MIXTURES OF NEURAL OPERATORS REDUCE ACTIVE COMPLEXITY IN OPERATOR LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Operator-learning systems are not governed solely by total parameter count; for one query, the relevant bottleneck can be the model that must be loaded and evaluated. We study this distinction for classical neural operators on compact Sobolev subsets through a constructive comparison between routed mixtures of neural operators (MoNOs) and a fixed single-neural-operator construction. A MoNO routes each input function through a tree to one expert. Our main theorem shows that every scalar uniformly continuous nonlinear operator with bounded output Sobolev radius on the approximation set admits a MoNO approximation whose active expert has smaller depth and width scaling, and no larger rank scaling, than the analyzed single-neural-operator construction; for Lipschitz targets these expert quantities are bounded by $\mathcal{O}(\varepsilon^{-1})$. A coefficient-quantization tree also gives polynomial routing work for Lipschitz targets under an explicit coefficient-access model, while storage remains exponential. For bounded $C^{1,1}$ functionals, affine experts improve the exponential storage power over constant lookup under target-independent routing. We also prove a quantitative universal approximation theorem for the underlying neural-operator architecture, with explicit dependence on compact-set diameter and modulus of continuity.

1 INTRODUCTION

Neural operators (NOs) learn maps between function spaces (Chen & Chen, 1995; Lu et al., 2021; Li et al., 2021; Kovachki et al., 2023), including PDE solution maps, dynamical rollouts, and inverse problems (Raonic et al., 2023; McCabe et al., 2023; de Hoop et al., 2022; Molinaro et al., 2023). Universal approximation results establish their broad expressivity (Kovachki et al., 2021; Korolev, 2022; Herrmann et al., 2024; Adcock et al., 2025); quantitative bounds relate that expressivity to the model size needed for a prescribed accuracy.

The required size can grow rapidly as the accuracy increases. In finite dimensions, general smooth-function approximation has rates whose exponent depends on dimension (Yarotsky, 2018; Petersen & Voigtlaender, 2018). On compact subsets of function space, there are Lipschitz functionals for which a single ReLU-based PCA-Net, DeepONet, or NOMAD requires exponentially many parameters (Lanthaler & Stuart, 2026), with the Lipschitz specialization established in Corollary 4.5; an activation-independent lower bound holds for the total bits of a family covering the bounded Lipschitz class (Lanthaler, 2024). These global representation costs motivate distinguishing the size of a stored model from the expert parameters loaded and evaluated on an individual query.

Sparse mixture-of-experts (MoE) architectures make this distinction explicit: a router activates one or a few experts for each input, while the remaining experts are stored for other inputs (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Zhou et al., 2022; Jiang et al., 2024). Such localization is also familiar in scientific computing, where a surrogate may cover several physical regimes. The *hp* reduced-basis method partitions the parameter domain with a binary tree and builds a local basis on each cell (Eftang et al., 2010), while local reduced-order models assign bases to clusters of the solution manifold (Amsallem et al., 2012). Mixture-of-experts DeepONets and operator transformers extend this approach to operator learning (Sharma & Shankar, 2025; Wang et al., 2025), motivating a quantitative analysis of the expert size required on each local region.

We develop this analysis for a *mixture of neural operators* (MoNO), whose experts approximate the target on local subsets of a compact Sobolev set. Routing selects one leaf of a tree and evaluates its

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.

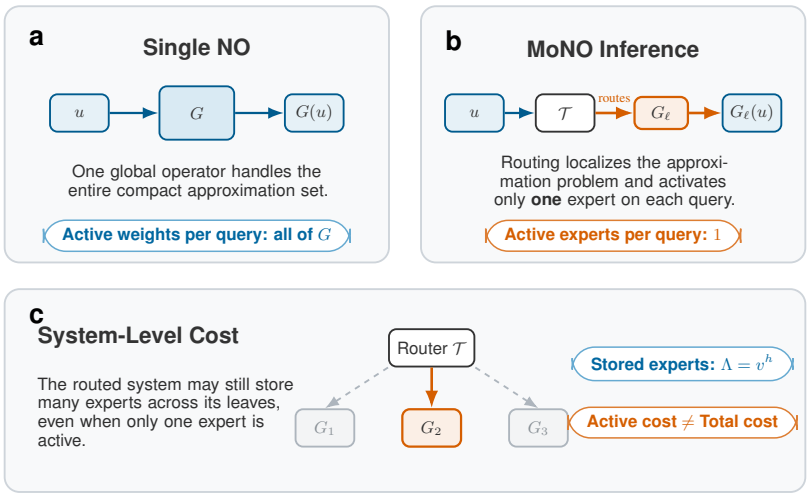

Figure 1: **Resources in operator approximation. a**, A single NO evaluates one global model. **b**, A MoNO selects one expert G_ℓ per query. **c**, The tree and inactive experts remain stored. Section 3.3.1 specifies the accounting for experts, routing work, and storage.

neural operator (Figure 1). Our main theorem covers every uniformly continuous scalar operator with bounded output Sobolev radius (Theorem 4.1). For Lipschitz targets, the selected expert has depth zero, bounded width, and rank $O(\varepsilon^{-d_2/s_2})$, while expert count and combined parameters are at most $\exp(C\varepsilon^{-d_1/s_1})$.

Two further results address storage and search. For bounded $C^{1,1}$ functionals, affine experts reduce the exponential storage power from ε^{-d_1/s_1} for constant lookup to $\varepsilon^{-d_1/(2s_1)}$, under target-independent routing and including represented centers (Proposition 4.7). A coefficient-quantization tree implements greedy routing in $O(\varepsilon^{-d_1/s_1}\log(1/\varepsilon))$ scalar operations for Lipschitz targets under an explicit coefficient-access model (Theorem 4.8). Combined with Lanthaler & Stuart (2026), this gives a fixed Lipschitz functional requiring exponentially many parameters in every single ReLU model of neural-network type, yet only polynomial routing work and a constant-size MoNO expert (Corollary 4.5). We also prove a quantitative universal approximation theorem for the underlying finite-rank NO, retaining explicit coefficient-dimension, compact-set diameter, and modulus dependence (Proposition 4.4).

Controlled Burgers experiments use four experts with a fixed backbone (Section 5). Specialization improves accuracy on prescribed viscosity regimes within the tested training budgets; the gain depends on conditioning and routing accuracy. A fixed-viscosity control shows no gain.

2 Related Work

Approximation theory for mixtures of experts. Mixtures of experts originate with Jacobs et al. (1991); hierarchical mixtures use trees of soft gates (Jordan & Jacobs, 1994). Classical work establishes approximation rates and universality (Zeevi et al., 1998; Jiang & Tanner, 1999; Nguyen et al., 2016). More recent results concern sparse approximation (Baykal et al., 2022; Kratsios et al., 2025), memorization versus reasoning (Jelassi et al., 2025), expressivity (Boix-Adsera & Rigollet, 2025; Wang & E, 2025), and learned clustering (Chen et al., 2022). For operators, the approximation domain is infinite-dimensional, requiring represented covers or controlled coefficient truncation; the single-ReLU lower bound is exponential in a power of ε^{-1}.

Mixtures and localization in operator learning. The name *Mixture of Neural Operators* (MoNO) was used by Majid & Tudisco (2024) for combining historical predictions to improve rollouts. Deighan et al. (2025) use spatial mixtures of operator experts for boundary conditions and model selection. Related work includes spatially localized DeepONet ensembles with universality guarantees (Sharma & Shankar, 2025) and sparse operator transformers for heterogeneous PDE data (Wang

et al., 2025). Our analysis selects one expert per input function and counts its size, routing work, and storage. Experiments use FNO experts; the theory uses Definition 3.2, without establishing backbone-independent empirical gains.

Complexity of single neural operators. Algebraic total-complexity bounds are available under additional structure: Barron-type or holomorphic regularity, low-dimensional or representation-adapted structure, kernel and random-feature methods, emulation of numerical schemes, or problem-specific physics (Barron, 1993; Adcock et al., 2025; Herrmann et al., 2024; Marcati & Schwab, 2023; Lanthaler, 2023; Batlle et al., 2024; Nelsen & Stuart, 2024; Lanthaler & Stuart, 2026; Lara Benitez et al., 2025); enlarging the activation class to super-expressive families weakens the barrier in a different way (Yarotsky & Zhevnerchuk, 2020; Yarotsky, 2021; Shen et al., 2021; Jiao et al., 2023). Without such structure, PCA-Net cannot achieve algebraic rates for general C^r operators (Lanthaler, 2023), and the required complexity is exponential for ReLU models of neural-network type and for bit-encoded models that are uniform over the Lipschitz class (Lanthaler & Stuart, 2026; Lanthaler, 2024). Our comparison is complementary: we assume no analytic, low-dimensional, or representation-adapted structure and use standard activations, so any gain comes from routing.

3 Problem Setup and Architectures

We fix a compact Sobolev approximation set, the single neural-operator architecture used in the comparison, and the routed MoNO model; all symbols are collected in Section A.

Let $d_1, d_2, d_{in}, d_{out} \in \mathbb{N}$ and define $D_i \overset{\text{def.}}{=} [0,1]^{d_i}$ for $i = 1, 2$; here d_1 and d_2 are the dimensions of the input and output domains, and d_{in} and d_{out} are the numbers of input and output channels. We work with Sobolev spaces embedded in ambient L^2 spaces, and approximation errors are measured in the relevant L^2 norm. For a uniformly continuous map $G : (H, \|\cdot\|_H) \to (H', \|\cdot\|_{H'})$ between separable Hilbert spaces, we write $\omega : [0,\infty) \to [0,\infty)$ for a monotone modulus of continuity satisfying

$$\|G(u) - G(\tilde{u})\|_{H'} \le \omega\big(\|u - \tilde{u}\|_H\big) \qquad \text{for all } u, \tilde{u} \in H.$$

3.1 Sobolev Approximation Sets

Let $D = [0,1]^d$, let $s > 0$, and let $H^s(D)$ denote the usual Sobolev space. For integer s, we use the standard weak-derivative Sobolev norm with derivatives up to order s; for noninteger s, we use the equivalent Sobolev–Slobodeckij norm (7) recalled in Section A. For vector-valued inputs we write $H^s(D)^{d_{in}}$ and endow it with the ambient norm $\|\cdot\|_{L^2(D)^{d_{in}}}$.

The compact approximation sets considered throughout the paper are either the centered Sobolev ball of radius $R > 0$,

$$K \overset{\text{def.}}{=} \left\{ f \in H^s(D)^{d_{in}} : \|f\|_{H^s(D)^{d_{in}}} \le R \right\}, \tag{1}$$

or, more generally, any closed nonempty subset

$$K = \overline{K}^{L^2} \subseteq \left\{ f \in H^s(D)^{d_{in}} : \|f - f_0\|_{H^s(D)^{d_{in}}} \le R \right\}, \tag{2}$$

for some $f_0 \in H^s(D)^{d_{in}}$. By the Rellich–Kondrashov theorem, such sets are compact in $L^2(D)^{d_{in}}$.

3.2 Neural Network and Neural Operator Models

Definition 3.1 (Multilayer perceptron)**.** Let $\sigma \in C(\mathbb{R})$ and let $p, q \in \mathbb{N}$. For a depth $J \in \mathbb{N}_{\ge 0}$ and widths $[n] \overset{\text{def.}}{=} (n_0 \overset{\text{def.}}{=} p, n_1, \dots, n_J, n_{J+1} \overset{\text{def.}}{=} q)$, an MLP with activation σ is a finite composition of affine maps and coordinatewise activations,

$$\hat{f}_\theta(x) = A^{(J)} x^{(J)} + c, \qquad x^{(j+1)} = \sigma \bullet \left(A^{(j)} x^{(j)} + b^{(j)}\right), \quad j = 0, \dots, J-1, \tag{3}$$

where $(A^{(j)}, b^{(j)}) \in \mathbb{R}^{n_{j+1} \times n_j} \times \mathbb{R}^{n_{j+1}}$ for $j < J$, $A^{(J)} \in \mathbb{R}^{q \times n_J}$, $c \in \mathbb{R}^q$, the input is $x^{(0)} = x \in \mathbb{R}^p$, and $\sigma \bullet y \overset{\text{def.}}{=} (\sigma(y_i))_i$ denotes coordinatewise activation.

We write $\mathcal{NN}^{\sigma}_{[n]}$ for this class of networks with widths $[n]$, and $\mathcal{NN}^{\sigma}_{\Delta,W:p,q}$ for the union over all such networks $\mathbb{R}^p \to \mathbb{R}^q$ with depth at most Δ and width at most W.

The single-model baseline compared against MoNOs is the following finite-rank neural-operator architecture.

Definition 3.2 (Neural operator)**.** Let $\boldsymbol{d} \overset{\text{def.}}{=} [d_1, d_2, d_{in}, d_{out}]$ and let $L, W, \Delta, N \in \mathbb{N}_{\geq 0}$ with $\max(d_1, d_2, d_{in}, d_{out}) \leq W$. Fix L^2-orthonormal families $\{\varphi_m\}_{m=1}^N \subset H^{s_1}(D_1)$ and $\{\psi_n\}_{n=1}^N \subset H^{s_2}(D_2)$. A neural operator

$$G : (H^{s_1}(D_1)^{d_{in}}, \|\cdot\|_{L^2(D_1)^{d_{in}}}) \to (H^{s_2}(D_2)^{d_{out}}, \|\cdot\|_{L^2(D_2)^{d_{out}}})$$

is given by

$$\begin{aligned} (G(u))(x) &= (K_N^{(L+1)} u_{L+1})(x) + b^{(L+1)}(x), \qquad u_1(x) = (K_N^{(0)} u)(x) + b^{(0)}(x), \\ u_{\ell+1}(x) &= \tanh \bullet \big(A^{(\ell)} u_\ell(x) + b^{(\ell)}\big), \qquad \ell = 1, \dots, L, \end{aligned} \tag{4}$$

where $q_1, \dots, q_{L+1} \leq W$ are channel widths, $u_\ell(x) \in \mathbb{R}^{q_\ell}$, $A^{(\ell)} \in \mathbb{R}^{q_{\ell+1} \times q_\ell}$, $b^{(\ell)} \in \mathbb{R}^{q_{\ell+1}}$, and the nonlocal maps

$$K_N^{(0)} u(x) = \sum_{m,n=1}^{N} C_{m,n}^{(0)} (u, \varphi_m)_{L^2(D_1)} \psi_n(x), \tag{5a}$$

$$K_N^{(L+1)} u'(x) = \sum_{m,n=1}^{N} C_{m,n}^{(L+1)} (u', \psi_m)_{L^2(D_2)} \psi_n(x) \tag{5b}$$

use N scalar basis modes (the rank parameter), with coefficient matrices $C_{m,n}^{(0)} \in \mathbb{R}^{q_1 \times d_{in}}$ and $C_{m,n}^{(L+1)} \in \mathbb{R}^{d_{out} \times q_{L+1}}$. Here $b^{(0)}$ and $b^{(L+1)}$ are coordinate networks on D_2 with $\tanh$ activation, width at most W, depth at most Δ, and output dimensions q_1 and d_{out}, respectively.

We denote this class by $\mathcal{NO}^{\tanh}_{N,W,L,\Delta,\boldsymbol{d}}$. For a neural operator G in this class, let $P(G)$ denote its number of trainable parameters. Then

$$P(G) \leq \underbrace{L\,W(W+1)}_{A^{(\ell)},\, b^{(\ell)}} + \underbrace{2N^2W^2}_{K_N^{(0)},\, K_N^{(L+1)}} + \underbrace{2(\Delta+1)\,W(W+1)}_{b^{(0)},\, b^{(L+1)}} \overset{\text{def.}}{=} \bar{P}_{\Delta,W,L,N,\boldsymbol{d}}. \tag{6}$$

We use $\bar{P}_{\Delta,W,L,N,\boldsymbol{d}}$ as the architecture-level upper bound. The $\Delta + 1$ counts the affine output layer of each coordinate network.

Remark 3.3 (Nonlocal operators). The maps in (5) arise by expanding integral kernels in the fixed orthonormal families and truncating to N modes. Nonlocal lifting and output maps suffice for our universality construction; related minimal nonlocal architectures are studied by Lanthaler et al. (2025).

3.3 Mixture of Neural Operators

3.3.1 Rooted tree router

We use a rooted tree $\mathcal{T} = (V, E, f_{1,1})$ with nodes V, edges E, and root $f_{1,1}$. Each node carries an $L^2(D_1)^{d_{in}}$ center, represented either by a ReLU MLP (Theorem 4.1 and Proposition 4.7) or by coefficients in a fixed smooth orthonormal family (Theorem 4.8). Distinct nodes may carry identical centers. Every internal node has $v \geq 2$ children and every leaf has depth h, giving $\Lambda = v^h$ leaves.

