# REPORT ON THE PAPER "NONLOCALITY AND NONLINEARITY IMPLIES UNIVERSALITY IN OPERATOR LEARNING"

DEVANSH TRIPATHI
ETH Zürich

Abstract. Neural operator architecture approximate operators between infinite dimensional Banach spaces of function. The paper discuss the basic questions about the requirements for universal approximation of neural operator and provide conditions under which neural operators are universal approximators. Author argued that the general approximation of operators between spaces of functions must be both *nonlocal* and *nonlinear*.

A popular variant of neural operators is the Fourier neural operator (FNO). Proving universal approximation theorem for FNOs is based on using unbounded number of Fourier modes, this work challenges this point of view and provide a novel minimal architecture called "averaging neural operator" (ANO) and its analysis showed that ANO is also a universal approximator. Only spatial average is taken as nonlocal ingredient which corresponds to retaning only a single Fourier mode in the case of FNO contrasts to unbounded number of modes.

## 1. Introduction

The task that we are trying to achieve with the help of neural networks is to approximate the underlying operator, which defines a mapping between two infinite-dimensional Banach spaces of functions. Neural operator generalizs the underlying framework of neural network to infinite-dimensional setting and learn such operators from the data. The wide range of neural operators introduced in [1, 4] are defined in analogy with neural networks, but appends weight matrices in the hidden layers with additional linear integral operators acting on the input functions. Special cases of this framework includes Fourier neural network (FNO) [6] in which the reliance on a Fourier basis limits the basic form of the FNO to periodic geometries, although, in that setting, use of fast Fourier transform (FFT) allows for efficient computations with total number of Fourier components limited by the grid resolution. Extension of FNO, called Neural Operator on Riemannian Manifold (NORM) [2], generalizes FNO to use arbitrary orthogonal eigenfunctions of the Laplace-Beltrami operator on any given spatial domain. The approach mentioned in the paper is closely related to low rank neural operator [4] which, however, has a more complicated architecture that proposed in this paper.

Seminar für Angewandte Mathematik, HG E 62.2, Rämistrasse 101, 8092 Zürich, Switzerland
devansh.tripathi@sam.math.ethz.ch.

In the paper author argued that universal approximation can be obtained in general geometries, with nonlocality along with nonlinearity. Nonlocality has been introduced using only a low-rank operator of fixed finite rank, and is not restricted to periodic domains.

Fourier neural operators are already nonlocal and they introduce it via the addition of a nonlocal operator in each hidden layer layer, which acts on the Fourier modes of the input function by matrix multiplication. FNOs are generally implemented with a first layer which lifts the input, a scalar or vector-valued function, to a vector-valued function where the vector dimension (also called model width) is much higher than that of input function itself. It has been showed in literature that increasing the number of channels (model width) rather than to retain more Fourier modes in the architecture is more beneficial in certain circumstances [5].

The author has proposed a underlying framework to many neural operators called averaging neural operator (ANO). The ANO is build upon two minimal ingredients, nonlinearity by composition of shallow neural networks, and nonlocality via a spatial average. Author has deduced many universal approximation theorems for the ANO.

**What does it mean that a neural network is universal?** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be continuous, $d, L \in \mathbb{N}$ and $K \subset \mathbb{R}^d$ be compact. Denote by $\mathrm{MLP}(\sigma, d, L)$ the set of all MLPs with $d$- dimensional input, $L$ layers, and activation function $\sigma$. We say that $\mathrm{MLP}(\sigma, d, L)$ is universal, if $\mathrm{MLP}(\sigma, d, L)$ is dense in $C(K)$.

1.1. **Neural Operator.** Let $\Omega \subset \mathbb{R}^d$ denote a bounded domain (or potentially a manifold) and let $\mathcal{X}(\Omega;\mathbb{R}^o)$,$\mathcal{Y}(\Omega;\mathbb{R}^o)$ and $\mathcal{V}(\Omega;\mathbb{R}^o)$ denote Banach spaces of $\mathbb{R}^o-$ valued functions over $\Omega$. The *nonlocal neural operator* (NNO) is defined as a mapping

$$\Psi \colon \mathcal{X}(\Omega;\mathbb{R}^k) \to \mathcal{Y}(\Omega;\mathbb{R}^k)$$

which can be written as composition of the form $\Psi = \mathcal{Q} \circ \mathcal{L}_L \circ \dots \mathcal{L}_1 \circ \mathcal{R}$ where $\mathcal{R}$ is lifting layer, $\mathcal{L}_l, l = 1, \dots L$ are hidden layers and $\mathcal{Q}$ is projection layer. Given a channel dimension $d_c$, the **lifting layer** $\mathcal{R}$ and **projection layer** $\mathcal{Q}$ are given by a mapping respectively:

$$\mathcal{R} \colon \mathcal{X}(\Omega;\mathbb{R}^k) \to \mathcal{V}(\Omega;\mathbb{R}^{d_c}),\ u(x) \mapsto R(u(x), x), \tag{1}$$

$$\mathcal{Q} : \mathcal{V}(\Omega;\mathbb{R}^{d_c}) \to \mathcal{Y}(\Omega;\mathbb{R}^{d_c}),\ v(x) \mapsto Q(v(x), x) \tag{2}$$

where $R \colon \mathbb{R}^k \times \Omega \to \mathbb{R}^{d_c}$ and $Q : \mathbb{R}^{d_c} \times \Omega \to \mathbb{R}^{k'}$ are learnable neural network acting between finite dimensional Euclidean spaces. For $l = 1, \dots, L$ (the number of **hidden layers**) and for $m = 0, \dots, M$ (the number of modes) choose functions $\psi_{l,m}, \phi_{l,m} \colon \Omega \to \mathbb{R}^{d_c}$. For $l = 1, \dots, L$, each hidden layer $\mathcal{L}_l$ is the mapping $\mathcal{V}(\Omega;\mathbb{R}^{d_c}) \to \mathcal{V}(\Omega;\mathbb{R}^{d_c})$ of the form:

$$(\mathcal{L}_l v)(x) := \sigma\left(W_l v(x) + b_l + \sum_{m=0}^{M} \langle T_{l,m} v, \psi_{l,m} \rangle_{L^2(\Omega;\mathbb{R}^{d_c})} \phi_{l,m}(x)\right)$$

where

(1) $W_l, T_{l,m} \in \mathbb{R}^{d_c \times d_c}$ and bias $b_l \in \mathbb{R}^{d_c}$ are the learnable parameters/matices.
(2) $\langle T_{l,m} v, \psi_{l,m} \rangle_{L^2(\Omega;\mathbb{R}^{d_c})}$ is a inner product in $L^2-$ function space. It measure the contribution of $\psi_{l,m}$ in $T_{l,m}$.

**Note.**

(1) The $L^2$ space is space of square integrable functions such that $L^2(\Omega; \mathbb{R}^{d_c}) := \left\{ f \mid \int_\Omega \|f(x)\|^2 dx < \infty \right\}$.
(2) The form of lifting and projection layer allows for *positional encoding.*
(3) This general framework reduces to the FNO in a periodic geometry and if the expansion function are choosen as Fourier basis functions, indexed by $m$ and independent of $l$.

The central questions that author posses: "which minimal assumptions have to be imposed on the expansion functions $\psi_{l,m}$ and $\phi_{l,m}$, to ensure universal approximation of the resulting architecture?"

**Assumptions on activation function:** $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed to be smooth, $\sigma \in C^\infty(\mathbb{R})$, nonpolynomial and Lipschitz continuous. It acts as a Nemitskii-operator, component-wise on inputs.

## 2. Averaging suffices for Universal Approximation

Author proposes a new architecture called averaging neural operator (ANO) and shows that only a simple averaging as nonlocality suffices for universality. The ANO is the subclass of many instantiations of general NNO architecture, and hence implies universality for general NNOs as corollaries. When the domain is periodic, the resulting ANO becomes a special case of FNO when only zeroth Fourier mode is retained.

### 2.1. Nonlinearity and Nonlocality.

In case of ordinary neural networks, nonlinearity alone suffices for universality but in case of neural operators, it does not alone suffies. To support this, we have a following example: a neural network with a single layer give rise to nonlinear operator, which maps an input function to an output function by composition,

$$u(x) \mapsto \sigma(Wu(x) + b).$$

Despite being nonlinear, it can be shown that they are not universal. For example, such mappings are not able to approximate even simple operators $\Psi^\dagger$ with a *nonlocal* dependence on the input, such as the shift operator $\Psi^\dagger(u)(x) := u(x + h)$ for fixed $h \neq 0$. This shows the need of nonlocality.

### 2.2. Averaging Neural Operator: a Special Subclass of the NNO.

The author has defined a special subclass of the NNO, which combines nonlinearity by composition with nonlocality by averaging. We define a special subclass of hidden layers of the form:

$$\mathcal{L} : \mathcal{V}(\Omega; \mathbb{R}^{d_c}) \to \mathcal{V}(\Omega; \mathbb{R}^{d_c}),\ \mathcal{L}(v)(x) := \sigma\left( Wv(x) + b + \fint_\Omega v(y) dy \right) \tag{3}$$

The author has taken up the case of *single hidden layer* hence the following architecture:

$$\Psi : \mathcal{V}(\Omega; \mathbb{R}^{d_c}) \to \mathcal{V}(\Omega; \mathbb{R}^{d_c}),\ \Psi(u) = \mathcal{Q} \circ \mathcal{L} \circ \mathcal{R}(u),$$

where lifting and projection maps $\mathcal{R}$ and $\mathcal{Q}$ are given as in above equations with $R$ and $Q$ be single-hidden layer neural networks respectively of width $d_c$.

