Refinement Does Not Amplify Discretization Mismatch in Fourier Neural Operators

Anonymous authors
Paper under double-blind review

Abstract

Neural operators provide a powerful framework for learning mappings between function spaces, including solution operators for partial differential equations (PDEs). Fourier neural operators (FNOs) are particularly attractive because their parameters are independent of the spatial resolution, enabling a model trained on one discretization to be evaluated on another. However, resolution-independent parameters do not imply discretization-invariant predictions: the outputs of an FNO can differ when the same input function is represented on different discretizations. This discretization mismatch is important for assessing the reliability of FNOs across discretizations, yet a precise theoretical characterization of how the output discrepancy depends on the two discretization sizes remains lacking. In this work, we derive an explicit upper bound for the pointwise discretization mismatch error (DME) of FNOs. Our bound characterizes the dependence of the DME on the size of the base discretization and the relative refinement between the two discretizations. In particular, for any fixed base discretization, the bound remains uniformly controlled as the finer discretization becomes arbitrarily fine. Thus, refining the finer discretization does not amplify the mismatch beyond a bound determined by the base discretization. Moreover, the bound converges to zero as the base discretization is refined. These results provide a quantitative characterization of how discretization refinement affects pointwise mismatch in FNOs. Numerical experiments on representative FNO settings support our theoretical results and demonstrate the refinement-uniform behavior predicted by the bound.

1 Introduction

1.1 Background

Many physical systems in science and engineering are governed by known physical laws such as partial differential equations (PDEs), and physics-informed machine learning incorporates such laws or prior physical knowledge into learning (Karniadakis et al., 2021; Meng et al., 2025; Hao et al., 2023). When the goal is to learn the mapping from PDE conditions to the corresponding solutions, neural operators provide a natural framework for learning solution operators between function spaces. Following early work such as the Graph Kernel Network (Li et al., 2020), this idea was formalized as a general framework for learning mappings between infinite-dimensional function spaces (Kovachki et al., 2023), and architectures such as Fourier neural operators (FNOs) (Li et al., 2021) and DeepONet (Lu et al., 2021) are now among the most widely used approaches for operator learning.

FNOs apply global convolution-type operations in Fourier space, and their Fourier-domain parameterization can be fixed independently of the spatial resolution of the data. This makes FNOs attractive in scientific computing, where high-resolution simulations are expensive but high-resolution predictions are desired: the same parameters can be used across discretizations, and an FNO trained on a coarse grid is often expected to retain its performance on a finer one. This expectation has been one of the main practical motivations for using FNOs, and they have been applied to tasks including global weather forecasting (Kurth

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.

et al., 2023), biomedical shape optimization (Zhou et al., 2024), and homogenization-based constitutive modeling (Bhattacharya et al., 2024); see Azizzadenesheli et al. (2024) for a broader overview.

Their theoretical properties have also been studied. Kovachki et al. (2021) showed that FNOs are universal approximators of continuous operators between infinite-dimensional function spaces and obtained quantitative approximation error estimates for representative PDE solution operators such as those of the stationary Darcy flow and the Navier–Stokes equations. These results concern the expressive power of FNOs at the continuous level, and do not address how a fixed FNO behaves when the same input function is represented on different discretizations.

1.2 Motivation

FNOs are widely expected to behave consistently across spatial resolutions, since their parameters are decoupled from the discretization of the data. In practice, a model trained on data at one resolution is often evaluated on another, and such changes in discretization introduce errors that are not captured by the continuous-level operator approximation error alone. From a structural viewpoint, Bartolucci et al. (2023) formalized this through Representation equivalent Neural Operators (ReNOs), based on continuous-discrete equivalence, and showed that standard FNOs may fail to be ReNOs because pointwise nonlinear activations destroy band-limitedness and introduce aliasing errors. Changing the discretization of the same input function can therefore change the output of an FNO even when the model parameters are fixed.

Two lines of theoretical work quantify this phenomenon. Lanthaler et al. (2024) studied the gap between a continuous FNO and its discretized implementation, and related this error to aliasing, Sobolev regularity, and the regularity-preserving properties of nonlinear activation functions. Gao et al. (2025) instead compared two discrete implementations directly. They questioned the common interpretation that discretization-invariant neural operators automatically maintain consistent performance across discretizations, formalized the resulting discrepancy as the discretization mismatch error (DME), derived an upper bound for it, and proposed a cross-resolution operator-learning pipeline (CROP) designed to avoid aliasing and discretization mismatch errors.

These concerns are also supported empirically. The zero-shot super-resolution ability reported for FNOs (Li et al., 2021, Section 5.4) has been questioned by Sakarvadia et al. (2026), who argued that the ability of machine-learned operators to be evaluated on arbitrary grids does not imply accurate super-resolution, and showed empirically that such models can be brittle under resolution changes due to aliasing. From the perspective of nonlinear aliasing, Colagrande et al. (2026) further reported that direct fine-grid inference can be worse than a baseline that performs inference on the coarse grid and subsequently upsamples the output.