Definition 3.4 (Mixture of neural operators)**.** Let $\boldsymbol{d} \overset{\text{def.}}{=} [d_1, d_2, d_{in}, d_{out}]$ and let $L, W, \Delta, N, h, v \in \mathbb{N}_{\geq 0}$ with $v \geq 2$ and $\max(d_1, d_2, d_{in}, d_{out}) \leq W$. A *mixture of neural operators* (MoNO) is a pair $(\mathcal{T}, \mathcal{NO})$ in which

$$\mathcal{T} = (V, E, f_{1,1}) \quad \text{and} \quad \mathcal{NO} = \{G_\ell\}_{\ell=1}^{\Lambda},$$

where $\mathcal{T}$ is a complete rooted tree of height h and valency v as above, with leaves $\hat{f}_1, \dots, \hat{f}_\Lambda$, and where the leaf $\hat{f}_\ell$ carries the expert $G_\ell \in \mathcal{NO}^{\tanh}_{N,W,L,\Delta,\boldsymbol{d}}$.

Each expert approximates the target on its local region. A classical neural operator is the one-expert MoNO (Section H).

Realization of a MoNO. For scalar inputs and outputs, a routing rule selects a leaf $\ell(u)$ and the realization is $G(u) = G_{\ell(u)}(u)$. The entropy cover in Theorem 4.1 uses exhaustive nearest-leaf search, with ties broken by index (Algorithm 1); its internal nodes only index the experts. The coefficient tree in Theorem 4.8 instead descends to the nearest child at each level, and this greedy rule finds a nearest leaf. Both rules satisfy the covering requirement in Remark 4.2.

The *expert-active complexity* and the *total distributed complexity* of $(\mathcal{T}, \mathcal{NO})$ are

$$\text{Activ-Cpl}(\mathcal{T}, \mathcal{NO}) \overset{\text{def.}}{=} \max_{1 \le \ell \le \Lambda} P(G_\ell) \qquad \text{and} \qquad \text{Dist-Cpl}(\mathcal{T}, \mathcal{NO}) \overset{\text{def.}}{=} \sum_{\ell=1}^{\Lambda} P(G_\ell),$$

so (6) bounds these by $\bar{P}_{\Delta,W,L,N,\boldsymbol{d}}$ and $\Lambda \bar{P}_{\Delta,W,L,N,\boldsymbol{d}}$, respectively.

Resource model. Expert parameter counts exclude routing; total system storage also includes centers and tree records. We report search work separately: Λ exact L^2 distances for the entropy cover, or coefficient queries and scalar operations for the quantized tree. The latter assumes access to the stated input coefficients and unit-cost random access to stored tree nodes and experts; Theorem 4.8 specifies sufficient coefficient accuracy. Fixed basis descriptions, the cost of acquiring coefficients or distances, output discretization, and physical loading latency are additional resources. Storage bounds count real parameter slots unless bits are specified. A comparison with single-model work uses conventional feedforward evaluation, processing every nonzero weight and bias; the cited lower bound itself is a representation-size bound. Learned routers inherit a covering guarantee only when they select an admissible leaf.

4 Main Results

We approximate a nonlinear operator $G^+ : (H^{s_1}(D_1)^{d_{in}}, \|\cdot\|_{L^2(D_1)^{d_{in}}}) \to (H^{s_2}(D_2)^{d_{out}}, \|\cdot\|_{L^2(D_2)^{d_{out}}})$ uniformly on a compact subset K. We choose a continuous, strictly increasing, unbounded concave majorant ω of its modulus, with $\omega(0) = 0$; ω^{-1} denotes its inverse. Such a majorant exists for a globally uniformly continuous map on this normed space. The main theorem treats scalar functions ($d_{in} = d_{out} = 1$).

Construction parameters. Let C_R be the constant from Lemma D.1 and $R_{G,K} \overset{\text{def.}}{=} \sup_{u \in K} \|G^+(u)\|_{H^{s_2}(D_2)}$. Choose $N_\varepsilon \ge 1$ for both projections in (9), with $N_\varepsilon \in \mathcal{O}\big(1 + [\omega^{-1}(\varepsilon/4)]^{-d_1/s_1} \vee \varepsilon^{-d_2/s_2}\big)$. The experts need only the output projection, of dimension $1 \le M_\varepsilon \le N_\varepsilon$ with $M_\varepsilon \in \mathcal{O}(1 + \varepsilon^{-d_2/s_2})$. Constants depend on R and $R_{G,K}$. Set

$$\rho_\varepsilon \overset{\text{def.}}{=} \min\big\{1, \omega^{-1}(\varepsilon/2)\big\}, \qquad \delta_\varepsilon \overset{\text{def.}}{=} \frac{\rho_\varepsilon}{2C_R},$$

$$h_\varepsilon \overset{\text{def.}}{=} \left\lceil \log_v\big(2^{\lceil \delta_\varepsilon^{-d_1/s_1} \rceil} - 1\big) \right\rceil, \qquad \Lambda \overset{\text{def.}}{=} v^{h_\varepsilon}.$$

Theorem 4.1 (Universal approximation for MoNOs)**.** *Assume the setting of Section 3.3. Let K be as in* (1)*, suppose $s_i > d_i$ for $i = 1, 2$, fix $\varepsilon > 0$, and choose a valency $v \in \mathbb{N}_{\ge 2}$. For every uniformly continuous operator $G^+ : (H^{s_1}([0,1]^{d_1}), \|\cdot\|_{L^2(D_1)}) \to (H^{s_2}([0,1]^{d_2}), \|\cdot\|_{L^2(D_2)})$ with modulus of continuity ω and $R_{G,K} < \infty$, there exists a MoNO $(\mathcal{T}, \mathcal{NO})$ whose realization $G : (H^{s_1}([0,1]^{d_1}), \|\cdot\|_{L^2(D_1)}) \to (H^{s_2}([0,1]^{d_2}), \|\cdot\|_{L^2(D_2)})$ satisfies*

$$\sup_{u \in K} \|G^+(u) - G(u)\|_{L^2([0,1]^{d_2})} \le \varepsilon \qquad \textit{and} \qquad \text{Activ-Cpl}(\mathcal{T}, \mathcal{NO}) \in \mathcal{O}\big(M_\varepsilon^2\big).$$

The experts are constant maps of the input, with depth zero, width $W_* = \max\{d_1, d_2, 1\}$, and rank parameter M_ε. The tree has height h_ε and $\Lambda = v^{h_\varepsilon}$ leaves. The balanced tree construction is given in Lemma D.1 and Section B.2, and Table 1 summarizes the complexity bounds.

Proof. See Section B.4. □

Remark 4.2 (Routing coverage). The proof builds each expert G_ℓ to be ε-accurate on the whole local set $K_\ell \overset{\text{def.}}{=} \{u \in K : \|u - \hat{f}_\ell\|_{L^2(D_1)} \le \rho_\varepsilon/2\}$, and these sets overlap. Hence the error bound of

Table 1: Constructive bounds, with $N = N_\varepsilon$ and $M = M_\varepsilon$. In the generic Lipschitz bounds, the rank order improves if $d_1/s_1 > d_2/s_2$ and agrees otherwise. ‡Fixed hard functional, with $\gamma = 1/(\alpha + 1 + \delta)$ (Corollary 4.5). †Family covering all 1-Lipschitz functionals with $\sup_K |S| \le 1$ (Lanthaler, 2024, Thm. 2.8 and Ex. 2.9).

Resource	**MoNO active expert**	**Constructive single NO**
Depth	0	(8)
Width; rank parameter	$\mathcal{O}(1)$; M	$\mathcal{O}(N)$; N
Expert count	v^{h_ε}	1
Entropy-cover routing	v^{h_ε} distances	none
Quantized routing (Theorem 4.8)	J_ε coefficients; $O(vJ_\varepsilon b_\varepsilon)$ operations	none
Known worst-case lower bounds		
Single ReLU model‡	$\exp(c\varepsilon^{-\gamma})$ nonzero parameters	
Any class-uniform family†	$\exp(c\varepsilon^{-d_1/s_1})$ bits in total	

Theorem 4.1 holds for *any* routing rule that sends $u \in K$ to some leaf ℓ with $\|u - \hat{f}_\ell\|_{L^2(D_1)} \le \rho_\varepsilon/2$; the nearest-leaf rule is one such rule. In particular, if a perturbation $\tilde{u} \in K$ of u is sent to a different admissible leaf, then $\|G(\tilde{u}) - G^+(u)\|_{L^2(D_2)} \le \varepsilon + \omega(\|\tilde{u} - u\|_{L^2(D_1)})$, so crossing a routing boundary does not by itself degrade accuracy.

Corollary 4.3 (Lipschitz targets)**.** *Assume the setting of Theorem 4.1, with $\omega(t) = L_G t$, $L_G > 0$. Its active expert has depth zero, bounded width, and rank $\mathcal{O}(\varepsilon^{-d_2/s_2})$, so all three are bounded by $\mathcal{O}(\varepsilon^{-1})$. For fixed v and small ε, the tree height is $\mathcal{O}(\varepsilon^{-d_1/s_1})$; expert count and total expert parameters are bounded by* $\exp(C\varepsilon^{-d_1/s_1})$.

Proof. The expert bounds follow from Theorem 4.1 and $d_2/s_2 < 1$. For small ε, $\delta_\varepsilon = \varepsilon/(4C_R L_G)$ gives the height and leaf bounds. Multiplying the leaf count by $\mathcal{O}(M_\varepsilon^2)$ preserves the exponential upper bound. □

We now record the constructive single-NO baseline used throughout the comparison.

Proposition 4.4 (Approximation rates for classical neural operators)**.** *Let $D_i = [0,1]^{d_i}$, let $s_i > 0$, and let K be as in* (2)*. Let $G^+ : H^{s_1}(D_1)^{d_{in}} \to H^{s_2}(D_2)^{d_{out}}$ be uniformly continuous in L^2 with concave modulus ω. Write $R_{K,\mathrm{in}} = \sup_K \|u\|_{H^{s_1}}$ and $R_{K,\mathrm{out}} = \sup_K \|G^+(u)\|_{H^{s_2}} < \infty$. For every $\varepsilon > 0$ there is a rank parameter $N \ge 1$ and $G \in \mathcal{NO}^{\tanh}_{N,W,L,1,\boldsymbol{d}}$, with smooth orthonormal basis elements, satisfying $\sup_K \|G^+ - G\|_{L^2} \le \varepsilon$ and the architecture bounds in* (8)*. In particular, $N \in \mathcal{O}(1 + [\omega^{-1}(\varepsilon/4)]^{-d_1/s_1} \vee \varepsilon^{-d_2/s_2})$ and $W = \max\{d_1, d_2, Nd_{in} + Nd_{out} + 3\}$. Constants depend on the two Sobolev radii. The depth bound retains the full coefficient-dimension dependence; basis evaluation and weight precision are separate.*

Proof. See Section B.1. The proof retains the polynomial degree and dimension dependence explicitly rather than using a fixed-dimensional depth asymptotic. □

4.1 Lower Bounds for Single Models

We next compare the routed construction with lower bounds for single ReLU models of neural-network type. These models require exponentially many parameters on certain Lipschitz targets, while the MoNO expert evaluated after routing has polynomial size. Following Lanthaler & Stuart (2026), a functional $S : L^2(D_1) \to \mathbb{R}$ is *of neural-network type* if $S(u) = \Phi(\mathcal{L}u)$ for a linear map $\mathcal{L} : L^2(D_1) \to \mathbb{R}^m$ (not required to be bounded) and a ReLU MLP Φ; the representation is needed only on K, and its complexity is the smallest number of nonzero weights and biases of such a Φ. For a fixed output point y, an operator $\mathcal{S}$ implemented by a ReLU PCA-Net (Bhattacharya et al., 2021), DeepONet (Lu et al., 2021), or NOMAD (Seidman et al., 2022) yields a functional of this type through point evaluation $u \mapsto \mathcal{S}(u)(y)$. Its parameter count is bounded below by a constant multiple of the resulting functional's complexity (Lanthaler & Stuart, 2026, Sec. 2.4, Lemmas 2.18, 2.20, and 2.22). This point-evaluation argument requires pointwise output accuracy; it does not imply

an operator lower bound in L^2. The corollary instead compares scalar functionals, using an integral readout that is bounded on L^2.

Corollary 4.5 (Active-complexity separation)**.** *Let K be the Sobolev ball* (1) *with $d_{in} = d_{out} = 1$ and $s_i > d_i$ for $i = 1, 2$, and fix $\alpha > 1 + s_1/d_1$ and $\delta > 0$. There exist a Lipschitz functional $S^\dagger : L^2(D_1) \to \mathbb{R}$ and constants $c, \bar{\varepsilon} > 0$ such that, for every $\varepsilon \in (0, \bar{\varepsilon}]$:*

(i) every functional S of neural-network type with $\sup_{u \in K} |S^\dagger(u) - S(u)| \le \varepsilon$ has complexity at least $\exp\big(c\,\varepsilon^{-1/(\alpha+1+\delta)}\big)$;

(ii) the operator $G^+(u) \overset{\text{def.}}{=} S^\dagger(u)\,\mathbf{1}_{D_2}$ admits a MoNO realization G with $\sup_{u \in K} \|G^+(u) - G(u)\|_{L^2(D_2)} \le \varepsilon$, whose constant experts have depth zero, bounded width, rank 1, and $\text{Activ-Cpl}(\mathcal{T}, \mathcal{NO}) \in \mathcal{O}(1)$. The scalar readout $\int_{D_2} G(u)(x)\,\mathrm{d}x$ approximates $S^\dagger$ within ε. For fixed v, Theorem 4.8 supplies routing with $O(\varepsilon^{-d_1/s_1})$ coefficient queries and $O(\varepsilon^{-d_1/s_1} \log(1/\varepsilon))$ scalar operations under the resource model of Section 3.3.1.

Proof. See Section B.5. □

Remark 4.6 (Interpretation of the separation). The hard scalar target uses one stored output value per expert, realizable with ReLU or $\tanh$ (Proposition B.2), paralleling the active/storage trade-off for memorization in Jelassi et al. (2025, Thms. 3.5–3.6). The entropy construction has the exponential power ε^{-d_1/s_1} of the class-uniform bit lower bound (Lanthaler, 2024, Thm. 2.8 and Ex. 2.9), without asserting a matching bit count. The quantized construction includes routing in polynomial per-query work, compared with conventional feedforward evaluation of the stated single-ReLU class; its storage exponent has an additional logarithmic factor.

4.2 Nonconstant Experts and Storage

Additional target regularity lets each expert represent variation within a larger cell. The following result quantifies this effect for routers chosen independently of the target, as in the covering construction above.

Proposition 4.7 (Affine experts versus lookup)**.** *Let $X = L^2([0,1]^d)$ and let K be the radius-R ball of $H^s([0,1]^d)$, with $s > d$ and $R > 0$. Let $\mathcal{F}$ consist of $S \in C^1(X; \mathbb{R})$ satisfying $\sup_K |S| \le 1$, $\sup_X \|DS\|_{X^*} \le 1$, and $\mathrm{Lip}(DS) \le 1$. A router is fixed across $\mathcal{F}$, while its experts may depend on S. The minimum number $\Lambda_0(\varepsilon)$ of constant experts needed to approximate every $S \in \mathcal{F}$ uniformly within ε satisfies, for small ε,*

$$\exp(c\varepsilon^{-d/s}) \le \Lambda_0(\varepsilon) \le \exp(C\varepsilon^{-d/s}).$$

There is a common nearest-center MoNO router with at most $\exp(C\varepsilon^{-d/(2s)})$ leaves for which affine experts approximate every $S(u)\mathbf{1}_{D_2}$ within ε. These experts have depth zero, bounded width, rank $O(\varepsilon^{-d/s})$, and $O(\varepsilon^{-2d/s})$ dense parameters. Including the ReLU center networks and tree records, total stored size is at most $\exp(C\varepsilon^{-d/(2s)})$ real parameters and records. Fixed basis descriptions, coefficient precision, and quadrature costs are excluded.