**Parameters of ANO.** Due to its minimal structure, the ANO depends on only one *hyperparameter;* the lifting dimension $d_c$. The *tunable parameters* of ANO are represented by the weight matrix $W \in \mathbb{R}^{d_c \times d_c}$ and bias $b \in \mathbb{R}^{d_c}$ in the hidden layer $\mathcal{L}$, and the internal weights and biases of the ordinary neural network $R$ and $Q$ defined in lifting and projection layers, respectively.

### 2.3. Universal Approximation.

**Theorem 1.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For given integers $s, s' \geq 0$ let $\Psi : C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ be continuous operator, and fix a compact set $K \subset C^s(\overline{\Omega}; \mathbb{R}^k)$. Then for any $\epsilon > 0$, there exists an averaging neural operator $\Psi : K \subset C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ such that*

$$\sup_{u \in K} \|\Psi^\dagger(u) - \Psi(u)\|_{C^{s'}} \leq \epsilon.$$

The above result is in the space of continuously differentiable functions. Author also has a corresponding result in the scale of Sobolov spaces $W^{s,p}$:

**Theorem 2.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary. For given integers $s, s' \geq 0$, and reals $p, p' \in [1, \infty)$, let $\Psi^\dagger : W^{s,p}(\Omega; \mathbb{R}^k) \to W^{s',p'}(\Omega; \mathbb{R}^{k'})$ be a continuous operator. Fix a compact set $K \subset W^{s,p}(\Omega; \mathbb{R}^k)$ of bounded functions, $\sup_{u \in K} \|u\|_{L^\infty} < \infty$. Then for any $\epsilon > 0$, there exists an averaging neural operator $\Psi : W^{s,p}(\Omega; \mathbb{R}^k) \to W^{s',p'}(\Omega; \mathbb{R}^{k'})$ such that*

$$\sup_{u \in K} \|\Psi^\dagger(u) - \Psi(u)\|_{W^{s',p'}} \leq \epsilon.$$

The consequences of the results is that any more general NNP architecture hidden layers of the form (3), is universal as long as an average can be represented; then the resulting NNO reduces to ANO with a specific setting of tunable weights (by setting certain parameters to zero).

**Remark 1.** Above two theorem show that ANO is universal i approximating a large class of operators, uniformly over a compact set of input functions $K$. But in practice, neural operators are trained by minimizing an empirical loss:

$$\mathcal{L}(\Psi) = \frac{1}{N} \sum_{n=1}^{N} \|\Psi^\dagger(u_n) - \Psi(u_n)\|^2_{W^{s',p'}(\Omega)}, \ u_1, \dots u_N \sim \mu,$$

where data are iid random samples form an underlying input probability measure $\mu$. A popular choice is sampling input function from a Gaussian random field. In this case, the set of input functions is no longer bounded, and hence not compact.

In this case when $\mu$ does not have a compact support, we can derive these results with a cut-off argument. For given $\epsilon > 0$, it is possible to show that the existence of $\Psi$, such that

$$\mathbb{E}_{u \sim \mu}[\|\Psi(u) - \Psi^\dagger(u)\|^2_{W^{s',p'}}] < \epsilon.$$

### 2.4. Intuition.

**Encoder-Decoder Structure.** The ANO architecture has hidden encoder-decoder structure. This structure can be obtained when setting the matrix $W$ and bias $b$ in the hidden layer (3) zero, in which case we note that the mapping $\mathcal{L} \circ \mathcal{R} : \mathcal{X} \to \mathbb{R}^{d_c}, u \mapsto \mathcal{L} \circ \mathcal{R}(u)$ can be though of as a nonlinear encoding of the input function and encode it to a (constant) vector $v \in \mathbb{R}^{d_c}$ while the projection mapping that takes $v \mapsto Q(v, .)$ is a nonlinear decoding of the corresponding output function $\Psi^{\dagger}(u) = Q(u, x)$.

**The Role of Positional Encoding.** Author has argued that some positional encoding ("explicit $x$-dependence") is necessary for universality. Positional encoding helps in breaking translational equivariance i.e. FNO $\Psi$ does not commute with the shift operator. Breaking translational equivariance is necessary because the PDEs with non-constant coefficients have solution operator which are non-translational equivariant. There is one more way to break translational equivariance that is to have bias function dependent on $x, b = b(x)$. It was used in previous work by the author [3] but in practice positional encoding is explicitly added in the input layer $\mathcal{R}$.

Explicit $x$-dependence in the output layer is not necessary since there are pointwise matrix multiplications involved in the hidden layers and they provide a mechanism (same as "skip connections") to forward positional encoding information to the output layer.

**Remark 2** (Skip Connections)**.** Skipping connections basically means bypass one or more layer of the neural network. It adds the output of layer($l$) to the output of layer($l + n$):

$$y = F(x) + x$$

where:

- $x$ is the input.
- $F(x)$ is the transformation applied to skip layer(s).
- $y$ is the output after the skip.

**Explanation and definition of terms:**

**Lipschitz Domain.** $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain and $\overline{\Omega}$ is compact (Heine-Borel theorem). The boundary $\delta\Omega = \overline{\Omega} \backslash \Omega$ is at least "Lipschitz regular"(locally it is the graph of a Lipschitz function). In particular, any bounded domain with piecewise smooth boundary is a Lipschitz domain.

**Function Extension for Lipschitz Domains.** We will be using this extension property in the proof. Given a function $u : \Omega \to \mathbb{R}$, defined on a domain $\Omega \subset \mathbb{R}^d$, to a function $u : \mathbb{R}^d \to \mathbb{R}$, defined on all of $\mathbb{R}^d$. The following lemma shows that this is possible, while preserving smoothness of $u$.

**Lemma 1** (A.1)**.** *[Periodic extension operator] Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. There exists a continuous, linear operator $\mathcal{E} : W^{s,p}(\Omega) \to W_{per}^{(}s, p)(B)$ for any $s \geq 0$ and $p \in [1, \infty]$, where $B \subset \mathbb{R}^d$ is a bounded hypercube containing $\Omega \subset B$, such that for any $u \in W^{s,p}(\Omega)$:*

(1) $\mathcal{E}(u) \mid_{\Omega}= u$;
(2) $\mathcal{E}(u) \in W^{s,p}_{per}(B)$ *is periodic on $B$ (including it's derivatives).*

*Furthermore, $\mathcal{E}$ maps continuously differentiable functions to continuously differentiable functions, i.e. $\mathcal{E}(C^s(\overline{\Omega})) \subset C^s_{per}(B)$ and hence defines a continuous mapping $\mathcal{E} : C^s(\overline{\Omega}) \to C^s_{per}(B)$.*

**Remark 3.** An operator $A$ is said to be continuous if it preserves the limit i.e. if a sequence $\{x_n\}$ converges to the limit $x$ then we have $Ax_n \to Ax$. For linear operator between normed spaces, continuity is equivalent to boundedness.

**Mollification (Smoothing) of Functions om Lipschitz Domains.** There exists a smooth mapping (a mollifier) $\rho : \mathbb{R}^d \to \mathbb{R}$, with the properties $p(x) \in [0,1]$ for all $x \in \mathbb{R}^d$.

$$\rho(x) = \begin{cases} = 1, & \text{if } x = 0 \\ = 0, & \text{if } |x| \geq 1 \\ \in [0,1], & \text{if } 0 < x < 1 \end{cases}$$

Normalization of $\rho$ can be done so that $\int_{\mathbb{R}^d} \rho(y)dy = 1$. With any such $\rho$, we can define a family of functions $\rho_\delta := \delta^{-d}\rho(x/\delta)$, with support in a $\delta-$ball around origin. Fixing such a family, $\epsilon$-mollification of a function $u : \mathbb{R}^d \to \mathbb{R}, u \in L^1(\mathbb{R}^d)$ is defined by a convolution $u_\delta(x) := (u * \rho_\delta)(x)$, i.e.

$$u_\delta(x) := \int_{\mathbb{R}^d} u(x-y)\rho_\delta(y)dy.$$

$u_\delta$ is smooth function for $\delta > 0$ (by lemma 2) and that

**Lemma 2.** *If $g \in C_c^\infty(\mathbb{R}^n)$ and*

**Lemma 3** (A.2)**.** *(Adapted mollification in a Lipschitz domain) Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. There exists a one-parameter family $\mathcal{M}_\delta$ of "mollification" operators, indexed by $\delta \geq 0$ and defining a linear mapping $\mathcal{M}_\delta : L^1(\Omega;\mathbb{R}^k) \to L^1(\Omega;\mathbb{R}^k)$ for $\delta > 0$, such that:*

(1) *For $\delta = 0$, we have $\mathcal{M}_0 u = u$ for all $u \in L^1(\Omega\mathbb{R}^k)$; for $\delta > 0$, $\mathcal{M}_\delta$ is smoothing, in the sense that it defines a mapping $\mathcal{M}_\delta : L^1(\Omega;\mathbb{R}^k) \to C^\infty(\overline{\Omega};\mathbb{R}^k)$.*

(2) *For any fixed $\delta > 0$, integer $s, r \geq 0$, the mapping,*

$$\mathcal{M}_\delta : C^s(\overline{\Omega};\mathbb{R}^k) \to C^r(\overline{\Omega};\mathbb{R}^k),$$

*is continuous. Furthermore, if $r \leq s$, then the operator norm is uniformly bounded, that is*

$$\sup_{\delta>0} \|\mathcal{M}_\delta\|_{C^s \to C^r} < \infty.$$

(3) *If $s \geq 0$ and $K \subset C^s(\overline{\Omega};\mathbb{R}^k)$ is a compact subset, then for any $\delta_0 \geq 0$,*

$$\lim_{\delta \to \delta_0} \sup_{u \in K} \|\mathcal{M}_{\delta_0} u - \mathcal{M}_\delta u\|_{C^s} = 0$$

.

