Lemma 28, on the support of ψ_θ, we have

$$\|\Phi_{0,K_0}(\mathbf{x}) - T_\theta(\mathbf{x})\|_2 \le \frac{B_0}{s!}\|\mathbf{x}-\theta\delta\|_2^\beta \le \frac{B_0}{s!}(2K_0+1)^{\beta/2}\delta^\beta = \frac{B_0}{s!}\eta^\beta.$$

Therefore, for any $\mathbf{x} \in \Omega_{\alpha,R,\delta}$,

$$\begin{aligned}
\big\|\phi^\circ(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x})[-N:N]\big\|_2 &= \Big\| \sum_{\theta\in\Theta} \psi_\theta(\mathbf{x}) \cdot \big\{T_\theta(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x})\big\}[-N:N] \Big\|_2 \\
&\le \sum_{\theta\in\Theta} \psi_\theta(\mathbf{x}) \big\| \big\{T_\theta(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x})\big\}[-N:N] \big\|_2 \\
&\le \sum_{\theta\in\Theta} \psi_\theta(\mathbf{x}) \big\| T_\theta(\mathbf{x}) - \Phi_{0,K_0}(\mathbf{x}) \big\|_2 \le \frac{B_0}{s!}\eta^\beta.
\end{aligned}$$

Step 1.6: Derive overall approximation error bound.
Since $\{\mathbf{a}[-K_0:K_0] : \mathbf{a} \in \mathcal{B}_R^{2,\alpha}\} \subset \Omega_{\alpha,R,\delta}$, it follows for any $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$ that

$$\begin{aligned}
&\big\|\phi^\dagger(\mathbf{a}[-K_0:K_0]) - \Phi_0(\mathbf{a}_{\le K_0})[-N:N]\big\|_2 \\
&= \big\|\phi^\dagger(\mathbf{a}[-K_0:K_0]) - \Phi_{0,K_0}(\mathbf{a}[-K_0:K_0])[-N:N]\big\|_2 \\
&\le \big\|\phi^\dagger(\mathbf{a}[-K_0:K_0]) - \phi^\circ(\mathbf{a}[-K_0:K_0])\big\|_2 \\
&\quad + \big\|\phi^\circ(\mathbf{a}[-K_0:K_0]) - \Phi_{0,K_0}(\mathbf{a}[-K_0:K_0])[-N:N]\big\|_2 \\
&\le sB_0\sqrt{2N+1}(2K_0+1)^s\big(3|\Theta|\epsilon_1 + \epsilon_2 + 6|\Theta|\epsilon_3\big) + \frac{B_0}{s!}\eta^\beta \\
&:= \mathcal{E}_2 + \frac{B_0}{s!}\eta^\beta.
\end{aligned}$$

Take $\eta = s^{-1/(\beta-s)}(2K+1)^{-s/(\beta-s)}\{\epsilon/(8B_0)\}^{1/\beta}$, we have $B_0\eta^\beta \le \epsilon/24$. Let $C_0 = s^2B_0\sqrt{2N+1}(2K_0+1)^s(2K+1)^s$ and take

$$\epsilon_1 = \frac{\epsilon}{24C_0|\Theta|}, \qquad \epsilon_2 = \frac{\epsilon}{8C_0}, \qquad \epsilon_3 = \frac{\epsilon}{48C_0|\Theta|},$$

we have $\mathcal{E}_2 \le C_0\big(3|\Theta|\epsilon_1 + \epsilon_2 + 6|\Theta|\epsilon_3\big)/3 \le \epsilon/8$. Consequently, the approximation error of $\phi^\dagger$ satisfies

$$\big\|\phi^\dagger(\mathbf{a}[-K_0:K_0]) - \Phi_0(\mathbf{a}_{\le K_0})[-N:N]\big\|_2 \le \frac{\epsilon}{4}$$

Next, we bound the network architecture parameters of $\phi^\dagger$. Note that

$$|\Theta| = \prod_{i=-K_0}^{K_0}(2J_i+1) \le (2J_0+1)^{2K_0+1} \le \left\{2R(2K+1)^{\beta/(\beta-s)}\left(\frac{8B_0}{\epsilon}\right)^{1/\beta} + 3\right\}^{2K_0+1}.$$

Using $K_0 \asymp (R/\epsilon^{1/\beta})^{1/\alpha}$, we have

$$\log|\Theta| \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon}, \qquad L_1 \lesssim \left\{\left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N\right\} \log\frac{R}{\epsilon},$$
$$L_2 \asymp \log\frac{R}{\epsilon} + \log N, \qquad L_3 \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N.$$

Consequently, the depth can be bounded as

$$L = (L_1 \vee L_2) + L_3 \lesssim \left\{\left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N\right\} \log\frac{R}{\epsilon}.$$

Moreover, the width can be bounded as

$$\log W \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N.$$

Since $(1+J_0)\delta \asymp R$, $M_T \asymp (R/\epsilon^{1/\beta})^{s/\alpha}$, and $1/\delta \asymp (R/\epsilon)^{1/(2\alpha\beta)}\epsilon^{-1/\beta}$, the weight-norm bound satisfies

$$\log B \lesssim \log\frac{R}{\epsilon}.$$

To bound the sparsity of $\phi^\dagger$, note that $\psi_{\theta,\times}$ has sparsity $\mathcal{O}(K^2L_1)$, $\tilde{T}_{j,\theta}$ has sparsity $\mathcal{O}(K^sL_2)$, and ϕ_j has sparsity $\mathcal{O}(|\Theta|\{K^2L_1 + K^sL_2 + L_3\})$. Consequently, the sparsity of $\phi^\dagger$ satisfies

$$\log S \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha} \log\frac{R}{\epsilon} + \log N.$$

Step 2: Approximate $T_0(\mathbf{a})$.
Let

$$\tilde{T}_0(\mathbf{a}_{\le K}) = \sum_{k=1}^{s} \frac{1}{k!}\mathcal{D}^k\Phi_0(\mathbf{a}_{\le K_0})\big[(\mathbf{a}_{K_0<\cdot\le K})^k\big].$$

For each $1 \le k \le s$, define

$$\Delta_k = \mathcal{D}^k\Phi_0(\mathbf{a}_{\le K_0})\big[(\mathbf{a}_{>K_0})^k\big] - \mathcal{D}^k\Phi_0(\mathbf{a}_{\le K_0})\big[(\mathbf{a}_{K_0<\cdot\le K})^k\big].$$

Then, $T_0(\mathbf{a}) - \tilde{T}_0(\mathbf{a}_{\le K}) = \sum_{k=1}^{s}\Delta_k/k!$. By the telescoping expansion

$$(\mathbf{a}_{>K_0})^k - (\mathbf{a}_{K_0<\cdot\le K})^k = \sum_{m=1}^{k} \mathbf{a}_{>K} \otimes (\mathbf{a}_{K_0<\cdot\le K})^{m-1} \otimes (\mathbf{a}_{>K_0})^{k-m}$$

and the multilinearlity of Fréchet derivative,

$$\|\Delta_k\|_2 \le B_0 \sum_{m=1}^{k} \|\mathbf{a}_{>K}\|_2 \|\mathbf{a}_{K_0<\cdot\le K}\|_2^{m-1} \|\mathbf{a}_{>K_0}\|_2^{k-m} \le B_0 k \|\mathbf{a}_{>K}\|_2,$$

where we have used $\|\mathcal{D}^k\Phi_0(\mathbf{a}_{\le K_0})\|_{B^k(\ell^2(\mathbb{Z}),\ell^2(\mathbb{Z}))} \le B_0$ and $\|\mathbf{a}_{>K_0}\|_2 \le \{\epsilon/(4B_0)\}^{1/\beta} \le 1$. Consequently,

$$\big\|T_0(\mathbf{a}) - \tilde{T}_0(\mathbf{a}_{\le K})\big\|_2 \le \sum_{k=1}^{s} \frac{1}{k!}\|\Delta_k\|_2 \le \sum_{k=1}^{s} \frac{B_0}{(k-1)!} \cdot \frac{\epsilon}{12B_0} \le \frac{e}{12}\epsilon < \frac{\epsilon}{4},$$

where we have used $\|\mathbf{a}_{>K}\|_2 \le \epsilon/(12B_0)$. Therefore, it suffices to approximate $\tilde{T}_0(\mathbf{a}_{\le K})$. If $K_0 = K$, we have $\tilde{T}_0(\mathbf{a}_{\le K}) = 0$. In the following, we consider $K_0 < K$. For each $j \in \mathbb{Z}$, define $h_j = \pi_j \circ \Phi_0 \circ \Upsilon_K$. Then,

$$\tilde{T}_j(\mathbf{a}_{\le K}) \coloneqq \pi_j \circ \tilde{T}_0(\mathbf{a}_{\le K}) = \sum_{k=1}^{s} \sum_{\substack{\|\mathbf{m}\|_1=k,\\ \mathbf{m}[-K_0:K_0]=0}} \frac{1}{\mathbf{m}!}\partial^{\mathbf{m}}h_j(\mathbf{a}_{\le K_0}[-K:K])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}},$$

where $\mathbf{m} = (m_{-K}, \ldots, m_K) \in \mathbb{N}^{2K+1}$. For each multi-index $\mathbf{m}$, the function

$$h^{(\mathbf{m})}(\mathbf{x}) \coloneqq [\partial^{\mathbf{m}} h_{-N}(\Upsilon_{K_0}(\mathbf{x})[-K:K]), \ldots, \partial^{\mathbf{m}} h_N(\Upsilon_{K_0}(\mathbf{x})[-K:K])]$$

belongs to $\mathcal{H}^{\beta-\|\mathbf{m}\|_1}(\mathbb{R}^{2K_0+1}, \mathbb{R}^{2N+1}; B_0)$. Using the partition of unit and localized Taylor polynomial in Step 1, for each $\mathbf{m}$ such that $1 \le \|\mathbf{m}\|_1 < s$, we can construct an FNN $\tilde{h}^{(\mathbf{m})}$ such that for any $\mathbf{x} \in \Omega_{\alpha,R,\delta}$,

$$\begin{aligned}
&\big\|\tilde{h}^{(\mathbf{m})}(\mathbf{x}) - h^{(\mathbf{m})}(\mathbf{x})\big\|_2 \\
&\le (s - \|\mathbf{m}\|_1) B_0 \sqrt{2N+1} (2K_0+1)^{s-\|\mathbf{m}\|_1} \big(3|\Theta|\epsilon_1 + \epsilon_2 + 6|\Theta|\epsilon_3\big) + \frac{B_0}{(s-\|\mathbf{m}\|_1)!} \eta^{\beta-\|\mathbf{m}\|_1} \\
&\le s^{-1}(2K+1)^{-s} \left\{1 + \left(\frac{8B_0}{\epsilon}\right)^{\|\mathbf{m}\|_1/\beta}\right\} \frac{\epsilon}{8}.
\end{aligned}$$

Moreover, $\tilde{h}^{(\mathbf{m})}$ has depth L, width $\mathcal{O}(W)$, weight-norm bound $\mathcal{O}(B)$, and sparsity $\mathcal{O}(S)$. If $\|\mathbf{m}\|_1 = s$, $h^{(\mathbf{m})}$ is $(\beta - s)$-Hölder continuous and the Taylor polynomials reduces to constants. We can construct $\tilde{h}^{(\mathbf{m})}$ using a linear combination of the partition of unity (refer to the case of $0 < \beta \le 1$ discussed later), such that it has same order of depth, width, weight-norm bound and sparsity and

$$\big\|\tilde{h}^{(\mathbf{m})}(\mathbf{x}) - h^{(\mathbf{m})}(\mathbf{x})\big\|_2 \le s^{-1}(2K+1)^{-s} \left\{1 + \left(\frac{8B_0}{\epsilon}\right)^{s/\beta}\right\} \frac{\epsilon}{8}, \qquad \forall\, \mathbf{x} \in \Omega_{\alpha,R,\delta}.$$

It follows that $\big\|\tilde{h}^{(\mathbf{m})}(\mathbf{x})\big\|_2 \le 3B_0$. Also note that $\mathbf{m}[-K_0 : K_0] = 0$ and $\|\mathbf{a}_{>K_0}\|_2 \le \{\epsilon/(8B_0)\}^{1/\beta} \le 1$. According to Lemma 27, for any $\mathbf{m}$ and $\epsilon_4 \in (0,1)$, there exists an FNN $P_{\times,\mathbf{m}}$ with depth $\lceil \log_2((s+1)^2/\epsilon_4) + s + 6 \rceil \lceil \log_2(s+1) \rceil + 1$, width $6(2N+1)(s+1)2^{s+1}$, and weight-norm bound 1 such that for any $\mathbf{u} \in [-1,1]^{2N+1}$ and $\mathbf{v} \in [-1,1]^{2K+1}$,

$$\|P_{\times,\mathbf{m}}(\mathbf{u}, \mathbf{v}) - \mathbf{u}\mathbf{v}^{\mathbf{m}}\|_\infty \le \epsilon_4.$$

Define

$$\phi^{\diamond}_{\mathbf{m}}(\mathbf{a}[-K:K]) = 3B_0 P_{\times,\mathbf{m}}\left(\frac{\tilde{h}^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])}{3B_0}, \mathbf{a}[-K:K]\right).$$

Then, for any $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$,

$$\begin{aligned}
&\big\|\phi^{\diamond}_{\mathbf{m}}(\mathbf{a}[-K:K]) - h^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}}\big\|_2 \\
&\le \big\|\phi^{\diamond}_{\mathbf{m}}(\mathbf{a}[-K:K]) - \tilde{h}^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}}\big\|_2 \\
&\quad + \big\|\tilde{h}^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}} - h^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}}\big\|_2 \\
&\le 3B_0\sqrt{2N+1}\epsilon_4 + s^{-1}(2K+1)^{-s}\big(\|\mathbf{a}_{\ge K_0}\|_\infty\big)^{\|\mathbf{m}\|_1} \left\{1 + \left(\frac{8B_0}{\epsilon}\right)^{\|\mathbf{m}\|_1/\beta}\right\} \frac{\epsilon}{8} \\
&\le 3B_0\sqrt{2N+1}\epsilon_4 + s^{-1}(2K+1)^{-s}\epsilon/4.
\end{aligned}$$

Take $\epsilon_4 = s^{-1}(2K+1)^{-s}(2N+1)^{-1/2}(24B_0)^{-1}\epsilon$, we have

$$\big\|\phi^{\diamond}_{\mathbf{m}}(\mathbf{a}[-K:K]) - h^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}}\big\|_2 \le s^{-1}(2K+1)^{-s}\frac{3\epsilon}{8}.$$

Further define

$$\phi^{\diamond}(\mathbf{a}[-K:K]) = \sum_{k=1}^{s} \sum_{\substack{\|\mathbf{m}\|_1 = k, \\ \mathbf{m}[-K_0:K_0]=0}} \frac{1}{\mathbf{m}!} \phi^{\diamond}_{\mathbf{m}}(\mathbf{a}[-K:K]).$$

Then,

$$\begin{aligned}
&\big\|\phi^{\diamond}(\mathbf{a}[-K:K]) - \tilde{T}_0(\mathbf{a}_{\le K})[-N:N]\big\|_2 \\
&\le \sum_{k=1}^{s} \sum_{\substack{\|\mathbf{m}\|_1 = k, \\ \mathbf{m}[-K_0:K_0]=0}} \big\|\phi^{\diamond}_{\mathbf{m}}(\mathbf{a}[-K:K]) - h^{(\mathbf{m})}(\mathbf{a}[-K_0:K_0])\big(\mathbf{a}[-K:K]\big)^{\mathbf{m}}\big\|_2 \\
&\le s\{2(K-K_0)\}^s \cdot s^{-1}(2K+1)^{-s}\frac{3\epsilon}{8} \le \frac{3\epsilon}{8}.
\end{aligned}$$

Note that $\phi^\diamond$ can be implemented by paralleling $s\{2(K-K_0)\}^s = \mathcal{O}((R/\epsilon)^{s/\alpha})$ subnetworks $\phi^\diamond_{\mathbf{m}}$, which yields depth $\mathcal{O}(L)$, logarithmic width $\mathcal{O}(W \vee (R/\epsilon)^{s/\alpha})$, weight-norm bound $\mathcal{O}(B)$, and sparsity $\mathcal{O}(S)$.

Step 3. Combine the approximation of $\Phi_0(\mathbf{a}_{\le K_0})$ and $T_0(\mathbf{a})$.
Define

$$\phi(\mathbf{a}[-K:K]) = \phi^\dagger(\mathbf{a}[-K_0,K_0]) + \phi^\diamond(\mathbf{a}[-K:K]).$$

Then, for any $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$,

$$\begin{aligned}
&\big\|\phi(\mathbf{a}[-K:K]) - \Phi_0(\mathbf{a})[-N:N]\big\|_2 \\
&\le \big\|\phi^\dagger(\mathbf{a}[-K_0:K_0]) - \Phi_0(\mathbf{a}_{\le K_0})[-N:N]\big\|_2 + \big\|\phi^\diamond(\mathbf{a}[-K:K]) - \tilde{T}_0(\mathbf{a}_{\le K})[-N:N]\big\|_2 \\
&\quad + \big\|\tilde{T}_0(\mathbf{a}_{\le K}) - T_0(\mathbf{a})\big\|_2 + \|R_s(\mathbf{a})\|_2 \\
&\le \frac{\epsilon}{4} + \frac{3\epsilon}{8} + \frac{\epsilon}{4} + \frac{\epsilon}{8} = \epsilon.
\end{aligned}$$

Moreover, ϕ can be implemented by an FNN with depth $\mathcal{O}(L)$, width $\mathcal{O}(W)$, weight-norm bound $\mathcal{O}(B)$, and sparsity $\mathcal{O}(S)$. This completes the proof for $\beta > 1$.

Lastly, if $0 < \beta \le 1$, Φ_0 is β-Hölder continuous. Set $K_0 = K = \lceil \{R(12B_0/\epsilon)^{1/\beta}\}^{1/\alpha} \rceil - 1$, then

$$\begin{aligned}
&\big\|\{\Phi_{0,K}(\mathbf{a}[-K:K]) - \Phi_0(\mathbf{a})\}[-N:N]\big\|_2 \le \|\Phi_0(\Upsilon_K(\mathbf{a}[-K:K]) - \Phi_0(\mathbf{a})\|_2 \\
&\le B_0\|\Upsilon_K(\mathbf{a}[-K:K]) - \mathbf{a}\|_2^\beta = B_0\|\mathbf{a}_{>K}\|_2^\beta \le \epsilon/12.
\end{aligned}$$

We approximate $f_j = \pi_j \circ \Phi_{0,K}$ using a linear combination of the partition of unity

$$\phi_j(\mathbf{x}) = \sum_{\theta\in\Theta} f_j(\theta\delta)\psi_{\theta,\times}(\mathbf{x}),$$

which can be implemented by paralleling $|\Theta|$ subnetworks $\{\psi_{\theta,\times} : \theta \in \Theta\}$. That said, ϕ_j is an FNN with depth L_1, width $\mathcal{O}(|\Theta|K)$, weight-norm bound $\mathcal{O}(R/\delta)$. For any $\mathbf{x} \in \Omega_{\alpha,R,\delta}$, it follows from the Hölder continuity of $\Phi_{0,K}$ that

$$\begin{aligned}
&\big\|\phi(\mathbf{x}) - \Phi_{0,K}(\mathbf{x})[-N:N]\big\|_2 = \Big\|\sum_{\theta\in\Theta}\big\{\Phi_{0,K}(\theta\delta)\psi_{\theta,\times}(\mathbf{x}) - \Phi_{0,K}(\mathbf{x})\psi_\theta(\mathbf{x})\big\}[-N:N]\Big\|_2 \\
&\le \sum_{\theta\in\Theta}|\psi_{\theta,\times}(\mathbf{x}) - \psi_\theta(\mathbf{x})| \cdot \|\Phi_{0,K}(\theta\delta)\|_2 + \sum_{\theta\in\Theta}\psi_\theta(\mathbf{x})\|\Phi_{0,K}(\theta\delta) - \Phi_{0,K}(\mathbf{x})\|_2 \\
&\le B_0|\Theta|\epsilon_1 + B_0(2K+1)^{\beta/2}\delta^\beta = B_0|\Theta|\epsilon_1 + B_0\eta^\beta.
\end{aligned}$$

Take $\eta = \{\epsilon/(4B_0)\}^{1/\beta}$ and $\epsilon_1 = \epsilon/(2B_0|\Theta|)$, we have

$$\big\|\phi(\mathbf{a}[-K:K]) - \Phi_0(\mathbf{a})[-N:N]\big\|_2 \le \epsilon.$$

Using $K \asymp (R/\epsilon^{1/\beta})^{1/\alpha}$, the architecture parameters L, W, B, and S can be bounded similarly as the case of $\beta > 1$. Specifically,

$$\begin{aligned}
L = L_1 &\lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha}\left(\log\frac{R}{\epsilon}\right)^2, \qquad \log W \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha}\log\frac{R}{\epsilon} + \log N, \\
\log B &\lesssim \log\frac{R}{\epsilon}, \qquad \log S \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha}\log\frac{R}{\epsilon} + \log N.
\end{aligned}$$

This concludes the proof of Lemma 2. □

B.3 Proof of Theorem 1

Proof. With $\tilde{\Psi}_0 = J_{\mathcal{Y}} \circ \Psi_0 \circ J_{\mathcal{X}}^{-1}$, we have

$$\sup_{x\in\Omega_R^\alpha}\big\|J_{\mathcal{Y}}^{-1} \circ \varphi \circ J_{\mathcal{X}}(x) - \Psi_0(x)\big\|_{\mathcal{Y}} = \sup_{\mathbf{a}\in\mathcal{B}_R^{2,\alpha}}\big\|\varphi(\mathbf{a}) - \tilde{\Psi}_0(\mathbf{a})\big\|_2.$$

That said, it suffices to approximate $\tilde{\Psi}_0$ with a Transformer $\bar{\varphi}$ on $\mathcal{B}_R^{2,\alpha}$. We define $Z^{(l)}$ $(l = 0, 1, 2, 3)$ by

$$Z^{(0)} = \{z_i^{(0)}\}_{i\in\mathbb{Z}} \coloneqq \mathrm{Enc}_P(\mathbf{a}),$$
$$Z^{(1)} = \{z_i^{(1)}\}_{i\in\mathbb{Z}} \coloneqq g_1 \circ \mathrm{Enc}_P(\mathbf{a}),$$
$$Z^{(2)} = \{z_i^{(2)}\}_{i\in\mathbb{Z}} \coloneqq f_1 \circ g_1 \circ \mathrm{Enc}_P(\mathbf{a}),$$
$$Z^{(3)} = \{z_i^{(3)}\}_{i\in\mathbb{Z}} \coloneqq g_2 \circ f_1 \circ g_1 \circ \mathrm{Enc}_P(\mathbf{a}).$$

Our proof consists of five steps.

Step 1. Construct the embedding layer Enc_P.
For any $\epsilon \in (0, 1)$, let

$$N_1 = \left\lceil \left\{ R\left(\frac{48B_0}{\epsilon}\right)^{1/(\beta\wedge 1)} \right\}^{1/\alpha} \right\rceil - 1,$$

$N_2 = \lceil (4B_{0,\gamma}/\epsilon)^{1/\gamma} \rceil - 1$, and $U = N_1 \vee N_2$, Let the embedding dimension be $D = 2U + 4$, and the embedding vector be $\eta = [1, 0, 0, \ldots, 0]^\top \in \mathbb{R}^D$. Define the positional encoding $P \in \mathbb{R}^{D\times\infty}$ by $P_i = [0, 0, \ldots, 0, \cos(i\phi), \sin(i\phi)] \in \mathbb{R}^D$ $(i \in \mathbb{Z})$, where $\phi = (2U + 1)^{-1}$. Then, for any $\mathbf{a} \in \mathbb{R}^\infty$, we have

$$z_i^{(0)} = [\mathrm{Enc}_P(\mathbf{a})]_i = [a_i, 0, 0, \ldots, 0, \cos(i\phi), \sin(i\phi)]^\top, \qquad i \in \mathbb{Z}.$$

Step 2. Construct the self-attention layer g_1.
Let $H_1 = 2N_1 + 1$. For each attention head indexed by $h \in \{-N_1, -N_1 + 1, \ldots, N_1\}$, define the parameter matrices $Q_1^{(h)}, K_1^{(h)} \in \mathbb{R}^{2\times D}$, and $V_1^{(h)} \in \mathbb{R}^{D\times D}$ as

$$Q_1^{(h)} = \chi_1 \begin{bmatrix} 0 & 0 & \cdots & 0 & \cos(h\phi) & -\sin(h\phi) \\ 0 & 0 & \cdots & 0 & \sin(h\phi) & \cos(h\phi) \end{bmatrix},$$
$$K_1^{(h)} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix},$$
$$V_1^{(h)} = E_{N_1+h+2,1},$$

where $\chi_1 > 0$, and $E_{ij} \in \mathbb{R}^{D\times D}$ is the matrix with its (i, j) element being 1 and all others being 0. Then, the key, query, and value vectors for a given head h are

$$Q_1^{(h)} z_i^{(0)} = \chi_1 [\cos((i + h)\phi), \sin((i + h)\phi)]^\top,$$
$$K_1^{(h)} z_i^{(0)} = [\cos(i\phi), \sin(i\phi)]^\top,$$
$$V_1^{(h)} z_i^{(0)} = a_i e_{N_1+h+2}.$$

where $e_i \in \mathbb{R}^D$ is the vector with its i-th element being 1 and all others being 0. Define $\tilde{Z}^{(1)} = \{\tilde{z}_i^{(1)}\}_{i\in\mathbb{Z}}$ by

$$\tilde{z}_i^{(1)} = z_i^{(0)} + \sum_{h=-N_1}^{N_1} V_1^{(h)} Z^{(0)}[i - U : i + U] e_{U+h+1} = z_i^{(0)} + \sum_{h=-N_1}^{N_1} V_1^{(h)} z_{i+h}^{(0)}.$$

Intuitively, $\tilde{Z}^{(1)}$ corresponds to the output of attention layer applied to $Z^{(0)}$, where the standard softmax function replaced by a *hardmax* opertation. Then, we have

$$\tilde{z}_i^{(1)} = [a_i, a_{i-N_1}, a_{i-N_1+1}, \ldots, a_{i+N_1}, 0, \ldots, 0, \cos(i\phi), \sin(i\phi)]^\top.$$

In particular, $\tilde{z}_0^{(1)}$ contains important features $(a_{-N_1}, a_{-N_1+1}, \ldots, a_{N_1})$. Let $s_1^{(h)} = \big(K_1^{(h)} Z^{(0)}[-U : U]\big)^\top Q_1^{(h)} z_0^{(0)}$. Then, for $j \in [-U, U]$,

$$\big[s_1^{(h)}\big]_j = \chi_1 \{\cos(j\phi)\cos(h\phi) + \sin(j\phi)\sin(h\phi)\} = \chi_1 \cos((h - j)\phi),$$

which implies $\big[s_1^{(h)}\big]_h \geq \big[s_1^{(h)}\big]_j + \chi_1/(19U^2)$ for all $j \neq h$. According to Lemma 25,

$$\begin{aligned}\big\|z_0^{(1)} - \tilde{z}_0^{(1)}\big\|_\infty &= \bigg\| \sum_{h=-N_1}^{N_1} V_1^{(h)} Z^{(0)}[-U:U]\Big\{\mathrm{Softmax}\big(s_1^{(h)}\big) - e_{U+h+1}\Big\}\bigg\|_\infty \\ &\overset{\text{(i)}}{\leq} \max_{-U\leq i\leq U} |a_i| \sum_{h=-N_1}^{N_1} \Big\|\mathrm{Softmax}\big(s_1^{(h)}\big) - e_{U+h+1}\Big\|_1 \\ &\leq 2R(2U+1)(2N_1+1)e^{-\chi_1/(19U^2)},\end{aligned}$$

where we have used $V_1^{(h)} z_i^{(0)} = a_i e_{N_1+h+2}$ in (i).