Proof. A router serving every bounded linear functional must have cells of diameter $O(\varepsilon)$ if its experts are constant. Affine Taylor experts have quadratic local error, permitting radius $O(\sqrt{\varepsilon})$. Sobolev entropy gives the two powers; Section B.8 proves the projection, architecture, and center-storage bounds. □

The lookup lower bound is uniform over the class with a common router; it does not cover target-adapted routing. This construction uses exhaustive search.

4.3 Routing by Quantized Coefficients

A Cartesian cover of a finite-dimensional input projection permits greedy tree descent. It uses more stored cells than the entropy cover, while retaining the same expert bounds.

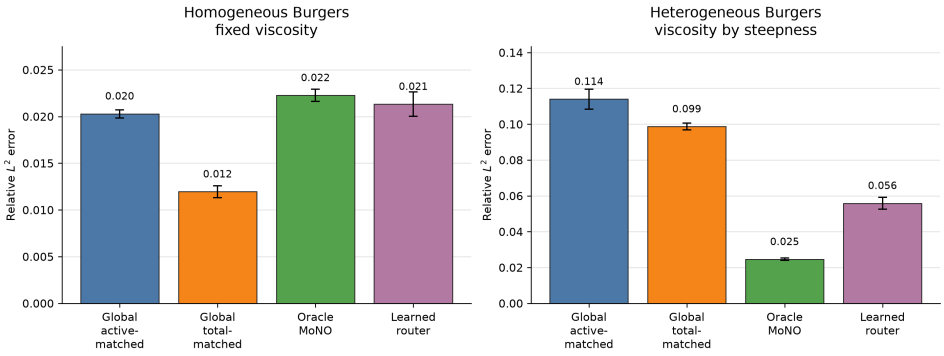

Figure 2: **Controlled Burgers experiments** (40 epochs, CUDA; mean relative L^2 error $\pm$ standard deviation over three seeds). Left: fixed viscosity. Right: four viscosity regimes. Expert architecture and training budget are shared.

Theorem 4.8 (Polynomial routing work)**.** *Under the hypotheses of Theorem 4.1, there is a smooth orthonormal input family of size $J_\varepsilon = O(1+\rho_\varepsilon^{-d_1/s_1})$ with projection error at most $\rho_\varepsilon/8$ on K. Set $B = \lceil \max\{1, R\} \rceil$ and*

$$b_\varepsilon = \max\left\{1, \left\lceil \log_v \frac{8B\sqrt{J_\varepsilon}}{\rho_\varepsilon} \right\rceil \right\}, \qquad h_\varepsilon^{\mathrm{q}} = J_\varepsilon b_\varepsilon, \qquad \Lambda_{\mathrm{q}} = v^{h_\varepsilon^{\mathrm{q}}}.$$

There is a complete v-ary MoNO tree with centers represented in this family and constant experts of rank M_ε, whose uniform error is at most ε. Greedy nearest-child descent finds a nearest leaf for exact coefficients. The error guarantee also holds when each queried coefficient has absolute error at most $\rho_\varepsilon/(8\sqrt{J_\varepsilon})$. Routing uses J_ε coefficient queries, $O(vJ_\varepsilon b_\varepsilon)$ scalar operations, and $O(J_\varepsilon b_\varepsilon)$ node accesses, followed by one expert evaluation. Total storage, including centers, experts, and tree records, is $O(\Lambda_{\mathrm{q}}(J_\varepsilon + M_\varepsilon^2 + 1))$ real slots and records. For fixed v and Lipschitz targets, routing takes $O(\varepsilon^{-d_1/s_1}\log(1/\varepsilon))$ scalar operations and total storage is at most $\exp(C\varepsilon^{-d_1/s_1}\log(1/\varepsilon))$ for small ε.

Proof. Successive tree levels read base-v digits of one projected coefficient. Orthonormality reduces nearest-child comparisons to that coordinate; inflated leaf neighborhoods accommodate coefficient errors. Section B.6 gives the construction, error estimates, and finite-precision accounting. □

5 Controlled Burgers Experiments

We examine specialization on two discrete Burgers solution maps. Inputs are periodic initial conditions $u_0(x) = \sum_{j=1}^{8} a_j \sin(2\pi jx) + b_j \cos(2\pi jx)$ on 64 grid points; targets come from the released finite-volume solver at $T = 0.4$, without claiming grid convergence. Route labels are empirical quartiles of initial steepness $\|\partial_x u_0\|_\infty$. The homogeneous control uses viscosity 0.01; the heterogeneous family assigns $(0.05, 0.02, 0.008, 0.003)$ to the quartiles. These hard switches can make the target discontinuous, so this tests specialization rather than the uniform-continuity theorem. Section F specifies the input distribution and solver.

Each experiment compares four models built from the same 1D Fourier-style neural operator (Li et al., 2021) with 10 Fourier modes and 4 spectral blocks. The active-matched global NO (width 16) has the size of one MoNO expert; the total-matched global NO (width 33) matches the stored size of the learned-router MoNO; the oracle MoNO dispatches each input to the expert indexed by $r(u_0)$; and the learned-router MoNO trains its router by supervised classification of $r(u_0)$ and evaluates only the selected expert at test time (details in Section F).

Figure 2 reports the 40-epoch CUDA runs. Heterogeneous errors are 0.1140 and 0.0986 for the active- and total-matched global NOs, versus 0.0245 for oracle and 0.0558 for learned routing. The larger global model is best on the homogeneous control. Table 3 gives variability and parameter counts.

Accuracy and active size. Each expert has $21{,}905$ parameters; the learned MoNO evaluates $27{,}477$ including its router, compared with $92{,}863$ for the total-matched global model. A separate 40-epoch MPS width sweep gives global errors near 0.073–0.075 at widths 128–256, versus 0.060 for learned routing (Section F.1). Learning curves and the conditioning controls below prevent interpreting this as an intrinsic global capacity requirement.

On MPS, global models given the router's features or the oracle label attain 0.0917 and 0.0362, versus 0.0598 and 0.0245 for the corresponding MoNOs (Section F.2). Feature conditioning matches inference inputs but leaves the MoNO with additional label supervision during training. Input-space nearest-center routing performs poorly (0.1381). Thus information supplied during training and inference materially affects the comparison.

A separate 100-epoch MPS ablation uses training-only thresholds and shares one supervised router (Section F.3). Route accuracy is 89.97%: FNO experts attain 0.0388 versus 0.0452 for a global FNO receiving the same predicted labels, a $1.16\times$ error ratio at comparable active size. Evaluating these unchanged checkpoints with true labels gives 0.0142 versus 0.0215 ($1.51\times$); this is an oracle diagnostic. Misrouted samples have mean relative error about 0.26 for both models. The MoNO stores $93{,}448$ values versus $27{,}797$ for the global predictor. Under input k-means routing, four FNO experts attain 0.1117, versus 0.3412 for 1024 constant experts and 0.3176 for full-training-set lookup. Figure 5 reports the full expert and storage comparison.

6 Discussion and Limitations

MoNOs separate the resources needed for approximation: constant experts give small active models, affine experts reduce storage under additional regularity, and coefficient quantization permits polynomial routing work under the stated access model. The hard-functional comparison includes routing, while retaining exponential storage.

Limitations and further directions. The experiments learn one-dimensional discrete Burgers maps with hard regime boundaries; they do not test the smooth-target storage rate. Validation curves still improve near the training limit, and routing errors materially affect accuracy. Converged training, continuously varying regimes, higher-dimensional PDEs, and measured latency and memory are needed to assess practical scaling. Transfer of the theory beyond cubes requires controlled Sobolev extension: the zero extension in Proposition E.1 preserves the L^2 modulus but need not preserve Sobolev regularity.

AI Use Statement

Generative AI tools assisted with editing, reference checks, mathematical drafting and checking, experiment code, and checkpoint diagnostics. This included work on the proofs of Theorems 4.1 and 4.8 and Proposition 4.7. The authors are responsible for all mathematical claims, experimental results, code, and references.

Ethics Statement

This work is theoretical and uses only procedurally generated PDE data. It involves no human subjects or personal data, and we are not aware of specific ethical concerns.

Reproducibility Statement

Assumptions are stated in Sections 3 and 4, and complete proofs are given in Sections B, D, and E. The Burgers experiments use procedurally generated data; the data generation, architectures, training protocol, random seeds, and hardware are described in Section 5 and Section F, and the scripts that produce all reported numbers and figures are included in the supplementary material.

REFERENCES

Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Moraga. Near-optimal learning of Banach-valued, high-dimensional functions via deep neural networks. *Neural Networks*, 181: 106761, 2025. doi: 10.1016/j.neunet.2024.106761.

David Amsallem, Matthew J. Zahr, and Charbel Farhat. Nonlinear model order reduction based on local reduced-order bases. *International Journal for Numerical Methods in Engineering*, 92(10): 891–916, 2012. doi: 10.1002/nme.4371.

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

Pau Batlle, Matthieu Darcy, Bamdad Hosseini, and Houman Owhadi. Kernel methods are competitive for operator learning. *Journal of Computational Physics*, 496:112549, 2024. doi: 10.1016/j.jcp.2023.112549.

Cenk Baykal, Nishanth Dikkala, Rina Panigrahy, Cyrus Rashtchian, and Xin Wang. A theoretical view on sparsely activated networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yoav Benyamini and Joram Lindenstrauss. *Geometric nonlinear functional analysis. Vol. 1*, volume 48 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2000. ISBN 0-8218-0835-4. doi: 10.1090/coll/048. URL https://doi.org/10.1090/coll/048.

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. doi: 10.5802/smai-jcm.74.

Mikhail Shlemovich Birman and Mikhail Zakharovich Solomyak. Piecewise-polynomial approximations of functions of the classes W_p^α. *Matematicheskii Sbornik*, 73(115)(3):331–355, 1967.

Enric Boix-Adsera and Philippe Rigollet. The power of fine-grained experts: Granularity boosts expressivity in mixture of experts. *arXiv preprint arXiv:2505.06839*, 2025.

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. doi: 10.1109/72.392253.

Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jin Cheng, Mourad Choulli, and Junshan Lin. Stable determination of a boundary coefficient in an elliptic equation. *Math. Models Methods Appl. Sci.*, 18(1):107–123, 2008. ISSN 0218-2025. doi: 10.1142/S0218202508002620. URL https://doi.org/10.1142/S0218202508002620.

Maarten V de Hoop, Matti Lassas, and Christopher A Wong. Deep learning architectures for nonlinear operator functions and nonlinear inverse problems. *Mathematical Statistics and Learning*, 4(1): 1–86, 2022.

Dwyer Deighan, Jonas A. Actor, and Ravi G. Patel. Mixture of neural operator experts for learning boundary conditions and model selection. *arXiv preprint arXiv:2502.04562*, 2025. URL https://arxiv.org/abs/2502.04562.

Jens L. Eftang, Anthony T. Patera, and Einar M. Rønquist. An "hp" certified reduced basis method for parametrized elliptic partial differential equations. *SIAM Journal on Scientific Computing*, 32 (6):3170–3200, 2010. doi: 10.1137/090780122.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

David Gilbarg and Neil S. Trudinger. *Elliptic partial differential equations of second order*, volume Vol. 224 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, Berlin-New York, 1977. ISBN 3-540-08007-4.

Lukas Herrmann, Christoph Schwab, and Jakob Zech. Neural and spectral operator surrogates: unified construction and expression rate bounds. *Advances in Computational Mathematics*, 50(4): 72, 2024. doi: 10.1007/s10444-024-10171-2.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Samy Jelassi, Clara Mohri, David Brandfonbrener, Alex Gu, Nikhil Vyas, Nikhil Anand, David Alvarez-Melis, Yuanzhi Li, Sham M. Kakade, and Eran Malach. Mixture of parrots: Experts improve memorization more than reasoning. In *International Conference on Learning Representations*, 2025.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Wenxin Jiang and Martin A. Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *The Annals of Statistics*, 27(3): 987–1011, 1999. doi: 10.1214/aos/1018031265.

Yuling Jiao, Yanming Lai, Xiliang Lu, Fengru Wang, Jerry Zhijian Yang, and Yuanyuan Yang. Deep neural networks with ReLU-sine-exponential activations break curse of dimensionality in approximation on Hölder class. *SIAM Journal on Mathematical Analysis*, 55(4):3635–3649, 2023.

Michael I. Jordan and Robert A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994. doi: 10.1162/neco.1994.6.2.181.

Yury Korolev. Two-layer neural networks with values in a Banach space. *SIAM Journal on Mathematical Analysis*, 54(6):6358–6389, 2022.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *J. Mach. Learn. Res.*, 22:Paper No. [290], 76, 2021. ISSN 1532-4435,1533-7928.

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.

Anastasis Kratsios and Léonie Papon. Universal approximation theorems for differentiable geometric deep learning. *Journal of Machine Learning Research*, 23(196):1–73, 2022. URL `https://jmlr.org/papers/v23/21-0716.html`.

Anastasis Kratsios, Haitz Sáez de Ocáriz Borde, Takashi Furuya, and Marc T. Law. Approximation rates and VC-dimension bounds for (P)ReLU MLP mixture of experts. *Transactions on Machine Learning Research*, 2025.

Samuel Lanthaler. Operator learning with PCA-Net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023. URL `http://jmlr.org/papers/v24/23-0478.html`.

Samuel Lanthaler. Operator learning of Lipschitz operators: An information-theoretic perspective. *arXiv preprint arXiv:2406.18794*, 2024.

Samuel Lanthaler and Andrew M. Stuart. The parametric complexity of operator learning. *IMA Journal of Numerical Analysis*, 46(2):647–712, 2026. doi: 10.1093/imanum/draf028. Published online 2025. Theorem, lemma, and equation numbers cited in this paper follow arXiv:2306.15924v4.

Samuel Lanthaler, Zongyi Li, and Andrew M Stuart. Nonlocality and nonlinearity implies universality in operator learning. *Constructive Approximation*, 62(2):261–303, 2025. doi: 10.1007/s00365-025-09718-3.

J. Antonio Lara Benitez, Kareem Hegazy, Junyi Guo, Ivan Dokmanić, Michael W. Mahoney, and Maarten V. de Hoop. Neural equilibria for long-term prediction of nonlinear conservation laws. *arXiv preprint arXiv:2501.06933*, 2025.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Zongyi Li, Nikola Borislavov 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. URL https://openreview.net/forum?id=c8P9NQVtmnO.

George G. Lorentz, Manfred v. Golitschek, and Yuly Makovoz. *Constructive approximation*, volume 304 of *Grundlehren der mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1996. ISBN 3-540-57028-4. doi: 10.1007/978-3-642-60932-9. URL https://doi.org/10.1007/978-3-642-60932-9. Advanced problems.

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.

Harris Abdul Majid and Francesco Tudisco. Mixture of neural operators: Incorporating historical information for longer rollouts. In *ICLR 2024 Workshop on AI4DifferentialEquations In Science*, 2024. URL https://openreview.net/forum?id=rYZjCmIlmx.

Carlo Marcati and Christoph Schwab. Exponential convergence of deep operator networks for elliptic partial differential equations. *SIAM Journal on Numerical Analysis*, 61(3):1513–1545, 2023.

Michael McCabe, Peter Harrington, Shashank Subramanian, and Jed Brown. Towards stability of autoregressive neural operators. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=RFfUUtKYOG.

Roberto Molinaro, Yunan Yang, Björn Engquist, and Siddhartha Mishra. Neural inverse operators for solving PDE inverse problems. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25105–25139. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/molinaro23a.html.

Nicholas H. Nelsen and Andrew M. Stuart. Operator learning using random features: A tool for scientific computing. *SIAM Review*, 66(3):535–571, 2024. doi: 10.1137/24M1648703.

Hien D. Nguyen, Luke R. Lloyd-Jones, and Geoffrey J. McLachlan. A universal approximation theorem for mixture-of-experts models. *Neural Computation*, 28(12):2585–2593, 2016. doi: 10.1162/NECO_a_00892.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

Bogdan Raonic, 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. *Advances in Neural Information Processing Systems*, 36, 2023.

Jacob H. Seidman, Georgios Kissas, Paris Perdikaris, and George J. Pappas. NOMAD: Nonlinear manifold decoders for operator learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ramansh Sharma and Varun Shankar. Ensemble and mixture-of-experts DeepONets for operator learning. *Transactions on Machine Learning Research*, 2025.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network with approximation error being reciprocal of width to power of square root of depth. *Neural Computation*, 33(4):1005–1036, 2021.

Hong Wang, Haiyang Xin, Jie Wang, Xuanze Yang, Fei Zha, Huanshuo Dong, and Yan Jiang. Mixture-of-experts operator transformer for large-scale PDE pre-training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 35566–35595, 2025.