Taken together, these studies have identified structural inconsistencies, analyzed the error between continuous models and their discrete approximations, demonstrated the limitations of cross-resolution inference empirically, and proposed architectures designed to mitigate these issues. The direct comparison of two discretizations in Gao et al. (2025) measures the DME in a whole-domain function-space norm, and their bound contains a term that may grow with the finer discretization size M. Because this whole-domain norm does not isolate the discrepancy at the grid points shared by the two discretizations, it is unclear whether this M-dependent growth reflects an actual increase in the discrepancy at these common grid points.

> *Does refinement of the finer discretization amplify the discrepancy between FNO outputs at the common grid points?*

We therefore ask how the outputs of a fixed FNO differ at the common grid points of two discretizations of sizes N and M. Such a comparison measures only the discrepancy between the two outputs where the discretizations agree and is unaffected by how the two reconstructions differ in between.

1.3 Contribution

This paper makes the following contributions.

1.3.1 Explicit distinction between continuous and discrete operators

We explicitly distinguish continuous operators from their discrete counterparts in the FNO formulation in Section 2.1. In existing analyses (Gao et al., 2025; Kovachki et al., 2023; Lanthaler et al., 2024), the lifting and projection operators are often written in the same form as their continuous counterparts, without explicitly distinguishing the continuous operators $\mathcal{P}$ and $\mathcal{Q}$ from their discrete implementations $\mathcal{P}^N$ and $\mathcal{Q}^N$. This notation may reflect the practical implementation of an FNO, but it can obscure the dependence of the operators on the spatial resolution. In contrast, we rigorously distinguish these operators and define $\mathcal{P}^N$ and $\mathcal{Q}^N$ separately from their continuous counterparts in Section 2.2. This formulation makes the dependence on the discretization explicit and provides a precise mathematical basis for analyzing the output discrepancy between different discretizations.

1.3.2 Pointwise DME

We introduce the pointwise discretization mismatch error (DME) to quantify the discrepancy between the outputs of an FNO evaluated on two different discretizations of the same input function. We then derive an explicit upper bound for the pointwise DME of a single-layer FNO model in Theorem 3.5. For the same input function a and the same parameter $\boldsymbol{\theta}$, we compare the outputs obtained from two different discretizations of a, of sizes $N = 2^n$ and $M = 2^{n+r}$, at their common grid points (see Table 2 about $\Psi_{\boldsymbol{\theta}}^N$). We show that the pointwise DME satisfies

$$\left|\left(\Psi_{\boldsymbol{\theta}}^M(a)\right)(x) - \left(\Psi_{\boldsymbol{\theta}}^N(a)\right)(x)\right| = O\left(\left(1 - \frac{1}{2^r}\right)\frac{1}{2^n}\right) \quad (x \in \Lambda_N^1).$$

This estimate provides an explicit characterization of the dependence of the output discrepancy on the two discretization sizes. In particular, it shows that the upper bound on the pointwise DME is of order $O(1/2^n)$ as the base discretization is refined.

1.3.3 Comparison with existing DME bounds

We show that the refinement-dependent growth term appearing in the existing DME bound does not appear in our pointwise estimate. Gao et al. (2025) define the DME as the $L^2(\mathbb{T}^d)$-norm of the difference between the outputs of an FNO evaluated on two different discretizations of the same input function. They consider a uniform discrete refinement $(\Omega_J)_{J=1}^{\infty}$ of the domain $\Omega = \mathbb{T}^d$, consisting of nested equidistant grids $\Omega_1 \subset \Omega_2 \subset \cdots \subset \Omega$. For a function $a : \Omega \to \mathbb{R}^m$, $a|_{\Omega_J}$ denotes its pointwise discretization on Ω_J. Here, $\mathcal{G}_\theta$ denotes the discrete FNO with parameters θ. The DME is defined by

$$E_{N,M}^{\text{Gao}} := \left\|\mathcal{G}_\theta\left(a|_{\Omega_M}\right) - \mathcal{G}_\theta\left(a|_{\Omega_N}\right)\right\|_{\mathcal{U}}.$$

Then, the structure of their upper bound can be schematically represented as

$$E_{N,M}^{\text{Gao}} \leq \underbrace{A_N}_{\substack{A_N \to 0 \\ \text{as } N \to \infty}} + \underbrace{B_M}_{\text{may grow with } M}.$$

The term B_M originates from the way the two discretizations are compared. In their analysis, the discrepancy between the two discretized inputs is estimated through the difference of the corresponding discrete Fourier coefficients, which contains a sum over the sample points that are present only in the finer discretization. The number of terms in this sum grows with M, and since the $L^2(\mathbb{T}^d)$-norm aggregates the discrepancy over the whole domain $\mathbb{T}^d$, this contribution is retained in the resulting bound.