Step 3. Construct the FNN layer f_1.
According to Lemma 2, there exists an FNN $\phi_1 : \mathbb{R}^{2N_1+1} \to \mathbb{R}^{2N_2+1}$ in $\mathcal{F}(L_\phi, W_\phi, S_\phi, B_\phi)$ such that for all $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$,

$$\big\|\phi_1(\mathbf{a}[-N_1:N_1]) - \tilde{\Psi}_0(\mathbf{a})[-N_2:N_2]\big\|_2 \leq \frac{\epsilon}{4}.$$

Since $N_2 \asymp \epsilon^{-1/\gamma}$, we have

$$L_\phi \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha}\left(\log\frac{R}{\epsilon}\right)^2, \qquad \log W_\phi \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha}\log\frac{R}{\epsilon},$$
$$\log B_\phi \lesssim \log\frac{R}{\epsilon}, \qquad \log S_\phi \lesssim \left(\frac{R}{\epsilon^{1/\beta}}\right)^{1/\alpha}\log\frac{R}{\epsilon}.$$

Define one-dimensional and $(2N_2+1)$-dimensional clipping functions

$$h_{B_{0,\gamma}}(y) = \sigma(y + B_{0,\gamma}) - \sigma(y - B_{0,\gamma}) - B_{0,\gamma}, \qquad y \in \mathbb{R},$$
$$\mathcal{C}_{B_{0,\gamma}}(\mathbf{y}) = [h_{B_{0,\gamma}}(y_{-N_2}), h_{B_{0,\gamma}}(y_{-N_2+1}), \ldots, h_{B_{0,\gamma}}(y_{N_2})], \qquad \mathbf{y} \in \mathbb{R}^{2N_2+1}.$$

Then, $h_{B_{0,\gamma}}(y) = \{-B_{0,\gamma} \vee (y \wedge B_{0,\gamma})\}$, and $\mathcal{C}_{B_{0,\gamma}}$ can be implemented by an FNN with depth 1, width $2(2N_2+1)$, sparsity $\mathcal{O}(N_2)$, and weight-norm bound $\max\{1, B_{0,\gamma}\}$. Since $\sup_{x\in\mathcal{X}} \|\Psi_0(x)\|_{\mathcal{Y}^\gamma} \leq B_{0,\gamma}$, we have $\tilde{\Psi}_0(\mathbf{a}) \in \mathcal{B}_{B_{0,\gamma}}^{2,\gamma}$ and $\|\tilde{\Psi}_0(\mathbf{a})[-N_2:N_2]\|_\infty \leq B_{0,\gamma}$. Therefore, for all $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$,

$$\big\|\mathcal{C}_{B_{0,\gamma}} \circ \phi_1(\mathbf{a}[-N_1:N_1]) - \tilde{\Psi}_0(\mathbf{a})[-N_2:N_2]\big\|_2 \leq \big\|\phi_1(\mathbf{a}[-N_1:N_1]) - \tilde{\Psi}_0(\mathbf{a})[-N_2:N_2]\big\|_2 \leq \frac{\epsilon}{4}.$$

Define

$$\phi_2(\rho) = \sigma\bigg(1 - \frac{1-\rho}{1-\cos(\phi)}\bigg), \qquad \rho \in \mathbb{R}.$$

Then, $\phi_2(\cos(0\phi)) = 1$, and $\phi_2(\cos(i\phi)) = 0$ for all $i \in [-2U, 2U] \setminus \{0\}$. In other words, $\phi_2(\cos(i\phi)) = \mathbb{1}_{\{0\}}(i)$ for $i \in [-U, U]$. Moreover, ϕ_2 can be implemented by an FNN with depth 1, width 1, sparsity $\mathcal{O}(1)$, and weight-norm bound $\{1-\cos(\phi)\}^{-1} = \mathcal{O}(U^2)$. Further define

$$h_\times(y, \theta) = \sigma(y + B_{0,\gamma}\{\theta - 1\}) - \sigma(-y + B_{0,\gamma}\{\theta - 1\}), \qquad y, \theta \in \mathbb{R},$$
$$\phi_3(\mathbf{y}, \theta) = [h_\times(y_{-N_2}, \theta), \ldots, h_\times(y_{N_2}, \theta)], \qquad \mathbf{y} \in \mathbb{R}^{2N_2+1}, \theta \in \mathbb{R}.$$

Then, for $y \in [-B_{0,\gamma}, B_{0,\gamma}]$ and $\theta \in \{0, 1\}$, $h_\times(y, \theta) = y\theta$. Moreover, ϕ_3 can be implemented by an FNN with depth 1, width $2(2N_2+1)$, sparsity $\mathcal{O}(N_2)$, and weight-norm bound $\max\{1, B_{0,\gamma}\}$.

With ϕ_1, $\mathcal{C}_{B_{0,\gamma}}$, ϕ_2, and ϕ_3 at hand, we define the FNN layer $f_1 : \mathbb{R}^\infty \to \mathbb{R}^\infty$ by

$$f_1(z) = \big[0, \phi_3\big(\mathcal{C}_{B_{0,\gamma}} \circ \phi_1(z_{2:2N_1+2}), \phi_2(z_{D-1})\big), 0, \ldots, 0, z_{D-1}, z_D\big]^\top, \qquad z \in \mathbb{R}^D.$$

For $Z^{(1)} \in \mathbb{R}^{D\times\infty}$, $f_1\big(Z^{(1)}\big) := \big(f_1\big(z_i^{(1)}\big)\big)_{i\in\mathbb{Z}}$. By definition of g_1, $\big[z_i^{(1)}\big]_{D-1} = \cos(i\phi)$ and $\big[z_i^{(1)}\big]_D = \sin(i\phi)$. Therefore, for all $i \in [-2U, 2U] \setminus \{0\}$, $\phi_2\big(\big[z_i^{(1)}\big]_{D-1}\big) = 0$ and $f_1\big(z_i^{(1)}\big) = [0, \ldots, 0, \cos(i\phi), \sin(i\phi)]^\top$. For $i = 0$,

$$f_1\big(z_0^{(1)}\big) = \Big[0, \mathcal{C}_{B_{0,\gamma}} \circ \phi_1\big(\big[z_0^{(1)}\big]_{2:2N_1+2}\big), 0, \ldots, \cos(0\phi), \sin(0\phi)\Big]^\top.$$

Since the ReLU activation function σ and clipping function $\mathcal{C}_{B_{0,\gamma}}$ are 1-Lipschitz continuous, and $\|A_i\xi + b_i - (A_i\xi' + b_i)\|_\infty \le B_\phi W_\phi \|\xi - \xi'\|_\infty$ ($0 \le i \le L_\phi$, $\xi, \xi' \in \mathbb{R}^{d_i}$), we have

$$\left\|\mathcal{C}_{B_{0,\gamma}} \circ \phi_1(\xi) - \mathcal{C}_{B_{0,\gamma}} \circ \phi_1(\xi')\right\|_\infty \le (B_\phi W_\phi)^{L_\phi+1} \|\xi - \xi'\|_\infty, \qquad \xi, \xi' \in \mathbb{R}^{2N_1+1}.$$

Therefore,

$$\begin{aligned}
&\left\|\left[f_1\big(z_0^{(1)}\big)\right]_{2:2N_2+2} - \tilde{\Psi}_0(\mathbf{a})[-N_2 : N_2]\right\|_2 \\
&\le \left\|\left[f_1\big(z_0^{(1)}\big)\right]_{2:2N_2+2} - \mathcal{C}_{B_{0,\gamma}} \circ \phi_1(\mathbf{a}[-N_1 : N_1])\right\|_2 \\
&\quad + \left\|\mathcal{C}_{B_{0,\gamma}} \circ \phi_1(\mathbf{a}[-N_1 : N_1]) - \tilde{\Psi}_0(\mathbf{a})[-N_2 : N_2]\right\|_2 \\
&\le \left\|\left[\mathcal{C}_{B_{0,\gamma}} \circ \phi_1\big(\big[z_0^{(1)}\big]_{2:2N_1+2}\big) - \mathcal{C}_{B_{0,\gamma}} \circ \phi_1\big(\big[\tilde{z}_0^{(1)}\big]_{2:2N_1+2}\big)\right\|_2 + \frac{\epsilon}{4} \\
&\le \sqrt{2N_2+1}(B_\phi W_\phi)^{L_\phi+1} \big\|z_0^{(1)} - \tilde{z}_0^{(1)}\big\|_\infty + \frac{\epsilon}{4}.
\end{aligned}$$

Take $\chi_1 = 19U^2 \log\big(8R(2U+1)^{5/2}(B_\phi W_\phi)^{L_\phi+1}/\epsilon\big)$, we have

$$\left\|\left[f_1\big(z_0^{(1)}\big)\right]_{1:2N_2+1} - \tilde{\Psi}_0(\mathbf{a})[-N_2 : N_2]\right\|_2 \le \frac{\epsilon}{2}.$$

In addition, note that

$$U \asymp \left(\frac{R}{\epsilon^{1/(\beta\wedge 1)}}\right)^{1/\alpha} \vee \epsilon^{-1/\gamma},$$

we have that f_1 is an FNN with depth L_1, width W_1, sparisty S_1, and weight-norm bound B_1 satisfying

$$L_1 \lesssim L_\phi, \qquad \log W_1 \lesssim \log W_\phi, \qquad \log S_1 \lesssim \log S_\phi, \qquad \log B_1 \lesssim \log B_\phi.$$

Step 4. Construct the self-attention layer g_2.
Let $H_2 = 2N_2 + 1$. For each attention head indexed by $h \in \{-N_2, -N_2+1, \ldots, N_2\}$, define the parameter matrices $Q_2^{(h)}, K_2^{(h)} \in \mathbb{R}^{2\times D}$, and $V_2^{(h)} \in \mathbb{R}^{D\times D}$ as

$$\begin{aligned}
Q_2^{(h)} &= \chi_2 \begin{bmatrix} 0 & 0 & \cdots & 0 & \cos(h\phi) & \sin(h\phi) \\ 0 & 0 & \cdots & 0 & -\sin(h\phi) & \cos(h\phi) \end{bmatrix}, \\
K_2^{(h)} &= K_1^{(h)}, \\
V_2^{(h)} &= E_{1,N_2+h+2},
\end{aligned}$$

where $\chi_2 > 0$. Then, the key, query, and value vectors for a given head h are

$$\begin{aligned}
Q_2^{(h)} z_i^{(2)} &= \chi_2[\cos((i-h)\phi), \sin((i-h)\phi)]^\top, \\
K_2^{(h)} z_i^{(2)} &= [\cos(i\phi), \sin(i\phi)]^\top, \\
V_2^{(h)} z_i^{(2)} &= \big[z_i^{(2)}\big]_{N_2+h+2} e_1.
\end{aligned}$$

Define $\tilde{Z}^{(3)} = \big\{\tilde{z}_i^{(3)}\big\}_{i\in\mathbb{Z}}$ by

$$\tilde{z}_i^{(3)} = z_i^{(2)} + \sum_{h=-N_2}^{N_2} V_2^{(h)} Z^{(2)}[i-U : i+U] e_{U-h+1} = z_i^{(2)} + \sum_{h=-N_2}^{N_2} V_2^{(h)} z_{i-h}^{(2)}.$$

Then, for $i \in [-N_2, N_2]$, $\big[z_i^{(3)}\big]_1 = \big[z_0^{(2)}\big]_{N_2+i+2}$. That is, the first row of $\tilde{Z}^{(3)}[-N_2, N_2]$ is $\big[f_1\big(z_0^{(1)}\big)\big]_{2:2N_2+2}$. Let $s_{2,i}^{(h)} = \big(K_2^{(h)} Z^{(2)}[i-U : i+U]\big)^\top Q_2^{(h)} z_i^{(0)}$. Then, for $j \in [i-U, i+U]$,

$$\big[s_{2,i}^{(h)}\big]_j = \chi_2\{\cos(j\phi)\cos((i-h)\phi) + \sin(j\phi)\sin((i-h)\phi)\} = \chi_2 \cos((i-h-j)\phi),$$

which implies $\big[s_{2,i}^{(h)}\big]_{i-h} \geq \big[s_{2,i}^{(h)}\big]_j + \chi_2/(19U^2)$ for all $j \neq i-h$. According to Lemma 25,

$$\begin{aligned}
\big\|z_i^{(3)} - \tilde{z}_i^{(3)}\big\|_\infty &= \Big\| \sum_{h=-N_2}^{N_2} V_2^{(h)} Z^{(2)}[i-U:i+U]\Big\{\mathrm{Softmax}\big(s_{2,i}^{(h)}\big) - e_{U-h+1}\Big\}\Big\|_\infty \\
&\leq \max_{-N_2\leq h\leq N_2} \big|\big[z_0^{(2)}\big]_{N_2+h+2}\big| \sum_{h=-N_2}^{N_2} \Big\|\mathrm{Softmax}\big(s_{2,i}^{(h)}\big) - e_{U-h+1}\Big\|_1 \\
&\leq 2B_{0,\gamma}(2U+1)(2N_2+1)e^{-\chi_2/(19U^2)}.
\end{aligned}$$

Step 5. Construct the FNN layer f_2 and define clipping function $\mathrm{Clip}_{N,\Gamma}$.
Lastly, let $f_2 : \mathbb{R}^D \to \mathbb{R}$ be the function defined by

$$f_2(z) = z_1, \qquad z \in \mathbb{R}^D.$$

Then, f_2 can be implemented by an FNN with depth 1, width 2, sparsity 4, and weight-norm bound 1. For $Z^{(3)} \in \mathbb{R}^{D\times\infty}$, $f_2\big(Z^{(3)}\big) \coloneqq \big(f_2\big(z_i^{(3)}\big)\big)_{i\in\mathbb{Z}}$. Furthermore, we set $N = N_2$ and $\Gamma \geq B_{0,\gamma}$ in the clipping function $\mathrm{Clip}_{N,\Gamma}$. We define the Transformer $\bar{\varphi} : \mathbb{R}^\infty \to \mathbb{R}^\infty$ by

$$\bar{\varphi} = \mathrm{Clip}_{N,\Gamma} \circ f_2 \circ g_2 \circ f_1 \circ g_1 \circ \mathrm{Enc}_P.$$

Let $T_N : \mathbb{R}^\infty \to \mathbb{R}^\infty$ be the truncation operator that keeps entries with indices $|i| \leq N$ unchanged and sets all other entries to zero. That is, $T_N\mathbf{y} = \big(y_i \mathbb{1}_{\{|i|\leq N\}}\big)_{i\in\mathbb{Z}}$. For any $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$, the approximation error can be bounded as

$$\begin{aligned}
&\big\|\bar{\varphi}(\mathbf{a}) - \tilde{\Psi}_0(\mathbf{a})\big\|_2 \leq \big\|\bar{\varphi}(\mathbf{a}) - T_N \circ \tilde{\Psi}_0(\mathbf{a})\big\|_2 + \big\|T_N \circ \tilde{\Psi}_0(\mathbf{a}) - \tilde{\Psi}_0(\mathbf{a})\big\|_2 \\
&\overset{(ii)}{\leq} \big\|\big\{f_2\big(Z^{(3)}\big) - \tilde{\Psi}_0(\mathbf{a})\big\}[-N_2:N_2]\big\|_2 + \frac{\epsilon}{4} \\
&\leq \big\|\big\{f_2\big(Z^{(3)}\big) - f_2\big(\tilde{Z}^{(3)}\big)\big\}[-N_2:N_2]\big\|_2 + \big\|\big\{f_2\big(\tilde{Z}^{(3)}\big) - \tilde{\Psi}_0(\mathbf{a})\big\}[-N_2:N_2]\big\|_2 + \frac{\epsilon}{4} \\
&\leq \sqrt{2N_2+1} \max_{-N_2\leq i\leq N_2} \big\|z_i^{(3)} - \tilde{z}_i^{(3)}\big\|_\infty + \big\|\big[f_1\big(z_0^{(1)}\big)\big]_{2:2N_2+2} - \tilde{\Psi}_0(\mathbf{a})[-N_2:N_2]\big\|_2 + \frac{\epsilon}{4} \\
&\leq 2B_{0,\gamma}(2U+1)(2N_2+1)^{3/2}e^{-\chi_2/(19U^2)} + \frac{3\epsilon}{4},
\end{aligned}$$

where we have used $\sup_{x\in\mathcal{X}} \|\Psi_0(x)\|_{\mathcal{Y}^\gamma} \leq B_{0,\gamma}$ in (ii). Take $\chi_2 = 19U^2\log(8B_{0,\gamma}(2U+1)^{5/2}/\epsilon)$, we have for $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$,

$$\big\|\bar{\varphi}(\mathbf{a}) - \tilde{\Psi}_0(\mathbf{a})\big\|_2 \leq \epsilon.$$

It remains to bound the architecture parameters of the Transformer network. By definition of N_1 and N_2, we have

$$N \lesssim D \asymp U \asymp H \asymp \max\left\{\left(\frac{R}{\epsilon^{1/(\beta\wedge 1)}}\right)^{1/\alpha}, \epsilon^{-1/\gamma}\right\}.$$

Note that

$$\chi_1 \lesssim U^2\left(\frac{R}{\epsilon^{1/\beta}}\right)^{2/\alpha}\left(\log\frac{R}{\epsilon}\right)^3,$$

we have $\log\ \chi_1 \lesssim \log(R/\epsilon)$. Moreover, $\log\chi_2 \lesssim \log U \lesssim \log(R/\epsilon)$. Consequently, $\log B \lesssim \log(R/\epsilon)$. This concludes the proof of Theorem 1. □

B.4 Overall approximation error bound

Corollary 1 (Overall approximation error bound)**.** *Suppose that $\nu \in \mathcal{P}_\alpha(\kappa)$ and $\bar{\varphi}$ is the Transformer in Theorem 1. Then,*

$$\|J_{\mathcal{Y}}^{-1} \circ \bar{\varphi} \circ J_{\mathcal{X}} - \Psi_0\|_{L^2(\nu)}^2 \leq \epsilon^2 + 2(\sqrt{2N+1}\Gamma + B_0)^2\exp(-R/\kappa).$$

Proof. By Theorem 1 and the definition of $\mathcal{T}_{N,\Gamma}(M, D, U, H, L, W, S, B)$, we have $\bar{\varphi} = \mathrm{Clip}_{N,\Gamma} \circ \varphi$ for some $\varphi \in \mathcal{T}(M, D, U, H, L, W, S, B)$, where N and $\Gamma \geq B_{0,\gamma}$ are as in Theorem 1. Write $\bar{\Psi} = J_{\mathcal{Y}}^{-1} \circ \bar{\varphi} \circ J_{\mathcal{X}}$. The map $x \mapsto \|x\|_{\mathcal{X}^\alpha} \in [0, \infty]$ is the supremum over $K \in \mathbb{N}$ of the continuous maps $x \mapsto (\sum_{|i| \leq K} (1 + |i|)^{2\alpha} \langle x, \phi_i \rangle_{\mathcal{X}}^2)^{1/2}$, so it is Borel and Ω_R^α is a Borel set. The operator $\bar{\Psi}$ is continuous. Indeed, the embedding and the feedforward layers act position-wise, and at position i each of the M attention layers reads only the positions $i-U, \ldots, i+U$, in its values and in its softmax weights. Hence output coordinate k of $\bar{\varphi}(\mathbf{a})$ is obtained from the finitely many input coordinates a_j with $|j - k| \leq MU$ by finitely many additions, multiplications, applications of ReLU and softmax, and clipping. Each map $x \mapsto \langle x, \phi_j \rangle_{\mathcal{X}}$ is continuous, and only the $2N+1$ coordinates with $|k| \leq N$ are nonzero, so $x \mapsto \bar{\varphi}(J_{\mathcal{X}} x)$ is continuous from $\mathcal{X}$ to $\ell^2(\mathbb{Z})$, and $J_{\mathcal{Y}}^{-1}$ is an isometry. The operator Ψ_0 is continuous, being Hölder continuous if $\beta \leq 1$ and Fréchet differentiable if $\beta > 1$. Hence $x \mapsto \|\bar{\Psi}(x) - \Psi_0(x)\|_{\mathcal{Y}}^2$ is Borel.

By Theorem 1, $\|\bar{\Psi}(x) - \Psi_0(x)\|_{\mathcal{Y}} \leq \epsilon$ for $x \in \Omega_R^\alpha$. For every $x \in \mathcal{X}$, the sequence $\bar{\varphi}(J_{\mathcal{X}} x) = \mathrm{Clip}_{N,\Gamma}(\varphi(J_{\mathcal{X}} x))$ has at most $2N + 1$ nonzero coordinates, each bounded by Γ in absolute value, so $\|\bar{\Psi}(x)\|_{\mathcal{Y}} = \|\bar{\varphi}(J_{\mathcal{X}} x)\|_2 \leq \sqrt{2N+1}\,\Gamma$. Moreover, $\|\Psi_0(x)\|_{\mathcal{Y}} \leq B_0$ by the case $k = 0$ in the definition of $\mathcal{H}^\beta(\mathcal{X}, \mathcal{Y}; B_0)$. Therefore, $\|\bar{\Psi}(x) - \Psi_0(x)\|_{\mathcal{Y}} \leq \sqrt{2N+1}\,\Gamma + B_0$ for all $x \in \mathcal{X}$.

Let $Z = \|x\|_{\mathcal{X}^\alpha}$ with $x \sim \nu$. Since $\|x\|_{\psi_1, \mathcal{X}^\alpha} \leq \kappa$, every $\lambda > \kappa$ exceeds some λ' with $\mathbb{E} e^{Z/\lambda'} \leq 2$, and then $\mathbb{E} e^{Z/\lambda} \leq \mathbb{E} e^{Z/\lambda'} \leq 2$ because $Z \geq 0$. Letting $\lambda \downarrow \kappa$, monotone convergence gives $\mathbb{E} e^{Z/\kappa} \leq 2$, and Markov's inequality gives $\nu(\mathcal{X} \setminus \Omega_R^\alpha) = \nu(Z > R) \leq e^{-R/\kappa}\, \mathbb{E} e^{Z/\kappa} \leq 2e^{-R/\kappa}$.

Splitting the integral as in Lemma 1, we obtain

$$\begin{aligned} \|\bar{\Psi} - \Psi_0\|_{L^2(\nu)}^2 &\leq \epsilon^2 \nu(\Omega_R^\alpha) + (\sqrt{2N+1}\,\Gamma + B_0)^2\, \nu(\mathcal{X} \setminus \Omega_R^\alpha) \\ &\leq \epsilon^2 + 2(\sqrt{2N+1}\,\Gamma + B_0)^2 \exp\left(- \frac{R}{\kappa} \right). \end{aligned}$$

This concludes the proof of Corollary 1. □

C Generalization error analysis

This section proves Theorem 2 and 4.. Appendix C.1 bounds the stochastic error of least-squares estimators under sub-exponential noise and the covering number of clipped Transformer classes. Appendix C.2 bounds the output coefficients beyond the cutoff and the approximation term in $L^2(\nu)$, including the contribution of inputs outside the Sobolev ball; Appendix C.3 constructs a single class that contains an approximating network for every target operator and bounds its covering number; Appendix C.4 calibrates the accuracy and the radius against the sample size; and Appendix C.5 combines these results. Appendix C.6 proves Theorem 4 with the same results.

C.1 Stochastic error under sub-exponential noise

This subsection bounds the stochastic error of least-squares estimators over bounded classes when the noise satisfies Definition 5. We follow the proof of Theorem 5.2 of Takakura & Suzuki (2023), which extends Lemma 4 of Schmidt-Hieber (2020) to several outputs. That proof bounds the expected empirical error of the estimator and, separately, the gap between the empirical and the population error. The gap does not involve the noise, and we cite their bound for it. The empirical error is where they use Gaussian noise; Lemma 4 bounds it under a Bernstein condition, which Lemma 3 derives from Definition 5.

Notations. For $N \in \mathbb{N}_0$ let $p \coloneqq 2N + 1$, let $V_p \coloneqq \mathrm{span}\{e_k : |k| \leq N\} \subset \ell^2(\mathbb{Z})$, let Π_p be the orthogonal projection of $\ell^2(\mathbb{Z})$ onto V_p, and let $\iota_p : V_p \to \mathbb{R}^p$ list the coordinates with indices $-N, \ldots, N$. For a measurable space A and a map $g : A \to \mathbb{R}^p$, let $\|g\|_\infty \coloneqq \sup_{\mathbf{c} \in A} \max_k |g(\mathbf{c})_k|$. For a class $\mathcal{G}$ of such maps and $\delta > 0$, a δ-cover of size m is a family $g_1, \ldots, g_m \in \mathcal{G}$ such that every $g \in \mathcal{G}$ satisfies $\|g - g_k\|_\infty \leq \delta$ for some k; repetitions are allowed. The smallest such m is the covering number $\mathcal{N}(\delta, \mathcal{G})$, as in Definition 5.1 of Takakura & Suzuki (2023). For inputs $\mathbf{c}_1, \ldots, \mathbf{c}_n$ with law μ, we write $\|g\|_n^2 \coloneqq n^{-1} \sum_i \|g(\mathbf{c}_i)\|_2^2$ and $\|g\|_{L^2(\mu)}^2 \coloneqq \int \|g\|_2^2\, d\mu$.

Lemma 3 (Sub-exponential noise)**.** *Let $\mathcal{H}$ be $\mathcal{Y}$ or $\mathbb{R}$, and let ε be a random element of $\mathcal{H}$ whose law belongs to $\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$ (Definition 5). Then $\mathbb{E}\|\varepsilon\|^k \le k!\,\kappa_\varepsilon^k$ for every $k \ge 1$, and for every $w \in \mathcal{H}$ with $\|w\| \le 1$ and every $|\lambda| \le 1/(2\kappa_\varepsilon)$,*

$$\mathbb{E}\exp(\lambda\langle\varepsilon, w\rangle) \le \exp(2\lambda^2\kappa_\varepsilon^2).$$

Proof. Since $\|\varepsilon\|_{\psi_1} \le \kappa_\varepsilon$, monotone convergence gives $\mathbb{E}e^{\|\varepsilon\|/\kappa_\varepsilon} \le 2$. As $e^s - 1 \ge s^k/k!$ for $s \ge 0$, we get $\mathbb{E}\|\varepsilon\|^k \le k!\,\kappa_\varepsilon^k\,\mathbb{E}(e^{\|\varepsilon\|/\kappa_\varepsilon} - 1) \le k!\,\kappa_\varepsilon^k$. Let $X = \langle\varepsilon, w\rangle$. Then $|X| \le \|\varepsilon\|$, and $\mathbb{E}X = \langle\mathbb{E}\varepsilon, w\rangle = 0$. For $|\lambda| \le 1/(2\kappa_\varepsilon)$ the series $\sum_k |\lambda|^k\mathbb{E}|X|^k/k!$ converges, so

$$\mathbb{E}e^{\lambda X} = 1 + \sum_{k\ge2}\frac{\lambda^k\mathbb{E}X^k}{k!} \le 1 + \sum_{k\ge2}(|\lambda|\kappa_\varepsilon)^k = 1 + \frac{(\lambda\kappa_\varepsilon)^2}{1 - |\lambda|\kappa_\varepsilon} \le 1 + 2\lambda^2\kappa_\varepsilon^2 \le e^{2\lambda^2\kappa_\varepsilon^2}.$$

□

In the fully observed model, the observed noise vector $\iota_p\Pi_p J_{\mathcal{Y}}\varepsilon$ satisfies $\langle\iota_p\Pi_p J_{\mathcal{Y}}\varepsilon, \mathbf{w}\rangle = \langle\varepsilon, J_{\mathcal{Y}}^{-1}\iota_p^{-1}\mathbf{w}\rangle_{\mathcal{Y}}$ for $\mathbf{w} \in \mathbb{R}^p$. By Lemma 3, it therefore satisfies the condition (3) below with $\sigma = b = 2\kappa_\varepsilon$, independently of the input.