Mingze Wang and Weinan E. On the expressive power of mixture-of-experts for structured complex tasks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Conference on learning theory*, pp. 639–649. PMLR, 2018.

Dmitry Yarotsky. Elementary superexpressive activations. In *International Conference on Machine Learning*, pp. 11932–11940. PMLR, 2021.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Assaf J. Zeevi, Ronny Meir, and Vitaly Maiorov. Error bounds for functional approximation and estimation using mixtures of experts. *IEEE Transactions on Information Theory*, 44(3):1010–1025, 1998. doi: 10.1109/18.669150.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Zhao, Andrew M Dai, Zhifeng Chen, Quoc V Le, and James Laudon. Mixture-of-experts with expert choice routing. *Advances in Neural Information Processing Systems*, 35:7103–7114, 2022.

A NOTATION

For noninteger $s > 0$ and the cube $D = [0,1]^d$, writing $m = \lfloor s \rfloor$, we use the equivalent Sobolev–Slobodeckij norm

$$\|f\|_{H^s(D)} \overset{\text{def.}}{=} \Big(\sum_{|\alpha| \le m} \|\partial^\alpha f\|^2_{L^2(D)} \Big)^{1/2} + |f|_{H^s(D)}, \tag{7}$$

$$|f|^2_{H^s(D)} \overset{\text{def.}}{=} \sum_{|\alpha| = m} \int_{D \times D} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|^2}{|x - y|^{2(s-m)+d}} \, \mathrm{d}x \, \mathrm{d}y.$$

The remaining notation is collected in Table 2.

Table 2: Notation used throughout the paper, in order of appearance.

Symbol	Meaning
d_1, d_2; $D_i = [0,1]^{d_i}$	dimensions of the input and output domains; the domains
d_{in}, d_{out}	numbers of input and output channels
s_1, s_2	Sobolev regularity of input and output functions
ω, ω^{-1}	chosen strictly increasing concave modulus of the target operator and its inverse
K, R, f_0	compact approximation set, Sobolev radius, and center ((1), (2))
$\mathcal{NN}^{\sigma}_{\Delta,W:p,q}$	MLPs $\mathbb{R}^p \to \mathbb{R}^q$ with activation σ, depth $\le \Delta$, width $\le W$ (Definition 3.1)
$\{\varphi_m\}$, $\{\psi_n\}$	orthonormal families in $L^2(D_1)$ and $L^2(D_2)$; smooth perturbations of approximation spaces in the proofs
N	number of scalar basis modes in the nonlocal maps $K_N^{(0)}$, $K_N^{(L+1)}$ (Definition 3.2)
L, W, Δ	NO depth, NO width, and depth of the coordinate bias networks $b^{(0)}$, $b^{(L+1)}$
$P(G)$, $\bar{P}_{\Delta,W,L,N,\boldsymbol{d}}$	number of parameters of a NO, and its architecture-level upper bound (6)
$\mathcal{NO}^{\tanh}_{N,W,L,\Delta,\boldsymbol{d}}$	class of neural operators of Definition 3.2
$\mathcal{T} = (V, E, f_{1,1})$	routing tree with nodes V, edges E, and root $f_{1,1}$
v, h, $\Lambda = v^h$	valency, height (selected path length), and number of leaves (experts) of $\mathcal{T}$
$\hat{f}_\ell$, c_ℓ, G_ℓ	leaf centers (ReLU networks or smooth-basis coefficient vectors) and attached expert
Activ-Cpl, Dist-Cpl	parameters of the largest expert (active) and of all experts (total)
G^+	target operator
$R_{G,K}$ (also $R_{K,\mathrm{out}}$), $R_{K,\mathrm{in}}$	Sobolev radius of $G^+(K)$ and of K
N_ε, M_ε, W_*	single-NO rank, output-only expert rank, and fixed expert width in Theorem 4.1
$\rho_\varepsilon, \delta_\varepsilon, h_\varepsilon$	local diameter bound, cover parameter, and tree height used in Theorem 4.1
C_R	covering constant of Lemma D.1
p, q, a, η, n, b	coefficient dimensions, diameter, resolution, tensor degree, and total degree in Proposition 4.4
Q_α, $S^\dagger$	infinite-dimensional hypercube and hard functional of Corollary 4.5
$\mathcal{F}$, $\Lambda_0(\varepsilon)$	bounded $C^{1,1}$ class and minimum constant-expert count with a target-independent router (Proposition 4.7)
J_ε, b_ε, $h^{\mathrm{q}}_\varepsilon$, Λ_{q}	input projection dimension, digits per coefficient, tree height, and leaf count in Theorem 4.8
$e_k(K)$	entropy numbers (Section B)

B Detailed Proofs

The proofs rely on several key notions from constructive approximation. We begin by recalling the concept of *metric entropy* for a compact subset of $L^2(D)$. For each $k \in \mathbb{N}$, this is quantified by its kth entropy number, defined as

$$e_k(K) \overset{\text{def.}}{=} \inf\left\{\delta > 0 : \exists f_1, \dots, f_{2^k-1} \in L^2(D), \quad \max_{f\in K} \min_{i=1,\dots,2^k-1} \|f - f_i\|_{L^2(D)} < \delta\right\}.$$

See Lorentz et al. (1996, Chapter 15) for further details.

The single-NO proof combines Sobolev projection with a finite-dimensional polynomial approximation and a narrow network realization, retaining the dependence on the growing coefficient dimension. The routed proof combines the same output projection spaces with a Sobolev cover.

B.1 Proof of Proposition 4.4

The explicit architecture bounds retained in Proposition 4.4 are

$$N \in \mathcal{O}\left(1 + \left(\frac{R_{K,\mathrm{in}}}{\omega^{-1}(\varepsilon/4)}\right)^{d_1/s_1} \vee \left(\frac{R_{K,\mathrm{out}}}{\varepsilon}\right)^{d_2/s_2}\right),$$
$$W = \max\{d_1, d_2, p+q+3\}, \qquad L \le Cq(b+1)\binom{p-1+b}{b}. \tag{8}$$

Here $p = Nd_{in}$, $q = Nd_{out}$, $a = \mathrm{diam}_{L^2}(K)$, $\eta = \omega^{-1}(\varepsilon/(4\sqrt{q}))$, $n = \max\{1, \lceil pa^2/(4\eta^2)\rceil\}$, and $b = pn$. The constant C is independent of these construction parameters. Basis evaluation and weight precision are outside this parameter bound.

Sobolev projection. The approximation spaces of Birman & Solomyak (1967) give scalar N-dimensional subspaces with L^2 projection errors bounded by $C_i N^{-s_i/d_i}$ on the corresponding unit Sobolev balls. Their raw piecewise-polynomial elements need not belong to H^{s_i}. We therefore first replace each finite orthonormal family by sufficiently close smooth functions and orthonormalize. This can change the projection operator by an arbitrarily small amount in operator norm: the Gram matrix and its inverse square root depend continuously on a linearly independent finite family. Choosing this perturbation smaller than N^{-s_i/d_i} preserves the rate. Include the constant function by enlarging the space by at most one dimension. Smooth density also permits completion to an orthonormal basis. The bases may depend on the chosen N; no assertion of nested optimal spaces is needed.

Write E_i for coefficient extraction and D_i^{coef} for reconstruction in these bases, applied componentwise. Both have norm one and $P_i = D_i^{\mathrm{coef}} E_i$ is an orthogonal projection. Choose N large enough that

$$\sup_{u\in K} \|u - P_1 u\|_{L^2} \le \omega^{-1}(\varepsilon/4), \qquad \sup_{u\in K} \|G^+(u) - P_2 G^+(u)\|_{L^2} \le \varepsilon/4. \tag{9}$$

This gives the stated rank bound. The factor $1/4$ must remain inside ω^{-1} for a general modulus; it cannot be absorbed into an outside constant without an additional scaling assumption on ω.

Finite-dimensional approximation with dimension dependence retained. Set $p = Nd_{in}$, $q = Nd_{out}$ and $F = E_2 G^+ D_1^{\mathrm{coef}} : \mathbb{R}^p \to \mathbb{R}^q$. The reconstructed inputs are smooth, so G^+ is defined on them. The map F has modulus at most ω, and $A = E_1(K)$ has diameter at most $a = \mathrm{diam}_{L^2}(K)$. Moreover,

$$\sup_{u\in K} \|G^+(u) - D_2^{\mathrm{coef}} F(E_1 u)\|_{L^2} \le \varepsilon/2.$$

If $a = 0$, a constant output suffices. Otherwise enclose A in a translated cube of side a and rescale it to $[0,1]^p$. For each component of F, its tensor Bernstein polynomial of degree n in each variable has error at most

$$\omega\left(a\sqrt{p/(4n)}\right).$$

Indeed, the Bernstein polynomial is the expectation at independent binomial grid points; their expected squared distance from the evaluation point is at most $p/(4n)$. Jensen's inequality for the concave ω, followed by Cauchy–Schwarz, gives this bound. Taking $\eta = \omega^{-1}(\varepsilon/(4\sqrt{q}))$ and $n = \max\{1, \lceil pa^2/(4\eta^2)\rceil\}$ makes the Euclidean error at most $\varepsilon/4$.

Each component polynomial has total degree at most $b = pn$. By the polynomial decomposition in Kratsios & Papon (2022, Lemma 58), it is a sum of at most $\binom{p-1+b}{b}$ univariate polynomials of degree at most b applied to linear forms. Each such polynomial can be approximated on a compact interval to arbitrary accuracy by $b+1$ $\tanh$ units, using finite differences of the activation; the finite-difference construction and its serialization into a narrow network are described in Kratsios & Papon (2022, proof of Proposition 59). Allocating a further $\varepsilon/4$ to these approximations gives a network $\Phi : \mathbb{R}^p \to \mathbb{R}^q$ of width $p+q+2$ and depth

$$L \le Cq(b+1)\binom{p-1+b}{b}, \qquad \sup_{z\in A} \|F(z) - \Phi(z)\|_2 \le \varepsilon/2.$$

Here C is an absolute serialization constant, independent of p, q, n, ε; the weights and the accuracy of the near-identity channels can depend on all these quantities. The displayed combinatorial count retains dimension dependence that a fixed-dimensional asymptotic estimate can hide. In particular the total degree is pn, not n.

Realization as a neural operator. The input lifting map writes $E_1 u$ as a spatially constant channel vector, using the constant output basis mode. The pointwise layers implement Φ, and the final finite-rank map reconstructs its output coefficients. Its affine output matrix is absorbed into the final finite-rank map. One additional constant hidden channel carries the affine output bias; a nonzero constant can be maintained exactly through $\tanh$ by a fixed bias. Thus $W = \max\{d_1, d_2, p+q+3\}$ suffices. The coordinate-bias networks may be zero. Since reconstruction has norm one, the projection and network errors sum to at most ε. This proves the proposition.

B.2 Proof of Lemma D.1

The Sobolev entropy bound of Birman & Solomyak (1967) gives $e_k(K) \le Ck^{-s_1/d_1}$. Centers may be chosen in K: replace any covering center whose ball meets K by a point of that intersection, increasing the radius by at most a factor two. Thus there are $M \le 2^k - 1$ centers $f_1, \dots, f_M \in K$ covering K to radius $C' k^{-s_1/d_1}$.

Set $k = \lceil \delta^{-d_1/s_1} \rceil$ and $h = \lceil \log_v(2^k - 1) \rceil$. Construct vertices (j, i) for $0 \le j \le h$ and $1 \le i \le v^j$, with parent $(j-1, \lceil i/v \rceil)$ for every $j \ge 1$. Every internal vertex then has exactly v children. Assign the M cover centers to the first M leaves and duplicate centers on unused leaves. Internal centers can all be the zero function. This indexing, rather than a nearest-parent assignment, guarantees the prescribed valency.

Since $s_1 > d_1$, the finitely many f_i have continuous representatives. Uniform ReLU approximation on the cube gives center networks $\hat{f}_i$ with $\|f_i - \hat{f}_i\|_{L^2} < \delta$. Consequently

$$\sup_{u \in K} \min_i \|u - \hat{f}_i\|_{L^2} \le C' k^{-s_1/d_1} + \delta \le (C'+1)\delta < C_R \delta$$

for a fixed $C_R > C' + 1$. Duplicating leaves preserves this estimate. The height, leaf count, and node count follow directly from the indexed tree. No property of the internal centers is needed for exhaustive nearest-leaf routing.

B.3 Proof of Proposition E.1

Since $H^{s_1}(D_1)$ is dense in $L^2(D_1)$ and $L^2(D_2)$ is complete, the uniformly continuous map G^+ has a unique continuous extension $\tilde{G} : L^2(D_1) \to L^2(D_2)$ with the same modulus of continuity. Indeed, for $u \in L^2(D_1)$ choose $u_n \in H^{s_1}(D_1)$ with $u_n \to u$ in $L^2(D_1)$ and set $\tilde{G}(u) = \lim_n G^+(u_n)$; uniform continuity makes this limit independent of the approximating sequence and preserves the modulus. For $i = 1, 2$, consider the *restriction* $\rho_{(i)} : L^2([0,1]^{d_i}) \to L^2(D_i)$ and *extension by zero* $E^{(i)} : L^2(D_i) \to L^2([0,1]^{d_i})$ operators defined by

$$\rho_{(i)}(u) \overset{\text{def.}}{=} u|_{D_i} \quad \text{and} \quad E^{(i)}(u) \overset{\text{def.}}{=} u \mathbf{1}_{D_i}. \tag{10}$$

Define

$$\bar{G} \overset{\text{def.}}{=} E^{(2)} \circ \tilde{G} \circ \rho_{(1)} : L^2([0,1]^{d_1}) \to L^2([0,1]^{d_2}).$$

Then $\bar{G}$ is uniformly continuous with modulus of continuity ω. Moreover, by construction, for each $u \in H^{s_1}(D_1)$

$$\bar{G}(\bar{u})|_{D_2} \overset{\text{def.}}{=} \bar{G}(E^{(1)}(u))|_{D_2} = \rho_{(2)} \circ E^{(2)} \circ \tilde{G} \circ \rho_{(1)} \circ E^{(1)}(u) = \tilde{G}(u) = G^+(u). \tag{11}$$

The right-hand side of (11) holds since $\tilde{G}$ extends G^+, and the left-hand side uses $E^{(1)}(u) = \bar{u}$.

Now, observe that, by construction, for $i = 1, 2$, $\rho_{(i)} \circ E^{(i)}$ is the identity on $L^2(D_i)$. Moreover, both maps are 1-Lipschitz since: for each $i = 1, 2$, every $u \in L^2(D_i)$, and each $v \in L^2([0,1]^{d_i})$

$$\|E^{(i)}(u)\|^2_{L^2([0,1]^{d_i})} = \int_{x \in D_i} \|u(x)\|^2 dx + \int_{x \in [0,1]^{d_i} \setminus D_i} 0 \, dx$$

$$
\begin{aligned}
&= \int_{x\in D_i} \|u(x)\|^2 dx = \|u\|^2_{L^2(D_i)} \\
\|\rho_{(i)}(v)\|^2_{L^2(D_i)} &= \int_{x\in D_i} \|v(x)\|^2 dx \\
&\le \int_{x\in D_i} \|v(x)\|^2 dx + \int_{x\in[0,1]^{d_i}\setminus D_i} \|v(x)\|^2 dx \\
&= \|v\|^2_{L^2([0,1]^{d_i})}.
\end{aligned}
$$

B.4 Proof of Theorem 4.1

Choose a smooth output space of dimension $M = M_\varepsilon$, containing $\mathbf{1}_{D_2}$, whose orthogonal projection P_2 satisfies the output estimate in (9). The Sobolev projection argument gives $M \in \mathcal{O}(1+\varepsilon^{-d_2/s_2})$, independently of the input resolution. Let $\rho = \rho_\varepsilon \le 1$, $\delta = \delta_\varepsilon$, and take the balanced neural cover from Lemma D.1. Its leaves satisfy

$$
\sup_{u\in K} \min_i \|u - \hat{f}_i\|_{L^2} < C_R\delta = \rho/2. \tag{12}
$$

Set $K_i = \{u \in K : \|u - \hat{f}_i\|_{L^2} \le \rho/2\}$. Every nonempty K_i is compact and has diameter at most ρ. Choose $u_i \in K_i$ and define the input-independent expert

$$
\hat{G}_i(u) = P_2 G^+(u_i).
$$

For $u \in K_i$, the diameter bound and the choice $\rho \le \omega^{-1}(\varepsilon/2)$ give

$$
\|G^+(u) - \hat{G}_i(u)\|_{L^2} \le \omega(\rho) + \varepsilon/4 \le \frac{\varepsilon}{2} + \frac{\varepsilon}{4} < \varepsilon.
$$

Empty cells may carry arbitrary zero experts; the nearest-leaf rule never selects one on K.