(4) *For fixed $\delta > 0$, integer $s, r \geq 0$, and $p \in [1, \infty)$, the mapping*

$$\mathcal{M}_\delta : W^{s,p}(\Omega;\mathbb{R}^k) \to W^{r,p}(\Omega;\mathbb{R}^k)$$

*is continuous.*

(5) *If $s \geq 0, p \in [1,\infty)$ and $K \subset W^{s,p}(\Omega;\mathbb{R}^k)$ is a compact subset, and $\delta_0 \geq 0$, then*

$$\lim_{\delta \to \delta_0} \sup_{u \in K} \|\mathcal{M}_{\delta_0} u - \mathcal{M}_\delta u\|_{W^{s,p}} = 0.$$

**Lemma 4** (A.4)**.** *Fix $s \geq 0$. Let $K \subset C^s(\overline{\Omega}; \mathbb{R}^k)$ be compact. Then for any $\delta > 0$, the set*

$$K_\delta := \bigcup_{0 \leq \delta' \leq \delta} \{\mathcal{M}_{\delta'} u \mid u \in K\},$$

*is also compact in $C^s(\overline{\Omega}; \mathbb{R}^k)$.*

*Proof.* Since $C^s(\overline{\Omega}; \mathbb{R}^k)$ is a normed space with respect to $C^s$ norm, it will be metric space with respect to induced metric from norm. Hence, it is enough to show that subset $K_\delta$ is sequentially compact.

Let $v_1, v_2, \ldots$ be an arbitrary sequence in $K_\delta$. It suffices to show that $v_j$ posses a convergent subsequence $v_{j_l} \to v \in K_\delta$. By definition, there exist a sequence $u_j \in K$ and $\delta_j \in (0, \delta]$ such that $\mathcal{M}_{\delta_j} u_j = v_j$ for all $j \in \mathbb{N}$. Since, $K$ is compact, there exists a convergent subsequence $u_{j_l} \to u \in K$. Furthermore, we may assume that $\delta_{l_j} \to \delta_\infty \in [0, \delta]$ converges to a limit. (but we need to show its convergence. TODO.) Let $v := M_{\delta_\infty} u$ and $v \in K_\delta$. We claim that $v_{j_l} \to v$.

$$\begin{aligned}
\limsup_{l \to \infty} \|v_{j_l} - v\|_{C^s} &= \limsup_{l \to \infty} \|\mathcal{M}_{\delta_{j_l}} u_{j_l} - \mathcal{M}_{\delta_\infty} u\|_{C^s} \\
(u \in C^s \text{ is fixed}) &= \limsup_{l \to \infty} \|\mathcal{M}_{\delta_{j_l}} u_{j_l} - \mathcal{M}_{\delta_{j_l}} u + \mathcal{M}_{\delta_{j_l}} u - \mathcal{M}_{\delta_\infty} u\|_{C^s} \\
&\leq \limsup_{l \to \infty} \|\mathcal{M}_{\delta_{j_l}} u_{j_l} - \mathcal{M}_{\delta_{j_l}} u\|_{C^s} + \limsup_{l \to \infty} \|\mathcal{M}_{\delta_{j_l}} u - \mathcal{M}_{\delta_\infty} u\|_{C^s} \\
&= \limsup_{l \to \infty} \|\mathcal{M}_{\delta_{j_l}}\|_{C^s \to C^s} \|u_{j_l} - u\|_{C^s} \\
&\qquad + \limsup_{l \to \infty} \|\mathcal{M}_{\delta_{j_l}} u - \mathcal{M}_{\delta_\infty} u\|_{C^s}
\end{aligned}$$

From Lemma 3 (part 3), we can say second part converges to 0. Again from Lemma 3 (part 2), $\mathcal{M}_\delta$ is uniformly bounded

$$\sup_{l \in \mathbb{N}} \|\mathcal{M}_{\delta_{j_l}}\|_{C^s \to C^s} \leq \sup_{0 \leq \delta' \leq \delta} \|\mathcal{M}_{\delta'}\|_{C^s \to C^s} < \infty$$

(why so? since $\delta_j \in (0, \delta]$); $\delta_{j_l}$ is the subsequence of $\delta_j$ and if the subsequence has the supremum of whole interval $[0, \delta]$ then its equality otherwise it has to be less than because supremum is taken in whole $[0, \delta]$. Also, since $K$ is compact, $u_{j_l} \to u$ which implies first part also tends to 0. Hence, we have $v_{j_l} \to v$ and $K_\delta$ is sequentially compact hence compact. ■

**Universal Approximation Of Neural Networks.** From [[7], Thm. 4.1], neural network architecture is universal in the class of $C^s$- function between Euclidean vector spaces ($C^s(\mathbb{R}^n; \mathbb{R}^m)$) when the activation function $\sigma$ is nonpolynomial and sufficiently smooth, $\sigma \in C^s$. This implies universality of neural networks in Sobolev spaces $W^{s,p}$ of functions between Euclidean vector spaces. (Why? TODO).

**Lemma 5** (A.5)**.** *Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then for any function $u : \Omega \to \mathbb{R}^k$, where $u$ belongs to either the space of continuously differentiable functions $C^s(\overline{\Omega}; \mathbb{R}^k)$, or the Sobolov space $W^{s,p}(\Omega; \mathbb{R}^k)$, for integer $s \geq 0$ and $p \in [1, \infty)$, and for any $\epsilon > 0$, there exists a neural network $\tilde{u} : \Omega \to \mathbb{R}^k$ with activation function $\sigma$, such that*

$$\sup_{x \in \Omega} |u(x) - \tilde{u}(x)| \leq \epsilon.$$

*Proof.* **Step 1.** We first assume that $u \in C^s(\overline{\Omega};\mathbb{R}^k)$. By Lemma 1, there exists a (periodic) extension $U : \mathbb{R}^d \to \mathbb{R}$, such that $u(x) = U(x)$ for $x \in \Omega$, and $U \in C^s(\mathbb{R}^d;\mathbb{R}^k)$. It follows from [7](Thm 4.1) that for any $\epsilon > 0$, there exists a neural network $\tilde{u} : \mathbb{R}^d \to \mathbb{R}$, such that

$$\|u - \tilde{u}\|_{C^s(\overline{\Omega};\mathbb{R}^k)} = \sup_{|\alpha|\le s} \sup_{x\in\overline{\Omega}} |D^\alpha U(x) - D^\alpha \tilde{u}(x)| \le \epsilon.$$

**Step 2.** If $u \in W^{s,p}(\Omega;\mathbb{R}^k)$, then we note that by Lemma 3, the boundary-adapted mollification $u_\delta : \mathcal{M}_\delta$ converges to $u$ as $\delta \to 0$. Let $\epsilon > 0$ be given. Choose $\delta > 0$ sufficiently small, such that

$$\|u - u_\delta\|_{W^{s,p}(\Omega;\mathbb{R}^k)} \le \epsilon/2.$$

We note that $C^{s+1}(\overline{\Omega};\mathbb{R}^k) \hookrightarrow W^{s,p}(\Omega;\mathbb{R}^k)$ has a continuous embedding, and hence there exists a constant $C_0 > 0$, such that

$$\|\cdot\|_{W^{s,p}(\Omega;\mathbb{R}^k)} \le C_0 \|\cdot\|_{C^{s+1}(\Omega;\mathbb{R}^k)}.$$

Now note that there exists a neural network $\tilde{u}$, such that $\|u_\delta - \tilde{u}\|_{C^{s+1}(\overline{\Omega};\mathbb{R}^k)} \le \epsilon/2C_0$, this follows since $u_\delta \in C^{s+1}(\overline{\Omega};\mathbb{R}^k)$. and Step 1. Combining these estimate, we obtain

$$\begin{aligned}\|u - \tilde{u}\|_{W^{s,p}(\Omega;\mathbb{R}^K)} &\le \|u - u_\delta\|_{W^{s,p}(\Omega;\mathbb{R}^K)} + \|u_\delta - \tilde{u}\|_{W^{s,p}(\Omega;\mathbb{R}^K)} \\ &\le \epsilon/2 + C_0\|u_\delta - \tilde{u}\|_{C^{s+1}(\overline{\Omega};\mathbb{R}^k)} \\ &\le \epsilon.\end{aligned}$$

This is true for all $x \in \Omega$. This concludes our proof. ■

**A Dense Subset Of Operators.** In order to prove universal approximation for averaging neural operator,we ned to use the fact that it is possible to reduce the problem for general operator $\Psi^\dagger : C^s(\overline{\Omega};\mathbb{R}^k) \to C^{s'}(\overline{\Omega};\mathbb{R}^{k'})$(or $\Psi^\dagger : W^{s,p}(\overline{\Omega};\mathbb{R}^k) \to W^{s',p'}(\overline{\Omega};\mathbb{R}^{k'}$, respectively), to a simple class of operators which can be written in the form,

$$\tilde{\Psi}^\dagger(u) = \sum_{j=1}^J \alpha_j(u)\eta_j,$$

where $\eta_j$'s are functions in $C^{s'}(\overline{\Omega};\mathbb{R}^{k'})$(resp. in $W^{s',p'}(\overline{\Omega};\mathbb{R}^{k'})$), and $\alpha_1,\dots,\alpha_J : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$ are continuous nonlinear functionals, defined on the space of integrable functions. The following proposition summarizes this fact.