Lemma 4 (Noise term)**.** *Let $(\mathbf{c}_i, \boldsymbol{\zeta}_i)$, $i = 1, \dots, n$, be i.i.d. random pairs, where $\mathbf{c}_i$ takes values in a measurable space A, $\mathbf{c}_1 \sim \mu$ and $\boldsymbol{\zeta}_i \in \mathbb{R}^p$. Assume that there are $\sigma > 0$ and $b \ge 0$ such that, almost surely,*

$$\mathbb{E}\big[\exp(\lambda\langle\boldsymbol{\zeta}_1, \mathbf{w}\rangle)\,\big|\,\mathbf{c}_1\big] \le \exp(\lambda^2\sigma^2/2) \quad \text{for all unit } \mathbf{w} \in \mathbb{R}^p \text{ and all } |\lambda| \le 1/b, \tag{3}$$

where λ ranges over $\mathbb{R}$ if $b = 0$. Let $\mathcal{G}$ be a nonempty class of measurable maps $A \to \mathbb{R}^p$ and let $g_0 : A \to \mathbb{R}^p$ be measurable, with $\|g\|_\infty \le R_0$ for all $g \in \mathcal{G} \cup \{g_0\}$. Let $\hat{g}$ be a least-squares estimator: for every sample, $\hat{g} \in \mathcal{G}$ minimizes $g \mapsto \sum_i \|g(\mathbf{c}_i) - g_0(\mathbf{c}_i) - \boldsymbol{\zeta}_i\|_2^2$ over $\mathcal{G}$, and $\hat{g}(\mathbf{c})$ is measurable in $\mathbf{c}$ and the sample. If $\mathcal{G}$ has a δ-cover of size m, then

$$\mathbb{E}\|\hat{g} - g_0\|_n^2 \le 2\inf_{g\in\mathcal{G}}\|g - g_0\|_{L^2(\mu)}^2 + \frac{(12\sigma^2 + 16b\sqrt{p}R_0)(\log m + 1)}{n} + 2p\delta^2 + 8p\delta(\sigma + 2b).$$

Proof. Step 1: conditional tail bounds. Let Q be a regular conditional distribution of $\boldsymbol{\zeta}_1$ given $\mathbf{c}_1$. Applying (3) to a countable dense set of pairs $(\mathbf{w}, \lambda)$ and passing to limits with Fatou's lemma, we find a μ-null set outside of which $\int \exp(\lambda\langle\mathbf{z}, \mathbf{w}\rangle)\,Q(\mathbf{c}, d\mathbf{z}) \le \exp(\lambda^2\sigma^2/2)$ for all $(\mathbf{w}, \lambda)$ allowed in (3). Because the pairs are i.i.d., given $C \coloneqq (\mathbf{c}_1, \dots, \mathbf{c}_n)$ the vectors $\boldsymbol{\zeta}_1, \dots, \boldsymbol{\zeta}_n$ are independent with laws $Q(\mathbf{c}_i, \cdot)$. Fix C outside the null set and vectors $\mathbf{u}_1, \dots, \mathbf{u}_n \in \mathbb{R}^p$ that are functions of C. Put $S = \sum_i\langle\boldsymbol{\zeta}_i, \mathbf{u}_i\rangle$, $s^2 = \sigma^2\sum_i\|\mathbf{u}_i\|_2^2$ and $\tau = b\max_i\|\mathbf{u}_i\|_2$. If $s = 0$, then $S = 0$. Otherwise, for $0 < \lambda \le 1/\tau$ (any $\lambda > 0$ if $\tau = 0$), conditional independence and (3), applied to the unit vectors $\mathbf{u}_i/\|\mathbf{u}_i\|_2$ with parameters $\lambda\|\mathbf{u}_i\|_2 \le 1/b$ (the factors with $\mathbf{u}_i = 0$ equal 1), give $\mathbb{E}[e^{\lambda S} \mid C] \le e^{\lambda^2 s^2/2}$. For $t > 0$, Markov's inequality gives $\mathbb{P}(S \ge t \mid C) \le e^{-\lambda t + \lambda^2 s^2/2}$. If $t\tau \le s^2$, the choice $\lambda = t/s^2$ gives the bound $e^{-t^2/(2s^2)}$. Otherwise $\lambda = 1/\tau$ and $s^2 < t\tau$ give the bound $e^{-t/(2\tau)}$. With $t = s\sqrt{2x} + 2\tau x$ both exponents are at most $-x$, so for every $x \ge 0$,

$$\mathbb{P}\big(S > s\sqrt{2x} + 2\tau x\,\big|\,C\big) \le e^{-x}. \tag{4}$$

The same two bounds with $n = 1$ and $\mathbf{u}_1 = \pm e_k$ give $\mathbb{P}(|\langle\boldsymbol{\zeta}_1, e_k\rangle| > t \mid \mathbf{c}_1) \le 2e^{-t^2/(2\sigma^2)} + 2e^{-t/(2b)}$ (the last term is absent if $b = 0$). Integrating, $\mathbb{E}[\langle\boldsymbol{\zeta}_1, e_k\rangle^2 \mid \mathbf{c}_1] \le 4\sigma^2 + 16b^2$ for each k, and hence

$$\mathbb{E}\|\boldsymbol{\zeta}_1\|_2 \le \big\{p(4\sigma^2 + 16b^2)\big\}^{1/2} \le 2\sqrt{p}(\sigma + 2b). \tag{5}$$

Finally, $\mathbb{E}[\boldsymbol{\zeta}_1 \mid \mathbf{c}_1] = 0$. Indeed, Jensen's inequality and (3) give $\lambda\mathbb{E}[\langle\boldsymbol{\zeta}_1, \mathbf{w}\rangle \mid \mathbf{c}_1] \le \lambda^2\sigma^2/2$ for all small $|\lambda|$; dividing by λ and letting $\lambda \to 0$ from both sides gives $\mathbb{E}[\langle\boldsymbol{\zeta}_1, \mathbf{w}\rangle \mid \mathbf{c}_1] = 0$.

Step 2: the good event. Let $g_1, \dots, g_m$ be a δ-cover of $\mathcal{G}$, fix $x \ge 0$, and put $x' = x + \log m$. For each k, apply (4) with $\mathbf{u}_i = g_k(\mathbf{c}_i) - g_0(\mathbf{c}_i)$. Then $\|\mathbf{u}_i\|_2 \le 2\sqrt{p}R_0$, so $\tau \le 2b\sqrt{p}R_0$, and $s = \sigma\sqrt{n}\,\|g_k - g_0\|_n$. Integrating over C and taking a union bound over k, the event

$$E_x \coloneqq \Big\{\frac{1}{n}\sum_i\langle\boldsymbol{\zeta}_i, g_k(\mathbf{c}_i) - g_0(\mathbf{c}_i)\rangle \le \sigma\|g_k - g_0\|_n\sqrt{\frac{2x'}{n}} + \frac{4b\sqrt{p}R_0x'}{n} \text{ for all } k \le m\Big\}$$

has probability at least $1 - me^{-x'} = 1 - e^{-x}$.

Step 3: the empirical error on E_x. Fix $g^* \in \mathcal{G}$, and put $\theta = \sigma\sqrt{2x'/n}$ and $Q^* = (2/n)\sum_i \langle \boldsymbol{\zeta}_i, g^*(\mathbf{c}_i) - g_0(\mathbf{c}_i)\rangle$. Since $\hat{g}$ minimizes the criterion and $g^* \in \mathcal{G}$, expanding the squares gives

$$\|\hat{g} - g_0\|_n^2 \le \|g^* - g_0\|_n^2 + \frac{2}{n}\sum_i \langle \boldsymbol{\zeta}_i, \hat{g}(\mathbf{c}_i) - g_0(\mathbf{c}_i)\rangle - Q^*.$$

On E_x, choose k with $\|\hat{g} - g_k\|_\infty \le \delta$. Then $\|\hat{g}(\mathbf{c}_i) - g_k(\mathbf{c}_i)\|_2 \le \sqrt{p}\delta$ and $\|g_k - g_0\|_n \le \|\hat{g} - g_0\|_n + \sqrt{p}\delta$, so

$$\frac{2}{n}\sum_i \langle \boldsymbol{\zeta}_i, \hat{g}(\mathbf{c}_i) - g_0(\mathbf{c}_i)\rangle \le \frac{2\sqrt{p}\delta}{n}\sum_i \|\boldsymbol{\zeta}_i\|_2 + 2\theta\big(\|\hat{g} - g_0\|_n + \sqrt{p}\delta\big) + \frac{8b\sqrt{p}R_0 x'}{n}.$$

We insert $2\theta\|\hat{g} - g_0\|_n \le \|\hat{g} - g_0\|_n^2/2 + 2\theta^2$ and $2\theta\sqrt{p}\delta \le \theta^2 + p\delta^2$, and rearrange. Since $6\theta^2 = 12\sigma^2 x'/n$, this gives on E_x

$$\|\hat{g} - g_0\|_n^2 \le V + \frac{(12\sigma^2 + 16b\sqrt{p}R_0)\,x}{n},$$

where

$$V := 2\|g^* - g_0\|_n^2 - 2Q^* + \frac{(12\sigma^2 + 16b\sqrt{p}R_0)\log m}{n} + 2p\delta^2 + \frac{4\sqrt{p}\delta}{n}\sum_i \|\boldsymbol{\zeta}_i\|_2.$$

Step 4: integration. Let $W = (\|\hat{g} - g_0\|_n^2 - V)_+$, a measurable random variable. By Step 3, $\{W > (12\sigma^2 + 16b\sqrt{p}R_0)x/n\} \subset E_x^c$, so this event has probability at most e^{-x} for every $x \ge 0$, and $\mathbb{E}W \le (12\sigma^2 + 16b\sqrt{p}R_0)/n$. By Step 1 and the boundedness of $g^* - g_0$, Q^* is integrable with $\mathbb{E}Q^* = 0$. Moreover $\mathbb{E}\|g^* - g_0\|_n^2 = \|g^* - g_0\|_{L^2(\mu)}^2$. With (5),

$$\mathbb{E}\|\hat{g} - g_0\|_n^2 \le \mathbb{E}V + \mathbb{E}W \le 2\|g^* - g_0\|_{L^2(\mu)}^2 + \frac{(12\sigma^2 + 16b\sqrt{p}R_0)(\log m + 1)}{n} + 2p\delta^2 + 8p\delta(\sigma + 2b).$$

Taking the infimum over $g^* \in \mathcal{G}$ proves the lemma. □

Lemma 5 (Oracle inequality)**.** *Under the assumptions of Lemma 4, let $\delta \in (0, 1)$ and let $\mathcal{G}$ have a δ-cover of size $m \ge 3$. Then*

$$\mathbb{E}\|\hat{g} - g_0\|_{L^2(\mu)}^2 \le 4\inf_{g\in\mathcal{G}} \|g - g_0\|_{L^2(\mu)}^2 + \frac{336(pR_0^2 + \sigma^2 + b^2)(\log m + 1)}{n} + 32p(1 + R_0 + \sigma + b)\delta.$$

Proof. Let $\bar{R} = p^{-1}\mathbb{E}\|\hat{g} - g_0\|_{L^2(\mu)}^2$ and $\hat{R} = p^{-1}\mathbb{E}\|\hat{g} - g_0\|_n^2$ be the population and the expected empirical error, averaged over the p output coordinates as in Takakura & Suzuki (2023). In the proof of their Theorem 5.2 (Appendix G, eq. (20)), with p output coordinates, entry bound R_0 and a δ-cover of size m, they show that

$$|\hat{R} - \bar{R}| \le \frac{1}{2}\bar{R} + \frac{4R_0^2}{n}\Big(\frac{37}{9}\log m + 40\Big) + 10R_0\delta.$$

This part of their proof uses only that $\mathbf{c}_1, \dots, \mathbf{c}_n$ are i.i.d., that $\hat{g}$ takes values in $\mathcal{G}$, the entry bound R_0 and $\log m \ge 1$. It uses neither the noise nor the fact that $\hat{g}$ minimizes the criterion, so it applies here. Hence

$$\mathbb{E}\|\hat{g} - g_0\|_{L^2(\mu)}^2 \le 2\,\mathbb{E}\|\hat{g} - g_0\|_n^2 + \frac{8pR_0^2}{n}\Big(\frac{37}{9}\log m + 40\Big) + 20pR_0\delta.$$

We insert Lemma 4 and use $32b\sqrt{p}R_0 \le 16b^2 + 16pR_0^2$, $37/9 \le 40$ and $\delta^2 \le \delta$. □

Lemma 5 has the structure of the oracle inequality of Takakura & Suzuki (2023) for Gaussian noise: an approximation term, a complexity term of order $(\log m)/n$, and a term linear in the resolution δ of the cover. The only property of the noise that enters is the Bernstein condition (3). Gaussian noise satisfies this condition with $b = 0$. A sub-exponential tail ($b > 0$) adds the term $16b\sqrt{p}R_0(\log m + 1)/n$ to the bound of Lemma 4; this term comes from the linear part $2\tau x$ of the tail bound (4). The same tail also enlarges the coefficient of the term linear in δ from $8p\sigma$ to $8p(\sigma + 2b)$. Both changes affect constants, not the form of the bound. Since the condition is imposed conditionally on the input, the lemma also covers noise whose law depends on the input, which is the situation of discretized observations (Section 4.2).

Clipped inputs. For $\rho > 0$ let Π^ρ clip every coordinate of a sequence to $[-\rho, \rho]$, $[\Pi^\rho \mathbf{c}]_j = \max\{-\rho, \min\{c_j, \rho\}\}$, and let

$$\mathcal{T}^\rho_{N,\Gamma}(M, D, U, H, L, W, S, B) \coloneqq \big\{\varphi \circ \Pi^\rho : \varphi \in \mathcal{T}_{N,\Gamma}(M, D, U, H, L, W, S, B)\big\}.$$

Every $\mathbf{c} \in \mathcal{B}^{2,\alpha}_\rho$ has $|c_j| \le \|\mathbf{c}\|_2 \le \rho$, so $\Pi^\rho \mathbf{c} = \mathbf{c}$: the clip does not change a network on the Sobolev sequence ball of radius ρ. We identify the members of these classes with maps into $\mathbb{R}^p$, $p = 2N + 1$, through the retained output coordinates.

Lemma 6 (Covering number)**.** *Let $B \ge 1$ and $\rho \ge 1$, and let the fixed positional encoding satisfy $\|P\|_\infty \le B$. There is a universal constant C_{TS} such that, for every $\delta \in (0, 1]$,*

$$\log \mathcal{N}\big(\delta, \mathcal{T}^\rho_{N,\Gamma}(M, D, U, H, L, W, S, B)\big) \le C_{\mathrm{TS}} M^3 (L + 1)(S + HD^2) \log \frac{2\rho BDH(L + 1)\bar{W}}{\delta},$$

where $\bar{W} \coloneqq \max\{W, D\}$.

Proof. For $\mathbf{c} \in \ell^2(\mathbb{Z})$, let $X(\mathbf{c}) \in [0, 1]^{2\times\infty}$ have the two rows $\rho^{-1} \max\{\Pi^\rho \mathbf{c}, 0\}$ and $\rho^{-1} \max\{-\Pi^\rho \mathbf{c}, 0\}$, taken coordinatewise. Then $\eta\, \Pi^\rho \mathbf{c} + P = EX(\mathbf{c}) + P$ with $E = [\rho\eta, -\rho\eta] \in \mathbb{R}^{D\times 2}$ and $\|E\|_\infty \le \rho B$. Hence, for every $\varphi \in \mathcal{T}_{N,\Gamma}(M, D, U, H, L, W, S, B)$, we have $\varphi(\Pi^\rho \mathbf{c}) = \mathrm{Clip}_{N,\Gamma}(F(X(\mathbf{c})))$, where F keeps the attention layers and feedforward networks of φ and replaces its embedding by $X \mapsto EX + P$. In the notation of Takakura & Suzuki (2023), F belongs to their class $\mathcal{T}(M, U, D, H, L + 1, \bar{W}, S, \rho B)$ with $d = 2$ input channels. Indeed, a feedforward network of $\mathcal{F}(L, W, S, B)$ has $L + 1$ affine maps, with input, hidden and output dimensions at most $\bar{W}$ and at most S nonzero weights. Moreover, all weights of F are bounded by ρB, and $\|P\|_\infty \le B \le \rho B$; padding the key and query matrices $K^{(h)}, Q^{(h)} \in \mathbb{R}^{D'\times D}$, $D' \le D$, with zero rows to size $D \times D$ changes neither the attention scores nor the weight bound. Since $\mathrm{Clip}_{N,\Gamma}$ is 1-Lipschitz in each coordinate and $X(\mathbf{c}) \in [0, 1]^{2\times\infty}$, each F_k in a $\delta/2$-cover of their class gives a map $\mathbf{c} \mapsto \mathrm{Clip}_{N,\Gamma}(F_k(X(\mathbf{c})))$. Every member of our class lies within $\delta/2$ of one of these maps, in the norm $\|\cdot\|_\infty$. Keeping one member of our class within $\delta/2$ of each map, whenever one exists, gives a δ-cover of our class of the same size. Their Theorem 5.3 bounds the logarithm of that size by $C_{\mathrm{TS}} M^3 (L + 1)(S + HD^2) \log\{2DH(L + 1)\bar{W}\rho B/\delta\}$. □

The cover in Lemma 6 is uniform over all inputs, as Lemma 5 requires, and the input clip is what makes this possible. Without it, the tokens $\eta c_j + P_j$ are unbounded, and networks with nearby embedding vectors need not be close uniformly over all inputs. The bound is linear in $S + HD^2$, which bounds the number of nonzero parameters of one block up to a constant factor, and in the number $L + 1$ of affine maps of each feedforward network; it is cubic in the number M of blocks, which equals 2 here, and only logarithmic in the weight bound, the width, the resolution δ and the clip radius ρ.

For the fully observed estimator, Lemma 5 applies with $\mathbf{c}_i = J_{\mathcal{X}} x_i$, $\boldsymbol{\zeta}_i = \iota_p \Pi_p J_{\mathcal{Y}} \varepsilon_i$, $\sigma = b = 2\kappa_\varepsilon$, $\mathcal{G} = \mathcal{T}^\rho_{N,\Gamma}$ and $g_0 = \iota_p \Pi_p J_{\mathcal{Y}} \Psi_0 J_{\mathcal{X}}^{-1}$, and Lemma 6 bounds the size of the cover.

C.2 Truncation and the approximation term

Throughout this appendix, $q \coloneqq \beta \wedge 1$, $\Gamma \coloneqq \max\{1, B_{0,\gamma}\}$ and $\tilde{\Psi}_0 \coloneqq J_{\mathcal{Y}} \Psi_0 J_{\mathcal{X}}^{-1}$. The norm of $\mathcal{Y}^\gamma$ is defined like that of $\mathcal{X}^\alpha$, $\|y\|^2_{\mathcal{Y}^\gamma} = \sum_i (1 + |i|)^{2\gamma} \langle y, \psi_i\rangle^2_{\mathcal{Y}}$. We write $\tilde{\nu} = (J_{\mathcal{X}})_{\#}\nu$ as in Section 2, and $\mathcal{B}^{2,\alpha}_\rho = J_{\mathcal{X}}(\Omega^\alpha_\rho)$ for the Sobolev sequence ball of Section 4.1. Every $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$ satisfies

$\|\Psi_0(x)\|_{\mathcal{Y}} \le B_0$, and it is continuous: it is Hölder continuous if $\beta \le 1$ and Fréchet differentiable if $\beta > 1$.

For any clipping level $\Gamma' > 0$, every member φ of a class $\mathcal{T}_{N,\Gamma'}(M, D, U, H, L, W, S, B)$ is a continuous map from $\ell^2(\mathbb{Z})$ to V_p, $p = 2N + 1$, with $\|\varphi(\mathbf{a})\|_\infty \le \Gamma'$ and hence $\|\varphi(\mathbf{a})\|_2 \le \Gamma'\sqrt{p}$. Indeed, its output coordinate k is obtained from the finitely many input coordinates a_j with $|j-k| \le MU$ by finitely many additions, multiplications, applications of ReLU and softmax, and clipping, and each map $\mathbf{a} \mapsto a_j$ is continuous on $\ell^2(\mathbb{Z})$. The same holds for $\varphi \circ \Pi^\rho$, since the input clip Π^ρ of Appendix C.1 is 1-Lipschitz on $\ell^2(\mathbb{Z})$.

Lemma 7 (Tail bound)**.** *Let $\nu \in \mathcal{P}_\alpha(\kappa)$. Then $x \mapsto \|x\|_{\mathcal{X}^\alpha} \in [0, \infty]$ is Borel, and $\nu(\{x : \|x\|_{\mathcal{X}^\alpha} > \rho\}) \le 2e^{-\rho/\kappa}$ for every $\rho \ge 0$.*

Proof. The map is the supremum over $K \in \mathbb{N}$ of the continuous maps $x \mapsto (\sum_{|i|\le K}(1 + |i|)^{2\alpha}\langle x, \phi_i\rangle_{\mathcal{X}}^2)^{1/2}$, so it is Borel. Write $Z = \|x\|_{\mathcal{X}^\alpha}$ with $x \sim \nu$. Since the infimum in the definition of $\mathcal{P}_\alpha(\kappa)$ is at most κ, every $\lambda > \kappa$ exceeds some λ' with $\mathbb{E}e^{Z/\lambda'} \le 2$, and then $\mathbb{E}e^{Z/\lambda} \le \mathbb{E}e^{Z/\lambda'} \le 2$. Letting $\lambda \downarrow \kappa$, monotone convergence gives $\mathbb{E}e^{Z/\kappa} \le 2$, and Markov's inequality gives $\nu(Z > \rho) \le e^{-\rho/\kappa}\mathbb{E}e^{Z/\kappa} \le 2e^{-\rho/\kappa}$. □

Lemma 8 (Output cutoff)**.** *Let $\epsilon \in (0, 1)$, let $N = \lceil (4B_{0,\gamma}/\epsilon)^{1/\gamma} \rceil - 1$ be the output cutoff in the proof of Theorem 1, and let $p = 2N + 1$. Then:*

(i) $(N + 1)^{-\gamma} B_{0,\gamma} \le \epsilon/4$ and $p \le 3(4\Gamma/\epsilon)^{1/\gamma}$;

(ii) every $y \in \mathcal{Y}$ with $\|y\|_{\mathcal{Y}^\gamma} \le B_{0,\gamma}$ satisfies $|\langle y, \psi_k\rangle_{\mathcal{Y}}| \le B_{0,\gamma}$ for all k, and $\|(I - \Pi_p)J_{\mathcal{Y}} y\|_2 \le \epsilon/4$.

Proof. Put $u = (4B_{0,\gamma}/\epsilon)^{1/\gamma} > 0$. Then $N + 1 = \lceil u \rceil \in [u, u + 1)$, so $(N + 1)^{-\gamma} B_{0,\gamma} \le u^{-\gamma} B_{0,\gamma} = \epsilon/4$ and $N < u$. Hence $p < 2u + 1 \le 3(4\Gamma/\epsilon)^{1/\gamma}$, because $u \le (4\Gamma/\epsilon)^{1/\gamma}$ and $4\Gamma/\epsilon > 1$. For (ii), let $b_k = \langle y, \psi_k\rangle_{\mathcal{Y}}$. Then $|b_k| \le (1 + |k|)^{-\gamma}\|y\|_{\mathcal{Y}^\gamma} \le B_{0,\gamma}$, and $\sum_{|k|>N} b_k^2 \le (N + 2)^{-2\gamma} \sum_k (1 + |k|)^{2\gamma} b_k^2 \le (N + 1)^{-2\gamma} B_{0,\gamma}^2 \le (\epsilon/4)^2$ by (i). □

Lemma 9 (Approximation term)**.** *Let $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, $\nu \in \mathcal{P}_\alpha(\kappa)$, $R \ge 1$ and $\epsilon \in (0, 1)$. Let $\Gamma' > 0$, let $\bar\varphi \in \mathcal{T}_{N,\Gamma'}(M, D, U, H, L, W, S, B)$ satisfy $\sup_{x \in \Omega_R^\alpha} \|J_{\mathcal{Y}}^{-1} \circ \bar\varphi \circ J_{\mathcal{X}}(x) - \Psi_0(x)\|_{\mathcal{Y}} \le \epsilon$, and put $p = 2N + 1$ and $F = \Gamma'\sqrt{p} + B_0$. Then*

$$\text{(i)} \quad \|J_{\mathcal{Y}}^{-1} \circ \bar\varphi \circ J_{\mathcal{X}} - \Psi_0\|_{L^2(\nu)}^2 \le \epsilon^2 + 2F^2 e^{-R/\kappa},$$

$$\text{(ii)} \quad \int \|\bar\varphi(\Pi^R \mathbf{a}) - \Pi_p \tilde\Psi_0(\mathbf{a})\|_2^2 \, \tilde\nu(d\mathbf{a}) \le \epsilon^2 + 2F^2 e^{-R/\kappa}.$$

With Γ' the clipping level of Theorem 1, (i) is Corollary 1.

Proof. The integrands are Borel, because $\bar\varphi$, Π^R and Ψ_0 are continuous. For $x \in \Omega_R^\alpha$ and $\mathbf{a} = J_{\mathcal{X}} x$, every coordinate satisfies $|a_j| \le \|\mathbf{a}\|_2 \le \|x\|_{\mathcal{X}^\alpha} \le R$, so $\Pi^R \mathbf{a} = \mathbf{a}$ and $\|\bar\varphi(\mathbf{a}) - \tilde\Psi_0(\mathbf{a})\|_2 \le \epsilon$ by isometry. Since $\bar\varphi(\mathbf{a}) \in V_p$, also $\|\bar\varphi(\mathbf{a}) - \Pi_p\tilde\Psi_0(\mathbf{a})\|_2 = \|\Pi_p\{\bar\varphi(\mathbf{a}) - \tilde\Psi_0(\mathbf{a})\}\|_2 \le \epsilon$. For every $\mathbf{a} \in \ell^2(\mathbb{Z})$ and $\mathbf{c} \in \{\mathbf{a}, \Pi^R\mathbf{a}\}$, $\|\bar\varphi(\mathbf{c})\|_2 \le \Gamma'\sqrt{p}$ and $\|\Pi_p\tilde\Psi_0(\mathbf{a})\|_2 \le \|\tilde\Psi_0(\mathbf{a})\|_2 \le B_0$, so both integrands are at most F^2. Splitting the integrals over Ω_R^α and its complement and applying Lemma 7 gives both bounds. □

Lemma 9 splits the approximation error according to whether the input lies in the Sobolev ball Ω_R^α. On the ball, the error is at most the uniform accuracy ϵ of Theorem 1. Outside the ball, the only available information is that the network output is clipped and that Ψ_0 is bounded by B_0, so the squared error is at most F^2, and by Lemma 7 the input leaves the ball with probability at most $2e^{-R/\kappa}$. When N is the output cutoff of Lemma 8 and $\Gamma' = \Gamma$, we have $F^2 \le 2\Gamma^2 p + 2B_0^2 \le c_F \epsilon^{-1/\gamma}$, with c_F as in Appendix C.4, and the radius $R = \kappa\{(2 + 1/\gamma)\log(1/\epsilon) + \log(4c_F)\}$ makes the second term at most $\epsilon^2/2$. Because the input distribution has sub-exponential tails, a radius that grows only logarithmically in $1/\epsilon$ suffices.

C.3 THE ESTIMATOR CLASS AND ITS COVERING NUMBER

The approximation theorem in the form used. Fix $R \geq 1$ and $\epsilon \in (0, e^{-1}]$, so that $\log(R/\epsilon) \geq 1$. Let N_1, N_2, $U = N_1 \vee N_2$, $D = 2U + 4$ and the fixed positional encoding $P = (P_i)_{i\in\mathbb{Z}}$, $P_i = (0, \dots, 0, \cos(i\phi), \sin(i\phi))^\top$ with $\phi = (2U+1)^{-1}$, be those of Step 1 of the proof of Theorem 1, and let $N = N_2$ as in its Step 5. They are explicit functions of $(\alpha, \beta, \gamma, B_0, B_{0,\gamma}, R, \epsilon)$, and $\|P\|_\infty \leq 1$. Put

$$X \coloneqq (R\epsilon^{-1/\beta})^{1/\alpha}, \qquad X^+ \coloneqq \max\{(R\epsilon^{-1/q})^{1/\alpha}, \epsilon^{-1/\gamma}\}, \qquad \bar{\ell} \coloneqq 1 + \log(R/\epsilon),$$

so that $1 \leq X \leq X^+$ and $\bar{\ell} \geq 1$. On this range, the constants hidden in Theorem 1 depend only on $(\alpha, \beta, \gamma, B_0, B_{0,\gamma})$, as its proof shows. Since $\log(R/\epsilon) \leq \bar{\ell}$, there is a constant $c_A \geq 1$, depending only on these quantities, with the following property. For every $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, the theorem, applied with $\Gamma = \max\{1, B_{0,\gamma}\}$, provides $\bar{\varphi} \in \mathcal{T}_{N,\Gamma}(2, D, U, H, L, W, S, B)$ with positional encoding P, accuracy ϵ on Ω_R^α, and

$$\max\{D, U, H\} \leq c_A X^+, \qquad L \leq c_A X \bar{\ell}^2, \qquad \log \max\{W, S\} \leq c_A X \bar{\ell}, \qquad \log B \leq c_A \bar{\ell}. \tag{6}$$

Moreover $L \geq 1$, since both feedforward networks of the construction have a hidden layer.