These experts belong to Definition 3.2. Take $L = \Delta = 0$, zero input kernel, constant lifting bias $u_1(x) = 1$, and zero output bias. Choose the first output basis element $\psi_1 = \mathbf{1}_{D_2}$ and set $C^{(1)}_{1,n} = (G^+(u_i), \psi_n)_{L^2(D_2)}$ for $1 \le n \le M$, with the other output-kernel coefficients zero. Since $|D_2| = 1$, the output kernel reconstructs $P_2G^+(u_i)$ exactly from these M coefficients. Any smooth orthonormal input family of size M is allowed because the input kernel vanishes. The width $W_* = \max\{d_1, d_2, 1\}$ satisfies the architectural constraint, and (6) gives

$$
P(\hat{G}_i) \le 2M^2W_*^2 + 2W_*(W_*+1) \in \mathcal{O}(M^2).
$$

Thus the experts have depth zero, bounded width, and rank parameter M_ε. Routing selects a cell containing u by (12), establishing the uniform error.

B.5 Proof of Corollary 4.5

Numbered results of Lanthaler & Stuart (2026) refer to arXiv:2306.15924v4. We use the notion of an infinite-dimensional hypercube of Lanthaler & Stuart (2026, Def. 2.3). Let X be a Banach space, $A > 0$, $\alpha > 1$, and let $e_1, e_2, \dots \in X$ be linearly independent with $\|e_j\|_X = 1$ and with a biorthogonal sequence $e_1^*, e_2^*, \dots \in X^*$ satisfying $\sup_j \|e_j^*\|_{X^*} < \infty$. A set $K \subseteq X$ *contains the hypercube* $Q_\alpha(A; e_1, e_2, \dots)$ if it contains every $u = A\sum_{j\ge1} j^{-\alpha} y_j e_j$ with $y_j \in [0,1]$ for all j.

Lemma B.1 (Sobolev balls contain infinite-dimensional hypercubes)**.** *Let K be the Sobolev ball* (1) *with $d_{in} = 1$, viewed as a compact subset of $X = L^2(D_1)$. For every $\alpha > 1 + s_1/d_1$ there is $A > 0$ such that K contains a hypercube $Q_\alpha(A; e_1, e_2, \dots)$ whose biorthogonal functionals have norm one.*

Proof. Let $\{e_j\}_{j\ge1}$ be the real trigonometric orthonormal basis of $L^2(D_1)$ formed by the tensor products of the functions 1, $\sqrt{2}\cos(2\pi kt)$, and $\sqrt{2}\sin(2\pi kt)$, $k \in \mathbb{N}$, enumerated so that the frequency vector $k_j \in \mathbb{N}^{d_1}_{\ge0}$ of e_j has nondecreasing $|k_j|_\infty$. The e_j are linearly independent, $\|e_j\|_{L^2(D_1)} = 1$, and $e_j^* \overset{\text{def.}}{=} (\cdot, e_j)_{L^2(D_1)}$ is a biorthogonal sequence of norm-one functionals. Every frequency vector k with $|k|_\infty \le n-1$ contributes at least one basis function, and there are n^{d_1} such vectors, so $|k_j|_\infty \le j^{1/d_1}$ for all j.

We claim that $\|e_j\|_{H^{s_1}(D_1)} \le C_1\, j^{s_1/d_1}$ with C_1 depending only on d_1 and s_1. For every multi-index β, $\partial^\beta e_j$ is a trigonometric product scaled by $\prod_i (2\pi k_{j,i})^{\beta_i}$, so $\|\partial^\beta e_j\|_{L^2} \le (2\pi|k_j|_\infty)^{|\beta|}$, $\|\partial^\beta e_j\|_{L^\infty} \le 2^{d_1/2}(2\pi|k_j|_\infty)^{|\beta|}$, and $\|\nabla\partial^\beta e_j\|_{L^\infty} \le 2^{d_1/2}\sqrt{d_1}(2\pi|k_j|_\infty)^{|\beta|+1}$. This bounds the integer-order part of the norm by a constant multiple of $(1+|k_j|_\infty)^{\lfloor s_1\rfloor}$. If s_1 is not an integer, write $s_1 = m+\theta$ with $m = \lfloor s_1\rfloor$ and $\theta \in (0,1)$. For $g = \partial^\beta e_j$ with $|\beta| = m$ and $k_j \neq 0$, split the Gagliardo integral at $|x-y| = r \overset{\text{def.}}{=} |k_j|_\infty^{-1}$ and use $|g(x)-g(y)| \le \min\{2\|g\|_{L^\infty}, \|\nabla g\|_{L^\infty}|x-y|\}$:

$$\int_{D_1\times D_1} \frac{|g(x)-g(y)|^2}{|x-y|^{2\theta+d_1}}\,\mathrm{d}x\,\mathrm{d}y \lesssim \|\nabla g\|_{L^\infty}^2\, r^{2-2\theta} + \|g\|_{L^\infty}^2\, r^{-2\theta} \lesssim |k_j|_\infty^{2(m+\theta)},$$

where the implied constants depend only on d_1 and s_1 (for $k_j = 0$ the seminorm vanishes). Hence $\|e_j\|_{H^{s_1}(D_1)} \lesssim (1+|k_j|_\infty)^{s_1} \lesssim j^{s_1/d_1}$, proving the claim.

Now let $u = A\sum_{j\ge1} j^{-\alpha} y_j e_j$ with $y_j \in [0,1]$. Since $\alpha - s_1/d_1 > 1$, the series converges absolutely in $H^{s_1}(D_1)$ and

$$\|u\|_{H^{s_1}(D_1)} \le A\, C_1 \sum_{j\ge1} j^{s_1/d_1-\alpha} \overset{\text{def.}}{=} A\, C_2 < \infty.$$

Choosing $A \overset{\text{def.}}{=} R/C_2$ gives $u \in K$, that is, $Q_\alpha(A; e_1, e_2, \dots) \subseteq K$. □

Proof of Corollary 4.5. Part (i). By the Rellich–Kondrashov theorem, K is a compact subset of the infinite-dimensional Banach space $X = L^2(D_1)$, and by Lemma B.1 it contains a hypercube $Q_\alpha(A; e_1, e_2, \dots)$ with $\alpha > 1 + s_1/d_1 > 1$ and norm-one biorthogonal functionals. Lanthaler & Stuart (2026, Thm. 2.11), applied with $r = 1$ and the given $\delta > 0$, yields constants $c, \bar\varepsilon > 0$ and a Fréchet differentiable functional $S^\dagger$ such that every functional S of neural-network type with $\sup_{u\in K}|S^\dagger(u) - S(u)| \le \varepsilon \le \bar\varepsilon$ has complexity at least $\exp(c\,\varepsilon^{-1/(\alpha+1+\delta)})$. A single model of this type evaluates its network Φ on every query, so all of these parameters are active.

It remains to show that $S^\dagger$ is Lipschitz on X. The functional constructed in the proof of Lanthaler & Stuart (2026, Thm. 2.11) is, for $r = 1$,

$$S^\dagger(u) = \sum_{k\ge1} 2^{-\alpha^* k}\, f_{2^k}\Big(A^{-1}2^{(k+1)\alpha}\big(e^*_{2^k}(u), \dots, e^*_{2^{k+1}-1}(u)\big)\Big), \qquad \alpha^* \overset{\text{def.}}{=} 1+\alpha+\delta/2,$$

where each $f_D : \mathbb{R}^D \to \mathbb{R}$ is continuously differentiable with all partial derivatives of order at most one bounded by 1 (Lanthaler & Stuart, 2026, eq. (A.12) and Lemma A.7). Differentiating term by term, as in the proof of Lanthaler & Stuart (2026, Lemma A.7), for $u, v \in X$,

$$\begin{aligned}\big|dS^\dagger(u)[v]\big| &\le \sum_{k\ge1} 2^{-\alpha^* k} A^{-1} 2^{(k+1)\alpha} \sum_{j=2^k}^{2^{k+1}-1} |e_j^*(v)| \\ &\le A^{-1}2^\alpha\, \|v\|_{L^2(D_1)} \sum_{k\ge1} 2^{(1+\alpha-\alpha^*)k} = \frac{A^{-1}2^\alpha}{2^{\delta/2}-1}\,\|v\|_{L^2(D_1)}.\end{aligned}$$

Since X is convex, the mean value inequality shows that $S^\dagger$ is Lipschitz on X with constant $\mathrm{Lip}(S^\dagger) \le A^{-1}2^\alpha/(2^{\delta/2}-1)$.

Part (ii). Set $G^+(u) \overset{\text{def.}}{=} S^\dagger(u)\mathbf{1}_{D_2}$. Since $|D_2| = 1$, for $u, \tilde u \in H^{s_1}(D_1)$,

$$\|G^+(u) - G^+(\tilde u)\|_{L^2(D_2)} = |S^\dagger(u) - S^\dagger(\tilde u)| \le \mathrm{Lip}(S^\dagger)\,\|u-\tilde u\|_{L^2(D_1)},$$

so G^+ is uniformly continuous with the linear modulus $\omega(t) = \mathrm{Lip}(S^\dagger)\,t$. The constant function $\mathbf{1}_{D_2}$ has $\|\mathbf{1}_{D_2}\|_{H^{s_2}(D_2)} = 1$, because all of its derivatives and its Gagliardo seminorm vanish, so $R_{G,K} = \sup_{u\in K}|S^\dagger(u)| < \infty$ by continuity of $S^\dagger$ on the compact set K. Hence Theorem 4.1 applies with $d_{in} = d_{out} = 1$. All outputs lie in $\mathrm{span}\{\mathbf{1}_{D_2}\}$, so the output projection is exact with $M_\varepsilon = 1$. The expert construction above therefore has $L = \Delta = 0$, width W_*, rank parameter 1, and $\text{Activ-Cpl}(\mathcal{T}, \mathcal{NO}) \in \mathcal{O}(1)$ while attaining the required uniform error. Finally, by the Cauchy–Schwarz inequality and $|D_2| = 1$,

$$\Big|\int_{D_2} G(u)(x)\,\mathrm{d}x - S^\dagger(u)\Big| = \Big|\int_{D_2}\big(G(u) - G^+(u)\big)(x)\,\mathrm{d}x\Big| \le \|G(u) - G^+(u)\|_{L^2(D_2)} \le \varepsilon$$

for all $u \in K$. Applying Theorem 4.8 to this same target, whose output projection is exact at rank one, gives the routing bounds in part (ii). □

B.6 Proof of Theorem 4.8

Write $\rho = \rho_\varepsilon$, $J = J_\varepsilon$, $M = M_\varepsilon$, and $b = b_\varepsilon$. The smooth finite-dimensional approximation spaces constructed in Section B.1 give an orthonormal family $(\varphi_j)_{j=1}^J$ whose projection Π_J satisfies

$$\sup_{u\in K} \|u - \Pi_J u\|_{L^2(D_1)} \le \rho/8, \qquad J = O(1 + \rho^{-d_1/s_1}).$$

The family is fixed for this accuracy and need not be nested as ε varies. Write $a_j(u) = (u, \varphi_j)_{L^2(D_1)}$. Since $\|u\|_{L^2} \le R$, all coefficients lie in $[-B, B]$, where $B = \lceil \max\{1, R\} \rceil$.

Centers and greedy descent. Divide each coordinate interval $[-B, B]$ into v^b equal intervals of width

$$\tau = 2Bv^{-b} \le \frac{\rho}{4\sqrt{J}}.$$

Leaves are indexed by all J-tuples of these intervals, with centers $c_\ell = \sum_{j=1}^J q_{\ell,j}\varphi_j$, where $q_{\ell,j}$ is the midpoint of the chosen interval. Start from the whole box. At the first b levels, successively subdivide coordinate 1 into v equal subintervals; do the same for coordinates $2, \dots, J$ at the subsequent levels. At each node, the coefficient vector of its center is the midpoint of its current box. This gives a complete v-ary tree of height Jb.

Children of a node differ only in the coordinate being subdivided. Orthonormality gives

$$\left\| u - \sum_{j=1}^J q_j\varphi_j \right\|_{L^2}^2 = \|u - \Pi_J u\|_{L^2}^2 + \sum_{j=1}^J |a_j(u) - q_j|^2.$$

All terms except that coordinate cancel when comparing children. The nearest midpoint of equal subintervals is the midpoint of an interval containing the queried coordinate. At a boundary either adjacent interval may be selected; fix one tie rule. Thus greedy descent finds a closest midpoint in each coordinate, and hence a nearest leaf in L^2. In particular, its center satisfies $\|u - c_\ell\|_{L^2} \le \rho/8 + \sqrt{J}\,\tau/2 \le \rho/4$.

Inexact coefficients and output error. Suppose the queried coefficients $\widetilde{a}_j$ satisfy $|\widetilde{a}_j - a_j(u)| \le \rho/(8\sqrt{J})$. Clamp each to $[-B, B]$, which cannot increase its error. Greedy descent using these coefficients selects a midpoint vector q_ℓ with $\|\widetilde{a} - q_\ell\|_2 \le \sqrt{J}\,\tau/2 \le \rho/8$. Consequently

$$\|u - c_\ell\|_{L^2} \le \|u - \Pi_J u\|_{L^2} + \|a(u) - \widetilde{a}\|_2 + \|\widetilde{a} - q_\ell\|_2 \le 3\rho/8 < \rho/2.$$

For every leaf use the enlarged set $K_\ell = \{u \in K : \|u - c_\ell\|_{L^2} \le \rho/2\}$, rather than its exact routing cell. If K_ℓ is nonempty, choose $u_\ell \in K_\ell$ and attach the constant expert $G_\ell(u) = P_{\mathrm{out}} G^+(u_\ell)$, where P_{out} is the smooth output projection of dimension M with error at most $\varepsilon/4$. The main theorem's depth-zero construction represents this expert with $O(M^2)$ dense parameters. Empty neighborhoods receive zero experts and are never selected under the stated coefficient accuracy. For a selected leaf, $u, u_\ell \in K_\ell$ and therefore

$$\|G^+(u) - G_\ell(u)\|_{L^2(D_2)} \le \omega(\|u - u_\ell\|_{L^2(D_1)}) + \varepsilon/4 \le \omega(\rho) + \varepsilon/4 \le 3\varepsilon/4.$$

Work, storage, and precision. Query each input coefficient once. Each of Jb levels compares at most v scalar midpoint distances and follows one stored child pointer. The tree has $(v\Lambda_{\mathrm{q}} - 1)/(v - 1) \le 2\Lambda_{\mathrm{q}}$ nodes; storing J center coefficients and the child pointers per node, together with the experts, takes $O(\Lambda_{\mathrm{q}}(J + M^2 + 1))$ real slots and records. For fixed v,

$$b = O(\log(2/\rho)), \qquad Jb = O(\rho^{-d_1/s_1} \log(2/\rho)).$$

For a Lipschitz target, ρ is a constant multiple of ε at small accuracy, and $M = O(1 + \varepsilon^{-d_2/s_2})$ is polynomial. These estimates prove the stated work and storage bounds.

The input precision is finite: rounding coefficients on a dyadic mesh of width at most $\rho/(4\sqrt{J})$ suffices if they are otherwise exact; coefficient acquisition and rounding errors must together meet the stated tolerance. Choosing the largest such dyadic mesh uses $O(\log(2 + B\sqrt{J}/\rho))$ bits per

coefficient. Since B is an integer, all midpoint coordinates are rational with denominator dividing v^b and require $O(\log(2+B) + b\log v)$ bits. Comparisons may use exact rational arithmetic on these rounded values. Expert output coefficients have magnitude at most $R_{G,K}$; rounding them on the largest dyadic mesh of width at most $\varepsilon/(2\sqrt{M})$ adds at most $\varepsilon/4$ to the L^2 error, so the final guarantee remains ε. Each such coefficient requires $O(\log(2 + R_{G,K}\sqrt{M}/\varepsilon))$ bits.

Node addresses require $O(\log \Lambda_{\mathrm{q}}) = O(Jb \log v)$ bits. The stated operation count uses unit-cost random access, as specified in Section 3.3.1; exact arithmetic on these finite words has polynomial bit cost for Lipschitz targets, but physical access to the exponentially stored table and acquisition of the input coefficients are not bounded by this theorem. No claim of efficient preprocessing is made.