**Proposition 1** (A.6). *Let $\Psi^\dagger : C^s(\Omega;\mathbb{R}^k) \to C^{s'}(\overline{\Omega};\mathbb{R}^{k'})$ be a continuous operator. Let $K \subset C^s(\overline{\Omega}'\mathbb{R}^k)$ be a compact subset. Then for any $\epsilon > 0$, there exist $J \in \mathbb{N}$, functions $\eta_1,\dots\eta_J \in C^{s'}(\overline{\Omega};\mathbb{R}^k)$, and continuous functionals $\alpha_1,\dots,\alpha_J : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$, such that the operator $\tilde{\Psi}^\dagger : L^1(\Omega;\mathbb{R}^k) \to C^{s'}(\overline{\Omega};\mathbb{R}^{k'}), \tilde{\Psi}^\dagger(u) := \sum_{j=1}^J \alpha_j(u)\eta_j$, satisfies*

$$\sup_{u\in K} \|\Psi^\dagger(u) - \tilde{\Psi}^\dagger(u)\|_{C^{s'}} \le \epsilon$$

**Remark 4.** We note that the underline operator $\Psi^\dagger(u)$ is only defined for $u \in C^s(\Omega;\mathbb{R}^k)$, and will generally not possess anycontinuous extension to an operator defined $u \in L^1(\Omega;\mathbb{R}^k)$. But the operator constructed in the above expression $\tilde{\Psi}^\dagger$ :

$L^1(\Omega;\mathbb{R}^k) \to C^{s'}(\Omega;\mathbb{R}^{k'})$ when restricted to compact $K \subset C^s(\overline{\Omega};\mathbb{R}^k)$ provides a good approximation of $\Psi^\dagger|_K$.

*Proof.* It can be easily shown that $C^{s'}(\Omega;\mathbb{R}^k) \simeq [C^{s'}(\Omega;\mathbb{R})]^k$ are homeomorphic where the homeomorphism can be $u(x) \mapsto (u_1(x),\dots,u_k(x))$ since $[C^{s'}(\Omega;\mathbb{R})]^k$ is $k$-length tuples of scalar real valued functions, and thus it will suffice to approximate each component of the mapping $\Psi^\dagger : C^s(\overline{\Omega};\mathbb{R}^k) \to [C^{s'}(\Omega;\mathbb{R})]^k$, individually. The $j$-th component defines the mapping $\Psi_j^\dagger : C^s(\overline{\Omega};\mathbb{R}^k) \to C^{s'}(\overline{\Omega};\mathbb{R})$.

**Step 1: (construction of $\eta_1,\dots,\eta_j$).** We are going to show that the $L^2$-orthogonal (real) Fourier sine/cosine basis is a good choice for $\eta's$. Consider $K' := \Psi^\dagger(K) \subset C^{s'}(\overline{\Omega})$ for compact $K$ which implies $K'$ is also compact. Let $B \supset \Omega$ be a bounding box, containing $\overline{\Omega}$ in its interior. By Lemma 1, there exists a continuous extension mapping

$$\mathcal{E} : C^{s'}(\Omega) \to C^{s'}_{per}(B),$$

where $C^{s'}_{per}(B)$ denotes the space of continuously differentiable functions $w : B \to \mathbb{R}$ on the Cartesian domain $B$, having periodic derivates up to order $s'$. Since $K' \subset C^{s'}(\overline{\Omega})$ is compact, it follows $K^{'}_{per} := \mathcal{E}(K^{'})$ is also compact subset of $C^{s'}_{per}(B)$ (continuous image of a compact set). Since functions are periodic, we can identify $B \simeq \mathbb{T}^d$ with the periodic torus in a natural way. Let $\eta_1, \eta_2, \dots$ denote an enumeration of the $L^2$-orthogonal (real) Fourier sine/cosine basis in $L^2_per(B)$. Fix $\delta > 0$, and let $w_\delta : B \to \mathbb{R}$ be the standard mollification of the periodic function $w : B \to \mathbb{R}$. Since $K^{'}_{per} \subset C^{s'}_{per}(B)$ is compact, the set $\{w_\delta \mid w \in K^{'}_{per}\}$ is uniformly bounded in the $C^{r'}$-norm for any given $r' > s'$. In particular, it can be shown that approximation by Fourier series converges uniformly over $\{w_\delta \mid w \in K^{'})_{per}\}$, i.e.

$$\lim_{J\to\infty} \sup_{w\in K'_{per}} \left\| w_\delta - \sum_{j=1}^{J} \langle w_\delta, \eta_j\rangle_{L^2}\eta_j \right\|_{C^{s'}_{per}(B)} = 0$$

Furthur, $\lim_{\delta\to 0} \|w_\delta - w\|_{C^{s'}} = 0$ (3, part 3), uniformly over $K^{'}_{per}$. In particular, given $\epsilon > 0$, we can find $\delta = \delta(\epsilon) > 0$, such that $\sup_{w\in K'_{per}} \|w_\delta - w\|_{C^{s'}} \le \epsilon/2$, and then $J \in \mathbb{N}$, such that $\sup_{w\in K'_{per}} \left\|\sum_{j=1}^{J} \langle w_\delta, \eta_j\rangle_{L^2}\eta_j\right\|_{C^{s'}_{per}(B)} \le \epsilon/2$. It's straight forward follows from triangle inequality that

$$\sup_{w\in K'_{per}} \left\| w - \sum_{j=1}^{J} \langle w_\delta, \eta_j\rangle_{L^2}\eta_j \right\|_{C^{s'}_{per}(B)} \le \epsilon.$$

This defines the choices of $\eta_1,\dots,\eta_J$. Note that we have the identity

$$\sum_{j=1}^{J} \langle w_\delta, \eta_j\rangle_{L^2} = \sum_{j=1}^{J} \langle w, \eta_{j,\delta}\rangle_{L^2}.$$

Since $\Omega \subset B$, and $\mathcal{E}$ is an extension operator, so that $\mathcal{E}(v)|_\Omega = v$ for all $v \in K'$. Furthurmore, we have $K'_{per} = \mathcal{E}(K')$ and $K' = \Psi^\dagger(K)$, by definition. It follows that after using above identify,

$$\begin{aligned}
\sup_{u\in K}\|\Psi^{\dagger}(u)-\sum_{j=1}^{J}\langle\mathcal{E}(\Psi^{\dagger}(u)),\eta_{j,\delta}\rangle_{L^2}\eta_j\|_{C^{s'}}(\overline{\Omega}) &= \sup_{v\in K'}\left\|v-\sum_{j=1}^{J}\langle\mathcal{E}(v),\eta_{j,\delta}\rangle_{L^2}\eta_j\right\|_{C^{s'}(\overline{\Omega})}\\
\text{(property of extension operator)}\qquad &= \sup_{v\in K'}\left\|\mathcal{E}(v)-\sum_{j=1}^{J}\langle\mathcal{E}(v),\eta_{j,\delta}\rangle_{L^2}\eta_j\right\|_{C^{s'}(\overline{\Omega})}\\
\because C^{s'}_{per}(B)\subset C^{s'}(\overline{\Omega})\qquad &\le \sup_{v\in K'}\left\|\mathcal{E}(v)-\sum_{j=1}^{J}\langle\mathcal{E}(v),\eta_{j,\delta}\rangle_{L^2}\eta_j\right\|_{C^{s'}_{per}(B)}\\
K'_{per}=\mathcal{E}(K')\qquad &= \sup_{w\in K'_{per}}\left\|w-\sum_{j=1}^{J}\langle w,\eta_{j,\delta}\rangle_{L^2}\eta_j\right\|_{C^{s'}_{per}(B)}\\
&\le \epsilon
\end{aligned}$$

**Step 2. (construction of $\alpha_1,\dots,\alpha_J$).** Let us define nonlinear functional $\beta_j : C^s(\overline{\Omega};\mathbb{R}^k)\to\mathbb{R}$ by $\beta_j(u) := \langle\mathcal{E}(v),\eta_{j,\delta}\rangle_{L^2}$. Then from above we have,

$$\sup_{u\in K}\|\Psi^{\dagger}(u)-\sum_{j=1}^{J}\beta_j(u)\eta_j\|_{C^{s'}}\le\epsilon. \tag{4}$$

In above result, $\beta_j$ does not define a continuous functional $L^1(\Omega;\mathbb{R}^k)\to\mathbb{R}$. In order to get our desired result, we rely on mollification adapted to the bounded Lipschitz domain $\Omega$. Let $\delta>0$ denote a mollification parameter. By Lemma 3, *part* 3, there exists a continuous operator $\mathcal{M}_\delta : L^1(\Omega;\mathbb{R}^k)\to C^s(\overline{\Omega};\mathbb{R}^k)$, such that over the compact set $K\subset C^s(\overline{\Omega};\mathbb{R}^k)$, we have

$$\lim_{\delta\to 0}\sup_{u\in K}\|u-\mathcal{M}_\delta(u)\|_{C^s}=0,$$

We wish to define $\alpha_j : L^1(\Omega;\mathbb{R}^k)\to\mathbb{R}$ by $\alpha_j(u) := \beta_j(\mathcal{M}_\delta u)$ for suitably chosen $\delta>0$. Recall that for fixed $\delta_0>0$, the set $K_\delta$ defined by

$$K_\delta := \bigcup_{0\le\delta\delta_0}\mathcal{M}_\delta(K),$$

is a compact subset of $C^s(\overline{\Omega};\mathbb{R}^k)$, by Lemma 4. Note that for any $j=1,\dots,J$, we have a continuous mapping $\beta_j : K_\delta\to\mathbb{R}$ (proof for this fact can be found in the appendix A). Since $K_\delta$ is compact, $\beta_j$'s are uniformly continuous which implies there exists a *modulus of continuity* $\omega : [0,\infty)\to[0,\infty)$, satisfying $\omega(0)=0$, such that

$$|\beta_j(u)-\beta_j(u')|\le\omega(\|u-u'\|_{C^s}),\quad \forall u,u'\in K_\delta,$$

holds for all $j=1,\dots,J$. It follows that since for each $u\in K_\delta, \mathcal{M}_\delta(u)\in K_\delta$,

$$|\beta_j(u)-\beta_j(\mathcal{M}_\delta(u))|\le\omega(\|u-\mathcal{M}_\delta(u)\|_{C^s}),\quad \forall u\in K_\delta,$$

for any $0\le\delta\le\delta_0$. Also, $|\beta_j(u)-\beta_j(\mathcal{M}_\delta u)|\le\sup_{u\in K_\delta}(|\beta_j(u)-\beta_j(\mathcal{M}_\delta u)|)$ and there exists $u\in K_\delta$ for which supremum is attained since $K_\delta$ is compact. Therefore, for that $u$,

$$\sup_{u\in K_\delta}\left(|\beta_j(u)-\beta_j(\mathcal{M}_\delta u)|\right)\le\omega\|u-\mathcal{M}_\delta(u)\|_{C^s}\le\omega\sup_{u\in K_\delta}\left(\|u-\mathcal{M}_\delta(u)\|_{C^s}\right)$$