Table 1 summarizes the two settings. *In contrast,* we define the pointwise DME at a common grid point $x \in \Lambda_N^1$ by

$$E_{N,M}^{\text{ours}}(x) := \left|\left(\Psi_\theta^M(a)\right)(x) - \left(\Psi_\theta^N(a)\right)(x)\right|,$$

Table 1: Comparison between the DME bound of Gao et al. (2025) and our pointwise estimate.

	Gao et al. (2025)	Ours
Quantity compared	$L^2(\mathbb{T}^d)$-norm over $\mathbb{T}^d$	pointwise value at $x \in \Lambda_N^1$
Discretizations	nested equidistant grids	nested dyadic grids, $N = 2^n$, $M = 2^{n+r}$
Fourier layers	L	single
Upper bound	$A_N + B_M$	$C\left(1 - 2^{-r}\right) 2^{-n}$
Dependence on M	B_M may grow with M	bounded uniformly in r

where the two discretizations have sizes N and M. Our estimate takes the form,

$$E_{N,M}^{\text{ours}}(x) \leq C\left(1 - \frac{1}{2^r}\right)\frac{1}{2^n} \leq \frac{C}{N},$$

where $C > 0$ is independent of N and M.

Here, the counterpart of B_M does not appear. Since the two outputs are compared only at the common grid points, the sample points that exist only in the finer discretization never enter the estimate as a separate sum whose length grows with M. What remains is the factor $1 - 2^{-r}$, which is bounded by one for every refinement level r. Refining the discretization does change the outputs at the common grid points, but the resulting bound is controlled by the base discretization size alone and remains uniform as the finer discretization is refined. Moreover, this uniform bound converges to zero as $N \to \infty$.

2 Preliminaries

The notation used in this section is summarized in Table 2.

Table 2: Notation list for the architecture of the discretized FNO

Notation	Description
Discretized FNO architecture	
d_a	The input channel dimension
d_u	The output channel dimension
d_v	The hidden channel dimension
$\mathcal{A}$	The input function space $\mathcal{A} := H^s(\mathbb{T}^d; \mathbb{R}^{d_a})$
G	The target operator $G : \mathcal{A} \to H^s(\mathbb{T}^d; \mathbb{R}^{d_u})$
Λ_N^d	The uniform grid $\{0, 1/N, \ldots, (N-1)/N\}^d$ on $[0,1)^d$
$\Psi_{\boldsymbol{\theta}}^N$	The FNO with $\boldsymbol{\theta}$ on a discretization of size N, $\Psi_{\boldsymbol{\theta}}^N : \mathcal{A} \to (\mathbb{R}^{d_u})^{\Lambda_N^d}$
$\mathcal{P}^N$	The lifting operator $\mathcal{P}^N : \mathcal{A} \to (\mathbb{R}^{d_v})^{\Lambda_N^d}$
$\mathcal{Q}^N$	The projection operator $\mathcal{Q}^N : (\mathbb{R}^{d_v})^{\Lambda_N^d} \to (\mathbb{R}^{d_u})^{\Lambda_N^d}$
L_t^N	The t-th Fourier layer $\mathsf{L}_t^N : (\mathbb{R}^{d_v})^{\Lambda_N^d} \to (\mathbb{R}^{d_v})^{\Lambda_N^d}$
$\mathcal{K}_t^N$	The integral operator in the t-th Fourier layer
σ	The activation function $\sigma : \mathbb{R}^{d_v} \to \mathbb{R}^{d_v}$
K	The truncated frequency mode set, $K = \{\boldsymbol{k} \in \mathbb{Z}^d : \lvert k_j\rvert \leq k_{\max}\}$

Notation	Description
Learnable parameters	
P	The lifting matrix, $P \in \mathbb{R}^{d_v \times d_a}$
Q	The projection matrix, $Q \in \mathbb{R}^{d_u \times d_v}$
W_t	The local linear transformation in the t-th Fourier layer, $W_t \in \mathbb{R}^{d_v \times d_v}$
$\boldsymbol{b}_t$	The bias vector in the t-th Fourier layer, $\boldsymbol{b}_t \in \mathbb{R}^{d_v}$
$R_t^{\boldsymbol{k}}$	The weight matrix for frequency mode $\boldsymbol{k}$ in the t-th Fourier layer, $R_t^{\boldsymbol{k}} \in \mathbb{C}^{d_v \times d_v}$
$\boldsymbol{\theta}$	The vector of all learnable parameters
$d_{\boldsymbol{\theta}}$	The total number of real-valued learnable parameters
$\boldsymbol{p}$	The lifting vector, $\boldsymbol{p} \in \mathbb{R}^{d_v}$
$\boldsymbol{q}^\top$	The projection row vector, $\boldsymbol{q}^\top \in \mathbb{R}^{1 \times d_v}$
N	The size of the base discretization, $N = 2^n$
M	The size of the refined discretization, $M = 2^{n+r}$

2.1 Architecture of the discretized FNO

Let $d_a, d_u, d_v \in \mathbb{N}$. Here, d_a denotes the input channel dimension, d_u denotes the output channel dimension, and d_v denotes the hidden channel dimension of the FNO. We consider the approximation of an operator $G : H^s(\mathbb{T}^d; \mathbb{R}^{d_a}) \to H^s(\mathbb{T}^d; \mathbb{R}^{d_u}) : a \mapsto G(a)$. We write $\mathcal{A} := H^s(\mathbb{T}^d; \mathbb{R}^{d_a})$. An element $a \in \mathcal{A}$ is regarded as an input function $a : \mathbb{T}^d \to \mathbb{R}^{d_a}$.