B.7 A CONSTANT-EXPERT BENCHMARK

Proposition B.2 (Approximation with constant experts)**.** *Let K be a compact subset of a Hilbert space, and let $G^+ : K \to L^2(D_2)$ have modulus of continuity ω. Suppose $c_1, \dots, c_M \in K$ form an r-net, the implemented centers satisfy $\|c_i - \hat{c}_i\| \le \eta$, and output representatives v_i satisfy $\|G^+(c_i) - v_i\|_{L^2} \le \tau$. Routing to a nearest $\hat{c}_i$ and returning the constant expert $u \mapsto v_i$ gives*

$$\sup_{u \in K} \|G^+(u) - v_{i(u)}\|_{L^2} \le \omega(r + 2\eta) + \tau.$$

If $G^+(K)$ has bounded H^{s_2} radius, the v_i can be chosen in one smooth approximation space with $\mathcal{O}(1 + \tau^{-d_2/s_2})$ coefficients per expert. For the scalar Sobolev input ball, a cover with $M \le \exp(Cr^{-d_1/s_1})$ suffices for small r.

Proof. Choose j with $\|u - c_j\| \le r$. Nearest-center routing gives

$$\|u - c_{i(u)}\| \le \|u - \hat{c}_{i(u)}\| + \eta \le \|u - \hat{c}_j\| + \eta \le r + 2\eta.$$

The error follows by adding the output-representation error τ. The output coefficient count follows from the smooth projection spaces constructed in Section B.1. The input cover count follows from Sobolev metric entropy, with centers in K as in Section B.2. Approximate center MLPs and padding to a complete tree are available from that proof. In the neural-operator class, each constant-in-input output is implemented by a constant lifting channel and a finite-rank output map, as in Section B.4; coefficient storage should not be confused with the number of slots in a dense parameter tensor. □

For example, take $r = \omega^{-1}(\varepsilon/2)/3$, $\eta = r$, and $\tau = \varepsilon/2$. The resulting error is at most ε. For Lipschitz targets this gives a cover size of at most $\exp(C\varepsilon^{-d_1/s_1})$, the same exponential order as the cover used in Theorem 4.1.

For the hard scalar functional of Corollary 4.5, take $v_i = S^\dagger(c_i)\mathbf{1}_{D_2}$ and $\tau = 0$. Each expert stores a single output value in a fixed-size architecture, with no hidden activation needed. It is both a constant ReLU network and a constant $\tanh$ neural operator, giving the realization described in Remark 4.6.

B.8 AFFINE EXPERTS UNDER TARGET-INDEPENDENT ROUTING

We give the full statement and proof of Proposition 4.7. Write $X = L^2([0,1]^d)$ and let K be the radius-R ball of $H^s([0,1]^d)$, with $R > 0$ and $s > d$. All derivative norms below are operator norms induced by X. Define

$$\mathcal{F} = \left\{ S \in C^1(X;\mathbb{R}) : \sup_K |S| \le 1, \quad \sup_{u \in X} \|DS(u)\|_{X^*} \le 1, \quad \mathrm{Lip}_X(DS) \le 1 \right\}.$$

In particular, the derivative Lipschitz constant is bounded uniformly over the class. A router $q : K \to \{1, \dots, \Lambda\}$ is chosen before S; its expert coefficients may depend on S. Let

$$E_0(q) = \sup_{S \in \mathcal{F}} \inf_{a \in \mathbb{R}^\Lambda} \sup_{u \in K} |S(u) - a_{q(u)}|, \qquad \Lambda_0(\varepsilon) = \min\{\Lambda : \exists q \text{ with } E_0(q) \le \varepsilon\}.$$

The lower bound permits arbitrary partitions, without imposing a nearest-center or neural representation on q. The affine upper bound uses the nearest-center MoNO router and the neural-operator architecture of Definition 3.2. Constants depend on d, s, R and the fixed output domain, but not on S or ε.

Constant-expert lower and upper bounds. Set $B = \max\{1, R\}$. For u, v in the same nonempty routing cell and $u \neq v$, the functional

$$S_{u,v}(w) = B^{-1} \left\langle w, \frac{u - v}{\|u - v\|_X} \right\rangle_X$$

belongs to $\mathcal{F}$: it is bounded by $R/B \leq 1$ on K, its derivative has norm $B^{-1} \leq 1$, and its derivative is constant. Every constant assigned to that cell therefore has error at least $\|u - v\|_X/(2B)$ at one of the two points. Consequently $E_0(q) \leq \varepsilon$ implies that every cell has diameter at most $2B\varepsilon$. Choosing one representative per nonempty cell covers K with at most Λ balls of radius $2B\varepsilon$. If $\mathcal{N}(K, t)$ denotes the covering number in X, then

$$\Lambda \geq \mathcal{N}(K, 2B\varepsilon), \qquad \log \mathcal{N}(K, t) \geq ct^{-d/s}$$

for small t, by Sobolev metric entropy (Birman & Solomyak, 1967; Lanthaler, 2024). Here only the input-set entropy estimate of Lanthaler (2024, Ex. 2.9) is used, not its lower bound for the entire Lipschitz functional class. Conversely, every $S \in \mathcal{F}$ is 1-Lipschitz. A common ε-net with centers in K and the constants $S(c_i)$ give error at most ε, while Sobolev entropy bounds its size by $\exp(C\varepsilon^{-d/s})$. Thus

$$\exp(c\varepsilon^{-d/s}) \leq \Lambda_0(\varepsilon) \leq \exp(C\varepsilon^{-d/s}). \tag{13}$$

Affine experts and their finite-rank realization. Take $0 < \varepsilon \leq 1$, $r = \sqrt{\varepsilon}/2$, and $\eta = \sqrt{\varepsilon}/4$. Choose an r-net $c_1, \ldots, c_m \in K$ and center ReLU networks $\hat{c}_i$ satisfying $\|c_i - \hat{c}_i\|_X \leq \eta$. These choices depend on K, ε alone. Nearest-center routing selects an index $q(u)$ satisfying

$$\|u - c_{q(u)}\|_X \leq r + 2\eta = \sqrt{\varepsilon}, \qquad m \leq \exp(C\varepsilon^{-d/(2s)}).$$

Duplicate centers to obtain a complete v-ary tree, with fixed $v \geq 2$; at most a factor v more leaves is needed. Use the smooth Sobolev projection spaces from Section B.1 to fix an orthogonal projection P_N with

$$\sup_{u \in K} \|(I - P_N)u\|_X \leq \varepsilon/4, \qquad N \leq C\varepsilon^{-d/s}.$$

For each $S \in \mathcal{F}$, define

$$A_i^S(u) = S(c_i) + DS(c_i)[P_N(u - c_i)]. \tag{14}$$

The integral form of the first-order remainder gives

$$|S(u) - S(c_i) - DS(c_i)[u - c_i]| \leq \int_0^1 t\|u - c_i\|_X^2 \, \mathrm{d}t = \tfrac{1}{2}\|u - c_i\|_X^2.$$

Since both u and c_i belong to K, truncating the derivative adds at most

$$|DS(c_i)[(I - P_N)(u - c_i)]| \leq \|(I - P_N)u\|_X + \|(I - P_N)c_i\|_X \leq \varepsilon/2.$$

Thus $|S(u) - A_{q(u)}^S(u)| \leq \varepsilon$ uniformly on K.

For clarity, no spatial regularity of the Riesz representative of $DS(c_i)$ is assumed. Only its action on the fixed finite input basis is stored. If $\varphi_1, \ldots, \varphi_N$ span $P_N X$, write

$$a_{im} = DS(c_i)[\varphi_m], \qquad b_i = S(c_i) - \sum_{m=1}^{N} a_{im}\langle c_i, \varphi_m \rangle_X.$$

Then $A_i^S(u) = b_i + \sum_m a_{im}\langle u, \varphi_m \rangle_X$. Choose a smooth output family of size N with $\psi_1 = \mathbf{1}_{D_2}$, where $D_2 = [0, 1]^{d_2}$. In Definition 3.2, set $L = \Delta = 0$, take one lifting channel, set $C_{m,1}^{(0)} = a_{im}$ and $b^{(0)} = b_i$, and set $C_{1,1}^{(1)} = 1$ and $b^{(1)} = 0$, with all remaining kernel coefficients zero. This realizes $u \mapsto A_i^S(u)\mathbf{1}_{D_2}$ exactly. Width $W_* = \max\{d, d_2, 1\}$ suffices, and the dense parameter bound is

$$P(G_i) \leq 2N^2 W_*^2 + 2W_*(W_* + 1) \leq C\varepsilon^{-2d/s}.$$

There are only $N + 1$ target-dependent affine coefficients, but we retain the paper's dense architecture accounting.

Storage of the center networks. We make the center approximation quantitative so that the reduction in expert count also controls the represented router. Sobolev embedding, since $s > d \geq 1$, gives a constant $H \geq 1$ such that every $f \in K$ satisfies $\|f\|_\infty \leq H$ and

$$|f(x) - f(y)| \leq H\|x - y\|_2^{1/2}.$$

Fix $0 < \eta \leq 1$ and a regular grid with spacing $h \leq \eta^2/(16dH^2)$ and $Q \leq C\eta^{-2d}$ vertices z_j. Put $a = H^2/\eta$ and define the piecewise-affine function

$$f_\eta(x) = \max_{1 \leq j \leq Q} \{f(z_j) - a\|x - z_j\|_1\}.$$

For any j, with $t = \|x - z_j\|_1$, the excess over $f(x)$ is at most $H\sqrt{t} - at \leq H^2/(4a) = \eta/4$. A nearest grid vertex has $t \leq dh \leq \eta^2/(16H^2)$, so the deficit is at most $H\sqrt{t} + at \leq 5\eta/16$. Hence $\|f - f_\eta\|_\infty \leq \eta$.

Absolute values are exactly realized by $|t| = \mathrm{ReLU}(t) + \mathrm{ReLU}(-t)$, and pairwise maxima by $\max\{a, b\} = \mathrm{ReLU}(a - b) + b$. Computing all cones in parallel and then a balanced maximum tree gives a ReLU MLP of width $O(dQ)$ and depth $O(\log(2Q))$. Signed intermediate values can be carried as their positive and negative parts, changing only the constants in these bounds. Even with dense weight matrices, this uses at most

$$C\eta^{-4d}\log(2/\eta)$$

parameter slots. Thus each implemented center for the affine construction has at most $C\varepsilon^{-2d}\log(2/\varepsilon)$ parameters. Internal tree centers are zero networks, and the tree has $O(\Lambda)$ vertices. Combining the expert and center counts yields

$$P_{\text{experts}} + P_{\text{centers}} + |V| \leq C\Lambda\left(\varepsilon^{-2d/s} + \varepsilon^{-2d}\log(2/\varepsilon) + 1\right) \leq \exp(C'\varepsilon^{-d/(2s)}) \tag{15}$$

for small ε. Here $|V|$ counts tree records; coefficient precision, fixed basis descriptions, and numerical integration are separate resources. Routing still scans Λ centers. This is a bound on stored real parameters and tree records, not on inference time or bit complexity.

Scope of the comparison. The lower bound has the quantifiers $\forall q\, \exists S$: the difficult linear functional can depend on the router and the accuracy. It is not a single-target lower bound and does not apply to target-adapted routers. Both approximation families choose their router independently of S, while all expert coefficients may depend on S. With this common restriction, nonconstant experts improve the exponential power in the required expert count, and even the full represented affine construction has fewer parameter slots asymptotically than the number of experts required by lookup. No optimality claim for affine experts is needed. The mechanism is local first-order approximation; nonlinear hidden layers are unnecessary for this result.

C Additional Material

This appendix collects supplemental material that would otherwise interrupt the main narrative: the routing pseudocode used by MoNOs, the auxiliary domain-extension result, additional controlled-experiment diagnostics, an inverse-problem case study that motivates the distributed construction, and the explicit identification of a classical neural operator as a one-expert MoNO.

Routing pseudocode. The entropy-cover construction uses the following exhaustive nearest-leaf search. The distinct coefficient tree and its greedy descent are constructed in Section B.6.

Algorithm 1 Exhaustive routing for the entropy-cover MoNO

Input: $u \in H^{s_1}(D_1) \subset L^2(D_1)$
Set $i^\star = 1$ and $d^\star = +\infty$
for $i = 1, \dots, \Lambda$ **do**
  Load center $\hat{f}_i$ and compute $d = \|\hat{f}_i - u\|_{L^2(D_1)}$
  if $d < d^\star$ **then**
    Set $i^\star = i$ and $d^\star = d$
  end if
  Release the current center
end for
Follow the unique root-to-leaf path in $\mathcal{T}$ ending at leaf $\hat{f}_{i^\star}$
Output: $G_{i^\star}(u)$

D Routing Tree Lemma

Lemma D.1 (A neural cover indexed by a balanced tree)**.** *Fix $s_1 > d_1$, $R > 0$, and the scalar Sobolev ball K in* (1)*. There is $C_R > 0$ such that, for every $\delta > 0$ and integer $v \geq 2$, a complete v-ary tree with ReLU-MLP center functions has leaves satisfying*

$$\sup_{u \in K} \min_{1 \leq i \leq \Lambda} \|u - \hat{f}_i\|_{L^2(D_1)} < C_R \delta. \tag{16}$$

One can take $k = \lceil \delta^{-d_1/s_1} \rceil$, $h = \lceil \log_v(2^k - 1) \rceil$, $\Lambda = v^h$, and $(v^{h+1} - 1)/(v - 1)$ nodes. Nodes are indexed vertices, so distinct vertices may store the same center function. This lemma bounds the cover size and tree height, but does not provide a bound on center-network sizes or a fast nearest-center search.

E Domain Extension Auxiliary Result

Proposition E.1 (Domain extension)**.** *Assume the setting of Definition 3.2 with $d_{in} = d_{out} = 1$. Let*

$$G^+ : \big(H^{s_1}(D_1), \|\cdot\|_{L^2(D_1)}\big) \to \big(H^{s_2}(D_2), \|\cdot\|_{L^2(D_2)}\big)$$

be uniformly continuous with modulus ω, and suppose each D_i is a compact subset of $[0,1]^{d_i}$ of positive Lebesgue measure. Then there exists a uniformly continuous operator

$$\bar{G} : L^2([0,1]^{d_1}) \to L^2([0,1]^{d_2})$$

with the same modulus of continuity such that $\bar{G}(\bar{u})|_{D_2} = G^+(u)$ whenever $\bar{u}|_{D_1} = u$.

F Additional Experimental Diagnostics

This appendix records the protocol and exact parameter accounting for the Burgers experiments in Section 5, together with the older soft-mixture mechanism check used during development. These diagnostics support the controlled empirical checks in the main text.

Burgers protocol. For each seed, we generate 6144 periodic initial conditions on a grid of size 64 and split them into $4096/1024/1024$ train/validation/test samples. The reported numbers use seeds $7, 17, 27$. The route label $r(u_0) \in \{0, 1, 2, 3\}$ is the empirical quartile of $\|\partial_x u_0\|_\infty$ in the entire generated input pool, before splitting. Thus the regime thresholds depend on held-out inputs as part of this synthetic task definition; the runs do not evaluate thresholds estimated from training data alone. In the homogeneous experiment, every sample is evolved with viscosity $\nu = 0.01$; the same quartile labels are kept only to test whether a routed model helps when the PDE operator itself is not regime-dependent. In the heterogeneous experiment, samples in the four quartiles are evolved with viscosities $(0.05, 0.02, 0.008, 0.003)$. All targets are produced by the explicit finite-volume Burgers solver used in the released script. Inputs and outputs are normalized using statistics from the training split. All models are trained for 40 epochs with Adam, learning rate 10^{-3}, batch size 128, and validation-based checkpoint selection on a single NVIDIA RTX PRO 6000 Blackwell GPU.

Table 3: Exact parameter accounting and relative L^2 test error for the two Burgers cases. The homogeneous case uses fixed viscosity $\nu = 0.01$; the heterogeneous case assigns $(0.05, 0.02, 0.008, 0.003)$ by steepness quartile. The feature router has 17 inputs and two hidden layers; its additional hidden layer makes it larger than the soft-mixture diagnostic's one-hidden-layer router, despite having fewer inputs.