Lemma 10 (Enlarging the parameters)**.** *Let $H \leq H'$, $1 \leq L \leq L'$, $W \leq W'$, $B \leq B'$ with $B' \geq 1$, and $S' \geq S + (L' - L)W$. Then $\mathcal{T}_{N,\Gamma}(M, D, U, H, L, W, S, B) \subset \mathcal{T}_{N,\Gamma}(M, D, U, H', L', W', S', B')$, with the same positional encoding.*

Proof. Add $H' - H$ heads with $Q^{(h)} = K^{(h)} = 0$ and $V^{(h)} = 0$ to every attention layer; they contribute 0 to the layer. Let $f \in \mathcal{F}(L, W, S, B)$ have hidden widths $d_1, \dots, d_L$, and write $f = (A_L \boldsymbol{\sigma}(\cdot) + b_L) \circ g$. Because $\sigma \circ \sigma = \sigma$, we have $f = (A_L \boldsymbol{\sigma}(\cdot) + b_L) \circ (I\boldsymbol{\sigma}(\cdot) + 0)^{\circ(L'-L)} \circ g$ with I the $d_L \times d_L$ identity matrix. This is a network of depth L' with maximal hidden width W, at most $S + (L' - L)d_L \leq S'$ nonzero parameters, and parameters bounded by $\max\{B, 1\} \leq B'$. Adding $W' - W$ neurons with zero incoming and outgoing weights to a hidden layer of width W changes neither the network nor the number of nonzero parameters, and makes the maximal hidden width W'. We apply this to every feedforward layer, and $\|\eta\|_\infty \leq B \leq B'$. □

Definition 6 (Calibrated class)**.** *For $R \geq 1$ and $\epsilon \in (0, e^{-1}]$, with $N, U, D, P, X, X^+, \bar{\ell}$ and c_A as above, let*

$$\bar{H} \coloneqq \lfloor c_A X^+ \rfloor, \quad \bar{L} \coloneqq \lfloor c_A X \bar{\ell}^2 \rfloor, \quad \bar{W} \coloneqq \lfloor e^{c_A X \bar{\ell}} \rfloor, \quad \bar{S} \coloneqq (\bar{L} + 1)\bar{W}, \quad \bar{B} \coloneqq e^{c_A \bar{\ell}},$$

and let $\mathcal{T}(\epsilon, R) \coloneqq \mathcal{T}^R_{N,\Gamma}(2, D, U, \bar{H}, \bar{L}, \bar{W}, \bar{S}, \bar{B})$, with positional encoding P and the input clip Π^R of Appendix C.1.

Lemma 11 (Witness)**.** *Let $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, $R \geq 1$ and $\epsilon \in (0, e^{-1}]$. There is $\bar{\varphi} \in \mathcal{T}_{N,\Gamma}(2, D, U, \bar{H}, \bar{L}, \bar{W}, \bar{S}, \bar{B})$ with $\sup_{x\in\Omega_R^\alpha} \|J_{\mathcal{Y}}^{-1} \circ \bar{\varphi} \circ J_{\mathcal{X}}(x) - \Psi_0(x)\|_{\mathcal{Y}} \leq \epsilon$. The member $\bar{\varphi} \circ \Pi^R$ of $\mathcal{T}(\epsilon, R)$ coincides with $\bar{\varphi}$ on $\mathcal{B}_R^{2,\alpha}$.*

Proof. Take the network $\bar{\varphi}$ provided above, with parameters (H, L, W, S, B) satisfying (6). These are integers except B, so $H \leq \bar{H}$, $1 \leq L \leq \bar{L}$, $W \leq \bar{W}$ and $S \leq \bar{W}$; hence $S + (\bar{L} - L)W \leq (\bar{L} + 1)\bar{W} = \bar{S}$. Also $B \leq \bar{B}$ and $\bar{B} \geq 1$. Lemma 10 gives the first claim, and the second holds because $\Pi^R \mathbf{a} = \mathbf{a}$ on $\mathcal{B}_R^{2,\alpha}$. □

The class $\mathcal{T}(\epsilon, R)$ is determined by (ϵ, R) and $(\alpha, \beta, \gamma, B_0, B_{0,\gamma})$, not by Ψ_0, so the least-squares estimator over it is defined without knowledge of Ψ_0. The parameters of the network provided by Theorem 1 may depend on Ψ_0, but each is bounded by the corresponding parameter of $\mathcal{T}(\epsilon, R)$. Raising the depth from L to $\bar{L}$ with identity layers costs at most $(\bar{L} - L)W$ additional nonzero parameters, which the choice $\bar{S} = (\bar{L} + 1)\bar{W}$ accommodates, so Lemma 10 places the network in the class.

Lemma 12 (Covering number of the calibrated class)**.** *Let C_{TS} be the constant of Lemma 6, and put $c_Z \coloneqq 8 + 2c_A + \log(4c_A^4)$ and $K_0 \coloneqq 48 C_{\mathrm{TS}} c_A^4 c_Z + 1$. For all $R \geq 1$, $\epsilon \in (0, e^{-1}]$ and $n \in \mathbb{N}$,*

$$\log \mathcal{N}\big(1/n, \mathcal{T}(\epsilon, R)\big) + 1 \leq K_0 \, R \, (X^+)^5 \, \bar{\ell}^5 \, e^{c_A X \bar{\ell}} \, (1 + \log n).$$

Proof. Lemma 6 applies with $M = 2$, $\rho = R \geq 1$, $\delta = 1/n$, width $\bar{W}$ (so that its $\max\{W, D\}$ is $\max\{\bar{W}, D\}$) and weight bound $\bar{B} \geq 1 \geq \|P\|_\infty$. With $Z \coloneqq 2R\bar{B}D\bar{H}(\bar{L}+1)\max\{\bar{W}, D\}$, it gives

$$\log \mathcal{N}\big(1/n, \mathcal{T}(\epsilon, R)\big) \leq 8C_{\mathrm{TS}}(\bar{L}+1)(\bar{S} + \bar{H}D^2)\log(Zn).$$

Since $X, X^+, \bar{\ell} \geq 1$ and $c_A \geq 1$, we have $\bar{L}+1 \leq 2c_A X\bar{\ell}^2$, $\bar{W} \leq e^{c_A X\bar{\ell}}$ and $D, \bar{H} \leq c_A X^+$ by (6). Hence $\bar{S} + \bar{H}D^2 \leq 2c_A X\bar{\ell}^2 e^{c_A X\bar{\ell}} + c_A^3 (X^+)^3 \leq 3c_A^3 (X^+)^3 \bar{\ell}^2 e^{c_A X\bar{\ell}}$. With $\max\{\bar{W}, D\} \leq \bar{W}D$, also $Z \leq 4c_A^4 R(X^+)^3 X\bar{\ell}^2 e^{c_A\bar{\ell} + c_A X\bar{\ell}}$, and $\log u \leq u$ gives

$$1 + \log Z \leq 1 + \log(4c_A^4) + R + 4X^+ + 2\bar{\ell} + 2c_A X\bar{\ell} \leq c_Z R X^+ \bar{\ell},$$

because each summand is at most its coefficient times $RX^+\bar{\ell}$. As $Z \geq 1$, $\log(Zn) \leq (1+\log Z)(1+\log n)$. Multiplying the three bounds gives $\log \mathcal{N} \leq 48 C_{\mathrm{TS}} c_A^4 c_Z R(X^+)^5 \bar{\ell}^5 e^{c_A X\bar{\ell}}(1+\log n)$, and the right side of the lemma exceeds this by at least 1. □

Up to factors polynomial in R, X^+ and $\bar{\ell}$, the logarithm of the covering number is therefore at most of order $e^{c_A X\bar{\ell}}(1+\log n)$. The exponential factor enters through the sparsity $\bar{S} = (\bar{L}+1)\bar{W}$, because the class must contain the networks of Theorem 1, whose width and number of nonzero parameters are only known to be at most $e^{c_A X\bar{\ell}}$, where $X\bar{\ell} = (R\epsilon^{-1/\beta})^{1/\alpha}\{1+\log(R/\epsilon)\}$. The remaining parameters of the class contribute only factors polynomial in R, X^+ and $\bar{\ell}$, and the resolution $1/n$ contributes the factor $1+\log n$. Since $X\bar{\ell} \geq \epsilon^{-1/(\alpha\beta)}$, this bound on the metric entropy is exponential in a power of $1/\epsilon$, and a complexity term of order this bound divided by n is comparable with ϵ^2 only if $\epsilon^{-1/(\alpha\beta)}$ is at most of order $\log n$; this is why the rate of Theorem 2 is logarithmic in n.

C.4 Calibration

We fix the following constants and functions of $t \geq t_*$:

$$c_F \coloneqq 6\Gamma^2(4\Gamma)^{1/\gamma} + 2B_0^2, \qquad c_\rho \coloneqq \log(4c_F), \qquad \rho(t) \coloneqq \kappa\Big[\Big(2+\frac{1}{\gamma}\Big)t + c_\rho\Big],$$

$$t_* \coloneqq \max\Big\{1, \frac{1}{\kappa(2+1/\gamma)}\Big\}, \qquad c_R \coloneqq \max\{1, \kappa(2+1/\gamma+c_\rho)\}, \qquad h_1 \coloneqq 3 + \log c_R,$$

$$X_t \coloneqq \rho(t)^{1/\alpha} e^{t/(\alpha\beta)}, \qquad X_t^+ \coloneqq \max\{(\rho(t)e^{t/q})^{1/\alpha}, e^{t/\gamma}\}, \qquad H_t \coloneqq 1 + t + \log\rho(t),$$

$$u_1 \coloneqq \max\Big\{\frac{1}{\alpha q}, \frac{1}{\gamma}\Big\}, \qquad a_0 \coloneqq 6 + \frac{5}{\alpha} + 5u_1, \qquad K_1 \coloneqq 1008\, c_F K_0 c_R^{1+5/\alpha} h_1^5,$$

$$c_2 \coloneqq a_0 + \frac{1}{\gamma} + 2, \qquad c_4 \coloneqq \log K_1, \qquad c_6 \coloneqq c_A c_R^{1/\alpha} h_1 + c_2 + c_4,$$

$$c_7 \coloneqq c_A\big(\kappa(2+1/\gamma)\big)^{1/\alpha}, \qquad G(t) \coloneqq c_A X_t H_t + c_2 t + c_4, \qquad \Lambda_n \coloneqq \log n - \log(1+\log n).$$

They depend only on $(\alpha, \beta, \gamma, \kappa, B_0, B_{0,\gamma})$, through c_A and the universal constant C_{TS}, and not on the noise. Note that $c_F \geq 8$, so $c_\rho > 0$, and that $K_1 \geq 1$. At $\epsilon = e^{-t}$ and $R = \rho(t)$, the quantities X, X^+ and $\bar{\ell}$ of Appendix C.3 equal X_t, X_t^+ and H_t.

Lemma 13 (Calibration function)**.** *For every $t \geq t_*$:*

- *(i)* $1 \leq \rho(t) \leq c_R t$;
- *(ii)* $1 \leq X_t \leq c_R^{1/\alpha} t^{1/\alpha} e^{t/(\alpha\beta)}$, $X_t \geq (\kappa(2+1/\gamma))^{1/\alpha} e^{t/(\alpha\beta)}$ *and* $X_t^+ \leq c_R^{1/\alpha} t^{1/\alpha} e^{u_1 t}$;
- *(iii)* $1 \leq H_t \leq h_1 t$;
- *(iv)* $c_7 e^{t/(\alpha\beta)} \leq G(t) \leq c_6 t^{1+1/\alpha} e^{t/(\alpha\beta)}$, *and G is continuous and strictly increasing with $G(t) \to \infty$.*

Moreover, $\Lambda_n \leq \log n$ *for* $n \geq 1$, *and* $\Lambda_n \geq \frac{1}{2}\log n$ *if* $\log n \geq 3$.

Proof. (i) $\rho(t) \geq \kappa(2+1/\gamma)t_* \geq 1$ since $c_\rho > 0$, and $\rho(t) \leq \kappa(2+1/\gamma+c_\rho)t \leq c_R t$ since $t \geq 1$. (ii) follows from (i), from $\rho(t) \geq \kappa(2+1/\gamma)t \geq \kappa(2+1/\gamma)$, from $t^{1/\alpha} \geq 1$ and from

$e^{t/(\alpha q)}, e^{t/\gamma} \le e^{u_1 t}$. (iii) $H_t \ge 1$ since $\rho(t) \ge 1$, and $H_t \le 1 + t + \log c_R + \log t \le (3 + \log c_R)t$, since $1 \le t$, $\log t \le t$ and $\log c_R \ge 0$. (iv) By (ii) and (iii), $c_A X_t H_t \le c_A c_R^{1/\alpha} h_1 t^{1+1/\alpha} e^{t/(\alpha\beta)}$, and $c_2 t + c_4 \le (c_2 + c_4) t^{1+1/\alpha} e^{t/(\alpha\beta)}$ because $t \ge 1$ and $c_4 \ge 0$. Also $G(t) \ge c_A X_t \ge c_7 e^{t/(\alpha\beta)}$, using $H_t \ge 1$. The functions ρ, X_t and H_t are positive, continuous and strictly increasing, and $c_2 > 0$; this gives the rest. For the last claim, $\log(1 + \log n) \ge 0$, and $v/2 - \log(1 + v)$ is nondecreasing for $v \ge 1$ and positive at $v = 3$, so $\log(1 + \log n) \le \frac{1}{2} \log n$ if $\log n \ge 3$. □

Definition 7 (Tuning). *Let $n_1 \coloneqq \exp(\max\{3, 2G(t_*), 1/c_7\})$. For $n \ge n_1$, let t_n be the unique solution in $[t_*, \infty)$ of $G(t) = \Lambda_n$ (Lemma 14), and put*

$$\epsilon_n \coloneqq e^{-t_n}, \qquad \rho_n \coloneqq \rho(t_n), \qquad \mathcal{T}_n \coloneqq \mathcal{T}(\epsilon_n, \rho_n).$$

The estimator is $\hat{\Psi}_n = J_{\mathcal{Y}}^{-1} \circ \hat{\varphi}_n \circ J_{\mathcal{X}}$, where $\hat{\varphi}_n$ minimizes the empirical risk R_n defined before Theorem 2 over $\mathcal{T}_n$. We assume that $\hat{\varphi}_n(\mathbf{a})$ is a measurable function of $\mathbf{a}$ and of the observed coefficients $(J_{\mathcal{X}} x_i, (J_{\mathcal{Y}} y_i)[-N : N])_{i \le n}$, on which R_n depends.

The tuning depends only on n and $(\alpha, \beta, \gamma, \kappa, B_0, B_{0,\gamma})$, not on Ψ_0, ν or the noise. The single radius ρ_n is the radius of the approximation, of the input clip and of the covering bound.

Lemma 14 (Calibration root). *For every $n \ge n_1$, the equation $G(t) = \Lambda_n$ has exactly one solution $t_n \in [t_*, \infty)$, and with $C' \coloneqq (2c_6(2\alpha\beta)^{1+1/\alpha})^{2\alpha\beta}$,*

$$t_n \le 2\alpha\beta \log\log n, \qquad \epsilon_n^2 \le C' (\log n)^{-2\alpha\beta} (\log\log n)^{2\beta(\alpha+1)}.$$

Proof. Since $\log n \ge 3$, Lemma 13 gives $\Lambda_n \ge \frac{1}{2} \log n \ge G(t_*)$. As G is continuous and strictly increasing on $[t_*, \infty)$ with $G(t) \to \infty$, there is exactly one root $t_n \ge t_*$. By Lemma 13(iv), $c_7 e^{t_n/(\alpha\beta)} \le G(t_n) = \Lambda_n \le \log n$, so $t_n \le \alpha\beta\{\log\log n + \log(1/c_7)\}$. If $c_7 < 1$, then $\log\log n \ge \log(1/c_7)$ because $\log n \ge 1/c_7$; if $c_7 \ge 1$, then $\log(1/c_7) \le 0 < \log\log n$. In both cases $t_n \le 2\alpha\beta \log\log n$. By Lemma 13(iv) again, $\frac{1}{2} \log n \le \Lambda_n = G(t_n) \le c_6 t_n^{1+1/\alpha} e^{t_n/(\alpha\beta)}$, so

$$e^{t_n/(\alpha\beta)} \ge \frac{\log n}{2c_6 t_n^{1+1/\alpha}} \ge \frac{\log n}{2c_6 (2\alpha\beta)^{1+1/\alpha} (\log\log n)^{1+1/\alpha}}.$$

Raising both sides to the power $-2\alpha\beta$ gives the bound on $\epsilon_n^2 = (e^{t_n/(\alpha\beta)})^{-2\alpha\beta}$, since $2\alpha\beta(1 + 1/\alpha) = 2\beta(\alpha + 1)$. □

The calibration equation balances the squared accuracy $\epsilon^2 = e^{-2t}$ against the complexity term of the oracle inequality, whose logarithm is bounded by $c_A X_t H_t - \Lambda_n$ plus an affine function of t. At the root, $c_A X_{t_n} H_{t_n}$ equals $\Lambda_n = \log n - \log(1 + \log n)$ up to the affine part $c_2 t_n + c_4$, which is of lower order. Since $X_t H_t$ is of order $t^{1+1/\alpha} e^{t/(\alpha\beta)}$, the two inequalities in the proof of Lemma 14 give $t_n = \alpha\beta \log\log n\,(1 + o(1))$. The factor $t^{1+1/\alpha}$, which comes from the radius $\rho(t) \asymp t$ inside X_t and from $H_t \asymp t$, produces the power of $\log\log n$ in Lemma 14.

C.5 Proof of Theorem 2

We first collect the properties of the tuning that the proof uses.

Lemma 15 (Properties of the tuning). *Let $n \ge n_1$, $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, and let N, $p = 2N + 1$ be the output cutoff of $\mathcal{T}_n$. Put $m_n \coloneqq \max\{3, \mathcal{N}(1/n, \mathcal{T}_n)\}$. Then:*

- *(T1) there is $\bar{\varphi} \in \mathcal{T}_n$ with $\sup_{\mathbf{c} \in \mathcal{B}_{\rho_n}^{2,\alpha}} \|\bar{\varphi}(\mathbf{c}) - \tilde{\Psi}_0(\mathbf{c})\|_2 \le \epsilon_n$; moreover $(N+1)^{-\gamma} B_{0,\gamma} \le \epsilon_n/4$ and $p \le 3(4\Gamma)^{1/\gamma} e^{t_n/\gamma}$;*
- *(T2) $\rho_n = \kappa\{(2 + 1/\gamma) t_n + \log(4c_F)\}$;*
- *(T3) $c_F \ge 2B_0^2$, $(\Gamma\sqrt{p} + B_0)^2 \le c_F e^{t_n/\gamma}$, and $336\, c_F e^{t_n/\gamma} (\log m_n + 1)/n \le \epsilon_n^2$;*
- *(T4) $n \ge K_1 e^{(2+2/\gamma)t_n}$, and $K_1 \ge 1008\, c_F$;*
- *(T5) $\epsilon_n^2 \le C' (\log n)^{-2\alpha\beta} (\log\log n)^{2\beta(\alpha+1)}$.*

Proof. Write $t = t_n$, $\epsilon = \epsilon_n$ and $\rho = \rho_n$; then $t \geq t_* \geq 1$, $\epsilon \in (0, e^{-1}]$ and $\rho \geq 1$. (T1) is Lemma 11 at (ϵ, ρ), together with Lemma 8(i) and $(4\Gamma/\epsilon)^{1/\gamma} = (4\Gamma)^{1/\gamma} e^{t/\gamma}$. (T2) is the definition of ρ. For (T3), $(\Gamma\sqrt{p} + B_0)^2 \leq 2\Gamma^2 p + 2B_0^2 \leq c_F e^{t/\gamma}$ by (T1). Next, Lemma 12 at (ϵ, ρ) and Lemma 13(i)–(iii) give

$$\begin{aligned}\log \mathcal{N}(1/n, \mathcal{T}_n) + 1 &\leq K_0 c_R^{1+5/\alpha} h_1^5 \, t^{6+5/\alpha} e^{5u_1 t} e^{c_A X_t H_t} (1 + \log n) \\ &\leq K_0 c_R^{1+5/\alpha} h_1^5 \, e^{a_0 t} e^{c_A X_t H_t} (1 + \log n),\end{aligned}$$

using $t^b \leq e^{bt}$ for $b \geq 0$. Since $\mathcal{N} \geq 1$ and $\log 3 \leq 2$, $\log m_n + 1 \leq 3(\log \mathcal{N}(1/n, \mathcal{T}_n) + 1)$. Hence, with $(1 + \log n)/n = e^{-\Lambda_n}$ and $c_2 = a_0 + 1/\gamma + 2$,

$$\frac{336\, c_F e^{t/\gamma} (\log m_n + 1)}{n} \leq K_1 e^{(c_2 - 2)t} e^{c_A X_t H_t} e^{-\Lambda_n} = e^{-2t} e^{G(t) - \Lambda_n} = \epsilon^2.$$

(T4) By Lemma 13, $\log n \geq \Lambda_n = G(t) \geq c_2 t + c_4 \geq (2 + 2/\gamma)t + \log K_1$, since $a_0 \geq 5u_1 \geq 1/\gamma$; and $K_1 \geq 1008 c_F$ since $K_0, c_R, h_1 \geq 1$. (T5) is Lemma 14. □

Proof of Theorem 2. Fix $n \geq n_1$, $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$, $\nu \in \mathcal{P}_\alpha(\kappa)$ and $\nu_\varepsilon \in \mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$. Write $t = t_n$, $\epsilon = \epsilon_n$, $\rho = \rho_n$, and let N, $p = 2N + 1$ be the output cutoff of $\mathcal{T}_n$.

Step 1: reduction to Appendix C.1. Put $\mathbf{c}_i = J_{\mathcal{X}} x_i$, $\boldsymbol{\zeta}_i = \iota_p \Pi_p J_{\mathcal{Y}} \varepsilon_i$, $g_0 = \iota_p \Pi_p \tilde{\Psi}_0$, $\mathcal{G} = \{\iota_p \circ \varphi : \varphi \in \mathcal{T}_n\}$ and $\hat{g} = \iota_p \circ \hat{\varphi}_n$. The pairs $(\mathbf{c}_i, \boldsymbol{\zeta}_i)$ are i.i.d., and since ε_i is independent of x_i, they satisfy (3) with $\sigma = b = 2\kappa_\varepsilon$ by Lemma 3 and the remark following it. The members of $\mathcal{G}$ are continuous with entries bounded by Γ (Appendix C.2), and g_0 is continuous with $|g_0(\mathbf{c})_k| \leq B_{0,\gamma} \leq \Gamma$ by Lemma 8(ii). Since $\iota_p \Pi_p J_{\mathcal{Y}} y_i = g_0(\mathbf{c}_i) + \boldsymbol{\zeta}_i$ and every $\varphi \in \mathcal{T}_n$ takes values in V_p, we have $nR_n(\varphi) = \sum_i \|\iota_p \varphi(\mathbf{c}_i) - g_0(\mathbf{c}_i) - \boldsymbol{\zeta}_i\|_2^2$. Moreover $(J_{\mathcal{Y}} y_i)[-N : N] = \iota_p \Pi_p J_{\mathcal{Y}} y_i = g_0(\mathbf{c}_i) + \boldsymbol{\zeta}_i$, so by Definition 7, $\hat{g}(\mathbf{c})$ is a measurable function of $\mathbf{c}$ and the pairs $(\mathbf{c}_i, \boldsymbol{\zeta}_i)$. Thus $\hat{g}$ is a least-squares estimator in the sense of Lemma 4. A minimal $(1/n)$-cover of $\mathcal{G}$, padded with repetitions, has size m_n, which is finite by Lemma 12, and $1/n \in (0, 1)$. Lemma 5 with $R_0 = \Gamma$ and $\sigma^2 + b^2 = 8\kappa_\varepsilon^2$ gives

$$\mathbb{E}\|\hat{g} - g_0\|_{L^2(\tilde{\nu})}^2 \leq 4 \inf_{g \in \mathcal{G}} \|g - g_0\|_{L^2(\tilde{\nu})}^2 + \frac{336(p\Gamma^2 + 8\kappa_\varepsilon^2)(\log m_n + 1)}{n} + \frac{32p(1 + \Gamma + 4\kappa_\varepsilon)}{n}.$$

Step 2: risk decomposition. Since $\hat{\varphi}_n(\mathbf{a}) \in V_p$, isometry and Pythagoras' identity give

$$\mathbb{E}\|\hat{\Psi}_n - \Psi_0\|_{L^2(\nu)}^2 = \mathbb{E}\|\hat{g} - g_0\|_{L^2(\tilde{\nu})}^2 + \int \|(I - \Pi_p)\tilde{\Psi}_0(\mathbf{a})\|_2^2 \, \tilde{\nu}(d\mathbf{a}),$$

and the last term is at most $\epsilon^2/16$ by Lemma 8(ii).

Step 3: approximation. Let $\bar{\varphi}$ be the network of Lemma 11 at (ϵ, ρ), so that $\iota_p \circ \bar{\varphi} \circ \Pi^\rho \in \mathcal{G}$. By Lemma 9(ii), (T2) and (T3),

$$\inf_{g \in \mathcal{G}} \|g - g_0\|_{L^2(\tilde{\nu})}^2 \leq \epsilon^2 + 2(\Gamma\sqrt{p} + B_0)^2 e^{-\rho/\kappa} \leq \epsilon^2 + 2c_F e^{t/\gamma} \frac{e^{-(2+1/\gamma)t}}{4c_F} = \frac{3}{2}\epsilon^2.$$

Step 4: statistical error. By (T1), $p\Gamma^2 + 8\kappa_\varepsilon^2 \leq (c_F + 8\kappa_\varepsilon^2) e^{t/\gamma}$, so (T3) and $c_F \geq 8$ give

$$\frac{336(p\Gamma^2 + 8\kappa_\varepsilon^2)(\log m_n + 1)}{n} \leq \Big(1 + \frac{8\kappa_\varepsilon^2}{c_F}\Big)\epsilon^2 \leq (1 + \kappa_\varepsilon^2)\epsilon^2.$$

Step 5: discretization error. By (T1), (T4), $K_1 \geq 1008 c_F \geq 6048\, \Gamma^2 (4\Gamma)^{1/\gamma}$ and $\Gamma \geq 1$,

$$\frac{32p(1 + \Gamma + 4\kappa_\varepsilon)}{n} \leq \frac{96(4\Gamma)^{1/\gamma}(1 + \Gamma + 4\kappa_\varepsilon) e^{t/\gamma}}{K_1 e^{(2+2/\gamma)t}} \leq \frac{1 + \Gamma + 4\kappa_\varepsilon}{63\,\Gamma^2}\, \epsilon^2 \leq \frac{2 + 4\kappa_\varepsilon}{63}\, \epsilon^2.$$

Step 6: conclusion. Collecting Steps 1–5, and using $4\kappa_\varepsilon/63 \leq \kappa_\varepsilon^2 + 4/63^2$ and $7 + 2/63 + 4/63^2 + 1/16 < 8$,

$$\mathbb{E}\|\hat{\Psi}_n - \Psi_0\|_{L^2(\nu)}^2 \leq \Big(6 + 1 + \kappa_\varepsilon^2 + \frac{2 + 4\kappa_\varepsilon}{63} + \frac{1}{16}\Big)\epsilon^2 \leq (8 + 2\kappa_\varepsilon^2)\, \epsilon_n^2.$$

The right side does not depend on Ψ_0, ν or ν_ε. By (T5), the theorem holds with $C = (8 + 2\kappa_\varepsilon^2)C'$, because $(\log n)^{-2\alpha\beta}(\log\log n)^{2\beta(\alpha+1)} = \{\log n/(\log\log n)^{1+1/\alpha}\}^{-2\alpha\beta}$. The threshold n_1 does not depend on κ_ε. □

Where the rate comes from. In $G(t) = c_A X_t H_t + c_2 t + c_4$, the leading term $c_A X_t H_t$ is of exact order $t^{1+1/\alpha} e^{t/(\alpha\beta)}$. It is the exponent $c_A X \bar{\ell}$ of (6) at $\epsilon = e^{-t}$ and $R = \rho(t)$, which bounds the logarithms of the width and of the number of nonzero parameters, and it enters the covering number through $\bar{S} = (\bar{L}+1)\bar{W}$. The radius plays two roles. Inside the leading term, it appears through $X_t = \rho(t)^{1/\alpha} e^{t/(\alpha\beta)}$ and $H_t = 1 + t + \log \rho(t)$; since $\rho(t) \asymp t$, it contributes the factor $t^{1/\alpha}$ discussed after Lemma 14. As the prefactor R in Lemma 12, it contributes only a factor $e^{O(t)}$. The same holds for the depth, the weight bound, the heads, the window, the embedding dimension and the output cutoff p: they affect only c_2 and c_4. In particular, the input cutoff N_1, of order $e^{t/(\alpha q)}$ up to a power of t, enters only through u_1, and the exponent $2\alpha\beta$ of the rate is determined by the growth $e^{t/(\alpha\beta)}$ of X_t, which involves β rather than $q = \beta \wedge 1$.