By Lemma 3, $\mathcal{M}_\delta u\to u$ uniformly over $u\in K$, so we can conclude,

$$\lim_{\delta\to 0}\sup_{u\in K}(|\beta_j(u)-\beta_j(\mathcal{M}_\delta u)|)\le\omega\lim_{\delta\to 0}\sup_{u\in K}\left(\|u-\mathcal{M}_\delta(u)\|_{C^s}\right)=0$$

In particular, we can choose $\delta>0$ sufficiently small, to ensure that $\alpha_j(u):=\beta_j(\mathcal{M}_\delta u)$ satisfies

$$\sup_{u\in K}|\beta_j(u)-\alpha_j(u)|\le\frac{\epsilon}{J\max_{j=1,\dots,J}\|\eta_j\|_{C^{s'}}},\tag{5}$$

for all $j=1,\dots,J$. Since $\delta>0$, $\mathcal{M}_\delta:L^1(\Omega;\mathbb{R}^k)\to C^s(\overline{\Omega};\mathbb{R}^k)$ is continuous (since its linear and bounded, by definition), and hence $\alpha_j=\beta_j\circ\mathcal{M}_\delta$ is continuous as mapping $\alpha_j:L^1(\Omega;\mathbb{R}^k)\to\mathbb{R}$.

**Step 3: (Conclusion).** Combining equation 4 and 5, we have

$$\begin{aligned}\sup_{u\in K}\|\Psi^\dagger(u)-\sum_{j=1}^{J}\alpha_j(u)\eta_j\|_{C^{s'}}&\le\sup_{u\in K}\|\Psi^\dagger(u)-\sum_{j=1}^{J}\beta_j(u)\eta_j\|_{C^{s'}}\\&\quad+\sup_{u\in K}\|\sum_{j=1}^{J}[\beta_j(u)-\alpha_j(u)]\eta_j\|_{C^{s'}}\\&\le\sup_{u\in K}\|\Psi^\dagger(u)-\sum_{j=1}^{J}\beta_j(u)\eta_j\|_{C^{s'}}\\&\quad+J\max_{j=1,\dots,J}\|\eta_j\|_{C^{s'}}\max_{j=1,\dots,J}\sup_{u\in K}|\beta_j(u)-\alpha_j(u)|\\&\le 2\epsilon.\end{aligned}$$

Since, $\epsilon>0$ is arbitrary, we have the desired result. ■

**Proposition 2.** *Let $\Psi^\dagger:W^{s,p}(\Omega;\mathbb{R}^k)\to W^{s',p'}(\Omega;\mathbb{R}^{k'})$ be a continuous operator. Let $K\subset W^{s,p}(\Omega;\mathbb{R}^k)$ be a compact subset. Then for any $\epsilon>0$, there exists $\eta_1,\dots,\eta_j\in W^{s',p'}(\Omega;\mathbb{R}^k)$, and continuous functional $\alpha_1,\dots,\alpha_N:L^1(\Omega;\mathbb{R}^k)\to\mathbb{R}$, such that*

$$\sup_{u\in K}\|\Psi^\dagger(u)-\sum_{j=1}^{J}\alpha_j(u)\eta_j\|_{W^{s',p'}}\le\epsilon.$$

**Remark 5.** The proof for this proposition above is exactly similar to the previous proposition. The space $W^{s,p}(\Omega;\mathbb{R}^k)\simeq[W^{s,p}(\Omega;\mathbb{R})]^k$ and the homeomorphism is given by $u(x)\mapsto(u_1(x),\dots,u_k(x))$. Wherever we have used Lemma 3, part 3, for Sobolev spaces we can use Lemma 3, part $4,5$ which provides analogous results for Sobolev spaces.

A similar result as Lemma 4 can also be proved in the settings of Sobolev space with the exactly similar sequential compactness implies compactness arguement for metric spaces.

**Approximation Of Nonlinear Functionals By The Averaging Neural Operator.** In this section, we will show that any nonlinear functional $\alpha : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}, u \to \alpha(u)$, can be approximated by an averaging neural operator in a suitable sense.

**Lemma 6.** *Let $\alpha : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$ be a continuous nonlinear functional. Let $K \subset L^1(\Omega;\mathbb{R}^k)$ be a compact set, consisting of bounded functions, $\sup_{u\in K} \|u\|_{L^\infty} < \infty$. Then for any $\epsilon > 0$, there exists an averaging neural operator $\tilde{\alpha} : L^1(\Omega;\mathbb{R}^k) \to L^1(\Omega)$, all of whose output functions are constant so that we may also view $\tilde{\alpha}$ as a function $\tilde{\alpha} : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$(since output functions are constant), such that*

$$\sup_{u\in K} |\alpha(u) - \tilde{\alpha}(u)| \le \epsilon.$$

*Proof.* Let $\alpha : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$ be a continuous linear functional. Let $K \subset L^1(\Omega;\mathbb{R}^k)$ be a compact set. Fix $\epsilon > 0$. Our aim is to show that there exists an averaging neural operator $\tilde{\alpha} : L^1(\Omega;\mathbb{R}^k) \to L^1(\Omega)$ with constant output such that

$$\sup_{u\in K} |\alpha(u) - \tilde{\alpha}(u)| \le \epsilon.$$

Note that we can identify any $u \in L^1(\Omega;\mathbb{R}^k)$ with a function in $L^1(B;\mathbb{R}^k)$, via an extension of $u(x) := 0$ for $x \in B\backslash\Omega$. Using this identification compact subset $K \in L^1(\Omega;\mathbb{R}^k)$ can be identified with a compact subset of $L^1(\mathbb{R}^d;\mathbb{R}^k)$. Fix a smooth mollifier $\rho \in C^\infty$, and denote $\rho_\delta := \delta^{-d}\rho(x/\delta)$ for $\delta > 0$. We denote by $u_\delta = (u * \rho_\delta)(x)$ the mollification of $u$ (extended to all of $\mathbb{R}^d$ by 0 outside $\Omega$). Since $K \subset L^1(\Omega;\mathbb{R}^k)$ is compact, it follows from properties of mollification that

$$\lim_{\delta\to 0} \sup_{u\in K} \|u - u_\delta\|_{L^1(\Omega;\mathbb{R}^k)} = 0.$$

Since, $\alpha : L^1(\Omega;\mathbb{R}^k)$ is continuous, it follows that the mapping $\alpha_\delta : L^1(B;\mathbb{R}^k) \to \mathbb{R}$ defined by $\alpha_\delta(u) := \alpha(u_\delta)$, (well-defined, since $\alpha$ is well-defined) for $\delta > 0$, converges uniformly over $K$, as $\delta \to 0$:

$$\lim_{\delta\to 0} \sup_{u\in K} |\alpha(u) - \alpha_\delta(u)| = \lim_{\delta\to 0} \sup_{u\in K} |\alpha(u) - \alpha(u_\delta)| = 0 \tag{6}$$

Since $\alpha$ is continuous and $u_\delta \to u$ uniformly. Now, for any choice of orthogonal basis $\xi_1, \xi_2, \dots$ of $L^2(\Omega;\mathbb{R}^k)$, which we may additionally choose to be smooth, $\xi_j \in C^\infty(\Omega;\mathbb{R}^k)$, we have uniform convergence,

$$\sup_{u\in K} \|u_\delta - \sum_{j=1}^{J} \langle u_\delta, \xi_j\rangle_{L^2} \xi_j\|_{L^1(\Omega;\mathbb{R}^k)} \to 0, \tag{7}$$

as $J \to \infty$ (this is the Fourier expansion of $u_\delta \in K$, a subset of Hilbert space). We note that above expression is well-defined since $K \subset L^1 \cap L^\infty \subset L^2$(which is the Hilbert space). Furthurmore, $K \subset L^2$ is compact in the $L^2$-norm, this makes it uniformly convergent. With the convention that $u$ and $\xi_j$ are expanded by 0 outside of $\Omega$, we have

$$\begin{aligned} \langle u_\delta, \xi_j\rangle_{L^2} &= \int_\Omega u_\delta(y) \cdot \xi_\delta(y) dy = \int_{\mathbb{R}^d} (u * \rho_\delta)(y) \cdot \xi_j(y) dy \\ &= \int_{\mathbb{R}^d} u(y) \cdot (\xi_j * \rho_\delta)(y) dy = \fint_\Omega u(y) \cdot \xi_{j,\delta}(y) dy, \end{aligned} \tag{8}$$

where the change of convolution is due to the assumption that $\rho_\delta$ is even (which is standard for the mollifiers) and we have defined $\xi_{j,\delta}(y) := |\Omega|(\xi_j * \rho_\delta)(y)$. Hence, if we define $\alpha_{\delta,J} := L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$ by

$$\alpha_{\delta,J} := \alpha\left(\sum_{j=1}^{J}\left(\fint_\Omega u(y)\cdot\xi_{j,\delta}(y)dy\right)\xi_j\right)$$

then it follows from (6), (7) and (8), that for $\delta(K) > 0$ sufficiently small, and $J = J(\delta, K) \in \mathbb{N}$ sufficiently large, we have $\sup_{u\in K}|\alpha_\delta(u) - \alpha_{\delta,J}(u)| \leq \epsilon/2$. TO indicate the connection to averaging neural operators, we note that we can write $\alpha_{\delta,J}$ as the following composition:

$$u \mapsto \left(\fint_\Omega u(y)\cdot\xi_{1,\delta}(y)dy, \dots, \fint_\Omega u(y)\cdot\xi_{J,\delta}(y)dy\right) \mapsto \alpha\left(\sum_{j=1}^{J}\left(\fint_\Omega u(y)\cdot\xi_{j,\delta}(y)dy\right)\xi_j\right).$$