Then, let

$$\Lambda_N^d := \left\{0, \frac{1}{N}, \dots, \frac{N-1}{N}\right\}^d \subset [0,1)^d,$$

denote the uniform grid on $[0,1]$ with mesh size $1/N$. For each discretization size $N \in \mathbb{N}$, we denote the FNO associated with the parameter $\boldsymbol{\theta}$ and evaluated on a discretization of size N by $\Psi_{\boldsymbol{\theta}}^N : \mathcal{A} \to (\mathbb{R}^{d_u})^{\Lambda_N^d}$. More precisely, $\Psi_{\boldsymbol{\theta}}^N$ is written as the composition,

$$\Psi_{\boldsymbol{\theta}}^N = \mathcal{Q}^N \circ \mathsf{L}_{T-1}^N \circ \cdots \circ \mathsf{L}_0^N \circ \mathcal{P}^N.$$

Here, $\mathcal{P}^N : \mathcal{A} \to (\mathbb{R}^{d_v})^{\Lambda_N^d} : a \mapsto \mathcal{P}^N a$ denotes the lifting operator, $\mathcal{Q}^N : (\mathbb{R}^{d_v})^{\Lambda_N^d} \to (\mathbb{R}^{d_u})^{\Lambda_N^d} : u \mapsto \mathcal{Q}^N u$ denotes the projection operator and $\mathsf{L}_t^N : (\mathbb{R}^{d_v})^{\Lambda_N^d} \to (\mathbb{R}^{d_v})^{\Lambda_N^d} : u \mapsto \mathsf{L}_t^N u$ denotes the t-th Fourier layer. Thus, an input function $a \in \mathcal{A}$ is first lifted by $\mathcal{P}^N$, then updated successively by the Fourier layers $\mathsf{L}_0^N, \dots, \mathsf{L}_{T-1}^N$, and finally mapped to the output space $(\mathbb{R}^{d_u})^{\Lambda_N^d}$ by $\mathcal{Q}^N$.

For $a \in \mathcal{A}$, the lifting operator $\mathcal{P}^N$ is defined as:

$$\mathcal{P}^N a(\boldsymbol{x}) := Pa(\boldsymbol{x}),$$

where $\boldsymbol{x} \in \Lambda_N^d$, and $P \in \mathbb{R}^{d_v \times d_a}$ is the learnable parameter. For $u \in (\mathbb{R}^{d_v})^{\Lambda_N^d}$, the projection operator $\mathcal{Q}^N$ is defined as:

$$\mathcal{Q}^N u(\boldsymbol{x}) := Qu(\boldsymbol{x}),$$

where $\boldsymbol{x} \in \Lambda_N^d$, and $Q \in \mathbb{R}^{d_u \times d_v}$ is the learnable parameter. For $u \in (\mathbb{R}^{d_v})^{\Lambda_N^d}$, the t-th Fourier layer L_t^N is defined as:

$$\mathsf{L}_t^N u(\boldsymbol{x}) := \sigma\left(W_t u(\boldsymbol{x}) + \mathcal{K}_t^N u(\boldsymbol{x}) + \boldsymbol{b}_t\right),$$

where $\boldsymbol{x} \in \Lambda_N^d$, and $W_t \in \mathbb{R}^{d_v \times d_v}$ and $b_t \in \mathbb{R}^{d_v}$ are the learnable parameters, and $\sigma : \mathbb{R}^{d_v} \to \mathbb{R}^{d_v}$ is the activation function. $\mathcal{K}_t^N : (\mathbb{R}^{d_v})^{\Lambda_N^d} \to (\mathbb{R}^{d_v})^{\Lambda_N^d} : u \mapsto \mathcal{K}_t^N u$ is the discretized case of the Fourier integral operator (Li et al., 2021, Definition 3). The operator $\mathcal{K}_t^N$ is defined as:

$$\mathcal{K}_t^N u(\boldsymbol{x}) := \sum_{\boldsymbol{k} \in K} R_t^{\boldsymbol{k}} \mathsf{DFT}(u)(\boldsymbol{k}) \exp\left(2\pi i \boldsymbol{k}^\top \boldsymbol{x}\right),$$

where $\boldsymbol{x} \in \Lambda_N^d$, and the set of the truncated frequency mode K is given by:

$$K := \left\{\boldsymbol{k} = (k_1, \dots, k_d)^\top \in \mathbb{Z}^d \;\middle|\; \forall j \in \{1, \dots, d\},\, |k_j| \le k_{\max}\right\}.$$

Here, $k_{\max} \in \mathbb{N}$ is the maximum number of modes. The matrices $R_t^{\boldsymbol{k}} \in \mathbb{C}^{d_v \times d_v}$ denote the learnable complex weight matrices for each frequency mode $\boldsymbol{k}$. To guarantee that $\mathcal{K}_t^N u \in (\mathbb{R}^{d_v})^{\Lambda_N^d}$ holds for any real-valued input $u \in (\mathbb{R}^{d_v})^{\Lambda_N^d}$, the weight matrices must satisfy the following conjugate symmetry condition:

$$R_t^{-\boldsymbol{k}} = \left(R_t^{\boldsymbol{k}}\right)^*,$$

for all $\boldsymbol{k} \in K$ (see Kovachki et al. (2021, Section 2.3)).