C.6 Proof of Theorem 4

The proof reduces the discretized problem to a regression of the encoded outputs on the encoded inputs, whose noise depends on the input, and bounds its stochastic error with Lemma 5. The input discretization is controlled separately through the Hölder continuity of Ψ_0, and the calibration of Appendix C.4 is reused unchanged.

Throughout, $q \coloneqq \beta \wedge 1$, $p \coloneqq 2N+1$ and $V_p \coloneqq \operatorname{span}\{e_k : |k| \le N\} \subset \ell^2(\mathbb{Z})$, which we identify with $\mathbb{R}^p$. We write Π_p for the orthogonal projection of $\ell^2(\mathbb{Z})$ onto V_p, and $\Pi_p^{\mathcal{Y}} = J_{\mathcal{Y}}^{-1} \Pi_p J_{\mathcal{Y}}$. We set $c_\alpha \coloneqq (\sum_{i\in\mathbb{Z}} (1+|i|)^{-2\alpha})^{1/2}$, which is finite since $\alpha > 1/2$, define c_γ in the same way, and put $B_\infty \coloneqq C_\psi c_\gamma B_{0,\gamma}$. We also write $\Gamma \coloneqq \max\{1, B_{0,\gamma}\}$. For one observed function we write $\mathbf{a} = J_{\mathcal{X}} x$, $U = (u_1, \dots, u_{n_u})$, $V = (v_1, \dots, v_{n_v})$ and $\Xi = (\xi_1, \dots, \xi_{n_v})$. The encoded quantities are $\hat{\mathbf{a}} = \hat{\mathbf{a}}(x, U)$, $\tilde{\mathbf{a}} = P_R(\hat{\mathbf{a}})$ and $\hat{\mathbf{b}} = \hat{\mathbf{b}}(x, V, \Xi)$, as in (2). The target coefficients are $\mathbf{b}(x) \coloneqq \Pi_p J_{\mathcal{Y}} \Psi_0(x)$, which satisfy $\|\mathbf{b}(x)\|_2 \le \|\Psi_0(x)\|_{\mathcal{Y}} \le B_0$.

Lemma 16 (Point values and moments)**.** *Let $\alpha > 1/2$ and $\|x\|_{\mathcal{X}^\alpha} < \infty$. Then $\sum_i |a_i| \le c_\alpha \|x\|_{\mathcal{X}^\alpha}$, and the series $x(u) \coloneqq \sum_i a_i \phi_i(u)$ converges absolutely and uniformly in u to a version of x with $\sup_u |x(u)| \le C_\phi c_\alpha \|x\|_{\mathcal{X}^\alpha}$. The map $(x,u) \mapsto x(u)$ is jointly measurable on $\{\|x\|_{\mathcal{X}^\alpha} < \infty\} \times \Omega_{\mathcal{X}}$. Likewise, $\Psi_0(x)$ has such a version with $\sup_v |\Psi_0(x)(v)| \le B_\infty$. If $\nu \in \mathcal{P}_\alpha(\kappa)$, then $\mathbb{E}\|x\|_{\mathcal{X}^\alpha}^2 \le 2\kappa^2$ and $\nu(\|x\|_{\mathcal{X}^\alpha} > R) \le 2e^{-R/\kappa}$ for all $R \ge 0$.*

Proof. By the Cauchy–Schwarz inequality, $\sum_i |a_i| = \sum_i (1+|i|)^{-\alpha} \cdot (1+|i|)^\alpha |a_i| \le c_\alpha \|x\|_{\mathcal{X}^\alpha}$. Hence $\sup_u |\sum_{|i|>M} a_i \phi_i(u)| \le C_\phi \sum_{|i|>M} |a_i| \to 0$. The partial sums therefore converge uniformly, and they also converge to x in $L^2(\mu_u)$, so the uniform limit is a version of x bounded by $C_\phi c_\alpha \|x\|_{\mathcal{X}^\alpha}$. Each partial sum is jointly measurable in (x,u), being a finite sum of products of the continuous functionals $x \mapsto a_i$ and the measurable functions ϕ_i. The pointwise limit is therefore jointly measurable. The same argument applies to $\Psi_0(x)$, since $\|\Psi_0(x)\|_{\mathcal{Y}^\gamma} \le B_{0,\gamma}$ and $\gamma > 1/2$. For the moments, the infimum defining the Orlicz norm is attained (by monotone convergence), so $\mathbb{E} e^{\|x\|_{\mathcal{X}^\alpha}/\kappa} \le 2$. Using $e^s \ge 1 + s^2/2$ for $s \ge 0$ gives $\mathbb{E}\|x\|_{\mathcal{X}^\alpha}^2 \le 2\kappa^2$, and Markov's inequality applied to $e^{\|x\|_{\mathcal{X}^\alpha}/\kappa}$ gives the tail bound. □

Lemma 17 (Input coefficients)**.** *For every x with $\|x\|_{\mathcal{X}^\alpha} < \infty$ and every $K \in \mathbb{N}$,*

$$\mathbb{E}\big[\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2 \,\big|\, x\big] \le \frac{(2K+1) C_\phi^2 \|x\|_{\mathcal{X}}^2}{n_u} + (2+K)^{-2\alpha} \|x\|_{\mathcal{X}^\alpha}^2.$$

Consequently, if $\nu \in \mathcal{P}_\alpha(\kappa)$ and $K = \lceil c_K n_u^{1/(2\alpha+1)} \rceil$ with $c_K > 0$, then $\mathbb{E}\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2 \le 2\kappa^2 \{(2c_K + 3) C_\phi^2 + c_K^{-2\alpha}\} n_u^{-2\alpha/(2\alpha+1)}$.

Proof. The vectors $\hat{\mathbf{a}} - \mathbf{a}_{\le K}$ and $\mathbf{a}_{>K}$ have disjoint supports, so $\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2 = \sum_{|k|\le K} (\hat{a}_k - a_k)^2 + \|\mathbf{a}_{>K}\|_2^2$. Since $u_1, \dots, u_{n_u}$ are i.i.d. μ_u and independent of x, we have $\mathbb{E}[x(u_1)\phi_k(u_1) \mid x] = \langle x, \phi_k \rangle_{\mathcal{X}} = a_k$. Hence $\mathbb{E}[(\hat{a}_k - a_k)^2 \mid x] = n_u^{-1} \mathrm{Var}(x(u_1)\phi_k(u_1) \mid x) \le n_u^{-1} C_\phi^2 \mathbb{E}[x(u_1)^2 \mid x] = n_u^{-1} C_\phi^2 \|x\|_{\mathcal{X}}^2$. Summing over the $2K+1$ indices, and using $\|\mathbf{a}_{>K}\|_2^2 \le (2+K)^{-2\alpha} \sum_{|i|>K} (1+|i|)^{2\alpha} a_i^2$, gives the first claim. For the second claim, note that $c_K n_u^{1/(2\alpha+1)} \le K \le c_K n_u^{1/(2\alpha+1)} + 1$ and $n_u^{1/(2\alpha+1)} \ge 1$ give $2K+1 \le (2c_K+3) n_u^{1/(2\alpha+1)}$ and $(2+K)^{-2\alpha} \le c_K^{-2\alpha} n_u^{-2\alpha/(2\alpha+1)}$. Then use $\mathbb{E}\|x\|_{\mathcal{X}}^2 \le \mathbb{E}\|x\|_{\mathcal{X}^\alpha}^2 \le 2\kappa^2$ (Lemma 16). □

The two terms in Lemma 17 are the Monte Carlo variance of the $2K+1$ retained coefficients, which grows linearly in K, and the truncation error of the discarded ones, which decays like $K^{-2\alpha}$ by the Sobolev regularity of the input. The choice $K \asymp n_u^{1/(2\alpha+1)}$ balances them. Only the second moment of $\|x\|_{\mathcal{X}^\alpha}$ enters, not the tail of ν.

Lemma 18 (Projection onto the Sobolev sequence ball)**.** *The set $\mathcal{B}_R^{2,\alpha}$ is nonempty, closed and convex in $\ell^2(\mathbb{Z})$. Hence the metric projection P_R is well defined, $P_R(\mathbf{c}) = \mathbf{c}$ for $\mathbf{c} \in \mathcal{B}_R^{2,\alpha}$, and $\|P_R(\mathbf{c}) - P_R(\mathbf{c}')\|_2 \le \|\mathbf{c} - \mathbf{c}'\|_2$. Let $\hat{\mathbf{c}}$ be supported in $\{|k| \le K\}$ and let $\omega_k = (1+|k|)^{2\alpha}$. Then $[P_R(\hat{\mathbf{c}})]_k = \hat{c}_k/(1+\lambda\omega_k)$, where $\lambda = 0$ if $\hat{\mathbf{c}} \in \mathcal{B}_R^{2,\alpha}$, and otherwise $\lambda > 0$ is the unique solution of $\sum_k \omega_k \hat{c}_k^2 (1+\lambda\omega_k)^{-2} = R^2$. In particular, $\tilde{\mathbf{a}} \in \mathcal{B}_R^{2,\alpha}$ always, and $\|\tilde{\mathbf{a}} - \mathbf{a}\|_2 \le \|\hat{\mathbf{a}} - \mathbf{a}\|_2$ whenever $x \in \Omega_R^\alpha$.*

Proof. $\mathcal{B}_R^{2,\alpha}$ contains 0. It is convex because $\mathbf{c} \mapsto \sum_k \omega_k c_k^2$ is convex, and closed because it is the intersection over M of the closed sets $\{\sum_{|k|\le M} \omega_k c_k^2 \le R^2\}$. Existence, uniqueness and nonexpansiveness of the metric projection onto a nonempty closed convex subset of a Hilbert space are standard. The explicit form follows from the Karush–Kuhn–Tucker conditions of the convex problem $\min\{\|\mathbf{c}-\hat{\mathbf{c}}\|_2^2 : \sum_k \omega_k c_k^2 \le R^2\}$, which are sufficient here. The function $\lambda \mapsto \sum_k \omega_k \hat{c}_k^2 (1+\lambda\omega_k)^{-2}$ is continuous and strictly decreasing from $\sum_k \omega_k \hat{c}_k^2 > R^2$ to 0. Finally, $x \in \Omega_R^\alpha$ means $\mathbf{a} \in \mathcal{B}_R^{2,\alpha}$, so $\|\tilde{\mathbf{a}} - \mathbf{a}\|_2 = \|P_R(\hat{\mathbf{a}}) - P_R(\mathbf{a})\|_2 \le \|\hat{\mathbf{a}} - \mathbf{a}\|_2$. □

Lemma 19 (Output coefficients)**.** *For every x with $\|x\|_{\mathcal{X}^\alpha} < \infty$, $\mathbb{E}[\hat{\mathbf{b}} \mid x, U] = \mathbf{b}(x)$. Moreover, for every unit vector $\mathbf{w} \in V_p$ and every $|\lambda| \le 1/b_\eta$,*

$$\mathbb{E}\big[\exp(\lambda\langle \hat{\mathbf{b}} - \mathbf{b}(x), \mathbf{w}\rangle) \,\big|\, x, U\big] \le \exp(\lambda^2\sigma_\eta^2/2),$$

where $\sigma_\eta^2 := C_\psi^2 p(B_\infty^2 + 4\kappa_\varepsilon^2)/n_v$ and $b_\eta := 2\kappa_\varepsilon C_\psi \sqrt{p}/n_v$.

Proof. Fix x and U, and write $g = \Psi_0(x)$ (the bounded version of Lemma 16) and $h(v) = \sum_{|k|\le N} w_k \psi_k(v)$. Since V and Ξ are independent of (x, U), conditioning on (x, U) only fixes g. Now $\|h\|_\infty \le C_\psi \sum_k |w_k| \le C_\psi\sqrt{p}$, and $\mathbb{E}[g(v_1)h(v_1)] = \sum_k w_k \langle g, \psi_k\rangle_{\mathcal{Y}} = \langle \mathbf{b}(x), \mathbf{w}\rangle$. We have $\langle \hat{\mathbf{b}} - \mathbf{b}(x), \mathbf{w}\rangle = S_1 + S_2$, where $S_1 = n_v^{-1}\sum_j \{g(v_j)h(v_j) - \mathbb{E}[g(v_1)h(v_1)]\}$ and $S_2 = n_v^{-1}\sum_j \xi_j h(v_j)$. Both have mean zero, which proves the first claim. The summands of S_1 are i.i.d., centered, and take values in an interval of length $2B_\infty C_\psi\sqrt{p}$. By Hoeffding's lemma, $\mathbb{E}e^{\lambda S_1} \le \exp(\lambda^2 B_\infty^2 C_\psi^2 p/(2n_v))$ for all λ. Conditionally on V, $S_2 = \sum_j a_j \xi_j$ is a sum of independent terms, with weights $a_j = h(v_j)/n_v$ satisfying $|a_j| \le C_\psi\sqrt{p}/n_v$. For $|\lambda| \le 1/b_\eta$ we have $|\lambda a_j| \le 1/(2\kappa_\varepsilon)$, so Lemma 3 with $\mathcal{H} = \mathbb{R}$ gives $\mathbb{E}[e^{\lambda S_2} \mid V] \le \exp(2\lambda^2\kappa_\varepsilon^2 \sum_j a_j^2) \le \exp(2\lambda^2\kappa_\varepsilon^2 C_\psi^2 p/n_v)$. Since S_1 is V-measurable, $\mathbb{E}e^{\lambda(S_1+S_2)} = \mathbb{E}[e^{\lambda S_1}\mathbb{E}[e^{\lambda S_2} \mid V]] \le \exp(\lambda^2\sigma_\eta^2/2)$ for $|\lambda| \le 1/b_\eta$. □

Lemma 20 (Regression function)**.** *There is a Borel map $m : \ell^2(\mathbb{Z}) \to V_p$ with $\|m(\mathbf{c})\|_2 \le B_0$ and $\|m(\mathbf{c})\|_\infty \le B_{0,\gamma}$ for all $\mathbf{c}$, and $m(\tilde{\mathbf{a}}) = \mathbb{E}[\mathbf{b}(x) \mid \tilde{\mathbf{a}}]$ almost surely. Moreover, $\mathbb{E}[\hat{\mathbf{b}} \mid \tilde{\mathbf{a}}] = m(\tilde{\mathbf{a}})$, and $\boldsymbol{\zeta} := \hat{\mathbf{b}} - m(\tilde{\mathbf{a}})$ satisfies, for all unit vectors $\mathbf{w} \in V_p$ and all $|\lambda| \le 1/b_\eta$,*

$$\mathbb{E}\big[\exp(\lambda\langle \boldsymbol{\zeta}, \mathbf{w}\rangle) \,\big|\, \tilde{\mathbf{a}}\big] \le \exp(\lambda^2\sigma_\zeta^2/2), \qquad \sigma_\zeta^2 := B_0^2 + \sigma_\eta^2.$$

Proof. $\tilde{\mathbf{a}}$ is a random element of the Polish space $\ell^2(\mathbb{Z})$, and $\mathbf{b}(x)$ is a bounded random vector in V_p. By the Doob–Dynkin lemma, $\mathbb{E}[\mathbf{b}(x) \mid \tilde{\mathbf{a}}] = m(\tilde{\mathbf{a}})$ for a Borel map m. The vector $\mathbf{b}(x)$ takes values in the closed convex set $Q := \{\mathbf{v} \in V_p : \|\mathbf{v}\|_2 \le B_0, \|\mathbf{v}\|_\infty \le B_{0,\gamma}\}$, because $\|\mathbf{b}(x)\|_2 \le \|\Psi_0(x)\|_{\mathcal{Y}} \le B_0$ and $|\langle \Psi_0(x), \psi_k\rangle_{\mathcal{Y}}| \le B_{0,\gamma}$ by Lemma 8(ii). Hence $m(\tilde{\mathbf{a}}) \in Q$ almost surely, and composing m with the metric projection onto Q, which is continuous, changes $m(\tilde{\mathbf{a}})$ only on a null event; we may therefore take m with values in Q. Since $\tilde{\mathbf{a}}$ is a function of (x, U), the tower property and Lemma 19 give $\mathbb{E}[\hat{\mathbf{b}} \mid \tilde{\mathbf{a}}] = \mathbb{E}[\mathbf{b}(x) \mid \tilde{\mathbf{a}}] = m(\tilde{\mathbf{a}})$. Write $\langle \boldsymbol{\zeta}, \mathbf{w}\rangle = A + B$ with $A = \langle \mathbf{b}(x) - m(\tilde{\mathbf{a}}), \mathbf{w}\rangle$ and $B = \langle \hat{\mathbf{b}} - \mathbf{b}(x), \mathbf{w}\rangle$. Since A and $\tilde{\mathbf{a}}$ are functions of (x, U), Lemma 19 and the tower property give, for $|\lambda| \le 1/b_\eta$, $\mathbb{E}[e^{\lambda(A+B)} \mid \tilde{\mathbf{a}}] = \mathbb{E}[e^{\lambda A}\mathbb{E}[e^{\lambda B} \mid x, U] \mid \tilde{\mathbf{a}}] \le e^{\lambda^2\sigma_\eta^2/2}\,\mathbb{E}[e^{\lambda A} \mid \tilde{\mathbf{a}}]$. Conditionally on $\tilde{\mathbf{a}}$, A is centered and takes values in an interval of length $2B_0$, because $\langle \mathbf{b}(x), \mathbf{w}\rangle \in [-B_0, B_0]$. Hoeffding's lemma then gives $\mathbb{E}[e^{\lambda A} \mid \tilde{\mathbf{a}}] \le e^{\lambda^2 B_0^2/2}$. □

Lemmas 19 and 20 show that output discretization affects only the noise level of the regression, and hence only the constant C in Theorem 4. Conditionally on (x, U), the encoded target $\hat{\mathbf{b}}$ is an unbiased estimate of $\mathbf{b}(x)$, and its deviation satisfies a Bernstein condition with variance proxy of order p/n_v. Consequently, $\hat{\mathbf{b}} = m(\tilde{\mathbf{a}}) + \boldsymbol{\zeta}$ with a bounded regression function m and a noise $\boldsymbol{\zeta}$ whose conditional law given $\tilde{\mathbf{a}}$ may depend on $\tilde{\mathbf{a}}$ but always satisfies the Bernstein condition (3) with $\sigma = \sigma_\zeta$ and $b = b_\eta$; this is the setting of Lemma 4. The variance proxy σ_ζ^2 contains the additional term B_0^2 because the encoded input does not determine the input, so $\mathbf{b}(x)$ fluctuates around $m(\tilde{\mathbf{a}})$.

Lemma 21 (Reduction). *Let $\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$ and $\nu \in \mathcal{P}_\alpha(\kappa)$, with $B_0 \geq 1$. Let $\mathcal{F}$ be a class of Borel maps $\varphi : \ell^2(\mathbb{Z}) \to V_p$ with $\|\varphi(\mathbf{c})\|_\infty \leq \Gamma$ for all $\mathbf{c}$. Let $\hat{\varphi}$ be a least-squares estimator over $\mathcal{F}$ for the pairs $(\tilde{\mathbf{a}}_i, \hat{\mathbf{b}}_i)$: for every sample it minimizes $\varphi \mapsto \sum_i \|\varphi(\tilde{\mathbf{a}}_i) - \hat{\mathbf{b}}_i\|_2^2$ over $\mathcal{F}$, and $\hat{\varphi}(\mathbf{c})$ is a measurable function of $\mathbf{c}$ and $(\tilde{\mathbf{a}}_i, \hat{\mathbf{b}}_i)_{i \leq n}$. If $\delta \in (0,1)$ and $\mathcal{F}$ has a δ-cover of size $m_\delta \geq 3$ (Appendix C.1), then*

$$\mathcal{E}(\hat{\Psi}) \leq 8\mathcal{A} + 9\mathcal{D} + \mathcal{T} + \frac{336(p\Gamma^2 + \sigma_\zeta^2 + b_\eta^2)(\log m_\delta + 1)}{n} + 32p(1 + \Gamma + \sigma_\zeta + b_\eta)\delta,$$

where $\mathcal{E}$ is the risk defined before Theorem 4, and the three remaining terms are

$$\mathcal{A} \coloneqq \inf_{\varphi \in \mathcal{F}} \sup_{\mathbf{c} \in \mathcal{B}_R^{2,\alpha}} \|\varphi(\mathbf{c}) - \Pi_p \tilde{\Psi}_0(\mathbf{c})\|_2^2,$$

$$\mathcal{T} \coloneqq \sup_x \|(I - \Pi_p^{\mathcal{Y}})\Psi_0(x)\|_{\mathcal{Y}}^2,$$

$$\mathcal{D} \coloneqq \mathbb{E}\|\tilde{\Psi}_0(\tilde{\mathbf{a}}) - \tilde{\Psi}_0(\mathbf{a})\|_2^2 \;\leq\; B_0^2\big(\mathbb{E}\|\hat{\mathbf{a}} - \mathbf{a}\|_2^2\big)^q + 8B_0^2 e^{-R/\kappa}.$$

Proof. Let (x, U) be a test input with its locations, independent of the training sample, and let $\tilde{\mathbf{a}} = \tilde{\mathbf{a}}(x, U)$. Since $\hat{\varphi}(\tilde{\mathbf{a}}) \in V_p$, Pythagoras' identity gives

$$\mathcal{E}(\hat{\Psi}) = \mathbb{E}\|\hat{\varphi}(\tilde{\mathbf{a}}) - \mathbf{b}(x)\|_2^2 + \mathbb{E}\|(I - \Pi_p^{\mathcal{Y}})\Psi_0(x)\|_{\mathcal{Y}}^2.$$

The last term is at most $\mathcal{T}$. Because (x, U) is independent of the sample and $\hat{\varphi}$ is measurable, $\mathbf{b}(x) - m(\tilde{\mathbf{a}})$ has mean zero conditionally on $\tilde{\mathbf{a}}$ and the training sample (Lemma 20). Hence

$$\mathbb{E}\|\hat{\varphi}(\tilde{\mathbf{a}}) - \mathbf{b}(x)\|_2^2 = \mathbb{E}\|\hat{\varphi}(\tilde{\mathbf{a}}) - m(\tilde{\mathbf{a}})\|_2^2 + \mathbb{E}\|m(\tilde{\mathbf{a}}) - \mathbf{b}(x)\|_2^2.$$

Among $\sigma(\tilde{\mathbf{a}})$-measurable predictors of $\mathbf{b}(x)$, the conditional expectation minimizes the mean squared error. Since $\Pi_p \tilde{\Psi}_0(\tilde{\mathbf{a}})$ is such a predictor and Π_p is a contraction, $\mathbb{E}\|m(\tilde{\mathbf{a}}) - \mathbf{b}(x)\|_2^2 \leq \mathbb{E}\|\Pi_p\{\tilde{\Psi}_0(\tilde{\mathbf{a}}) - \tilde{\Psi}_0(\mathbf{a})\}\|_2^2 \leq \mathcal{D}$. For the first term, the pairs $(\tilde{\mathbf{a}}_i, \boldsymbol{\zeta}_i)$ with $\boldsymbol{\zeta}_i = \hat{\mathbf{b}}_i - m(\tilde{\mathbf{a}}_i)$ are i.i.d., and $\hat{\mathbf{b}}_i = m(\tilde{\mathbf{a}}_i) + \boldsymbol{\zeta}_i$. By Lemma 20, they satisfy (3) with $\sigma = \sigma_\zeta$ and $b = b_\eta$; moreover $\|m(\mathbf{c})\|_\infty \leq B_{0,\gamma} \leq \Gamma$. Since $\hat{\mathbf{b}}_i = m(\tilde{\mathbf{a}}_i) + \boldsymbol{\zeta}_i$ with m Borel, $\hat{\varphi}(\mathbf{c})$ is a measurable function of $\mathbf{c}$ and the pairs $(\tilde{\mathbf{a}}_i, \boldsymbol{\zeta}_i)$, and it minimizes $\sum_i \|\varphi(\tilde{\mathbf{a}}_i) - m(\tilde{\mathbf{a}}_i) - \boldsymbol{\zeta}_i\|_2^2$. Lemma 5 therefore applies with $A = \ell^2(\mathbb{Z})$, $\mathbf{c}_i = \tilde{\mathbf{a}}_i$, $\mathcal{G} = \mathcal{F}$, $g_0 = m$, $R_0 = \Gamma$, μ the law of $\tilde{\mathbf{a}}$, and the δ-cover of size m_δ. It remains to bound its approximation term. For $\varphi \in \mathcal{F}$, write $m(\tilde{\mathbf{a}}) - \Pi_p \tilde{\Psi}_0(\tilde{\mathbf{a}}) = \mathbb{E}[\mathbf{b}(x) - \Pi_p \tilde{\Psi}_0(\tilde{\mathbf{a}}) \mid \tilde{\mathbf{a}}]$ and apply Jensen's inequality:

$$\begin{aligned}\mathbb{E}\|\varphi(\tilde{\mathbf{a}}) - m(\tilde{\mathbf{a}})\|_2^2 &\leq 2\,\mathbb{E}\|\varphi(\tilde{\mathbf{a}}) - \Pi_p \tilde{\Psi}_0(\tilde{\mathbf{a}})\|_2^2 + 2\,\mathbb{E}\|m(\tilde{\mathbf{a}}) - \Pi_p \tilde{\Psi}_0(\tilde{\mathbf{a}})\|_2^2 \\ &\leq 2 \sup_{\mathbf{c} \in \mathcal{B}_R^{2,\alpha}} \|\varphi(\mathbf{c}) - \Pi_p \tilde{\Psi}_0(\mathbf{c})\|_2^2 + 2\mathcal{D}.\end{aligned}$$

Collecting terms gives the displayed inequality, since $4(2\mathcal{A} + 2\mathcal{D}) + \mathcal{D} = 8\mathcal{A} + 9\mathcal{D}$.