Case	Model	Active expert	Router	Total stored	Rel. L^2
Homogeneous	Global NO (active-matched)	21,905	0	21,905	0.0203 ± 0.0004
	Global NO (total-matched)	92,863	0	92,863	0.0119 ± 0.0006
	Oracle MoNO	21,905	0	87,620	0.0223 ± 0.0007
	Learned-router MoNO	21,905	5,572	93,192	0.0213 ± 0.0013
Heterogeneous	Global NO (active-matched)	21,905	0	21,905	0.1140 ± 0.0056
	Global NO (total-matched)	92,863	0	92,863	0.0986 ± 0.0019
	Oracle MoNO	21,905	0	87,620	0.0245 ± 0.0007
	Learned-router MoNO	21,905	5,572	93,192	0.0558 ± 0.0033

Input distribution and discrete solver. For each input, independently draw a spectral decay $p \sim \mathrm{Unif}[1, 2.25]$, a common amplitude $A \sim \mathrm{Unif}[0.65, 1.5]$, and independent standard Gaussian sine and cosine coefficients, scaled by A/j^p for modes $j = 1, \ldots, 8$. Subtract the grid mean and rescale to an independently sampled RMS in $[0.8, 1.5]$; this final normalization cancels the common amplitude. The solver uses a first-order spatial Rusanov flux, centered viscous differences, and second-order SSP Runge–Kutta time stepping. The step count enforces $\Delta t \leq 0.35\Delta x / \max |u_0|$ and $\Delta t \leq 0.20\Delta x^2/\nu$, as well as the requested minimum number of steps. Numerical diffusion and spatial resolution are part of this discrete target map; no grid-convergence claim is made.

Training convergence. All 24 saved validation-MSE curves for the original two-case, three-seed, four-model runs reach their minima at epochs 38–40. From epoch 30 to 40, validation MSE decreases by approximately 11–38%. These runs therefore establish performance within the fixed budget, not converged accuracy.

Model rows. The active-matched global NO is a single Fourier-style neural operator with width 16, 10 Fourier modes, and 4 spectral blocks. The total-matched global NO uses the same architecture with width 33, chosen to match the stored parameter count of the learned-router MoNO. The oracle MoNO consists of four width-16 experts and no learned router; both training and evaluation dispatch sample u_0 to expert $r(u_0)$. The learned-router MoNO uses the same four experts plus a width-64 feature router. During training, the experts are still updated under the oracle assignment $r(u_0)$, while the router is trained with cross-entropy to predict $r(u_0)$. Its feature vector contains the input mean, standard deviation, root mean square, maximum, minimum, mean absolute periodic finite difference, maximum absolute periodic finite difference, and the first ten nonzero real-FFT magnitudes of the normalized input. At test time, the learned router selects $\arg\max_i q_\theta(i \mid u_0)$ and only the selected expert is evaluated. Thus the learned-router row is a supervised-routing diagnostic. Active-including-router counts one expert plus the router; total stored parameters count all experts plus the router.

Soft-mixture diagnostic. The following table and figure record the earlier controlled soft-mixture operator-learning diagnostic. Its target is itself a soft mixture, so it is best read as a mechanism check; the Burgers experiment in the main text is the less synthetic empirical illustration.

Training-budget note. For the soft-mixture diagnostic, we also reran the same three-seed experiment for 100 epochs, keeping the architecture, synthetic operator, and parameter accounting fixed. In that longer-budget rerun, the active-matched global NO attains test MSE 0.000360 ± 0.000024, the total-matched global NO attains 0.000158 ± 0.000023, the oracle MoNO attains 0.000191 ± 0.000006, and the learned-router MoNO attains 0.000179 ± 0.000004 with routing accuracy 0.993 ± 0.001. Thus, the same fixed-active-size conclusion remains visible under longer training: routed models still outperform the active-matched global baseline. At the same time, a larger-active total-matched global baseline can improve enough to outperform the routed models, which is fully consistent with the

Table 4: Exact parameter accounting and test error for the controlled neural-operator illustration. The learned-router row reports both the active expert size and the slightly larger active count obtained when the router itself is included. The qualitative message is the same as in the theory: better approximation with a smaller active model can come at the price of a larger stored ensemble.

Model	Active expert	Router	Active incl. router	Total stored	Test MSE
Global NO (active-matched)	21,905	0	21,905	21,905	0.000785 ± 0.000102
Global NO (total-matched)	87,329	0	87,329	87,329	0.000456 ± 0.000020
Oracle MoNO	21,905	0	21,905	87,620	0.000224 ± 0.000004
Learned-router MoNO	21,905	2,212	24,117	89,832	0.000234 ± 0.000004

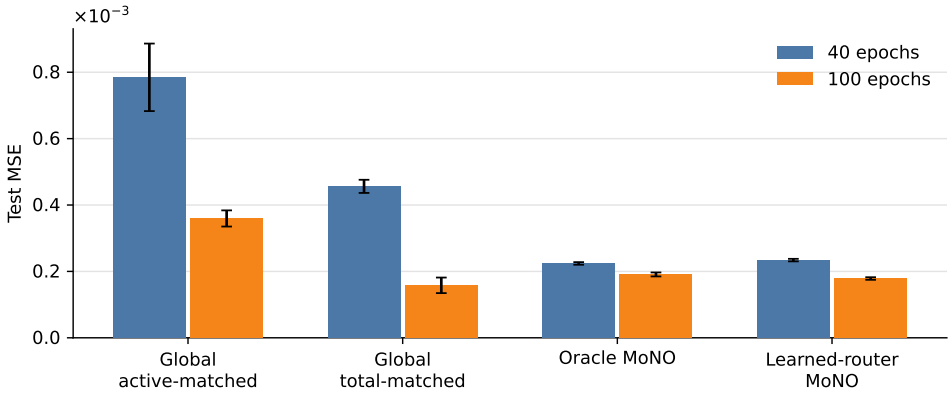

Figure 3: **Training budget in the soft-mixture diagnostic.** Test MSE at 40 and 100 epochs, with means and sample standard deviations over the same three seeds. The total-matched global model improves enough to overtake the routed models at the longer budget. All values are regenerated from the saved aggregate results; model sizes are unchanged.

paper's claim, since the theory concerns active complexity rather than universal empirical dominance over larger-active models.

Active-complexity reading of the soft-mixture diagnostic. As in the Burgers case, Table 4 fixes the active expert and compares errors. At the 40-epoch budget, the learned-router MoNO evaluates 24,117 parameters per query including its router and attains a lower test MSE than the total-matched global NO, which evaluates 87,329 parameters per query (3.6× more). At the 100-epoch budget the total-matched global NO overtakes the routed models. The observed ordering therefore depends on the training budget; these two widths do not identify an exact equal-accuracy parameter ratio.

F.1 Global-width sweep

To examine error as a function of active size within a fixed architecture and training protocol, we reran both Burgers cases with the protocol above while sweeping the width of the global NO over $\{16, 24, 33, 48, 64, 96, 128, 192, 256\}$ (21,905 to 5,572,865 parameters), and retrained the width-16 oracle and learned-router MoNOs alongside as same-run references. The sweep uses seeds $7, 17, 27$, so the data are generated with the same seeds as in Table 3, but it runs on an Apple M5 Max GPU (PyTorch MPS backend instead of CUDA) and initializes each model from its own seed. The absolute errors differ accordingly: the width-16 global NO on the heterogeneous family attains 0.128 ± 0.003 here versus 0.114 ± 0.006 there, and the learned router reaches route accuracy 0.838 versus 0.858. The qualitative conclusions are unchanged. Figure 4 shows the results. On the heterogeneous family, the global error decreases from 0.128 at width 16 to about 0.073–0.075 for widths 128 to 256, whereas the learned-router MoNO attains 0.060 ± 0.003 (route accuracy 0.838 ± 0.011) with 27,477 active parameters including its router, and the oracle MoNO attains 0.0245 ± 0.0015 with 21,905. No swept width reaches either routed model: under the same 40-epoch budget, even the widest global NO, with 5,572,865 parameters (203× the active count of the learned-router MoNO including its router and 254× that of the oracle MoNO), remains less accurate than both. In the homogeneous control,

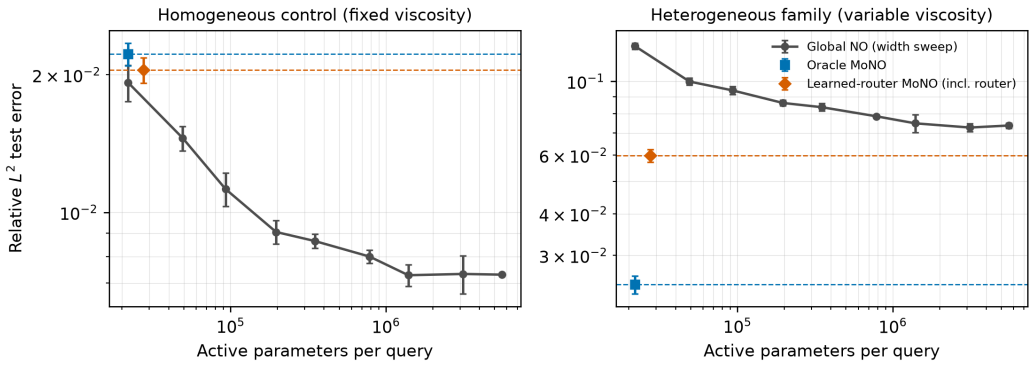

Figure 4: **Global-width sweep.** Relative L^2 test error versus active parameters per query (mean and standard deviation over three seeds). Gray: global NOs of width 16 to 256. Markers and dashed lines: width-16 oracle and learned-router MoNOs (the learned router is plotted at its active count including the router). Left: in the homogeneous control, a global NO of the same active size already matches the routed models and larger global NOs are better. Right: on the heterogeneous family, the global NO levels off above both routed models, even at $203\times$ the active size of the learned-router MoNO ($254\times$ that of the oracle MoNO).

the width-16 global NO (0.0192 ± 0.0017) already matches the learned-router (0.0205 ± 0.0013) and oracle (0.0222 ± 0.0013) MoNOs, and wider global NOs improve to about 0.0073 at width 128 and then level off, so routing provides no reduction there. The sweep does not optimize the learning rate, training duration, or backbone separately for each width and does not provide an equal-accuracy crossing point. The $203\times$ ratio describes the largest tested model, not a lower bound on optimal global parameter requirements. The sweep is produced by `toy_burgers_width_sweep.py` and `plot_burgers_width_sweep.py` in the supplementary material.

F.2 Routing and side information

The learned router of Section 5 sees 17 hand-crafted input features, including the maximal periodic finite difference, which is closely related to the steepness that defines the route label, and the oracle MoNO sees the route label itself. To separate the effect of routing from this side information, we reran both Burgers cases with the data, seeds, and hardware of Section F.1, retraining the width-16 global NO and the two routed MoNOs as same-run references, and added four models. A *feature-conditioned* global NO receives an MLP embedding of the same 17 features (with the router's architecture, trained end to end) as extra spatially constant input channels; its active size ($27{,}541$) matches that of the learned-router MoNO including its router ($27{,}477$). The feature-conditioned global model uses only the regression loss, whereas the learned MoNO also uses label-supervised cross-entropy and oracle expert assignments during training. Hence this comparison does not isolate routing from supervision. A *label-conditioned* global NO receives the one-hot route label, the information used by the oracle MoNO. Two unsupervised MoNOs route by nearest centroid after k-means clustering ($k = 4$, k-means++, 8 restarts, fitted on the training inputs) of either the normalized inputs in $\mathbb{R}^{64}$, a discrete analogue of the nearest-center router of Theorem 4.1, or the standardized router features. Table 5 reports the results; the routing column gives the route accuracy of the learned router and, for k-means, the best-permutation agreement between clusters and route labels (the fixed-label balanced reference is 0.25; maximizing over permutations can raise finite-sample chance agreement).

On the heterogeneous family, side information explains much, but not all, of the gain. With approximately matched active size and the stated inference inputs, the global-to-routed error ratio is about 1.5: 0.060 for the learned-router MoNO versus 0.092 for the feature-conditioned global NO, and 0.0245 for the oracle MoNO versus 0.036 for the label-conditioned global NO. Unsupervised k-means routing recovers the regimes poorly and helps little or hurts. In the homogeneous control, no variant improves on the global NO. The results are produced by `toy_burgers_baselines.py` in the supplementary material.

Table 5: Side-information baselines and unsupervised routing on the Burgers cases (relative L^2 test error, mean $\pm$ standard deviation over seeds $7, 17, 27$; same data, hardware, and 40-epoch budget as Section F.1). Active sizes include the router; the routing column is the route accuracy (learned router) or the best-permutation agreement with the route labels (k-means).

Case	Model	Active incl. router	Total stored	Rel. L^2	Routing
Heterogeneous	Global NO (active-matched)	21,905	21,905	0.1277 ± 0.0029	–
	Global NO + router features	27,541	27,541	0.0917 ± 0.0073	–
	Global NO + route label	21,969	21,969	0.0362 ± 0.0011	–
	Oracle MoNO	21,905	87,620	0.0245 ± 0.0015	–
	Learned-router MoNO	27,477	93,192	0.0598 ± 0.0026	0.838
	k-means MoNO (inputs, L^2)	22,161	87,876	0.1381 ± 0.0049	0.276
	k-means MoNO (router features)	22,007	87,722	0.1051 ± 0.0026	0.555
Homogeneous	Global NO (active-matched)	21,905	21,905	0.0192 ± 0.0017	–
	Global NO + router features	27,541	27,541	0.0216 ± 0.0027	–
	Global NO + route label	21,969	21,969	0.0206 ± 0.0018	–
	Oracle MoNO	21,905	87,620	0.0222 ± 0.0013	–
	Learned-router MoNO	27,477	93,192	0.0205 ± 0.0013	0.841
	k-means MoNO (inputs, L^2)	22,161	87,876	0.0231 ± 0.0004	0.276
	k-means MoNO (router features)	22,007	87,722	0.0209 ± 0.0009	0.555

F.3 Expert families with shared routing and supervision

This experiment compares constant, affine, and FNO experts while holding routing fixed within each comparison. It also gives a global FNO the same regime-label training supervision and predicted labels as the learned-router MoNO. The protocol is separate from the preceding 40-epoch experiments: it fits the steepness quartiles on the 4096 training inputs alone, freezes them for the 1024 validation and 1024 test inputs, and regenerates all PDE targets. The four physical viscosities remain $(0.05, 0.02, 0.008, 0.003)$ throughout, including when the number of approximation cells changes. We use seeds $7, 17, 27$, the same 64-point grid and initial-condition distribution, and 100 training epochs for each neural regressor on the MPS backend.

Shared supervised routing. For each seed, a single feature router is trained for 60 epochs by cross-entropy, with its checkpoint selected by validation cross-entropy. Its exact predicted labels are reused by all four supervised rows in Table 6. The constant and affine experts are fitted within the true training regimes. The four FNO experts and the conditioned global FNO also train with true regime labels; both use predicted labels at validation and test time. The global model receives the label as a one-hot input through the existing `ConditionedFNO1d` architecture. Thus both neural comparisons receive the same supervision and inference information. They use width 16, 10 modes, 4 blocks, Adam with learning rate 10^{-3}, batch size 128, and the checkpoint of minimum validation MSE within 100 epochs. Including the shared classifier, the active parameter counts are $27{,}477$ for the MoNO and $27{,}541$ for the global model; total storage differs.

Local regression and input-space routing. Each constant expert is the mean training output in its cell. Affine experts regress the 64 output values on the 16 real Fourier coordinates of the initial condition, with an intercept. These coordinates preserve discrete input L^2 distances up to a common factor. Centered ridge fits use a single penalty per partition, chosen on validation MSE from $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$; the zero-penalty fit uses an SVD pseudoinverse. Empty cells fall back to the global training mean. The unsupervised comparison fits k-means on training inputs only, with three initializations and at most 100 iterations, for $4, 8, \ldots, 1024$ cells. Constant and affine experts share each partition; four additional FNO experts use the same four-cell partition. The 1-nearest-neighbor predictor stores all 4096 training input/output pairs and uses the same Fourier metric.

Routing diagnostics. The shared classifier has test accuracy 89.97% averaged over seeds. With its predicted labels, the global-to-routed relative-error ratio is 1.16 (a 14.1% error reduction), while the MSE reduction is 2.8%. Evaluating the same checkpoints on CPU with true test labels, without retraining or checkpoint reselection, gives relative errors 0.01418 for the MoNO and 0.02148 for the

Table 6: Expert ablation under the separate training-threshold protocol. Relative L^2 error is the mean per-example relative norm, reported as mean $\pm$ sample standard deviation over three seeds; MSE is averaged over seeds. Stored values include expert coefficients, router or center values, and the normalization arrays required by each predictor. The global model shares the supervised router.