2.2 Parameterization

The learnable parameters are collected in a vector $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$. Here, $d_{\boldsymbol{\theta}} \in \mathbb{N}$ denotes the total number of real-valued learnable parameters. More explicitly, we write

$$\boldsymbol{\theta} = \left(\boldsymbol{p}, \boldsymbol{q}, \{W_t, \boldsymbol{b}_t, \{R_t^{\boldsymbol{k}}\}_{\boldsymbol{k} \in K}\}_{t=0}^{T-1}\right).$$

For each layer $t \in \{0, \dots, T-1\}$, W_t denotes the local linear transformation, b_t the bias term, and $R_t^{\boldsymbol{k}}$ the learnable Fourier weights used in the truncated Fourier integral operator.

Complex-valued Fourier weights are identified with their real and imaginary parts. The architecture itself, including the number of layers T, the channel widths, and the Fourier mode truncation, is fixed independently of the discretization size N. Therefore, the parameter dimension $d_{\boldsymbol{\theta}}$ does not depend on N. On the other hand, a minimizer of the loss may depend on N.

2.3 Pointwise discretization mismatch error

We now introduce the pointwise discrepancy between the outputs of an FNO evaluated on two different discretizations of the same input function. Let $N = 2^n$ and $M = 2^{n+r}$, and let $a \in \mathcal{A}$ and $\boldsymbol{\theta} \in \Theta$. Since the two discretizations are nested, we compare the outputs at the common grid points $x \in \Lambda_N^1$.

The pointwise discretization mismatch error (pointwise DME) is defined by

$$\left| \left(\Psi_{\boldsymbol{\theta}}^M(a)\right)(x) - \left(\Psi_{\boldsymbol{\theta}}^N(a)\right)(x) \right|,$$

for all $a \in \mathcal{A}$, $x \in \Lambda_N^1$ and $\boldsymbol{\theta} \in \Theta$.

3 Main results

In this section, we consider a single-layer FNO model instead of the general multi-layer FNO architecture. This simplification allows us to streamline the notation and present the main ideas of the paper more transparently. Similar simplifications have also been adopted in Kovachki et al. (2021, Appendix D). We consider the case $d = d_a = d_u = 1$, $d_v \geq 1$, and $t = 1$. More precisely, we consider an FNO of the form,

$$\Psi_{\boldsymbol{\theta}}^N := \mathcal{Q}^N \circ \mathsf{L}_0^N \circ \mathcal{P}^N,$$

where $N \in \mathbb{N}$ denotes the discretization size. Throughout this section, we restrict the discretization size to $N = 2^n$. This is a standard assumption when employing the fast Fourier transform (FFT). In this setting, the lifting matrix $P \in \mathbb{R}^{d_v \times d_a}$ is identified with a vector $\boldsymbol{p} \in \mathbb{R}^{d_v}$, and the projection matrix $Q \in \mathbb{R}^{d_u \times d_v}$ is identified with a row vector $\boldsymbol{q}^\top \in \mathbb{R}^{1 \times d_v}$. Therefore, for $\boldsymbol{x} \in \Lambda_N^1$, we write $\mathcal{P}^N a(\boldsymbol{x}) = \boldsymbol{p}\, a(\boldsymbol{x})$ and $\mathcal{Q}^N u(\boldsymbol{x}) = \boldsymbol{q}^\top u(\boldsymbol{x})$. Let $n, r \in \mathbb{N}$, and set $N = 2^n$ and $M = 2^{n+r}$. Since $M = 2^r N$, for all $\boldsymbol{x} = j/2^n \in \Lambda_N^1$, where $j \in \{0, 1, \dots, 2^n - 1\}$, we have

$$\boldsymbol{x} = \frac{j}{2^n} = \frac{2^r \cdot j}{2^r \cdot 2^n} = \frac{2^r j}{2^{n+r}} \in \Lambda_M^1.$$

Therefore, $\Lambda_N^1 \subset \Lambda_M^1$. Hence, for all $\boldsymbol{x} \in \Lambda_N^1$, both $(\Psi_{\boldsymbol{\theta}}^N(a))(\boldsymbol{x})$ and $(\Psi_{\boldsymbol{\theta}}^M(a))(\boldsymbol{x})$ are well defined, and we can compare the two outputs pointwise on the common grid Λ_N^1.