For the bound on $\mathcal{D}$, first note that $\|\Psi_0(x_1) - \Psi_0(x_2)\|_{\mathcal{Y}} \leq B_0 \|x_1 - x_2\|_{\mathcal{X}}^q$. For $\beta \leq 1$ this is the Hölder condition in the definition of $\mathcal{H}^\beta$; for $\beta > 1$ it follows from the mean value inequality and $\sup_x \|\mathcal{D}\Psi_0(x)\| \leq B_0$. Also $\|\Psi_0(x_1) - \Psi_0(x_2)\|_{\mathcal{Y}} \leq 2B_0$. On $\{x \in \Omega_R^\alpha\}$, Lemma 18 gives $\|\tilde{\mathbf{a}} - \mathbf{a}\|_2 \leq \|\hat{\mathbf{a}} - \mathbf{a}\|_2$. Hence $\mathcal{D} \leq B_0^2 \mathbb{E}\|\hat{\mathbf{a}} - \mathbf{a}\|_2^{2q} + 4B_0^2 \nu(\|x\|_{\mathcal{X}^\alpha} > R)$. The bound follows from Jensen's inequality ($q \leq 1$) and Lemma 16. □

Lemma 21 bounds the risk in terms of the approximation error $\mathcal{A}$ on the Sobolev sequence ball, the output truncation $\mathcal{T}$, the input discretization $\mathcal{D}$, and the stochastic error of the regression of Lemma 20. The input discretization enters twice: the regression function m differs from $\Pi_p\tilde{\Psi}_0$ because the input is observed only through $\tilde{\mathbf{a}}$, and the test input is encoded in the same way. In both places it appears only through $\|\tilde{\mathbf{a}}-\mathbf{a}\|_2$ and the Hölder continuity of Ψ_0, which is why the exponent $\beta\wedge 1$, rather than β, appears in our bound on its contribution; we do not prove a matching lower bound for this term.

Proof of Theorem 4. Let t_n, $\epsilon_n = e^{-t_n}$, ρ_n, n_1 and $\mathcal{T}_n$ be the tuning of Definition 7, and let $n \geq n_1$. By the assumption stated before Theorem 4, the minimizer $\hat{\varphi}_n$ exists and $\hat{\varphi}_n(\mathbf{c})$ is a measurable function of $\mathbf{c}$ and $(\tilde{\mathbf{a}}_i, \hat{\mathbf{b}}_i)_{i\leq n}$. Put $m_n \coloneqq \max\{3, \mathcal{N}(1/n, \mathcal{T}_n)\}$, and let N, $p = 2N+1$ be the output cutoff of $\mathcal{T}_n$. We use the properties (T1)–(T5) of Lemma 15, with the constants c_F and K_1 of Appendix C.4 and C' of Lemma 14. Apply Lemma 21 with $\mathcal{F} = \mathcal{T}_n$, whose members are continuous with entries bounded by Γ (Appendix C.2), $R = \rho_n$, $\delta = 1/n$ and $m_\delta = m_n$; a minimal δ-cover can be padded with repetitions to size m_n. Each term is bounded as follows.

Step 1: approximation and output truncation. Let $\bar{\varphi} \in \mathcal{T}_n$ be as in (T1). Since $\bar{\varphi}(\mathbf{c}) \in V_p$, we have $\|\bar{\varphi}(\mathbf{c}) - \Pi_p\tilde{\Psi}_0(\mathbf{c})\|_2 \leq \|\bar{\varphi}(\mathbf{c}) - \tilde{\Psi}_0(\mathbf{c})\|_2$, so (T1) gives $\mathcal{A} \leq \epsilon_n^2$. By Lemma 8(ii), $\mathcal{T} \leq \epsilon_n^2/16$.

Step 2: input discretization. By Lemmas 17 and 21, $9\mathcal{D} \leq C_u n_u^{-2\alpha q/(2\alpha+1)} + 72B_0^2 e^{-\rho_n/\kappa}$, with

$$C_u \coloneqq 9B_0^2\big[2\kappa^2\{(2c_K+3)C_\phi^2 + c_K^{-2\alpha}\}\big]^q.$$

By (T2), $e^{-\rho_n/\kappa} \leq e^{-2t_n}/(4c_F)$, and with $c_F \geq 2B_0^2$ from (T3), $72B_0^2 e^{-\rho_n/\kappa} \leq 72B_0^2\epsilon_n^2/(4c_F) \leq 9\epsilon_n^2$.

Step 3: noise level. By Lemmas 19 and 20, and since $n_v \geq 1$, $\sigma_\zeta^2 + b_\eta^2 = B_0^2 + C_\psi^2 p(B_\infty^2 + 4\kappa_\varepsilon^2)/n_v + 4\kappa_\varepsilon^2 C_\psi^2 p/n_v^2 \leq B_0^2 + C_\psi^2 p(B_\infty^2 + 8\kappa_\varepsilon^2)/n_v$. With the bound on p in (T1),

$$\sigma_\zeta^2 + b_\eta^2 \leq \bar{\sigma}^2 e^{t_n/\gamma}, \qquad \bar{\sigma}^2 \coloneqq B_0^2 + \frac{3(4\Gamma)^{1/\gamma} C_\psi^2 (B_\infty^2 + 8\kappa_\varepsilon^2)}{n_v}.$$

Since $p\Gamma^2 \leq (\Gamma\sqrt{p} + B_0)^2 \leq c_F e^{t_n/\gamma}$ by (T3), we obtain $p\Gamma^2 + \sigma_\zeta^2 + b_\eta^2 \leq (c_F + \bar{\sigma}^2)e^{t_n/\gamma}$. Moreover $1 + \Gamma + \sigma_\zeta + b_\eta \leq (1 + \Gamma + \sqrt{2}\,\bar{\sigma})e^{t_n/(2\gamma)}$, because $\sigma_\zeta + b_\eta \leq \{2(\sigma_\zeta^2 + b_\eta^2)\}^{1/2}$.

Step 4: calibration. By Step 3 and (T3), the statistical term of Lemma 21 is at most

$$\frac{c_F + \bar{\sigma}^2}{c_F} \cdot \frac{336\, c_F e^{t_n/\gamma}(\log m_n + 1)}{n} \leq \Big(1 + \frac{\bar{\sigma}^2}{c_F}\Big)\epsilon_n^2.$$

By Step 3, (T1), (T4), $K_1 \geq 1008 c_F \geq 6048\,\Gamma^2(4\Gamma)^{1/\gamma}$ and $\Gamma \geq 1$,

$$\begin{aligned}\frac{32p(1+\Gamma+\sigma_\zeta+b_\eta)}{n} &\leq \frac{96(4\Gamma)^{1/\gamma}(1+\Gamma+\sqrt{2}\,\bar{\sigma})e^{3t_n/(2\gamma)}}{K_1 e^{(2+2/\gamma)t_n}} \\ &\leq \frac{1+\Gamma+\sqrt{2}\,\bar{\sigma}}{63\,\Gamma^2}\,\epsilon_n^2 \leq \frac{2+\sqrt{2}\,\bar{\sigma}}{63}\,\epsilon_n^2.\end{aligned}$$

Collecting Steps 1–4, and using $8 + 9 + 1/16 + 1 + 2/63 < 18.1$,

$$\mathcal{E}(\hat{\Psi}_n) \leq C_0\,\epsilon_n^2 + C_u\, n_u^{-2\alpha q/(2\alpha+1)}, \qquad C_0 \coloneqq 18.1 + \frac{\bar{\sigma}^2}{c_F} + \frac{\sqrt{2}\,\bar{\sigma}}{63}.$$

The right side does not depend on Ψ_0, ν or the law of the noise within $\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$. By (T5), the theorem holds with $C = C_0 C'$, since $(\log n)^{-2\alpha\beta}(\log\log n)^{2\beta(\alpha+1)} = \{\log n/(\log\log n)^{1+1/\alpha}\}^{-2\alpha\beta}$. The tuning and n_1 do not depend on n_u, n_v or κ_ε, and K is the only encoding parameter that depends on n_u. The constant C_u does not depend on n_v or κ_ε, and $C = C_0 C'$ depends on them only through $\bar{\sigma}$, because C' and c_F are determined by the tuning. Since C_0 is increasing in $\bar{\sigma}$, and $\bar{\sigma}$ is decreasing in n_v and increasing in κ_ε, the constant C is nonincreasing in n_v and nondecreasing in κ_ε. □

The discretized estimator inherits the calibration of the fully observed one: the approximation, output truncation and complexity terms are controlled by the same properties (T1)–(T5) of the

tuning as in Appendix C.5. Because the projection keeps $\tilde{\mathbf{a}}$ in $\mathcal{B}^{2,\alpha}_{\rho_n}$, the approximation term is a supremum over this set, and the tail of ν enters through $\mathcal{D}$ instead, where (T2) and (T3) give $72B_0^2 e^{-\rho_n/\kappa} \le 9\epsilon_n^2$. Discretization therefore affects the bound only through the numerical constant, the noise level $\bar{\sigma}^2$, and the input term $C_u n_u^{-2\alpha q/(2\alpha+1)}$.

D MINIMAX LOWER BOUND

This section contains the proof of Theorem 3. We first derive several useful lemmas for establishing the minimax lower bound.

Lemma 22 (Varshamov–Gilbert packing)**.** *For any $\alpha, R > 0$ and $r \in (0, 24^{-\alpha}R/3]$, let $K = \lfloor 2^{-1}\{R/(3r)\}^{1/\alpha} \rfloor$ and $V_K = \mathrm{span}\{\phi_i : |i| \le K\}$, where $\lfloor a \rfloor := \max\{n \in \mathbb{Z} : n \le a\}$. Then, there exist an integer $m \ge 2^{(2K+1)/8} + 1 \ge 8$ and distinct points $x_1, x_2, \ldots, x_m \in \Omega^\alpha_R \cap V_K$ such that*

$$\|x_j - x_k\|_{\mathcal{X}} \ge r, \qquad \forall\, j \ne k.$$

Proof. Since $r \le 24^{-\alpha}R/3$, we have $K \ge 11$. According to the Varshamov–Gilbert bound (Tsybakov, 2009, Lemma 2.9), there exist an integer $M \ge 2^{(2K+1)/8}$ and distinct sequences $\theta_0, \theta_1, \ldots, \theta_M \in \{0,1\}^{2K+1}$ such that

$$\rho(\theta_j, \theta_k) \ge \frac{2K+1}{8}, \qquad \forall\, j \ne k,$$

where ρ is the Hamming distance. Take $m = M + 1$, we have $m \ge 2^{23/8} + 1 > 8$. For each $j \in \{1, 2, \ldots, m\}$, suppose that $\theta_{j-1} = (\theta_{j-1,-K}, \ldots, \theta_{j-1,K})$, where $\theta_{j-1,i} \in \{0,1\}$ for $-K \le i \le K$. Define $\tau_0 = r\sqrt{8/(2K+1)}$ and

$$x_j = \tau_0 \sum_{i=-K}^{K} \theta_{j-1,i}\phi_i.$$

Then, $x_j \in V_K$. Next, we show that $x_j \in \Omega^\alpha_R$. For each $i \in \mathbb{Z}$,

$$\langle \phi_i, x_j \rangle = \begin{cases} \tau_0\theta_{j-1,i}, & \text{if } |i| \le K, \\ 0, & \text{if } |i| > K. \end{cases}$$

Then,

$$\begin{aligned} \|x_j\|^2_{\mathcal{X}^\alpha} &= \sum_{i=-K}^{K} (1+|i|)^{2\alpha}\tau_0^2\theta^2_{j-1,i} \le \sum_{i=-K}^{K} (1+K)^{2\alpha}\tau_0^2 = (2K+1)(1+K)^{2\alpha}\tau_0^2 \\ &= 8(1+K)^{2\alpha}r^2. \end{aligned}$$

By definition of K, for $r \le 24^{-\alpha}R/3$,

$$(1+K)^\alpha \le \left\{1 + \frac{1}{2}\left(\frac{R}{3r}\right)^{1/\alpha}\right\}^\alpha \le \frac{R}{3r}.$$

Therefore,

$$\|x_j\|^2_{\mathcal{X}^\alpha} \le 8 \cdot \frac{R^2}{9r^2} \cdot r^2 = \frac{8}{9}R^2 < R^2,$$

which yields $x_j \in \Omega^\alpha_R \cap V_K$. It remains to show $\|x_j - x_k\|_{\mathcal{X}} \ge r$ for any $j \ne k$. To see this, note that

$$\begin{aligned} \|x_j - x_k\|^2_{\mathcal{X}} &= \tau_0^2 \sum_{i=-K}^{K} (\theta_{j-1,i} - \theta_{k-1,i})^2 = \tau_0^2 \sum_{i=-K}^{K} \mathbb{1}_{\{\theta_{j-1,i} \ne \theta_{k-1,i}\}} \\ &= \tau_0^2 \rho(\theta_{j-1}, \theta_{k-1}) \ge \tau_0^2 \frac{2K+1}{8} = r^2. \end{aligned}$$

This concludes the proof of Lemma 22. □

Lemma 23. *Suppose that $\alpha, R > 0$, and ν is a Borel probability measure on $\mathcal{X}$ with $\nu(\Omega_R^\alpha) = 1$. Then, $\nu \in \mathcal{P}_\alpha(\kappa)$ for any $\kappa \geq R/\log 2$.*

Proof. Note that

$$\mathbb{E}_{x\sim\nu}\big[\exp(\|x\|_{\mathcal{X}^\alpha}/\kappa)\big] = \int_{\Omega_R^\alpha} \exp(\|x\|_{\mathcal{X}^\alpha}/\kappa)\nu(dx) + \int_{\mathcal{X}\setminus\Omega_R^\alpha} \exp(\|x\|_{\mathcal{X}^\alpha}/\kappa)\nu(dx).$$

The second intergal equals zero since $\nu(\mathcal{X} \setminus \Omega_R^\alpha) = 0$. On Ω_R^α, we have $\exp(\|x\|_{\mathcal{X}^\alpha}/\kappa) \leq e^{R/\kappa}$. Therefore, for any $\kappa \geq R/\log 2$,

$$\mathbb{E}_{x\sim\nu}\big[\exp(\|x\|_{\mathcal{X}^\alpha}/\kappa)\big] \leq e^{R/\kappa}\nu(\Omega_R^\alpha) = e^{R/\kappa} \leq 2.$$

That said, $\nu \in \mathcal{P}_\alpha(\kappa)$. This concludes the proof of Lemma 23. □

Lemma 24. *Given $m \in \mathbb{Z}_+$ and $r \in (0, 1]$, suppose that $x_1, x_2, \ldots, x_m \in \mathcal{X}$ satisfies $\|x_j - x_k\|_{\mathcal{X}} \geq 3r$ for all $j \neq k$. For each $\theta = (\theta_1, \theta_2, \ldots, \theta_m) \in \{-1, 1\}^m$, define operator $\Psi_\theta : \mathcal{X} \to \mathcal{Y}$ by*

$$\Psi_\theta(x) = \Delta \sum_{j=1}^m \theta_j g\bigg(\frac{\|x - x_j\|_{\mathcal{X}}^2}{r^2}\bigg)\psi_0,$$

where $\Delta > 0$ and function $g : \mathbb{R} \to \mathbb{R}$ is given by

$$g(t) = \begin{cases} \exp\big(1 - \frac{1}{1-t}\big), & t < 1, \\ 0, & t \geq 1. \end{cases}$$

Then, for every $\theta \in \{-1, 1\}^m$,

(i) for any $\gamma > 0$, $\sup_{x\in\mathcal{X}} \|\Psi_\theta(x)\|_{\mathcal{Y}^\gamma} \leq \Delta$;

(ii) the operator Ψ_θ has arbitrary order Fréchet derivative;

(iii) for any $\beta, B_0 > 0$, if $\Delta \leq c_\beta B_0 r^\beta$, then $\Psi_\theta \in \mathcal{H}^\beta(\mathcal{X}, \mathcal{Y}; B_0)$, where $c_\beta > 0$ is a constant depending only on β.

Proof. For each $j \in \{1, 2, \ldots, m\}$, let $S_j = \{x \in \mathcal{X} : \|x - x_j\|_{\mathcal{X}} < r\}$ be the open ball centered at x_j and of radius r. Since $\|x_j - x_k\|_{\mathcal{X}} \geq 3r$ for all $k \neq j$, we have $S_j \cap S_k = \emptyset$. Let $u_j(x) = \|x - x_j\|_{\mathcal{X}}^2/r^2$. If $x \notin S_j$, then $u_j(x) \geq 1$ and $g^{(k)}(u_j(x)) = 0$ for every $k \in \mathbb{N}$. Consequently, for any $x \in \mathcal{X}$, there is at most one $j \in \{1, 2, \ldots, m\}$ such that $\mathcal{D}^k(g \circ u_j)(x) \neq 0$ for some $k \in \mathbb{N}$.

Fix $j \in \{1, 2, \ldots, m$, for any $x, h \in \mathcal{X}$, we have

$$\begin{aligned} u_j(x + h) &= \frac{\|x + h - x_j\|_{\mathcal{X}}^2}{r^2} = \frac{\|x - x_j\|_{\mathcal{X}}^2 + 2\langle x - x_j, h\rangle_{\mathcal{X}} + \|h\|_{\mathcal{X}}^2}{r^2} \\ &= u_j(x) + \frac{2}{r^2}\langle x - x_j, h\rangle_{\mathcal{X}} + \frac{\|h\|_{\mathcal{X}}^2}{r^2}. \end{aligned}$$

The map $h \mapsto 2\langle x - x_j, h\rangle_{\mathcal{X}}/r^2$ is a bounded linear functional and the remainder $\|h\|_{\mathcal{X}}^2/r^2 = o(\|h\|_{\mathcal{X}})$ as $\|h\|_{\mathcal{X}} \to 0$. Therefore, u_j is Fréchet differentiable at every x with

$$\mathcal{D}u_j(x)[h] = \frac{2}{r^2}\langle x - x_j, h\rangle_{\mathcal{X}}.$$

The map $x \mapsto \mathcal{D}u_j(x)$ is affine in x with linear part $x \mapsto (h \mapsto 2\langle x, h\rangle_{\mathcal{X}}/r^2)$, which is a bounded linear operator from $\mathcal{X}$ to $\mathcal{X}^*$. Therefore, u_j is twice Fréchet differentiable at every x with

$$\mathcal{D}^2u_j(x)[h_1, h_2] = \frac{2}{r^2}\langle h_1, h_2\rangle_{\mathcal{X}}, \qquad \forall\, h_1, h_2 \in \mathcal{X},$$

which is independent of x. Consequently, $\mathcal{D}^k u_j(x) \equiv 0$ for all $k \geq 3$. In particular, $u_j \in C^\infty(\mathcal{X})$, the class of infinitely Fréchet differentiable operators. By Cauchy-Schwarz inequality,

$|\mathcal{D}u_j(x)[h]|/\|h\|_{\mathcal{X}}$ attains its supremum in the direction $h = (x - x_j)/\|x - x_j\|_{\mathcal{X}}$ when $x \neq x_j$. Thus,

$$\|\mathcal{D}u_j(x)\|_{\mathcal{X}^*} = \frac{2\|x - x_j\|_{\mathcal{X}}}{r^2}.$$

This equation also holds when $x = x_j$. For $x \in \overline{S_l}$, we further have $\|\mathcal{D}u_j(x)\|_{\mathcal{X}^*} \leq 2/r$. Moreover, $\|\mathcal{D}^2 u_j(x)\|_{B^2(\mathcal{X},\mathbb{R})} = 2/r^2$.

Since $g \in C^\infty(\mathbb{R})$ and $u_j \in C^\infty(\mathcal{X})$, we have $g \circ u_j \in C^\infty(\mathcal{X})$. According to the Faà di Bruno Formula (Fraenkel, 1978, Formula A), for $k \geq 1$ and $h_1, h_2, \ldots h_k \in \mathcal{X}$,

$$\mathcal{D}^k(g \circ u_j)(x)[h_1, \ldots, h_k] = \sum_{\pi \in \mathcal{P}_k} g^{(|\pi|)}(u_j(x)) \prod_{B \in \pi} \mathcal{D}^{|B|} u_j(x) \bigotimes_{i \in B} h_i, \tag{7}$$

where $\mathcal{P}_k$ is the set of all partitions of $\{1, 2, \ldots, k\}$, $|\pi|$ is the number of blocks in the partition π, and $|B|$ is the cardinality of a block B. Note that $\mathcal{D}^p u_j(x) \equiv 0$ for all $p \geq 3$. If a partition π contains a block B with $|B| \geq 3$, the corresponding summand in the right-hand side of Eq. (7) equals zero. Therefore, it suffices to consider partitions composed of blocks of size 1 or 2. Suppose that such a partition has a blocks of size 1 and b blocks of size 2, then $a + 2b = k$ and $|\pi| = a + b = k - b$. The number of such partitions is

$$\frac{k!}{(k-2b)!b!2^b}.$$

For such a partition π, $x \in \overline{S_j}$, and $h_1, \ldots, h_k \in \{h \in \mathcal{X} : \|h\|_{\mathcal{X}} \leq 1\}$, we have

$$\left| \prod_{B \in \pi} \mathcal{D}^{|B|} u_j(x) \bigotimes_{i \in B} h_i \right| \leq \left(\frac{2}{r}\right)^a \left(\frac{2}{r^2}\right)^b = 2^{a+b} r^{-(a+2b)} = 2^{k-b} r^{-k}.$$

Let $G_k = \max_{1 \leq m \leq k} \left\|g^{(m)}\right\|_{L^\infty(\mathbb{R})}$. Then, for any $x \in \overline{S_j}$,

$$\|\mathcal{D}^k(g \circ u_j)(x)\|_{B^k(\mathcal{X},\mathbb{R})} \leq G_k \Lambda_k r^{-k}, \qquad \Lambda_k \coloneqq \sum_{b=0}^{\lfloor k/2 \rfloor} \frac{k! 2^{k-2b}}{(k-2b)!b!}.$$

For $x \in \mathcal{X} \setminus \overline{S_j}$, we have $\mathcal{D}^k(g \circ u_j)(x) \equiv 0$ since $g^{(p)}(u_j(x)) = 0$ for all $p \in \mathbb{N}$. For $k = 0$, we have $\sup_{x \in \mathcal{X}} \|g \circ u_j(x)\|_{\mathcal{X}} \leq 1$ since $u_j(x) \geq 0$. Define $G_0 = 1$. Then, for all $k \in \mathbb{N}$,

$$\|\mathcal{D}^k(g \circ u_j)(x)\|_{B^k(\mathcal{X},\mathbb{R})} \leq G_k \Lambda_k r^{-k},$$

Since each $g \circ u_j \in C^\infty(\mathcal{X})$, we have $\Psi_\theta \in C^\infty(\mathcal{X}, \mathcal{Y})$. This proves (ii). Furthermore,

$$\mathcal{D}^k \Psi_\theta(x)[h_1, \ldots, h_k] = \Delta \sum_{j=1}^m \theta_j \mathcal{D}^k(g \circ u_j)(x)[h_1, \ldots, h_k]\psi_0, \qquad \forall\, k \in \mathbb{N}.$$

For every $x \in \mathcal{X}$, at most one term in the summation is nonzero, say the one indexed by $j(x)$. If all terms equal zero, the following upper bounds hold trivially. Using $|\theta_{j(x)}| = 1$ and $\|\psi_0\|_{\mathcal{Y}} = 1$, we have

$$\|\mathcal{D}^k \Psi_\theta(x)\|_{B^k(\mathcal{X},\mathcal{Y})} = \Delta \|\mathcal{D}^k(g \circ u_{j(x)})(x)\|_{B^k(\mathcal{X},\mathbb{R})} \leq \Delta G_k \Lambda_k r^{-k}, \qquad \forall\, k \in \mathbb{N}. \tag{8}$$

To show (i), note that $\|\psi_0\|_{\mathcal{Y}^\gamma} = 1$ for any $\gamma > 0$, which yields

$$\|\Psi_\theta(x)\|_{\mathcal{Y}^\gamma} = |\Delta \theta_{j(x)} g \circ u_{j(x)}(x)| \leq \Delta, \qquad \forall\, x \in \mathcal{X}.$$

It remains to show (iii). Let $\beta = s + \rho > 0$ with $s \in \mathbb{N}$ and $\rho \in (0, 1]$. Taking $k = s + 1$ in Eq. (8), we have that the map $\mathcal{D}^s \Psi_\theta : \mathcal{X} \to B^s(\mathcal{X}, \mathcal{Y})$ is Fréchet differentiable with $\sup_{x \in \mathcal{X}} \|\mathcal{D}^{s+1} \Psi_\theta(x)\|_{B^{s+1}(\mathcal{X},\mathcal{Y})} \leq \Delta G_{s+1} \Lambda_{s+1} r^{-(s+1)}$. Since $\mathcal{X}$ is convex, according to the mean value inequality (Abraham et al., 1988, Inequality 2.4.8), for all $x_1, x_2 \in \mathcal{X}$,

$$\|\mathcal{D}^s \Psi_\theta(x_2) - \mathcal{D}^s \Psi_\theta(x_1)\|_{B^s(\mathcal{X},\mathcal{Y})} \leq \Delta G_{s+1} \Lambda_{s+1} r^{-(s+1)} \|x_2 - x_1\|_{\mathcal{X}}.$$

For $x_1 \neq x_2$, we consider two cases.

Case (a): $0 < \|x_2 - x_1\|_{\mathcal{X}} \le r$. Since $\rho \in (0, 1]$, we have $\|x_2 - x_1\|_{\mathcal{X}}^{1-\rho} \le r^{1-\rho}$. It follows that

$$\begin{aligned}\|\mathcal{D}^s\Psi_\theta(x_2) - \mathcal{D}^s\Psi_\theta(x_1)\|_{B^s(\mathcal{X},\mathcal{Y})} &\le \Delta G_{s+1}\Lambda_{s+1} r^{-(s+1)} r^{1-\rho}\|x_2 - x_1\|_{\mathcal{X}}^{\rho} \\ &= \Delta G_{s+1}\Lambda_{s+1} r^{-\beta}\|x_2 - x_1\|_{\mathcal{X}}^{\rho}.\end{aligned}$$

Case (b): $\|x_2 - x_1\|_{\mathcal{X}} > r$. Taking $k = s$ in Eq. 8, we have

$$\begin{aligned}&\|\mathcal{D}^s\Psi_\theta(x_2) - \mathcal{D}^s\Psi_\theta(x_1)\|_{B^s(\mathcal{X},\mathcal{Y})} \le 2\Delta G_s\Lambda_s r^{-s} \\ &\le 2\Delta G_s\Lambda_s r^{-s} r^{-\rho}\|x_2 - x_1\|_{\mathcal{X}}^{\rho} = 2\Delta G_s\Lambda_s r^{-\beta}\|x_2 - x_1\|_{\mathcal{X}}^{\rho}\end{aligned}$$

Combine both cases, we have

$$\sup_{x_1 \ne x_2} \frac{\|\mathcal{D}^s\Psi_\theta(x_2) - \mathcal{D}^s\Psi_\theta(x_1)\|_{B^s(\mathcal{X},\mathcal{Y})}}{\|x_2 - x_1\|_{\mathcal{X}}^{\rho}} \le \Delta \max\{G_{s+1}\Lambda_{s+1}, 2G_s\Lambda_s\} r^{-\beta}. \tag{9}$$

Moreover, for $0 \le k \le s$ and $r \in (0, 1]$, we have $r^{-k} \le r^{-\beta}$. Therefore, Eq. (8) yields

$$\max_{0\le k\le s} \sup_{x\in\mathcal{X}} \|\mathcal{D}^k\Psi_\theta(x)\|_{B^k(\mathcal{X},\mathcal{Y})} \le \Delta \max_{0\le k\le s}\{G_k\Lambda_k\} r^{-\beta}. \tag{10}$$

Define c_β by

$$c_\beta = \frac{1}{\max\{G_0\Lambda_0, G_1\Lambda_1, \ldots, G_{s+1}\Lambda_{s+1}, 2G_s\Lambda_s\}}.$$

Then, $c_\beta \in (0, 1]$ and depends only on β. If $\Delta \le c_\beta B_0 r^\beta$, the right-hand sides of both Eq. (9) and Eq. (10) are upper bounded by B_0. Therefore, $\Psi_\theta \in \mathcal{H}^\beta(\mathcal{X}, \mathcal{Y}; B_0)$. This proves (iii) and concludes the proof of Lemma 24. □

D.1 Proof of Theorem 3

Proof. The proof builds on Theorem 2.5 in Tsybakov (2009), a lower bound on the minimax probability of error for hypothesis testing problems. To apply this result, we construct $(M+1)$ hypotheses $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(M)}$ using Lemma 24 and the Varshamov–Gilbert bound. The proof consists of four steps.