First mapping requires calculation of average hence requires non-locality. Second mapping is of the type $\beta_J : \mathbb{R}^J \to \mathbb{R}, c = (c_1, \dots c_J) \mapsto \alpha(\sum_{j=1}^J c_j\xi_j)$. Approximating this mapping can be done by ordinary neural networks and requires nonlinearity and not nonlocality. Choose $M > 0$, such that the image of the compact set $K \subset L^1(\Omega;\mathbb{R}^k)$ under the mapping

$$L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}^J, u \mapsto \left(\fint_\Omega u(y)\cdot\xi_{1,\delta}(y)dy, \dots, \fint_\Omega u(y)\cdot\xi_{J,\delta}(y)dy\right),$$

is compact and hence contained in a box $[-M, M]^J$. This is required since the input of $\alpha$ need to be a element of a compact set for out analysis to make sense. Note that $c \to \beta(c) = \alpha\left(\sum_{j=1}^J c_j\xi_j\right)$ is continuous, by continuity of $\alpha$. By universality of conventional neural networks, there exists a neural network $\tilde{\beta} : \mathbb{R}^J \to \mathbb{R}$, such that

$$\sup_{c\in[-M,M]^J}|\tilde{\beta(c)} - \beta(c)| \leq \epsilon. \tag{9}$$

We can write $\tilde{\beta} : \mathbb{R}^J \to \mathbb{R}$ as composition of the hidden layers:

$$\begin{cases} v_0 := c, \\ v_l := \sigma\left(\tilde{A}_l v_{l-1} + b_l\right), l = 1, \dots, L, \\ \tilde{\beta}(c) := \tilde{A}_{L+1}v_L + \tilde{b}_{L+1}, \end{cases} \tag{10}$$

where $\tilde{A}_l, \tilde{b}_l$ are the weights and biases of the hidden layers. Parallelizing the neural networks constructed in Lemma 7, it follows that for any $\epsilon > 0$, there exists a neural network $\tilde{R} : \mathbb{R}^K \times \Omega \to \mathbb{R}^J, (v, x) \mapsto \tilde{R}(v, x) - ((\tilde{R}_1)(v, x), \dots, \tilde{R}_J(v, x))$, such that

$$\sup_{u\in K}\left|\fint_\Omega \tilde{R}_j(u(y), y)dy - \fint_\Omega u(y)\cdot\xi_{j,\delta}(y)dy\right| \leq \epsilon',$$

for each $j = 1, \dots, J$. Composing the output layer with an affine mapping, we can construc another network $R$, such that

$$R(u(x), x) = \tilde{A}_1\tilde{R}(u(x), x) + \tilde{b}_1,$$

where $\tilde{A}_1, \tilde{b}_1$ are the weights and biases of the input layer of $\tilde{\beta}$ defined by 10. In particular, it follows that for any input $u \in K$, and defining coefficients $c(u) :=$

$(c_1,\dots,c_J)$ by $c_j := \fint_\Omega u(y)\cdot \xi_{j,\delta}(y)dy$, we have

$$\sup_{u\in K}\left|\sigma\left(\fint_\Omega R(u(x),y)dy\right) - \sigma\left(\tilde{A}_1 c(u) + \tilde{b}_1\right)\right| \le \|\sigma\|_{Lip}\|\tilde{A}\|\epsilon' \tag{11}$$

where $\|\tilde{A}_1\|$ denotes the operator norm of $\tilde{A}_1$. Above inequality is derived using the last two inequalities before it. The second term in the above inequality is exactly the output of the first hidden layer of $\tilde{\beta}$ in equation 10. Let us decompose $\tilde{\beta}(c) = \tilde{\beta}_1 \circ \sigma(\tilde{A}_1 c + \tilde{b}_1)$, where $\tilde{\beta}_1$ denote the composition of the other hidden layers, $l = 2,\dots,L$, amd the output layer. Composing each of the terms appearing on the left-hand side of above equation with $\tilde{\beta}_1$ and choosing $\epsilon'$ sufficiently small (depending on $\tilde{\beta}_1, \tilde{A}_1$), we can ensure that

$$\sup_{u\in K}\left|\tilde{\beta}_1\left(\sigma\left(\fint_\Omega R(u(y),y)dy\right)\right) - \tilde{\beta}_1\left(\sigma\left(\tilde{A}_1 c(u) + \tilde{b}_1\right)\right)\right| \le \epsilon,$$

since $\tilde{\beta}_1$ is the composition of linear terms with nonlinearity activation function which is Lipschitz hence applying similar argument as equation 11 multiple times will give the above result. Above result is equivalent to,

$$\sup_{u\in K}\left|\tilde{\beta}_1\left(\sigma\left(\fint_\Omega R(u(y),y)dy\right)\right) - \tilde{\beta}\left(c(u)\right)\right| \le \epsilon \tag{12}$$

This is because $\tilde{\beta}(c) = \tilde{\beta}_1 \circ \sigma(\tilde{A}_1 c + \tilde{b}_1)$. From 9 and 12, it follows that the averaging neural operator $\tilde{\alpha}$, defined by the following composition,

$$\tilde{\alpha} : u \xmapsto{(1)} R(u(\cdot),\cdot) \xmapsto{(2)} \sigma\left(\fint_\Omega R(u(y),y)dy\right) \xmapsto{(3)} \tilde{\beta}_1 \circ \sigma\left(\fint_\Omega R(u(y),y)dy\right), \tag{13}$$

satisfies

$$\begin{aligned}
\sup_{u\in K}|\alpha(u) - \tilde{\alpha}(u)| &= \sup_{u\in K}|\beta(c(u)) - \tilde{\alpha}(u)| \\
&\le \sup_{u\in K}|\beta(c(u)) - \tilde{\beta}(c(u))| \\
&\qquad + \sup_{u\in K}\left|\tilde{\beta}(c(u)) - \tilde{\beta}_1 \circ \sigma\left(\fint_\Omega R(u(y),y)\right)\right| \\
&\le \sup_{c\in[-M,M]^N}|\beta(c) - \tilde{\beta}(c)| \\
&\qquad + \sup_{u\in K}\left|\tilde{\beta}(c(u)) - \tilde{\beta}_1 \circ \sigma\left(\fint_\Omega R(u(y),y)dy\right)\right| \\
&\le \epsilon + 0 = \epsilon.
\end{aligned}$$

■

**Lemma 7.** *Let $K \subset L^1(\Omega;\mathbb{R}^k)$ be compact, consisting of uniformly bounded function $\sup_{u\in K}\|u\|_{L^\infty} < \infty$. Let $\xi \in C^\infty(\overline{\Omega};\mathbb{R}^k)$ be given and fixed. Then for $\epsilon > 0$, there exists a nerual network $\tilde{R} : \mathbb{R}^k \times \Omega \to \mathbb{R}^k$, such that*

$$\sup_{u\in K}\left|\fint_\Omega \tilde{R}((u(y)),y)dy - \fint_\Omega u(y)\cdot\epsilon(y)dy\right| \le \epsilon.$$

*Proof.* Fix $\epsilon > 0$. In the following, we denote by

$$M_K := \sup_{u\in K}\|u\|_{L^\infty},$$

the upper $L^\infty$-bound on elements $u \in K$. By assumption, $M_K$ is finite. As $\xi \in C^\infty(\overline{\Omega}; \mathbb{R}^k)$ is fixed, and $\Omega \subset \mathbb{R}^d$ is bounded, there exists a neural network $\tilde{\xi} : \Omega \to \mathbb{R}^k$, such that

$$\sup_{u \in K} |\xi(x) - \tilde{\xi}(x)|_{l^\infty} \leq \epsilon/(2kM_K),$$

where $k$ is the number of components of $\xi$. Furthermore define

$$M_\xi := \max\{\|\xi\|_{L^\infty}, \|\tilde{\xi}\|_{L^\infty}\} < \infty.$$

By universality of ordinary neural networks, there exists a neural network $\tilde{\times}$ : $[-M_K, M_K]^k \times [-M_\epsilon, M_\epsilon]^k \to \mathbb{R}$, such that

$$\sup_{|v|_{l^\infty} \leq M_K, |w|_{l^\infty} \leq M_\epsilon} |v \cdot w - \tilde{\times}(u, w)| \leq \epsilon/2.$$

Defining a new neural network as the composition $\tilde{R}(v, x) := \tilde{\times}(v, \tilde{\epsilon}$, it now follows that for any $u \in K$:

$$\begin{aligned}
\sup_{x \in \Omega} |\tilde{R}(u(x), x) - u(x) \cdot \xi(x)| &= \sup_{x \in \Omega} |\tilde{\times}(u(x), \tilde{\times}(x)) - u(x) \cdot \xi(x)| \\
&\leq \sup_{x \in \Omega} |\tilde{\times}(u(x), \tilde{\xi}(x)) - u(x) \cdot \tilde{\xi}(x)| \\
&\qquad \sup_{x \in \Omega} |u(x) \cdot \tilde{\xi}(x) - u(x) \cdot \xi(x)| \\
&\leq \sup_{|v|_{l^\infty} \leq M_K, |w|_{l^\infty} \leq M_\epsilon} |\tilde{\times}(v, w) - v \cdot w| \\
&\qquad + kM \sup_{x \in \Omega} |\tilde{\xi}(x) - \xi(x)|_{l^\infty} \\
&\leq \epsilon.
\end{aligned}$$

It can easily be seen that this upper bound implies that

$$\sup_{u \in K} \left| \fint_\Omega \tilde{R}(u(y), y) dy - \fint_\Omega (u(y) \cdot \xi(y) dy \right| \leq \epsilon.$$

■

**Proof of Universal Approximation $C^s \to C^{s'}$, Theorem 1.** The main idea behind the proof is that using Proposition 1, it suffices to approximate operators of the form

$$\tilde{\Psi}^\dagger(u) = \sum_{j=1}^{J} \alpha_j(u) \eta_j,$$

where $\alpha_j : L^1(\Omega; \mathbb{R}^n) \to \mathbb{R}$ are the functionals, and $\eta_j \in C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ are fixed functions. Given averaging neural operators $\Psi_1, \dots, \Psi_J$, we can obtain a new averaging neural operator $\Psi$, such that $\Psi(u) = \sum_{j=1}^{J} \Psi_j(u)$ for all input functions $u$. Thus, it suffices to prove that any one mapping of the form

$$\tilde{\Psi}_j^\dagger : L^1(\Omega; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'}), u \mapsto \alpha_j(u) \eta_j,$$

can be approximated by an averaging neural operator.