We make the following Assumption 3.1 (A1)-(A3).

Assumption 3.1.

(A1) *Assume that $s > 3/2$. This implies that the Sobolev space $H^s(\mathbb{T}; \mathbb{R})$ can considered as a subspace of $C^1(\mathbb{T})$ from Theorem A.1.*

(A2) *The activation function $\sigma : \mathbb{R}^{d_v} \to \mathbb{R}^{d_v}$ is Lipschitz continuous. Namely, there exists a constant $L_\sigma > 0$ such that*

$$\|\sigma(\boldsymbol{x}) - \sigma(\boldsymbol{y})\|_2 \leq L_\sigma \|\boldsymbol{x} - \boldsymbol{y}\|_2,$$

for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{d_v}$.

(A3) *There exists a constant $D > 0$ such that $\|\boldsymbol{\theta}\|_2 \leq D$ for all $\boldsymbol{\theta} \in \Theta$.*

Remark 3.2. *Assumption 3.1 (A1) motivates the use of smooth activation functions, since $s > 3/2$ implies $H^s(\mathbb{T};\mathbb{R}) \hookrightarrow C^1(\mathbb{T})$. Neural operators typically consist of linear integral operator blocks followed by pointwise nonlinear activations, where smooth activations such as Gaussian Error Linear Units (GeLU) (Li et al., 2020) are commonly used, while ReLU is usually avoided due to its non-smoothness (Nair & Hinton, 2010).*

Assumption 3.1 (A2) is satisfied by standard smooth activation functions such as the sigmoid activation function and the GeLU activation function proposed by Hendrycks & Gimpel (2023). Verifications are deferred to Appendix B.1.

Moreover, the use of GeLU is consistent with the FNO discretization error analysis in Lanthaler et al. (2024, subsection 4.1), where smooth regularity-preserving activation functions are used to propagate Sobolev regularity through the FNO layers.

Assumption 3.1 (A3) allows us to obtain estimates that are uniform with respect to $\boldsymbol{\theta}$, as the constants in our estimates can be bounded independently of the particular choice of $\boldsymbol{\theta}$. Similar bounded-weight network classes have been considered in theoretical analyses of neural networks, where bounds on weight magnitudes are used to control the complexity of the hypothesis class (Bartlett, 1996).

To establish the discretization-dependent estimates below, we first derive uniform bounds on the individual parameters of the FNO from Assumption 3.1 (A3). The following lemma collects these bounds. Proofs are deferred to Appendix B.2.

Lemma 3.3. *Suppose that Assumption 3.1 (A3) holds. Then,*

$$\|\boldsymbol{p}\|_2 \le D, \quad \|\boldsymbol{q}\|_2 \le D,$$

for all $\boldsymbol{\theta} \in \Theta$, and,

$$\|W_t\|_{\mathrm{F}} \le D, \quad \|\boldsymbol{b}_t\|_2 \le D, \quad \sum_{k\in K} \left\|R_t^k\right\|_{\mathrm{F}} \le (2k_{\max}+1)D,$$

for all $t = 0, \dots, T-1$.

With these uniform bounds, we next estimate the difference between the discretized Fourier coefficients on the two grids. This is the key discretization-dependent quantity in the single-layer model, since $\boldsymbol{v}_0^M(x) = \boldsymbol{v}_0^N(x)$ on the common grid Λ_N^1, while the Fourier coefficients use different grid points. Proofs are deferred to Appendix B.3

Lemma 3.4. *Suppose that Assumption 3.1 (A1) holds. Let $n, r \in \mathbb{N}$, and set $N = 2^n$ and $M = 2^{n+r}$. We define functions $f_k : [0,1] \to \mathbb{C}$ as $f_k(x) := a(x)\exp(-2\pi i k x)$, where $x \in [0,1]$, $a \in \mathcal{A}$, $k \in K$ and $h = 1, \dots, N_d$. Moreover, let $C_f(a) \ge 0$ be a constant defined by*

$$C_f(a) := \max_{k\in K} \max_{x\in[0,1]} |f_k'(x)|.$$

Let $\boldsymbol{v}_0^M := \mathcal{P}^M(a)$ and $\boldsymbol{v}_0^N := \mathcal{P}^N(a)$. Then,

$$\left\|\mathsf{DFT}(\boldsymbol{v}_0^M)(k) - \mathsf{DFT}(\boldsymbol{v}_0^N)(k)\right\|_2 \le \frac{C_f(a)D}{2} \cdot \left(1 - \frac{1}{2^r}\right)\frac{1}{2^n},$$

for all $a \in \mathcal{A}$ and $k \in K$.