Step 1. Construction.
Let $R = \kappa \log 2$. For any $0 < r \le \min\{1, 24^{-\alpha}R/9\}$, let $K = \lfloor 2^{-1}\{R/(9r)\}^{1/\alpha} \rfloor$. According to Lemma 22, there exist an integer $m \ge 2^{(2K+1)/8} + 1 \ge 8$ and $x_1, x_2, \ldots, x_m \in \Omega_R^\alpha$ such that $\|x_j - x_k\|_{\mathcal{X}} \ge 3r$, for all $j \ne k$. Let $\nu = m^{-1}\sum_{j=1}^m \delta_{x_j}$. Then, $\nu(\Omega_R^\alpha) = 1$, and Lemma 23 gives $\nu \in \mathcal{P}_\alpha(\kappa)$.

Let $\{\Psi_\theta : \theta \in \{-1, 1\}^m\}$ be the operators defined in Lemma 24 based on $x_1, \ldots, x_m$, and c_β be the constant that depends only on β. Take $c_{\beta,B} = \min\{c_\beta B_0, B_{0,\gamma}\}$ and $\Delta = c_{\beta,B} r^\beta$, we have $\Psi_\theta \in \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma})$ for every $\theta \in \{-1, 1\}^m$.

Since $m \ge 8$, by the Varshamov–Gilbert bound (Tsybakov, 2009, Lemma 2.9), there exist integer $M \ge 2^{m/8}$ and a subset $\{\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(M)}\} \subseteq \{-1, 1\}^m$ such that

$$\rho\big(\theta^{(j)}, \theta^{(k)}\big) \ge \frac{m}{8}, \qquad \forall\, j \ne k.$$

Let $\varepsilon = Z\psi_0$, where $Z \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = \kappa_\varepsilon/2$. Then, $\mathbb{E}[\varepsilon] = 0$, and

$$\mathbb{E}\left[\exp\left(\frac{\|\varepsilon\|_{\mathcal{Y}}}{\kappa_\varepsilon}\right)\right] = \mathbb{E}\left[\exp\left(\frac{|Z|}{\kappa_\varepsilon}\right)\right] = 2\exp\left(\frac{\sigma^2}{2\kappa_\varepsilon^2}\right)\Phi\left(\frac{\sigma}{\kappa_\varepsilon}\right) = 2\exp(1/8)\Phi(1/2) < 2,$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. That said, $\nu_\varepsilon \coloneqq \mathrm{Law}(\varepsilon)$ belongs to the class of mean-zero sub-exponential distributions $\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$ on $\mathcal{Y}$.

Given a Borel measurable Ψ_θ, let P_θ be the Borel probability measure on $\mathcal{X} \times \mathcal{Y}$ defined by

$$P_\theta(dx, dy) = \nu(dx)\nu_\varepsilon(dy - \Psi_\theta(x)),$$

that is, the law of (x, y) with $x \sim \nu$, $y = \Psi_\theta(x) + \varepsilon$, $\varepsilon \sim \nu_\varepsilon$, and $\varepsilon \perp\!\!\!\perp x$. Then, n i.i.d. copies of (x, y) has law $P_\theta^{\otimes n}$ on $(\mathcal{X}, \mathcal{Y})^n$. For ease of notation, we abbreviate P_θ^n for $P_\theta^{\otimes n}$.

Step 2. Derive bounds for $\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}$ and $\mathrm{KL}(P_\theta^n \parallel P_{\theta'}^n)$.
In this step, we derive some useful inequalities for $\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}$ and $\mathrm{KL}(P_\theta^n \parallel P_{\theta'}^n)$, which are subsequently used to verify the conditions in Theorem 2.5 in Tsybakov (2009).

Note that $\nu = m^{-1}\sum_{j=1}^m \delta_{x_j}$ and $\Psi_\theta(x_j) = \Delta\theta_j\psi_0$, we have

$$\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}^2 = \frac{1}{m}\sum_{j=1}^m \|\Psi_\theta(x_j) - \Psi_{\theta'}(x_j)\|_{\mathcal{Y}}^2 = \frac{\Delta^2}{m}\sum_{j=1}^m |\theta_j - \theta_j'|^2.$$

Since $\theta_j, \theta_j' \in \{-1, 1\}$, we have $|\theta_j - \theta_j'|^2 = 4\mathbb{1}_{\{\theta_j \neq \theta_j'\}}$ and

$$\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}^2 = \frac{4\Delta^2}{m}\rho(\theta, \theta').$$

Therefore, $\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}^2 \le 4\Delta^2$. For $0 \le j < k \le M$, we have $\rho\big(\theta^{(j)}, \theta^{(k)}\big) \ge m/8$, which yields $\|\Psi_{\theta^{(j)}} - \Psi_{\theta^{(k)}}\|_{L^2(\nu)}^2 \ge \Delta^2/2$.

Next, we bound $\mathrm{KL}(P_\theta^n \parallel P_{\theta'}^n)$. Note that $P_\theta(dx, dy) = \nu(dx)P_\theta(dy \mid x)$ with $P_\theta(dy \mid x) = \nu_\varepsilon(dy - \Psi_\theta(x))$. Let

$$a_\theta(x) = \Delta\sum_{j=1}^m \theta_j g\bigg(\frac{\|x - x_j\|_{\mathcal{X}}^2}{r^2}\bigg),$$

where the function $g : \mathbb{R} \to \mathbb{R}$ is defined in Lemma 24. Then, $\Psi_\theta(x) = a_\theta(x)\psi_0$. Also note that $\varepsilon = Z\psi_0$ with $Z \sim \mathcal{N}(0, \sigma^2)$, we have $P_\theta(\cdot \mid x) = (t \mapsto t\psi_0)_\#\mathcal{N}(a_\theta(x), \sigma^2)$. Since the map $t \mapsto t\psi_0$ is a measurable bijection from $\mathbb{R}$ onto $\mathrm{span}\{\psi_0\}$, we have

$$\mathrm{KL}\big(P_\theta(\cdot \mid x) \parallel P_{\theta'}(\cdot \mid x)\big) = \mathrm{KL}\big(\mathcal{N}(a_\theta(x), \sigma^2) \parallel \mathcal{N}(a_{\theta'}(x), \sigma^2)\big)$$
$$= \frac{|a_\theta(x) - a_{\theta'}(x)|^2}{2\sigma^2} = \frac{\|\Psi_\theta(x) - \Psi_{\theta'}(x)\|_{\mathcal{Y}}^2}{2\sigma^2}.$$

By chain rule of KL divergence,

$$\mathrm{KL}(P_\theta \parallel P_{\theta'}) = \int_{\mathcal{X}} \mathrm{KL}\big(P_\theta(\cdot \mid x) \parallel P_{\theta'}(\cdot \mid x)\big)\nu(dx)$$
$$= \frac{1}{2\sigma^2}\int_{\mathcal{X}} \|\Psi_\theta(x) - \Psi_{\theta'}(x)\|_{\mathcal{Y}}^2\nu(dx) = \frac{\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}^2}{2\sigma^2}$$

Furthermore, by definition of P_θ^n, we have $\mathrm{KL}(P_\theta^n \parallel P_{\theta'}^n) = n\mathrm{KL}(P_\theta \parallel P_{\theta'})$. Using the upper bound $\|\Psi_\theta - \Psi_{\theta'}\|_{L^2(\nu)}^2 \le 4\Delta^2$ and $\sigma = \kappa_\varepsilon/2$, we obtain $\mathrm{KL}(P_\theta^n \parallel P_{\theta'}^n) \le 8n\Delta^2/\kappa_\varepsilon^2$. In particular,

$$\frac{1}{M}\sum_{j=1}^M \mathrm{KL}(P_{\theta^{(j)}}^n \parallel P_{\theta^{(0)}}^n) \le \frac{8n\Delta^2}{\kappa_\varepsilon^2}.$$

Step 3. Establish minimax lower bound.
Define

$$C_0 = \frac{1024c_{\beta,B}^2}{\kappa_\varepsilon^2 \log 2}, \qquad T_n = \log n + \log C_0.$$

If $n \ge n_1 \coloneqq C_0^{-1}2^{12/5}$, we have $T_n \ge (12/5)\log 2 > 1$. For such n, set

$$r = \frac{R}{9}\bigg(\frac{\log 2}{10T_n}\bigg)^\alpha.$$

Then, $r \le 24^{-\alpha}R/9$. If further

$$n \ge n_2 \coloneqq C_0^{-1}\exp\bigg(\frac{\log 2}{10}(R/9)^{1/\alpha}\bigg),$$

we have $r \leq 1$. That said, the choice of r is valid for all $n \geq n_1 \vee n_2$. In this circumstances,

$$K = \left\lfloor 2^{-1}\left(\frac{R}{9r}\right)^{1/\alpha} \right\rfloor = \left\lfloor \frac{5T_n}{\log 2} \right\rfloor \geq \frac{5T_n}{\log 2} - 1 > \frac{4T_n}{\log 2}.$$

It follows that

$$m \geq 2^{(2K+1)/8} + 1 > 2^{K/4} > e^{T_n} = C_0 n.$$

Also note that $M \geq 2^{m/8}$, we have $\log M \geq (m/8)\log 2$. Consequently,

$$\frac{8n\Delta^2}{\kappa_\varepsilon^2} \leq \frac{8nc_{\beta,B}^2}{\kappa_\varepsilon^2} \leq \frac{8mc_{\beta,B}^2}{C_0\kappa_\varepsilon^2} \leq \frac{64c_{\beta,B}^2}{C_0\kappa_\varepsilon^2 \log 2}\log M = \frac{1}{16}\log M.$$

Moreover, let $s_n = 2^{-3/2}\Delta$, then $\|\Psi_{\theta^{(j)}} - \Psi_{\theta^{(k)}}\|_{L^2(\nu)} \geq 2^{-1/2}\Delta = 2s_n$.

According to Theorem 2.5 in Tsybakov (2009), we have

$$\inf_{\hat{\Psi}} \max_{0\leq j\leq M} \mathbb{P}_{\theta^{(j)}}\left(\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)} \geq s_n\right) \geq p_0(M),$$

with

$$p_0(M) = \frac{\sqrt{M}}{1+\sqrt{M}}\left(1 - \frac{1}{8} - \sqrt{\frac{1}{8\log M}}\right).$$

Since both $\sqrt{M}/(1+\sqrt{M})$ and $\{1 - 1/8 - (8\log M)^{-1/2}\}$ are positive and increasing for $M \geq 2$, we have

$$p_0(M) \geq p_0(2) = \frac{\sqrt{2}}{1+\sqrt{2}}\left(1 - \frac{1}{8} - \sqrt{\frac{1}{8\log 2}}\right) > \frac{1}{4}.$$

For each j, applying Markov's inequality to $\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)}^2$ gives

$$\mathbb{E}_{P^n_{\theta^{(j)}}}\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)}^2 \geq s_n^2 \mathbb{P}_{\theta^{(j)}}\left(\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)} \geq s_n\right).$$

Taking maximum over j and infimum over possible estimators, we have

$$\inf_{\hat{\Psi}} \max_{0\leq j\leq M} \mathbb{E}_{P^n_{\theta^{(j)}}}\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)}^2 \geq s_n^2 \inf_{\hat{\Psi}} \max_{0\leq j\leq M} \mathbb{P}_{\theta^{(j)}}\left(\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)} \geq s_n\right) \geq \frac{s_n^2}{4}.$$

According to Step 1, for each j, the triple $(\nu, \Psi_{\theta^{(j)}}, \nu_\varepsilon)$ belongs to $\mathcal{P}_\alpha(\kappa) \times \mathcal{H}^{\beta,\gamma}(B_0, B_{0,\gamma}) \times \mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)$. Therefore,

$$\inf_{\hat{\Psi}} \sup_{\substack{\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0,B_{0,\gamma}),\\ \nu\in\mathcal{P}_\alpha(\kappa),\ \nu_\varepsilon\in\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)}} \mathbb{E}\left\|\hat{\Psi} - \Psi_{\theta^{(j)}}\right\|_{L^2(\nu)}^2 \geq \frac{s_n^2}{4}.$$

It remains to evaluate the order of s_n^2. If $n \geq C_0$, then $T_n \leq 2\log n$, which yields

$$\frac{s_n^2}{4} = \frac{\Delta^2}{32} = \frac{c_{\beta,B}^2}{32} r^{2\beta} \geq \frac{c_{\beta,B}^2}{32}\left(\frac{R}{9}\right)^{2\beta}\left(\frac{\log 2}{20}\right)^{2\alpha\beta}(\log n)^{-2\alpha\beta} \coloneqq c_0(\log n)^{-2\alpha\beta}.$$

Consequently, for any $n \geq n_0 \coloneqq \max\{n_1, n_2, C_0\}$,

$$\inf_{\hat{\Psi}} \sup_{\substack{\Psi_0 \in \mathcal{H}^{\beta,\gamma}(B_0,B_{0,\gamma}),\\ \nu\in\mathcal{P}_\alpha(\kappa),\ \nu_\varepsilon\in\mathcal{P}^{\mathrm{c}}(\kappa_\varepsilon)}} \mathbb{E}\left\|\hat{\Psi} - \Psi_0\right\|_{L^2(\nu)}^2 \geq c_0(\log n)^{-2\alpha\beta}.$$

Since $R = \kappa\log 2$, n_0 and c_0 depend only on α, β, κ, κ_ε, B_0, and $B_{0,\gamma}$. This concludes the proof of Theorem 3. □

E LEMMAS FOR APPROXIMATION ERROR ANALYSIS

This section presents several existing and derived results that assist our approximation error analysis.

Lemma 25 (Lemma C.1 in Takakura & Suzuki (2023)). *For $\theta \in \mathbb{R}^d$, suppose that there exist an index $i^* \in \{1, 2, \ldots, d\}$ and $\delta > 0$ such that $\theta_{i^*} \geq \theta_i + \delta$ for all $i \neq i^*$. Then,*

$$\|\mathrm{Softmax}(\theta) - e_{i^*}\|_1 \leq 2de^{-\delta}.$$

Lemma 26 (Lemma A.2 in Schmidt-Hieber (2020)). *For $d \in \mathbb{Z}_+$ and any $\epsilon > 0$, there exists a feedforward neural network $\phi_\times \in \mathcal{F}(L, W, S, B)$ with depth $L = \lceil \log_2(d^2/\epsilon) + 5 \rceil \lceil \log_2 d \rceil$, width $W = 6d$, and weight-norm bound $B = 1$ such that*

$$\sup_{\mathbf{x} \in [0,1]^d} \left| \phi_\times(\mathbf{x}) - \prod_{i=1}^d x_i \right| \leq \epsilon.$$

Moreover, $\phi_\times(\mathbf{x}) = 0$ whenever $\prod_{i=1}^d x_i = 0$.

Lemma 27 (Extension of Lemma 26). *For $d \in \mathbb{Z}_+$ and any $\epsilon > 0$, there exists a feedforward neural network $\phi_{\times,[-1,1]} \in \mathcal{F}(L, W, S, B)$ with depth $L = \lceil \log_2(d^2/\epsilon) + d + 5 \rceil \lceil \log_2 d \rceil + 1$, width $W = 6d2^d$, and weight-norm bound $B = 1$ such that*

$$\sup_{\mathbf{x} \in [-1,1]^d} \left| \phi_{\times,[-1,1]}(\mathbf{x}) - \prod_{i=1}^d x_i \right| \leq \epsilon.$$

Moreover, $\phi_{\times,[-1,1]}(\mathbf{x}) = 0$ whenever $\prod_{i=1}^d x_i = 0$.

Proof. Note that $x = \sigma(x) - \sigma(-x)$ ($x \in \mathbb{R}$), we have

$$\prod_{i=1}^d x_i = \prod_{i=1}^d \{\sigma(x_i) - \sigma(-x_i)\} = \sum_{\mathbf{s} \in \{-1,1\}^d} \left\{ a_{\mathbf{s}} \prod_{i=1}^d \sigma(s_i x_i) \right\},$$

where $a_{\mathbf{s}} \coloneqq \prod_{i=1}^d s_i \in \{-1, 1\}$. The vector $[\sigma(x_1), \sigma(-x_1), \ldots, \sigma(x_d), \sigma(-x_d)]$ can be computed by an FNN with depth 1, width $2d$, and weight-norm bound 1. For each $x_i \in [-1, 1]$ and $s_i \in \{-1, 1\}$, we have $\sigma(s_i x_i) \in [0, 1]$. According to Lemma 26, for any $\mathbf{s} \in \{-1, 1\}^d$ and $\epsilon > 0$, there exists an FNN with depth $\lceil \log_2(d^2/\epsilon) + d + 5 \rceil \lceil \log_2 d \rceil + 1$, width $6d$, and weight-norm bound 1 such that

$$\left| \phi_{\mathbf{s}}(\mathbf{x}) - \prod_{i=1}^d \sigma(s_i x_i) \right| \leq \frac{\epsilon}{2^d}.$$

Define

$$\phi_{\times,[-1,1]}(\mathbf{x}) = \sum_{\mathbf{s} \in \{-1,1\}^d} a_{\mathbf{s}} \phi_{\mathbf{s}}(\mathbf{x}),$$

then $\phi_{\times,[-1,1]}$ can be implemented by an FNN with depth $\lceil \log_2(d^2/\epsilon) + d + 5 \rceil \lceil \log_2 d \rceil + 1$, width $6d2^d$, and weight-norm bound 1. Moreover,

$$\left| \phi_{\times,[-1,1]}(\mathbf{x}) - \prod_{i=1}^d x_i \right| \leq \sum_{\mathbf{s} \in \{-1,1\}^d} \left| \phi_{\mathbf{s}}(\mathbf{x}) - \prod_{i=1}^d \sigma(s_i x_i) \right| \leq \epsilon.$$

In addition, if $x_i = 0$, then $\sigma(x_i) = \sigma(-x_i) = 0$. By Lemma 26, for all $\mathbf{s} \in \{-1, 1\}^d$, we have $\phi_{\mathbf{s}}(\mathbf{x}) = 0$. Consequently, $\phi_{\times,[-1,1]}(\mathbf{x}) = 0$ whenever there exists index i such that $x_i = 0$. This concludes the proof of Lemma 27. □

Lemma 28 (Taylor remainder bound for Fréchet-Hölder smooth operators). *Let $\mathcal{X}, \mathcal{Y}$ be Banach spaces, $\Omega \subseteq \mathcal{X}$ be an open convex set, and $\Psi \in \mathcal{H}^\beta(\Omega, \mathcal{Y}; B_0)$ for some $\beta = s + r > 0$ with $s = \lfloor \beta \rfloor \in \mathbb{N}$ and $r \in (0, 1]$. For any $x, x_0 \in \Omega$, we have*

$$\left\| \Psi(x) - \sum_{k=0}^s \frac{1}{k!} \mathcal{D}^k \Psi(x_0)\left[(x - x_0)^k\right] \right\|_{\mathcal{Y}} \leq \frac{B_0}{s!} \|x - x_0\|_{\mathcal{X}}^\beta.$$

Proof. The case of $s = 0$ is trivial. In the following, we consider $s \geq 1$. By Theorem 5.12 in Kriegl & Michor (1997), any s-times continuously Fréchet differentiable operator $\Psi : \mathcal{X} \to \mathcal{Y}$ admits the Taylor formula with integral remainder

$$\Psi(x) = \sum_{k=0}^{s} \frac{1}{k!} \mathcal{D}^k \Psi(x_0)\big[(x - x_0)^k\big] + R_s(x; x_0),$$

where

$$R_s(x; x_0) = \int_0^1 \frac{(1-t)^{s-1}}{(s-1)!} \big\{\mathcal{D}^s \Psi(x_0 + t(x - x_0)) - \mathcal{D}^s \Psi(x_0)\big\}\big[(x - x_0)^s\big] dt.$$

Applying the r-Hölder continuity of the s-th order Fréchet derivative $\mathcal{D}^s \Psi$, we have

$$\big\|\mathcal{D}^s \Psi(x_0 + t(x - x_0)) - \mathcal{D}^s \Psi(x_0)\big\|_{B^s(\mathcal{X},\mathcal{Y})} \leq B_0 \|t(x - x_0)\|_{\mathcal{X}}^r = B_0 t^r \|x - x_0\|_{\mathcal{X}}^r.$$

Therefore,

$$\begin{aligned}
\|R_s(x; x_0)\|_{\mathcal{Y}} &\leq \int_0^1 \frac{(1-t)^{s-1}}{(s-1)!} \big\|\mathcal{D}^s \Psi(x_0 + t(x - x_0)) - \mathcal{D}^s \Psi(x_0)\big[(x - x_0)^s\big]\big\|_{\mathcal{Y}} dt \\
&\leq \int_0^1 \frac{(1-t)^{s-1}}{(s-1)!} \big(B_0 t^r \|x - x_0\|_{\mathcal{X}}^r\big) \|x - x_0\|_{\mathcal{X}}^s dt \\
&= \frac{B_0 \|x - x_0\|_{\mathcal{X}}^{\beta}}{(s-1)!} \int_0^1 (1-t)^{s-1} t^r dt.
\end{aligned}$$

Based on the Beta function identity

$$\int_0^1 (1-t)^{s-1} t^r dt = B(s, r+1) = \frac{\Gamma(s)\Gamma(r+1)}{\Gamma(s+r+1)},$$

we have

$$\|R_s(x; x_0)\|_{\mathcal{Y}} \leq \frac{B_0 \Gamma(r+1)}{\Gamma(s+r+1)} \|x - x_0\|_{\mathcal{X}}^{\beta}.$$

For $r \in (0, 1]$, we have $\Gamma(r+1) \leq 1$ and $\Gamma(s+r+1) \geq \Gamma(s+1) = s!$. Consequently,

$$\|R_s(x; x_0)\|_{\mathcal{Y}} \leq \frac{B_0}{s!} \|x - x_0\|_{\mathcal{X}}^{\beta}.$$

This concludes the proof of Lemma 28. □

F Experimental details and additional results

All models were trained on a single NVIDIA RTX 4060 Laptop GPU (8 GB). Test functions are used only for evaluation, and checkpoints are selected by validation risk. Uncertainties are sample standard deviations over repeats, and what a repeat varies depends on the study:

- KS, the elliptic operator and reaction-diffusion: a repeat draws a new training pool and a new initialization, and, for discretized observations, new observation locations.
- Burgers at $n = 1000$: repeats share one training set and differ only in initialization.

F.1 Common implementation

Coefficient maps. For every task we fix orthonormal bases of the input and output spaces and compute $J_{\mathcal{X}} x$ and $J_{\mathcal{Y}} y$ by quadrature on the solver grid. Periodic problems use real Fourier bases; nonperiodic problems use cosine (or sine) bases, and trajectories use tensor products with Legendre polynomials in time. The network output is mapped back through the fixed output basis, so every prediction is an element of $\mathrm{span}\{\psi_i : |i| \leq N\}$. Training minimizes the mean squared error of the reconstructed function on the grid. The Fourier and cosine bases are orthonormal for the grid inner product, so for them this loss equals the retained-coefficient loss plus the φ-independent projection residual of the targets. This is the empirical form of the splitting used in Section 4. For

the space-time trajectories, the Legendre time basis is not orthonormal on the time grid. Coefficients are therefore computed by weighted least squares with the quadrature Gram matrix, and the same splitting holds in the quadrature inner product. With discretized observations (Appendix F.3) the functions are not available on a grid, and the loss is the squared error of the coefficients. Inputs are scaled with training statistics: coordinatewise for KS and the elliptic operator, and by a single scalar for the trajectory problems. Outputs are centered and scaled with training statistics (for the Transformers of Section 5, coordinatewise centering and one pooled scale; for DeepONet, pointwise on the grid), and all errors are computed after undoing the scaling.

KS networks. The Transformer embeds coefficient a_i as the token $a_i w_i + e_i \in \mathbb{R}^{96}$. Here w_i and e_i are learned for each coefficient index, and e_i plays the role of the positional encoding P_i. Three pre-LayerNorm blocks follow (width 96, four heads, GELU feedforward width 384), then a final LayerNorm, mean pooling over tokens, and a $96 \to 256 \to 127$ readout. The prediction adds the input mean, which the KS flow conserves. The model has 417,567 parameters. DeepONet uses a branch network on the 127 input coefficients (three hidden layers of width 256, GELU) and a trunk network on the periodic features $\{\cos(2\pi kx/64), \sin(2\pi kx/64)\}_{k=1}^{8}$ (two hidden layers of width 256), with rank 256. It adds the input field as a residual connection and has 366,080 parameters. The shared-embedding variant of Figure 1(a) and Appendix F.4 replaces $a_i w_i + e_i$ by $\eta a_i + P_i$, with one learned $\eta \in \mathbb{R}^{96}$ and the sinusoidal encoding P_i of Vaswani et al. (2017), with four heads, or sixteen heads of dimension 6 as in Figure 1(a); it has 393,279 parameters for either head count.

Optimization. All KS models use AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-3} and weight decay 10^{-4}, batch size 32, 500 warm-up steps, cosine decay to 2% of the initial rate, and gradient clipping at norm 1. The loss is the grid mean squared error divided by the variance of the training targets. Each minibatch applies a random joint periodic translation to the input and output. The KS operator commutes with translations. The validation risk on 512 fixed validation functions is computed every 1000 updates, and the best checkpoint is kept.

F.2 Kuramoto-Sivashinsky endpoint operator

Data. Initial fields are $\sum_{k=1}^{10} A_k \sin(2\pi f_k x/64 + \phi_k)$ with $A_k \sim \mathrm{U}(-\frac{1}{2}, \frac{1}{2})$, $f_k \sim \mathrm{U}\{1, 2\}$ and $\phi_k \sim \mathrm{U}(0, 2\pi)$, all independent, sampled on 256 grid points. They are evolved for 72 time units; the resulting states are the inputs x, and the states 4 time units later are the outputs $\Psi_0(x)$. The solver is ETDRK4 (Cox & Matthews, 2002; Kassam & Trefethen, 2005) in float64 with time step 0.05 and two-thirds dealiasing. On four functions, halving the time step changes the output by a relative 3.3×10^{-6} and doubling the grid by 4.4×10^{-12}, both far below the learning errors. Ten independent pools of 8192 training functions are drawn. The seven training sizes are nested subsets of each pool, with the same subsets used for both methods. A common set of 1024 test functions is used for all fits. This KS setting follows the chaotic regime used in neural PDE solver benchmarks (Brandstetter et al., 2022; Lippe et al., 2023). Tables 1 and 3 use the Transformer with learned tokens. The curve of Figure 1(a) and Table 4 uses five repeats, each with its own pool of 8192 training functions and nested training sets shared by both methods, and the 2048 test functions of Table 2.

Table 1: KS endpoint operator, Transformer with learned tokens: complete sample-size results (48k updates; mean $\pm$ SD over ten paired repeats; 1024 common test functions). Risk is the squared L^2 error with respect to $dx/64$.

	Test risk ($\times 10^{-3}$)		Relative L^2 error (%)	
n	Transformer	DeepONet	Transformer	DeepONet
128	26.56 ± 3.16	24.71 ± 1.25	11.86 ± 0.65	11.59 ± 0.30
256	8.53 ± 0.45	10.30 ± 0.44	6.75 ± 0.16	7.48 ± 0.16
512	4.91 ± 0.20	7.28 ± 0.25	5.18 ± 0.10	6.30 ± 0.10
1024	3.79 ± 0.21	6.31 ± 0.15	4.56 ± 0.12	5.86 ± 0.06
2048	3.33 ± 0.21	5.76 ± 0.15	4.29 ± 0.14	5.62 ± 0.07
4096	3.13 ± 0.18	5.58 ± 0.08	4.16 ± 0.11	5.53 ± 0.04
8192	3.07 ± 0.20	5.50 ± 0.13	4.13 ± 0.13	5.50 ± 0.06

Training budget. We retrained both methods from scratch for 96k updates on three of the training pools with $n \in \{512, 2048, 8192\}$, with the cosine schedule stretched to the longer horizon. All 36 models (the 18 original 48k fits and the 18 new ones) were evaluated on a second common set of 2048 test functions. Table 2 shows that the longer budget lowers the risk in all 18 paired comparisons. Under both budgets the risk decreases in n. Over the last 12.5% of updates the validation risk still decreases (by 12–21% for the Transformer and 6–9% for DeepONet). The achieved risk therefore includes optimization error at both budgets.