*Proof.* (Theorem 1) Let $\Psi^\dagger : C^s(\overline{\Omega};\mathbb{R}^k) \to C^{s'}(\overline{\Omega};\mathbb{R}^k)$ be a continuous operator. We aim to show that there exists an averaging neural operator $\Psi$ of the form $\Psi = \mathcal{Q} \circ \mathcal{L}_L \circ \ldots L_1 \circ \mathcal{R}$, such that

$$\sup_{u\in K} \|\Psi^\dagger(u) - \Psi(u)\|_{C^{s'}} \le \epsilon.$$

Fix $\epsilon > 0$. By Proposition 1, there exists functions $\eta_1, \ldots, \eta_J \in C^{s'}(\overline{\Omega};\mathbb{R}^{k'})$, and continuous functionals $\alpha_1, \ldots, \alpha_J : L^(\Omega;\mathbb{R}^k) \to \mathbb{R}$, such that

$$\sup_{u\in K} \|\Psi^\dagger(u) - \sum_{j=1}^{J} \alpha_j(u)\eta_j\|_{C^{s'}} \le \epsilon/2.$$

We now prove make the following claim:

**Claim A.12 .** Let $K \subset L^1(\Omega;\mathbb{R}^k)$ be a compact set, consisting of bounded functions $\sup_{u\in K}\|u\|_{L^\infty} < \infty$. Let $\eta_1, \ldots, \eta_J \in C^{s'}(\overline{\Omega};\mathbb{R}^k)$ be functions and let $\alpha_1, \alpha_J : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$ be continuous nonlinear functionals. Then for any $j = 1, \ldots, J$, there exists an averaging neural operator $\Psi_j : L^1(\Omega;\mathbb{R}^k) \to C^{s'}(\overline{\Omega};\mathbb{R}^k)$, such that

$$\sup_{u\in K} \|\alpha_j(u)\eta_j - \Psi_j(u)\|_{C^{s'}} \le \epsilon/2J.$$

Relying on above claim, it is easy to see that there exists an averaging neural operator, such that $\Psi(u) = \sum_{j=1}^J \Psi_j(u)$ for all input functions $u$. This operator satisfies, for any $u \in K$:

$$\begin{aligned}
\|\Psi^\dagger(u) - \Psi(u)\|_{C^{s'}} &\le \left\|\Psi^\dagger(u) - \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_j\right\|_{C^{s'}} + \left\|\sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_j - \Psi(u)\right\|_{C^{s'}} \\
&\le \left\|\Psi^\dagger(u) - \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_j\right\|_{C^{s'}} + \sum\nolimits_{j=1}^{J} \|[\alpha_j(u)\eta_j - \Psi_j(u)]\|_{C^{s'}} \\
&\le \epsilon/2 + J\epsilon/2J = \epsilon,
\end{aligned}$$

This is true for all $u \in K$ hence taking supremum over $K$, the above claim implies,

$$\sup_{u\in K} \|\Psi^\dagger(u) - \Psi(u)\|_{C^{s'}} \le \epsilon,$$

which gives the desired result of Theorem 1.

In order to prove above **Claim A.12**, fix $j \in \{1, \ldots, J\}$ for all the following. We first define

$$M_\eta := \|\eta_j\|_{C^{s'}(\Omega;\mathbb{R}^{k'})}. \tag{14}$$

We observe that by Lemma 6, there exists an averaging neural operator $\tilde{\alpha}_j : L^1(\Omega;\mathbb{R}^k) \to L^1(\Omega)$, with the constant output functions, such that

$$\sup_{u\in K} |\alpha_j(u) - \tilde{\alpha_j}(u)| \le \epsilon/6JM_\eta. \tag{15}$$

Choose $M_\alpha > 0$, such that

$$\sup_{u\in K} |\alpha_j(u)|, \sup_{u\in K} |\tilde{\alpha}_j(u) \le M_\alpha. \tag{16}$$

Since $\eta_j \in C^{s'}(\overline{\Omega};\mathbb{R}^{k'})$, there exists an ordinary neural network $\tilde{\eta}_j : \Omega \to \mathbb{R}^{k'}$ such that (Lemma 5):

$$\|\eta_j - \tilde{\eta}_j\|_{C^{s'}} \le \epsilon/6JM_\alpha. \tag{17}$$

Let us also define

$$\tilde{M}_\eta := \max\{M_\eta, \|\tilde{\eta}_j\|_{C^{s'}}\}. \tag{18}$$

Fix a small parameter $\delta > 0$. Since scalar multiplication $\times : [-M_\alpha, M_\alpha] \times [-\tilde{M}_\eta, \tilde{M}_\eta]^k \to \mathbb{R}^K, (a, v) \mapsto \times(a, v) := av$ defines a smooth mapping, there similarly exists a nerual network $\tilde{\times} : [-M_\alpha, M_\alpha] \times [-\tilde{M}_\eta, \tilde{M}_\eta]^k \to \mathbb{R}^K, (a, v) \mapsto \times(a, v)$, such that

$$\| \times (\cdot, \cdot) - \tilde{\times}(\cdot, \cdot)\|_{C^{s'}([-M_\alpha, M_\alpha]\times[-\tilde{M}_\eta, \tilde{M}_\eta]^{k'}; \mathbb{R}^{k'})} \leq \delta \tag{19}$$

We also know that composition of $C^{s'}$- functions is itself $C^{s'}$, and that there exists a constant $C_0 = C_0(s', k') > 0$, depending only on $s'$ and $k'$, such that

$$\begin{aligned}
\sup_{|\alpha| \leq M_\alpha} & \| \times (a, \tilde{\eta}_j(\cdot)) - \tilde{\times}(a, \tilde{\eta}_j(\cdot))\|_{C^{s'}(\Omega; \mathbb{R}^{k'})} \\
& \leq \| \times (\cdot, \tilde{\eta}_j(\cdot)) - \tilde{\times}(\cdot, \tilde{\eta}_j(\cdot))\|_{C^{s'}([-M_\alpha, M_\alpha]\times\Omega; \mathbb{R}^{k'})} \\
& \leq C_0 \| \times (\cdot, \cdot) - \tilde{\times}(\cdot, \cdot)\|_{C^{s'}([-M_\alpha, M_\alpha]\times[-\tilde{M}_\eta, \tilde{M}_\eta]^{k'}; \mathbb{R}^{k'})} \|\tilde{\eta}_j\|_{C^{s'}(\Omega; \mathbb{R}^{k'})} \\
& \leq C_0 \delta \tilde{M}_\eta.
\end{aligned}$$

Thus, for any $|\alpha|, |\tilde{\alpha}| \leq M_\alpha$, we obtain

$$\begin{aligned}
\| \times (a, \eta_j(\cdot)) - & \tilde{\times}(\tilde{a}, \tilde{\eta}_j(\cdot))\|_{C^{s'}(\Omega; \mathbb{R}^{k'})} \\
& \leq \| \times (a, \eta_j(\cdot)) - \times(\tilde{a}, \eta_j(\cdot))\|_{C^{s'}(\Omega; \mathbb{R}^{k'})} \\
& \qquad + \| \times (\tilde{a}, \eta_j(\cdot)) - \times(\tilde{a}, \tilde{\eta}_j)\|_{C^{s'}(\Omega; \mathbb{R}k')} \\
& \leq |a - \tilde{a}| M_\eta + M_\alpha \|\eta_j - \tilde{\eta}_j\|_{C^{s'}(\Omega; \mathbb{R}^{k'})} + C_0 \delta \tilde{M}_\eta.
\end{aligned}$$

From equation 15 and 17, we bound the first two terms, and choosing $\delta := \epsilon/(6NC_0\tilde{M}_\eta)$ in equation 19, it follows that

$$\| \times (a, \eta_j(\cdot)) - \tilde{\times}(\tilde{a}, \tilde{\eta}_j(\cdot))\|_{C^{s'}(\Omega; \mathbb{R}^k)} \leq \epsilon/2J,$$

for any $|a|, |\tilde{a}| \leq M_\alpha$. By definition of $M_\alpha$, given arbitrary $u \in K$, we have $|\alpha_j(u)| \leq M_\alpha$ and $|\tilde{\alpha}_j(u)| \leq M_\alpha$. Hence, the above estimates implies that

$$\sup_{u \in K} \|\alpha_j(u)\eta_j - \tilde{\times}(\tilde{\alpha}_j(u), \tilde{\eta}_j(\cdot))\|_{C^{s'}(\Omega; \mathbb{R}^k)} \leq \epsilon/2J. \tag{20}$$

The mapping $\mathbb{R} \times \Omega \to \mathbb{R}^{k'}, (a, x) \mapsto \tilde{\times}(a, \tilde{\eta}_j(x))$ is an ordinary neural network in $(a, x)$. Since $\tilde{\alpha}_j$ is an averaging neural operator by construction (Lemma 6), we can write it in the form $\tilde{Q} \circ \mathcal{L}_L \circ \cdots \circ \mathcal{L}_1 \circ \mathcal{R}$, in terms of a raising operator $\mathcal{R}$, hidden layer $\mathcal{L}_l$, and a projection layer $\tilde{Q}$, where the values $\tilde{Q}(v)(x) := \tilde{Q}(v)(x) \in \mathbb{R}$ are given in terms of an ordinary neural network $\tilde{Q}$. Let $\mathcal{Q}$ denote the composition $\mathcal{Q}(v)(x) := \tilde{\times}(\tilde{Q}(v(x), x), \tilde{\eta}_j(x))$. Then

$$\Psi(u) := \mathcal{Q} \circ \mathcal{L}_L \circ \circ \cdots \circ \mathcal{L}_1 \circ \mathcal{R}(u),$$

defined an averaging neural operator, for which

$$\Psi(u)(x) = \tilde{\times}(\tilde{\alpha}_j(u), \tilde{\eta}_j(x)).$$

Above relation is true since given $\tilde{\alpha}_j$ as composition, its output will be output of $\tilde{\mathcal{Q}}(v)(x)$ which is same as $\tilde{Q}(v(x), x)$, and $\Psi$'s output depends upon $\mathcal{Q}$ which is defined above. Considering all these facts, above relation comes out to be true.