The next theorem transfers the Fourier coefficient estimate to the output of the FNO. The key point is that, on the common grid Λ_N^1, the lifting terms agree exactly. Therefore, the difference between the two outputs comes only from the Fourier part of the layer. The Lipschitz continuity of the activation function in Assumption 3.1 (A2) is used to propagate this difference through the nonlinear activation. Proofs are deferred to Appendix B.4

Theorem 3.5. *Suppose that Assumptions 3.1 (A1), (A2) and (A3) hold. Let $n, r \in \mathbb{N}$, and set $N = 2^n$ and $M = 2^{n+r}$. Let $C_f(a) \ge 0$ be a constant defined in Lemma 3.4. Then,*

$$\left|\left(\Psi_{\boldsymbol{\theta}}^M(a)\right)(x) - \left(\Psi_{\boldsymbol{\theta}}^N(a)\right)(x)\right| \le \frac{L_\sigma C_f(a) D^3 (2k_{max}+1)}{2} \cdot \left(1 - \frac{1}{2^r}\right)\frac{1}{2^n},$$

for all $a \in \mathcal{A}$, $x \in \Lambda_N^1$ and $\boldsymbol{\theta} \in \Theta$.

4 Numerical Experiments

We numerically evaluate the quantity analyzed theoretically in Theorem 3.5. Specifically, we consider the pointwise DME of the two outputs at their common grid points,

$$\left|(\Psi_{\boldsymbol{\theta}}^{M}(a_h))(x) - (\Psi_{\boldsymbol{\theta}}^{N}(a_h))(x)\right|, \tag{1}$$

where $h = 1, 2$. Here, $a_1, a_2 : \mathbb{T} = [0, 1] \to \mathbb{R}$ are defined by $a_1 : \mathbb{T} = [0, 1] \to \mathbb{R}$ is defined by

$$a_1(x) := \sin(2\pi x) + 2\cos(4\pi x), \qquad a_2(x) := \begin{cases} 1, & \text{if } 0 \le x \le \dfrac{1}{2}, \\ 0, & \text{if } \dfrac{1}{2} < x \le 1. \end{cases}$$

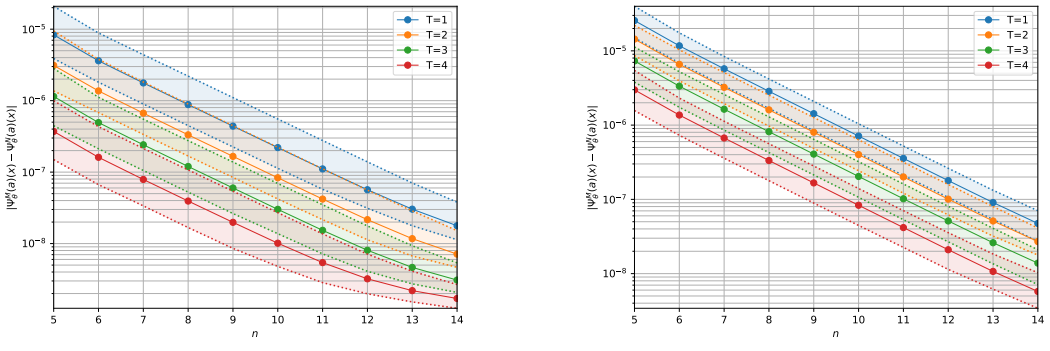
The parameters used throughout the numerical experiments are listed in Table 3.

Table 3: Parameters used in the numerical experiments.

Parameter	Description	Figure 1	Figure 2
n	Base discretization exponent	$5, \ldots, 14$	6
r	Refinement level	1	$1, \ldots, 8$
T	Number of Fourier layers	$1, \ldots, 4$	$1, \ldots, 4$
$k_{\max}$	Maximum number of Fourier modes	16	16
d_a	Input channel dimension	1	1
d_u	Output channel dimension	1	1
d_v	Hidden channel dimension	64	64

Let $N = 2^n$ and $M = 2^{n+r}$. We plot the quantity equation 1 as a function of n with r fixed, where $n = 5, 6, \ldots, 14$, and as a function of r with n fixed, where $r = 1, 2, \ldots, 8$. In each case the quantity is averaged over all N points in Λ_N^1. For the parameter $\boldsymbol{\theta}$, we performed the experiment by using 100 randomly generated values. The thick line plots the mean over the 100 trials, while the maximum and minimum values are shown as dotted lines. The area bounded by the maximum and minimum values is painted the same color as the corresponding line. The experiments were repeated for $T = 1, 2, 3$, and 4, where T denotes the number of Fourier layers in the discretized FNO. The experiments were conducted on a MacBook Pro equipped with an Apple M5 chip and 16 GB of unified memory, running the macOS Tahoe version 26.5.2 operating system. The algorithms were written in Python 3.14.2 with the PyTorch 2.10.0 package and the Matplotlib 3.10.8 package. The implementation of the FNO used in the numerical experiments is based on the implementation provided in Li et al. (2021), available at `https://github.com/neuraloperator/neuraloperator`.