Table 2: KS training-budget control (three paired repeats; 2048 common test functions). "Reduction" is the mean over repeats of the paired relative decrease of the test risk from 48k to 96k updates.

		Test risk ($\times 10^{-3}$)			Relative L^2 error (%)	
n	Method	48k	96k	Reduction	48k	96k
512	Transformer	4.84 ± 0.18	2.59 ± 0.15	46.5%	5.17 ± 0.09	3.76 ± 0.10
512	DeepONet	7.22 ± 0.18	4.06 ± 0.12	43.8%	6.30 ± 0.05	4.71 ± 0.05
2048	Transformer	3.29 ± 0.20	1.65 ± 0.12	49.8%	4.27 ± 0.13	3.02 ± 0.11
2048	DeepONet	5.82 ± 0.05	3.11 ± 0.02	46.6%	5.66 ± 0.03	4.14 ± 0.01
8192	Transformer	3.13 ± 0.12	1.51 ± 0.07	51.8%	4.17 ± 0.08	2.89 ± 0.07
8192	DeepONet	5.66 ± 0.06	3.03 ± 0.11	46.4%	5.58 ± 0.03	4.08 ± 0.07

Training loss and generalization gap. Table 3 splits the test risk of the Transformer into the training loss of the selected network and the generalization gap, the difference of the two. Over the ten repeats of Table 1, the training loss lies between 2.1 and 3.0×10^{-3} for every n. The gap falls from 2.4×10^{-2} at $n = 128$ to 7.9×10^{-5} at $n = 8192$, by a factor of about 2 to 4 at each doubling of n. For large n the test risk therefore consists mainly of the training loss. The 96k-update fits halve the training loss at every n. Their gap is also smaller and decreases in n. The same split for the elliptic operator (Appendix F.3; not plotted) gives a gap of 1.1×10^{-2}, 3.8×10^{-3}, 1.7×10^{-4} and 3.0×10^{-5} at $n = 256$, 1024, 4096 and 16,384, with training losses of at most 1.3×10^{-4}.

The shared-embedding Transformer of Figure 1(a) behaves in the same way (Table 4); its relative L^2 error falls from 13.3% at $n = 128$ to 3.9% at $n = 8192$. Its mean training loss lies between 1.8 and 2.9×10^{-3}, and its gap falls from 3.2×10^{-2} to 9.8×10^{-5} and decreases at every doubling of n in each of the five repeats. DeepONet's gap is of similar size, from 1.9×10^{-2} to 1.6×10^{-4}, and also decreases at every doubling in every repeat, while its mean training loss stays between 5.1 and 5.5×10^{-3}.

Table 3: KS Transformer: training loss of the selected network and generalization gap (test risk minus training loss), in units of 10^{-3} (mean $\pm$ SD over repeats). 48k: the fits of Table 1; 96k: the longer fits of Table 2, evaluated on their 2048 test functions.

Updates	n	Repeats	Training loss	Gap
48k	128	10	2.08 ± 0.12	24.479 ± 3.112
48k	256	10	2.65 ± 0.06	5.876 ± 0.421
48k	512	10	2.91 ± 0.07	2.002 ± 0.137
48k	1024	10	2.97 ± 0.14	0.817 ± 0.084
48k	2048	10	2.96 ± 0.19	0.365 ± 0.046
48k	4096	10	2.94 ± 0.15	0.185 ± 0.040
48k	8192	10	2.99 ± 0.19	0.079 ± 0.030
96k	512	3	1.40 ± 0.02	1.195 ± 0.136
96k	2048	3	1.42 ± 0.08	0.229 ± 0.041
96k	8192	3	1.45 ± 0.08	0.055 ± 0.009

Capacity and locality. Table 5 reports controls at $n = 2048$ with five paired repeats. They vary width and depth and restrict attention to a frequency window. In the local model, the token of frequency k attends only to tokens of frequencies k' with $|k - k'| \le 8$; each sine/cosine pair counts as one frequency. In the real Fourier ordering (constant, $\cos 1, \sin 1, \dots$) this is a band of about ± 16 token positions. It plays the role of the sliding window of Section 2. With the window, the relative

Table 4: KS endpoint operator: Transformer with the shared embedding $\eta a_i + P_i$ and sixteen heads (T) against DeepONet (D), in units of 10^{-3} (48k updates; mean $\pm$ SD over five paired repeats; 2048 common test functions). The gap is the test risk minus the training loss of the selected network.

	Test risk		Training loss		Gap	
n	T	D	T	D	T	D
128	33.52 ± 11.16	24.44 ± 1.57	1.82 ± 0.28	5.08 ± 0.27	31.698 ± 10.882	19.352 ± 1.395
256	7.05 ± 1.17	10.52 ± 0.34	2.21 ± 0.30	5.34 ± 0.17	4.846 ± 0.889	5.184 ± 0.324
512	4.17 ± 0.48	7.30 ± 0.18	2.51 ± 0.24	5.34 ± 0.08	1.658 ± 0.258	1.963 ± 0.243
1024	3.70 ± 1.18	6.36 ± 0.12	2.94 ± 0.96	5.46 ± 0.12	0.754 ± 0.226	0.905 ± 0.182
2048	2.84 ± 0.40	5.85 ± 0.05	2.52 ± 0.35	5.34 ± 0.08	0.320 ± 0.064	0.507 ± 0.085
4096	3.14 ± 0.66	5.67 ± 0.11	2.94 ± 0.61	5.39 ± 0.14	0.194 ± 0.055	0.284 ± 0.033
8192	2.73 ± 0.37	5.66 ± 0.06	2.63 ± 0.34	5.49 ± 0.06	0.098 ± 0.041	0.161 ± 0.042

error at the same width and depth is 3.51%, against 4.31% with global attention (Table 5). Token embeddings are compared in Appendix F.4.

Table 5: KS capacity and attention locality at $n = 2048$ (learned tokens, pooled readout; 48k updates; mean $\pm$ SD over five paired repeats; 1024 common test functions).

Variant	Width/depth	Parameters	Rel. L^2 error (%)	Risk ($\times 10^{-3}$)
global attention	64/2	165,631	6.56 ± 0.10	7.74 ± 0.21
global attention	96/3	417,567	4.31 ± 0.17	3.36 ± 0.27
global attention	128/4	891,519	3.86 ± 0.10	2.69 ± 0.14
window $\lvert k - k' \rvert \le 8$	96/3	417,567	3.51 ± 0.09	2.22 ± 0.11

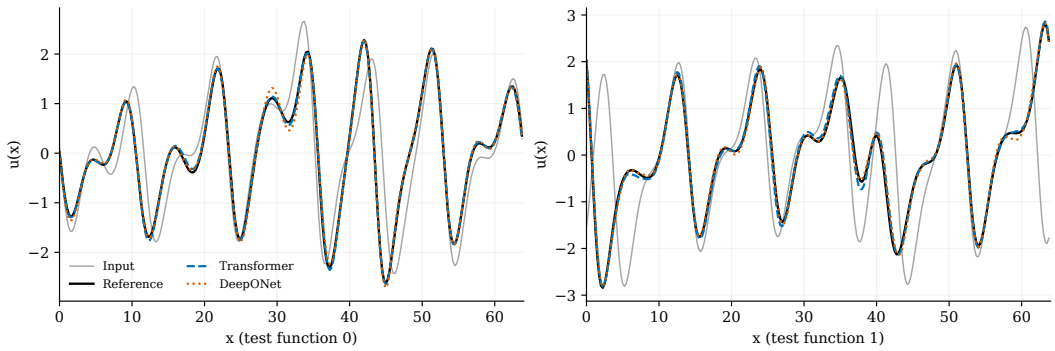


Figure 2: KS endpoint predictions at $n = 8192$ for the first two test functions (repeat 0). Grey: input $u(\cdot, 0)$; black: reference $u(\cdot, 4)$; dashed: Transformer with learned tokens (Table 1); dotted: DeepONet.

F.3 Discretized observations

Operators. Besides the KS operator we use a controlled operator whose input law has a known regularity. Let $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{T})$ on the unit torus with the real Fourier basis, and let $\Psi_0 : f \mapsto u$, where u solves

$$u - 0.03^2 u'' + 0.5 \tanh(u) = \tanh(Sf), \qquad S = (I - 0.02^2 \partial_{xx})^{-1}.$$

The Fourier coefficients of f up to frequency 64 are independent centered Gaussians with standard deviation 0.3 for the constant mode and $0.3/(1 + (k/6)^2)$ at frequency k. The norm $\|f\|_{\mathcal{X}^\alpha}$ is then a Gaussian quadratic form, so the input law lies in $\mathcal{P}_\alpha(\kappa)$ for every $\alpha < 3/2$ (in the limit of infinitely many modes), and we take $\alpha = 1.4$. The solution is computed on 512 points by a contractive fixed-point iteration, stopped when successive iterates differ by less than 10^{-12}. The network reads the input coefficients up to frequency 32 and returns the output coefficients up to frequency 16; the

output energy beyond frequency 16 is about 10^{-5} of the total and is included in the risk. Each repeat has its own pool of 16,384 training functions; 1024 validation and 2048 test functions are shared. For KS the domain $[0, 64)$ is rescaled to $[0, 1)$ and the data are those of Appendix F.2; their input spectra decay exponentially, and we take $\alpha = 2$.

Observation model and Monte Carlo encoder. We follow Section 4.2 with $\mu_u = \mu_v$ uniform on the torus, so that $C_\phi = C_\psi = \sqrt{2}$, and without observation noise ($\xi_{i,j} = 0$); the output coefficients then carry only Monte Carlo error. Every training, validation and test function receives its own locations, drawn once. The input coefficients are estimated up to frequency K ($2K+1$ real coefficients) and projected onto $\mathcal{B}_R^{2,\alpha}$ with weight $(1+k)^{2\alpha}$ at frequency k. The radius R is the 0.999 quantile of $\|x\|_{\mathcal{X}^\alpha}$ over the training inputs; the projection changes at most 1.6% of the encoded validation inputs. The output coefficients are estimated for all returned indices and are not projected. For each n_u, the cutoff K minimizes the input error $\mathbb{E}\|\tilde{x} - x\|_{\mathcal{X}}^2$ on the validation functions. This gives $K = 7, 12, 18$ for $n_u = 64, 256, 1024$ on the elliptic operator, that is, $K = \lceil c_K n_u^{1/(2\alpha+1)} \rceil$ with c_K between 2.3 and 2.9, and $K = 22, 24$ for $n_u = 16{,}384, 65{,}536$ on KS.

Model-free discretization error. Figure 3 evaluates the encoder without any learning. For the validation functions, it reports the input error $\mathbb{E}\|\tilde{x} - x\|_{\mathcal{X}}^2$ and the output error $\mathbb{E}\|\Psi_0(\tilde{x}) - \Psi_0(x)\|_{\mathcal{Y}}^2$, where $\tilde{x}$ is the encoded input and K is chosen as above. The output error bounds the Bayes risk of a regression on encoded inputs, because $\Psi_0(\tilde{x})$ is one predictor that uses $\tilde{x}$ only. Lemma 17 bounds the input error by a multiple of $n_u^{-2\alpha/(2\alpha+1)}$. For the elliptic operator the fitted slope over $n_u \in [2^6, 2^{14}]$ is -0.73, against $-2\alpha/(2\alpha+1) = -0.74$ at $\alpha = 1.4$ and -0.75 in the limit $\alpha \to 3/2$. For KS the slope over $n_u \in [2^8, 2^{16}]$ is -0.92, close to the limit -1 for inputs of unbounded Sobolev regularity. For the elliptic operator the output error decays faster than the input error, with slope -1.02, since Ψ_0 damps the high frequencies that carry most of the Monte Carlo error. For KS it is about five times the input error, since the flow amplifies perturbations of the unstable low modes. As a consequence, KS needs many locations: even $n_u = 65{,}536$ gives an output error of 6.6×10^{-3}, about twice the risk reached from exact observations.

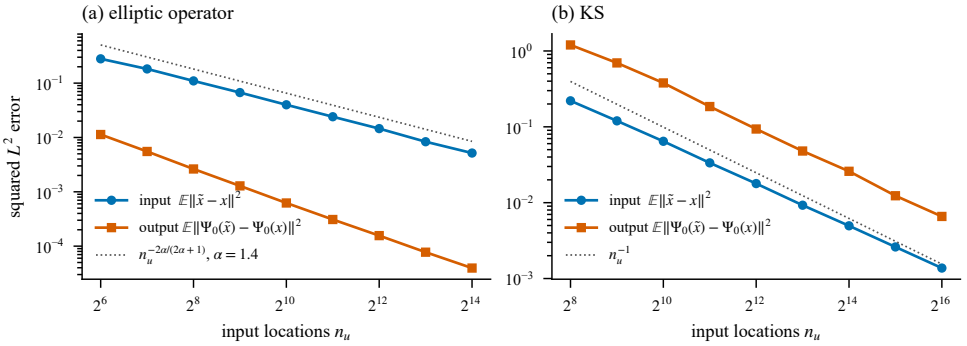


Figure 3: Model-free error of the Monte Carlo encoder with the Sobolev projection, on validation functions (elliptic: 1024; KS: 512), with K chosen as in the text. Dotted: the rate $n_u^{-2\alpha/(2\alpha+1)}$ of Lemma 17 with $\alpha = 1.4$ (a), and its limit n_u^{-1} as $\alpha \to \infty$ (b).

Training and evaluation. Both operators use the pooled Transformer with learned tokens of Appendix F.1. On the elliptic operator it has 65 input tokens, 33 outputs and 381,505 parameters, and is trained for 6000 AdamW updates with batch size 128, learning rate 10^{-3}, weight decay 10^{-4}, 200 warm-up steps and cosine decay, with validation every 250 updates. KS uses the 48k-update recipe of Appendix F.1. Checkpoints are selected on validation functions observed in the same way as the training functions. Test inputs are encoded from fresh locations, and the test risk is measured against exact outputs. A repeat draws a new training pool, new locations and a new initialization; the training sets are nested in n.

Results. Figure 4 shows the test risk of all fits. For the elliptic operator with sampled inputs, the risk decreases with n and approaches the model-free output error of Figure 3 from above, while

exact observations reach 1.5×10^{-4} at $n = 16{,}384$. With $n_u = 64$ the risk falls from 3.2×10^{-2} at $n = 256$ to 1.18×10^{-2} at $n = 16{,}384$, against an output error of 1.13×10^{-2}. With $n_u = 256$ it falls from 1.5×10^{-2} to 2.95×10^{-3} (output error 2.63×10^{-3}), and with $n_u = 1024$ from 1.0×10^{-2} to 8.9×10^{-4} (output error 6.2×10^{-4}). At $n = 16{,}384$ the risk exceeds the sum of the exact-observation risk (1.53×10^{-4}) and the output error by 3%, 6% and 14% for $n_u = 64$, 256 and 1024. The risk is thus close to the sum of the exact-observation risk and the model-free output error, consistent with the additive form of the bound in Theorem 4. With sampled outputs the risk keeps decreasing over the whole range of n: from 5.0×10^{-2} to 4.3×10^{-3} with $n_v = 64$, and from 2.4×10^{-2} to 1.25×10^{-3} with $n_v = 256$. The Monte Carlo noise of the output coefficients has mean squared norm 5.4×10^{-2} and 1.35×10^{-2}, half and one eighth of $\mathbb{E}\|\Psi_0(x)\|^2 = 0.108$. For $n \geq 1024$ the training loss stays near this noise level, and the noise raises the test risk at every n without creating a floor. Table 6 lists all values.

Table 6: Elliptic operator with discretized observations: test risk ($\times 10^{-3}$; mean $\pm$ SD over three repeats; 2048 test functions). n_u and n_v are the numbers of random input and output grids per function; "exact" uses exact coefficients. The last column is the model-free output error $\mathbb{E}\|\Psi_0(\tilde{x}) - \Psi_0(x)\|^2$ of the input encoder (Figure 3).

Observation	$n = 256$	$n = 1024$	$n = 4096$	$n = 16{,}384$	Encoder
exact	11.00 ± 0.56	3.84 ± 0.14	0.29 ± 0.02	0.15 ± 0.01	–
$n_u = 64$	31.61 ± 0.92	21.20 ± 0.52	15.66 ± 0.12	11.83 ± 0.09	11.33
$n_u = 256$	15.36 ± 0.11	8.50 ± 0.16	4.54 ± 0.06	2.95 ± 0.02	2.63
$n_u = 1024$	9.98 ± 0.53	4.30 ± 0.12	1.23 ± 0.03	0.89 ± 0.01	0.62
$n_v = 64$	49.76 ± 1.71	19.12 ± 0.46	10.09 ± 0.82	4.34 ± 0.13	–
$n_v = 256$	24.00 ± 0.47	10.03 ± 0.15	4.97 ± 0.31	1.25 ± 0.03	–

KS. Table 7 and the bottom row of Figure 4 report the KS fits for $n \in \{512, 2048, 8192\}$ (Transformer with learned tokens; three repeats; 2048 test functions). With sampled inputs the risk stays far above the risk from exact observations: with $n_u = 65{,}536$ it falls from 1.22×10^{-2} at $n = 512$ to 9.0×10^{-3} at $n = 8192$, and with $n_u = 16{,}384$ from 3.22×10^{-2} to 2.33×10^{-2}, against 3.13×10^{-3} from exact observations. From $n = 2048$ to 8192 the two sampled-input risks decrease by about 6% and the exact-observation risk by about 5%; quadrupling n_u lowers the sampled-input risks by a factor of 2.6 at every n. With $n_u = 16{,}384$ the learned risk for $n \geq 2048$ is slightly below the model-free output error (2.59×10^{-2}). With sampled outputs the risk keeps decreasing over the whole range, from 1.82×10^{-2} to 6.0×10^{-3} with $n_v = 1024$ and from 6.5×10^{-2} to 1.11×10^{-2} with $n_v = 256$. The Monte Carlo noise of the 127 output coefficients has mean squared norm about one eighth and one half of $\mathbb{E}\|\Psi_0(x)\|^2 = 1.72$.

Table 7: KS operator with discretized observations: test risk ($\times 10^{-3}$; mean $\pm$ SD over three repeats; 2048 test functions; Transformer with learned tokens). The last column is the model-free output error $\mathbb{E}\|\Psi_0(\tilde{x}) - \Psi_0(x)\|^2$ of the input encoder on 512 validation functions (Figure 3).

Observation	$n = 512$	$n = 2048$	$n = 8192$	Encoder
exact	4.84 ± 0.18	3.29 ± 0.20	3.13 ± 0.12	–
$n_u = 16{,}384$	32.21 ± 0.97	24.74 ± 0.27	23.35 ± 0.24	25.93
$n_u = 65{,}536$	12.18 ± 0.25	9.53 ± 0.07	9.00 ± 0.10	6.58
$n_v = 256$	65.46 ± 3.03	16.44 ± 0.45	11.08 ± 0.32	–
$n_v = 1024$	18.20 ± 0.19	7.75 ± 0.28	6.01 ± 0.16	–

F.4 Token embeddings

Embeddings of the KS Transformer. In $\mathcal{T}_{N,\Gamma}$, coefficient a_i enters as the token $\eta a_i + P_i$, with one vector η for all indices and a fixed positional encoding P. The KS Transformer instead learns $a_i w_i + e_i$, with w_i and e_i free for every index. Table 8 compares embeddings at $n = 2048$ with the recipe of Appendix F.1 otherwise unchanged, on the 2048 test functions of Table 2. The fixed encoding is the sinusoidal encoding of Vaswani et al. (2017). With a shared η, the value vector of head h at token j is $a_j V^{(h)} \eta + V^{(h)} P_j$, so the part that depends on the coefficients points in the single direction

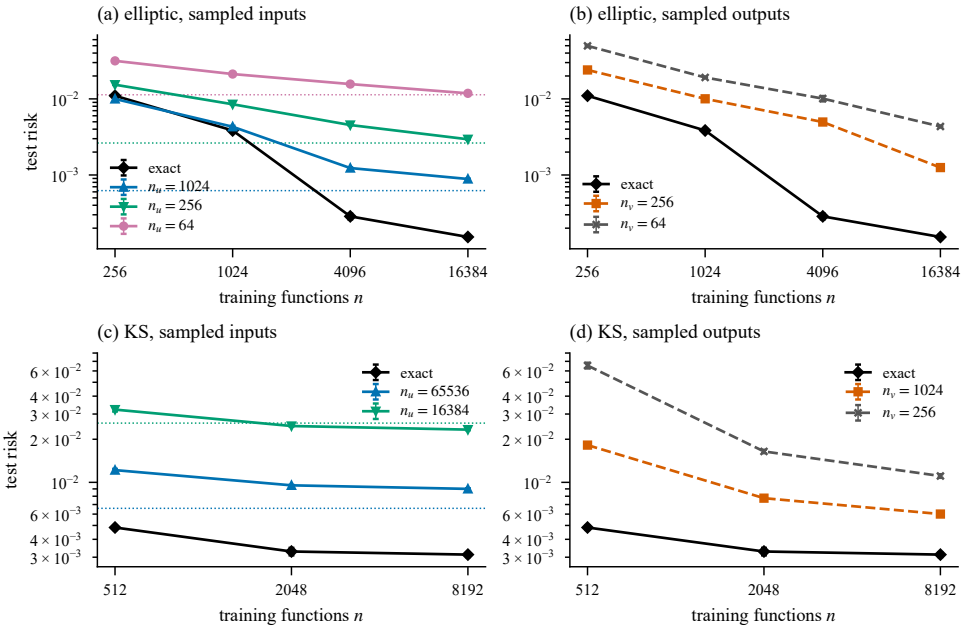


Figure 4: Test risk with discretized observations (mean ± SD over three repeats). Top: elliptic operator; bottom: KS (Transformer with learned tokens). Left: inputs observed at n_u random grids, with the model-free output error of the Monte Carlo encoder dotted; right: outputs observed at n_v random grids. Black: exact observations.

$V^{(h)}\eta$, and each head gathers one scalar mixture of the coefficients. The construction of Theorem 1 accordingly uses one head for each coordinate that it gathers into a token. We therefore also train the shared embedding with sixteen heads instead of four, at the same width. All embeddings act on coefficients standardized coordinatewise with training statistics (Appendix F.1).

With four heads, the shared embedding roughly doubles the risk, from 3.29×10^{-3} to 6.80×10^{-3}. The training loss rises in the same proportion, from 2.9×10^{-3} to 6.0×10^{-3}. Training twice as long does not remove it: with 96k updates the four-head model reaches $(3.09 \pm 0.05) \times 10^{-3}$, the level of learned tokens after 48k updates, whereas learned tokens reach $(1.65 \pm 0.12) \times 10^{-3}$. With sixteen heads, the shared embedding reaches $(2.84 \pm 0.40) \times 10^{-3}$ after 48k updates (five repeats; $(3.02 \pm 0.42) \times 10^{-3}$ over the three repeats of the other rows), as accurate as learned tokens.

Table 8: KS Transformer with different token embeddings at $n = 2048$ (48k updates unless stated; mean ± SD over three repeats, five for the sixteen-head row; 2048 test functions).

Token	Heads	Parameters	Test risk ($\times 10^{-3}$)	Relative L^2 error (%)
learned $a_i w_i + e_i$	4	417,567	3.29 ± 0.20	4.27 ± 0.13
shared $\eta a_i + P_i$	4	393,279	6.80 ± 0.73	6.14 ± 0.33
shared $\eta a_i + P_i$	16	393,279	2.84 ± 0.40	3.96 ± 0.27
learned $a_i w_i + e_i$, 96k updates	4	417,567	1.65 ± 0.12	3.02 ± 0.11
shared $\eta a_i + P_i$, 96k updates	4	393,279	3.09 ± 0.05	4.13 ± 0.03

F.5 Output truncation on trajectory operators

The estimator retains the output coefficients with index at most N, and the part of the target beyond this cutoff enters the risk as a truncation term (Lemma 8). We examine this term on two operators whose outputs are space-time trajectories.

Operators. The viscous Burgers equation $u_t + uu_x = 0.1u_{xx}$ on $[0, 2\pi)$ maps the initial function to the full trajectory on $[0, 2\pi) \times [0, 1]$. Initial functions are the public inputs of Li et al. (2021);

trajectories are recomputed with ETDRK4 on 512 spatial points and stored on a 128×101 space-time grid, with $1000/200/200$ training, validation and test functions. The reaction-diffusion equation $u_t = 0.01u_{xx} + 0.01u^2 + f(x)$ on $[0,1]\times[0,1]$, with zero initial and Dirichlet boundary data, maps the source f, a centered Gaussian field with squared-exponential covariance of length scale 0.2, to the solution u, following the setting of Lu et al. (2021). We use a 100×100 grid, three independent pools of 1000 training functions, and shared validation and test sets of 256 functions. The inputs are represented by 63 Fourier coefficients (Burgers) or 32 cosine coefficients (reaction-diffusion), and the outputs in tensor-product Fourier-Legendre or sine-Legendre bases.

Protocol and results. The Transformer has three four-head blocks of width 96 (Burgers) or 64 (reaction-diffusion) and is trained at $n = 1000$ for 48k updates with batch size 32, three repeats and validation-based selection. Table 9 and Figure 5 vary the number of retained output coordinates. Refining the output basis lowers the projection residual, the squared norm of the part of the reference outside the retained output space, by more than three orders of magnitude for Burgers and about thirtyfold for reaction-diffusion, and it lowers the test risk, consistent with the truncation term of the analysis.

Table 9: Output representation of the trajectory Transformer at $n = 1000$ (mean $\pm$ SD over three repeats). The residual is the squared L^2 norm of the part of the reference outside the retained output space; it does not depend on the fit. Default configurations are marked with *.

Problem	Output coords.	Test risk	Projection residual
Burgers	752^*	$(2.07 \pm 0.11) \times 10^{-4}$	6.7×10^{-8}
	372	$(2.17 \pm 0.17) \times 10^{-4}$	3.0×10^{-6}
	120	$(2.77 \pm 0.20) \times 10^{-4}$	1.7×10^{-4}
Reaction-diffusion	384^*	$(6.62 \pm 6.05) \times 10^{-6}$	2.3×10^{-7}
	240	$(8.00 \pm 4.76) \times 10^{-6}$	9.6×10^{-7}
	128	$(15.93 \pm 10.61) \times 10^{-6}$	6.9×10^{-6}

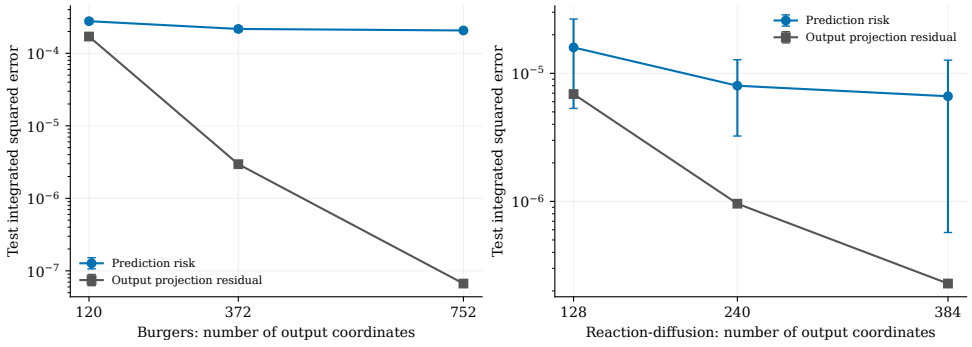


Figure 5: Output representation at $n = 1000$: prediction risk and output projection residual versus the number of retained output coordinates (mean $\pm$ SD over three repeats: initializations on a shared training set for Burgers, independent training pools for reaction-diffusion). Left: Burgers; right: reaction-diffusion.