Routing	Predictor	Active expert	Stored values	MSE	Relative L^2
Supervised	Constant (4)	64	5,956	0.22361	1.0001 ± 0.0005
	Affine (4)	1,088	10,052	0.06873	0.5208 ± 0.0072
	FNO (4)	21,905	93,448	0.00146	0.0388 ± 0.0022
	Global FNO + predicted label	21,969	27,797	0.00150	0.0452 ± 0.0016
Input k-means	Constant (4)	64	320	0.08941	0.5974 ± 0.0037
	Affine (4)	1,088	4,416	0.03107	0.3575 ± 0.0056
	FNO (4)	21,905	87,940	0.00307	0.1117 ± 0.0037
	Constant (1024)	64	81,920	0.03139	0.3412 ± 0.0036
	Affine (1024)	1,088	1,130,496	0.02692	0.3092 ± 0.0026
Nearest neighbor	Stored outputs (4096)	64	327,680	0.03156	0.3176 ± 0.0098

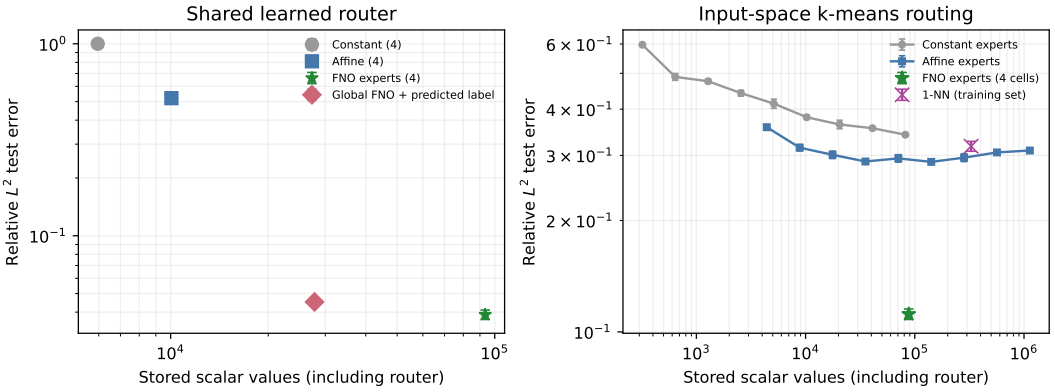

Figure 5: **Expert family and storage.** Left: constant, affine, and FNO experts share the learned router; the global FNO receives its predicted label. Right: constant and affine predictors share each input-space k-means partition, with 4–1024 cells; the four-expert FNO uses the four-cell partition. Error bars show the standard deviation over three seeds. The axes count all stored predictor and router values; all stored arrays use float32.

global model, a ratio of 1.51. This oracle diagnostic isolates the effect of substituting the correct routes and conditioning labels; it is not the learned router's test performance. Among misrouted examples, mean relative errors are 0.2570 and 0.2550, respectively, versus 0.01457 and 0.02186 among correctly routed examples (means of per-seed subgroup averages). A simple alternative that thresholds the maximum absolute periodic difference of the normalized input, using training-only quartiles, classifies 97.62% correctly. This is one of the existing router features; the learned classifier uses unstandardized features. The diagnostic script and JSON reproduce these evaluations from the saved checkpoints and record their hashes.

Results and scope. The MoNO stores 3.36 times as many values as the shared-router global predictor. FNO experts outperform the tested constant and affine predictors under both shared-router comparisons. Increasing lookup cells improves accuracy, but neither 1024-cell constants nor full-training-set nearest neighbors match the four-cell FNO at comparable or larger storage. The four supervised constant experts are close to a zero predictor because each regime averages many phases. Fine affine partitions have few training samples, and the selected ridge penalty is the largest tested value in 18 of the 30 supervised and input-partition fits. Validation minima for the nine neural fits occur at epochs 90–98 of 100. These are fixed-budget, finite-data comparisons, not lower bounds on optimal fitted predictors; the hard viscosity switches also preclude testing the uniform $C^{1,1}$ rate in Proposition 4.7.

Storage counts include all expert arrays, the shared 5572-parameter classifier or the 16-coordinate centers, and required input/output normalization statistics. Fixed analytic Fourier bases and the grid are excluded. This is a scalar-storage comparison, not a latency or peak-memory measurement. The supplement contains the generated data, fitted local models, neural checkpoints, learning curves, test predictions, and a checker that reconstructs metrics and predictions from these artifacts. The preceding experiments retain their original protocols and are not pooled with these runs.

G Applications to Inverse Problems

The following inverse-problem case study illustrates localization for a target operator with weak stability.

Let $\Omega \subset \mathbb{R}^2$ be an open bounded connected set with smooth boundary $\partial\Omega$. Consider the inverse problem of corrosion detection in electrostatics. We model the conductor by the domain Ω and assume that it has an inaccessible relatively open part of the boundary, denoted by $\Gamma \subset \partial\Omega$, that is affected by corrosion. We model this by the equation

$$\Delta u(x) = 0, \quad \text{in } x \in \Omega, \tag{17}$$

$$\partial_\nu u + qu = 0, \quad \text{on } x \in \Gamma, \tag{18}$$

$$\partial_\nu u = f, \quad \text{on } x \in \partial\Omega \setminus \Gamma. \tag{19}$$

We denote the solution of (17)-(19) by $u^q = u$. Here, $u = u(x)$, $u \in H^1(\Omega)$ is the electric voltage at the point $x \in \Omega$, ν is the exterior unit normal vector of $\partial\Omega$, $\partial_\nu u$ is the normal derivative of u on the boundary, the Robin coefficient $q(x)$ models the corrosion (causing contact impedance on the boundary), and $f(x)$ is the external normal current through the boundary.

Assume also that we are given a function q_0 that is close to the true Robin coefficient and that the solution of the equation

$$\Delta u_0(x) = 0, \quad \text{in } x \in \Omega, \tag{20}$$

$$\partial_\nu u_0 + q_0 u_0 = 0, \quad \text{on } x \in \Gamma, \tag{21}$$

$$\partial_\nu u_0 = f, \quad \text{on } x \in \partial\Omega \setminus \Gamma \tag{22}$$

satisfies

$$|u_0(x)| \geq c_0 > 0 \quad \text{on } \overline{\Gamma}. \tag{23}$$

Note that a condition analogous to (23) is necessary to determine the coefficient q_0. For example, if $u_0|_\Gamma$ vanishes, then q_0 does not influence the data $u_0|_\Sigma$.

When q is close to q_0, the inverse problem is to determine q from the measurement $g^q := u^q|_\Sigma$, where $\Sigma \subset \partial\Omega \setminus \overline{\Gamma}$ is an open, non-empty set. To study this inverse problem, assume $q_0 \in H_0^2(\Gamma)$, $q_0 \geq 0$, and $q_0 \not\equiv 0$, and that the true Robin coefficient q and the boundary current f satisfy

$$q \in H_0^2(\Gamma), \quad q \geq 0, \tag{24}$$

$$f \in C^{1,\alpha}(\partial\Omega),\ 0 < \alpha < 1, \quad \operatorname{supp} f \subset \partial\Omega \setminus \overline{\Gamma}, \tag{25}$$

$$\|q - q_0\|_{H^2(\partial\Omega)} < \varepsilon_0, \tag{26}$$

and $H_0^2(\Gamma)$ is the closure of $C_0^\infty(\Gamma)$ in the Sobolev space $H^2(\partial\Omega)$, while $C_0^\infty(\Gamma)$ is the set of smooth, compactly supported functions on Γ. The support assumption on f makes the piecewise boundary conditions equivalent to the globally smooth Robin condition $\partial_\nu u + qu = f$ on $\partial\Omega$. We take $\varepsilon_0 < \|q_0\|_{H^2(\partial\Omega)}/2$, so every admissible q is nonzero. The solution map of the inverse problem is

$$\mathcal{F}_0 : u^q|_\Sigma \mapsto q|_\Gamma. \tag{27}$$

Proposition G.1. *When $\varepsilon_0 > 0$ is small enough, the solution map* (27) *admits a continuous nonlinear extension*

$$\mathcal{F} : L^2(\Sigma) \to L^2(\Gamma), \tag{28}$$

$$\mathcal{F}(u^q|_\Sigma) = q|_\Gamma \tag{29}$$

that satisfies

$$\|\mathcal{F}(g_1) - \mathcal{F}(g_2)\|_{L^2(\Gamma)} \leq \omega(\|g_1 - g_2\|_{L^2(\Sigma)}), \tag{30}$$

where

$$\omega(t) = C_0\left(\log(1+\frac{C_1}{t})\right)^{-1}, \quad t>0, \qquad \omega(0)=0, \tag{31}$$

for some positive constants C_0, C_1 depending on Ω, Γ, Σ, f, and q_0. The extension can also be chosen to have uniformly bounded $H^2(\Gamma)$ output radius.

Proof. Let $\mathcal{Y} \subset H_0^2(\Gamma)$ be the set of functions that satisfy conditions (24)-(26). Moreover, let $\mathcal{X} \subset L^2(\Sigma)$ be the set of the functions $u^q|_\Sigma$, for which there exists q satisfying (24)-(26).

Fix $0 < \gamma < \min\{\alpha, 1/2\}$. Using Gilbarg & Trudinger (1977, Theorem 6.31) and the remark following it, we see that the solution $u^q \in C^{2,\gamma}(\overline{\Omega})$ exists and is unique. Moreover, by applying Gilbarg & Trudinger (1977, Theorem 6.30) to $v = u^q - u^{q_0}$ that satisfies

$$\Delta v(x) = 0, \quad \text{in } x \in \Omega, \tag{32}$$

$$\partial_\nu v + qv = (q_0 - q)u_0, \quad \text{on } x \in \Gamma, \tag{33}$$

$$\partial_\nu v = 0, \quad \text{on } x \in \partial\Omega \setminus \Gamma, \tag{34}$$

we see that $u^q \in C^{2,\gamma}(\overline{\Omega})$ depends continuously on $q \in C^{1,\gamma}(\partial\Omega)$. The Sobolev embedding $H^2(\partial\Omega) \hookrightarrow C^{1,\gamma}(\partial\Omega)$ is bounded, so, when $\varepsilon_0 > 0$ is small enough, it holds for all $q \in \mathcal{Y}$ that

$$|u^q(x)| > c_0/2 > 0, \quad \text{for all } x \in \overline{\Gamma}. \tag{35}$$

Then, the assumptions of Cheng et al. (2008, Corollary 2.2) are valid, and this result implies that there is a unique $q \in \mathcal{Y}$ such that $g = u^q|_\Sigma$. Thus, there is a well-defined map $\mathcal{F}_0 : \mathcal{X} \to \mathcal{Y}$ defined by $\mathcal{F}_0(u^q|_\Sigma) = q|_\Gamma$. Moreover, by Cheng et al. (2008, Corollary 2.2), for $q_1, q_2 \in \mathcal{Y}$ the functions $g_1 = u^{q_1}|_\Sigma$ and $g_2 = u^{q_2}|_\Sigma$ satisfy

$$\|\mathcal{F}_0(g_1) - \mathcal{F}_0(g_2)\|_{L^2(\Gamma)} \le \omega(\|g_1 - g_2\|_{L^2(\Sigma)}), \tag{36}$$

where ω is given in (31), after adjusting its constants. Indeed, all admissible coefficients have a uniform $C^{1,\gamma}$ bound, any pair is sufficiently close for that corollary when ε_0 is small, and the voltage is bounded away from zero on $\overline{\Gamma}$. The local logarithmic estimate extends to all data distances occurring in $\mathcal{X}$ by the uniform bound on $\mathcal{Y}$.

Consider the sets $\mathcal{X}$ and $\mathcal{Y}$ as subsets of Lebesgue spaces, $\mathcal{X} \subset L^2(\Sigma)$ and $\mathcal{Y} \subset L^2(\Gamma)$. The modulus in (31) is increasing and concave: its second derivative is nonpositive by $\log(1+x) \ge 2x/(2+x)$ for $x \ge 0$. Since $L^2(\Sigma)$ and $L^2(\Gamma)$ are Hilbert spaces, Benyamini & Lindenstrauss (2000, Theorem 1.12) gives an extension

$$\widetilde{\mathcal{F}} : L^2(\Sigma) \to L^2(\Gamma)$$

with the same modulus bound and $\widetilde{\mathcal{F}}|_{\mathcal{X}} = \mathcal{F}_0$. To preserve output regularity, let

$$\mathcal{C} = \{q|_\Gamma : q \in H_0^2(\Gamma),\ q \ge 0,\ \|q - q_0\|_{H^2(\partial\Omega)} \le \varepsilon_0\}.$$

This nonempty convex set is closed in $L^2(\Gamma)$: a convergent sequence has zero extensions bounded in $H^2(\partial\Omega)$, and a weakly convergent subsequence retains the support, nonnegativity, and norm bound. Its L^2 metric projection $\Pi_{\mathcal{C}}$ is therefore nonexpansive. Set $\mathcal{F} = \Pi_{\mathcal{C}} \circ \widetilde{\mathcal{F}}$. The projection fixes $\mathcal{F}_0(\mathcal{X}) \subset \mathcal{C}$, preserves (30), and gives the asserted uniform H^2 output bound. □

Regularity and interpretation. For $0 < \theta < \beta < \gamma < \min\{\alpha, 1/2\}$, elliptic regularity and the strict Hölder-to-Sobolev embedding give

$$u^q \in C^{2,\gamma}(\overline{\Omega}) \hookrightarrow H^{2+\beta}(\Omega).$$

The trace theorem and the one-dimensional Sobolev embeddings then imply

$$u^q|_{\partial\Omega} \in H^{3/2+\beta}(\partial\Omega) \hookrightarrow C^{1,\theta}(\partial\Omega), \tag{37}$$

$$\partial_\nu u^q|_{\partial\Omega} \in H^{1/2+\beta}(\partial\Omega) \hookrightarrow C^{0,\theta}(\partial\Omega). \tag{38}$$

The strict inequalities between exponents avoid endpoint embeddings, which do not hold in general. On physical data $g^q \in \mathcal{X}$, unique continuation determines the harmonic voltage, and the Robin condition gives

$$q|_\Gamma = -\frac{\partial_\nu u^q|_\Gamma}{u^q|_\Gamma}.$$

The denominator is bounded away from zero by (35). The quantitative logarithmic estimate for recovery of q is supplied directly by Cheng et al. (2008, Corollary 2.2). No optimality of this modulus on the restricted coefficient family, or corresponding bound in a stronger norm for harmonic continuation, is asserted. The example illustrates weak stability; it does not establish a storage or routing advantage for MoNOs.

H NEURAL OPERATOR AS (DISTRIBUTED) MIXTURES OF NEURAL OPERATORS

In this section, we show that a standard neural operator can be identified as a mixture of neural operators. Suppose that the assumptions in Definition 3.2 hold. Let $G \in \mathcal{NO}^{\tanh}_{N,W,L,\Delta,\boldsymbol{d}}$ so that

$$G : (H^{s_1}(D_1)^{d_{in}}, \|\cdot\|_{L^2(D_1)^{d_{in}}}) \to (H^{s_2}(D_2)^{d_{out}}, \|\cdot\|_{L^2(D_2)^{d_{out}}}).$$

Choose an arbitrary but fixed $f_{1,1} \in H^{s_1}(D_1)^{d_{in}}$ and define the trivial tree

$$\mathcal{T}_{\text{trivial}} = (\{f_{1,1}\}, \varnothing, f_{1,1}),$$

which consists solely of the root $f_{1,1}$ and no other nodes. The pair $(\mathcal{T}_{\text{trivial}}, \{G\})$ vacuously satisfies the conditions in Section 3.3.

Moreover,

$$\text{Activ-Cpl}((\mathcal{T}_{\text{trivial}}, \{G\})) = \text{Dist-Cpl}((\mathcal{T}_{\text{trivial}}, \{G\}));$$

that is, the active complexity coincides with the total complexity because there is only one computational node: the unique leaf, which is also the root. This single expert is responsible for all inputs in $H^{s_1}(D_1)^{d_{in}}$.

The tree has height zero: there is no branch decision, only the root/leaf evaluation. Furthermore, there exist infinitely many such pairs $(\mathcal{T}_{\text{trivial}}, \{G\})$, since the choice of the root $f_{1,1}$ is arbitrary.

However, there is exactly one realization of $(\mathcal{T}_{\text{trivial}}, \{G\})$, and it coincides with the classical centralized neural operator. Henceforth, we identify the classical neural operator with this trivial distributed counterpart.