By 20, it follows that

$$\|\alpha_j(u)\eta_j - \Psi(u)\|_{C^{s'}} \leq \epsilon/2J.$$

This concludes the proof of **Claim A.12**. ■

**Proof of Universal Approximation $C^s \to C^{s'}$, Theorem 2.** This section provides the proof in the scale of Sobolev spaces. The proof is similar to the case of spaces of continuity differentiable functions hence only the alteration to the proof have been mentioned here.

*Proof.* (Theorem 2 Let $\Psi^\dagger : W^{s,p}(\Omega;\mathbb{R}^k) \to W^{s',p'}(\Omega;\mathbb{R}^{k'})$ be a continuous operator with integer $s, s' \geq 0$ and $p, p' \in [1, \infty)$. Let $K \subset W^{s,p}(\Omega;\mathbb{R}^k)$ be compact, consisting of bounded functions, $\sup_{u\in K} \|u\|_{L^\infty} < \infty$. We aim to show that for any $\epsilon > 0$, there exists an averaging neural operator $\Psi$ of the form $\Psi = \mathcal{Q} \circ L_L \circ \cdots \circ L_1 \circ \mathcal{R}$, such that

$$\sup_{u\in K} \|\Psi^\dagger(u) - \Psi(u)\|_{W^{s',p'}} \leq \epsilon.$$

Fix $\epsilon > 0$. By Proposition 2, there exist functions $\eta_1, \dots, \eta_J \in W^{s',p'}(\Omega;\mathbb{R}^{k'})$, and continuous functionals $\alpha_1, \dots, \alpha_J : L^1(\Omega;\mathbb{R}^k) \to \mathbb{R}$, such that

$$\sup_{u\in K} \|\Psi^\dagger(u) - \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_j\|_{W^{s',p'}} \leq \epsilon/3.$$

Approximating each $\eta_j$ by its boundary-adapted mollification, $\eta_j, \delta := \mathcal{M}_\delta \eta_j \in C^\infty(\overline{\Omega};\mathbb{R}^{k'})$ with $\delta > 0$ choose sufficiently small (Lemma 3), we can ensure that

$$\sup_{u\in K} \|\Psi^\dagger(u) - \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_{j,\delta}\|_{W^{s',p'}} \leq 2\epsilon/3.$$

Also, we know that $C^{s+1}(\overline{\Omega};\mathbb{R}^k) \hookrightarrow W^{s',p'}(\Omega;\mathbb{R}^{k'})$ has a continuous embedding, hence there exists a constant $C_0 > 0$, such that

$$v\|_{W^{s',p'}} \leq C_0 \|v\|_{C^{s'+1}}, \forall v \in C^{s'+1}(\overline{\Omega};\mathbb{R}^{k'}). \tag{21}$$

We note that $u \mapsto \sum_{j=1}^{J} \alpha_j(u)\eta_{j,\delta}$ denotes a continuous operator $L^1(\omega;\mathbb{R}^k) \to C^{s'+1}(\overline{\Omega};\mathbb{R}^{k'})$ and recall that, by assumption, $K \subset W^{s,p}(\Omega;\mathbb{R}^k) \subset L^1(\Omega;\mathbb{R}^k)$ is a compact set consisting of bounded function, $\sup_{u\in K} \|U\|_{L^\infty} < \infty$. It thus follows from **Claim A.12** that there exists an averaging neural operator $\Psi : L^1(\Omega;\mathbb{R}^k) \to C^{s'+1}(\overline{\Omega};\mathbb{R}^{k'})$, such that

$$\sup_{u\in K} \| \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_{j,\delta} - \Psi(u)\|_{C^{s'+1}} \leq \epsilon/3C_0,$$

where $C_0 > 0$ denotes the embedding constant of 20. For this averaging neural operator $\Psi$, it follows that

$$\begin{aligned}
\sup_{u\in K} \|\Psi^\dagger(u) - \Psi(u)\|_{W^{s',p'}} &\leq \sup_{u\in K} \|\Psi^\dagger(u) - \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_{j,\delta}\|_{W^{s',p'}} \\
&\qquad + \sup_{u\in K} \| \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_{j,\delta} - \Psi(u)\|_{W^{s',p'}} \\
&\leq \sup_{u\in K} \|\Psi^\dagger(u) - \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_{j,\delta}\|_{W^{s',p'}} \\
&\qquad + C_0 \sup_{u\in K} \| \sum\nolimits_{j=1}^{J} \alpha_j(u)\eta_{j,\delta} - \Psi(u)\|_{C^{s'+1}} \\
&\leq \epsilon.
\end{aligned}$$

This concludes our proof. ■

## Appendix A. Detailed proofs

**Lemma 8.** *The mapping $\beta_j : K_\delta \to \mathbb{R}$ given by $u(x) \mapsto \langle \mathcal{E}(\Psi^\dagger(u)), \eta_{j,\delta} \rangle_{L^2}$ is continuous where $K_\delta$ is defined as in Lemma 4 and is compact. $\Psi^\dagger : C^s(\overline{\Omega}; \mathbb{R}^k) \to C^{s'}(\overline{\Omega}; \mathbb{R}^{k'})$ and $\mathcal{E}$ is the extension operator as in Lemma 1.*

*Proof.* Let $\{u_i\}, i = 1, \dots, n$ be the sequence in $K_\delta$ (compact) such that $u_n \to u \in K_\delta$. Our claim is to show that $\beta_j(u_n) \to \beta_j(u)$ as $n \to \infty$. For $u_n \in K_\delta$,

$$\begin{aligned}
\lim_{n\to\infty} \beta_j(u_n) &= \lim_{n\to\infty} \langle \mathcal{E}(\Psi^\dagger(u_n)), \eta_{j,\delta} \rangle_{L^2} \\
&= \lim_{n\to\infty} \langle \mathcal{E}(v_n), \eta_{j,\delta} \rangle_{L^2} \qquad \text{for } v_n \in K_\delta^{'} := \Psi^\dagger(K_\delta) \\
&= \lim_{n\to\infty} \langle w_n, \eta_{j,\delta} \rangle_{L^2} \qquad \text{for } w_n \in K'_{\delta,per} := \mathcal{E}(K'_\delta) \\
&= \lim_{n\to\infty} \int_\Omega w_n n_{j,\delta} dx \qquad \text{for } x \in \Omega \\
&= \int_\Omega \lim_{n\to\infty} w_n \eta_{j,\delta} dx \\
&= \int_\Omega w \eta_{j,\delta} dx = \langle \mathcal{E}(\Psi^\dagger(u)), \eta_{j,\delta} \rangle_{L^2} \\
&= \beta_j(u)
\end{aligned}$$

Since $u_n \to u$ and, $\Psi^\dagger$ and $\mathcal{E}$ are continuous operators hence we have $\mathcal{E}(\Psi^\dagger(u_n)) \to \mathcal{E}(\Psi^\dagger(u))$(a.k.a $w_n \to w$) over a compact set $K_\delta$ which implies uniform convergence over $K_\delta$. Hence, the set $\{w_n \mid n = 1, 2, \dots\}$ is uniformly bounded which implies $|w_n \eta_{j,\delta}| \le M |\eta_{j,\delta}| =: g \in L^2_{per}(B)$ for some constant $M$.

$$\begin{aligned}
\int_\Omega |g(x)| \cdot 1 dx &\le \left( \int_\Omega \|g(x)\|_{L^2} dx \right)^{1/2} \cdot \left( \int_\Omega 1 dx \right)^{1/2} \qquad \text{(Hölder inequality)} \\
&< \infty. \qquad \because \Omega \subset B \text{ which is bounded.}
\end{aligned}$$

This implies $g \in L^1(\Omega)$ and hence by Dominated convergence theorem, limit and integral can be interchange leading to the desired result. ■

## References

[1] A. Anandkumar, K. Azizzadenesheli, et al. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

[2] Guannan Chen, Xuan Liu, Qi Meng, Long Chen, Chunmei Liu, and Yulan Li. Learning neural operators on riemannian manifolds. 2023. Preprint.

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

[4] 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.

[5] Samuel Lanthaler, Roberto Molinaro, Philipp Hadorn, and Siddhartha Mishra. Nonlinear reconstruction for operator learning of pdes with discontinuities. *arXiv preprint arXiv:2210.01074*, 2022. URL `https://arxiv.org/abs/2210.01074`. Preprint.

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

[7] Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999. doi: 10.1017/S0962492900002919.