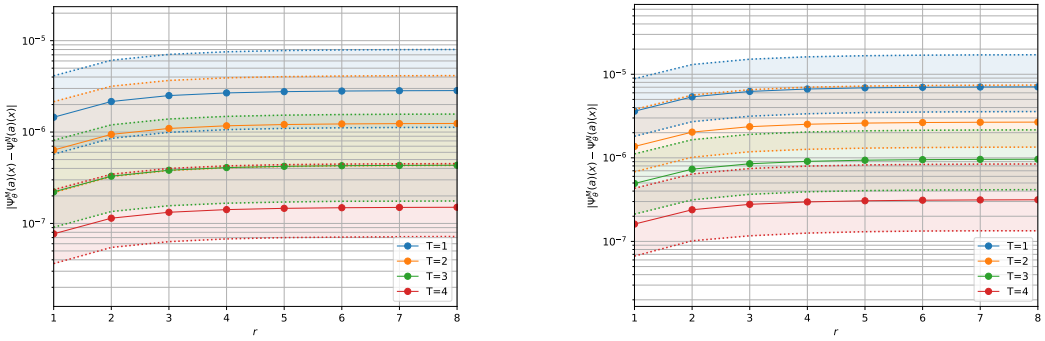
The numerical results for the input functions a_1 and a_2 are presented in Figure 1 and Figure 2.



(a) Input function a_1.

(b) Input function a_2.

Figure 1: Pointwise DME for the input functions a_1 and a_2 with $r = 1$ fixed and increasing n. Thick lines show the mean over 100 trials; the shaded region spans the maximum and minimum. T is the number of Fourier layers.



(a) Input function a_1.

(b) Input function a_2.

Figure 2: Pointwise DME for the input functions a_1 and a_2 with $n = 6$ fixed and varying r. Thick lines show the mean over 100 trials; the shaded region spans the maximum and minimum. T is the number of Fourier layers.

For both input functions, the discrepancy decreases as the base discretization size $N = 2^n$ increases, which is consistent with the theoretical estimate in Theorem 3.5,

$$\left|(\Psi_{\boldsymbol{\theta}}^{M}(a))(x) - (\Psi_{\boldsymbol{\theta}}^{N}(a))(x)\right| = O\left(\left(1 - \frac{1}{2^r}\right) \cdot \frac{1}{2^n}\right).$$

Since the refinement level r was fixed in the experiments, the estimate reduces to an $O(2^{-n})$ convergence rate with respect to N, which agrees well with the overall trends observed in both figures.

Figures 2a and 2b further examine the dependence on the refinement level r by fixing $n = 6$. For both input functions, the discrepancy increases initially and then levels off as r grows. This behavior is consistent with the factor $1 - 1/2^r$ in Theorem 3.5, which remains bounded by 1 and approaches 1 as r increases. Hence, for a fixed base discretization, the discrepancy remains uniformly bounded even as the finer discretization size $M = 2^{n+r}$ becomes arbitrarily large.

Finally, we note that the two input functions cover different regularity regimes. The smooth input a_1 satisfies Assumption 3.1 (A1) (see Appendix B.5), so the theoretical bound applies to this case. In contrast, the discontinuous input a_2 does not satisfy this assumption and therefore lies outside the regularity regime covered by our analysis. Nevertheless, Figures 1 and 2 exhibit the same qualitative behavior for a_2 as for a_1. This suggests that the discretization-mismatch behavior characterized by our bound may persist beyond the Sobolev regularity assumed in the theory, although establishing such an extension is beyond the scope of this work.

5 Conclusion

In this paper, we provided a quantitative analysis of discretization mismatch in a single-layer FNO. By directly comparing the outputs obtained from two discretizations of the same input function, we showed that the pointwise DME remains uniformly bounded with respect to the refinement level and does not diverge even when the finer discretization becomes arbitrarily large. This complements the analysis of Gao et al. (2025) by giving an explicit pointwise characterization. We also confirmed numerically that the pointwise DME behaves as predicted by the bound, both as the base discretization is refined and as the refinement level increases. These results indicate that the stability of an FNO across discretizations is governed by the resolution used as the base, rather than by how far apart the two discretizations are.

AI use statement

In this work, we used generative AI tools for research ideation. We have not used generative AI tools for writing code, deriving or proving theoretical results, or conducting the numerical experiments, and retrieval of literature and generation of figures or data are not applicable to this work. Additionally, we used generative AI tools for polishing the phrasing and grammar of the manuscript. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey, and the AI-assisted text was reviewed by the authors to confirm that no technical claim was introduced or altered. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

This work is a theoretical analysis of Fourier neural operators, accompanied by numerical experiments on synthetic input functions defined in closed form. It does not involve human subjects, personal data, or the release of new datasets, and we are not aware of any potential for harmful application arising from the results. The authors declare no conflicts of interest.

Reproducibility Statement

The notation for the FNO architecture and mathematical preliminaries is summarized in Tables 2 and 4, respectively. Further mathematical preliminaries are provided in Appendix A. Complete proofs of the theoretical results are provided in Appendix B. The numerical experiments are described in Section 4.

References

Robert A. Adams and John J. F. Fournier. *Sobolev Spaces*, volume 140. Elsevier, 2003.

Kamyar Azizzadenesheli, Nikola Kovachki, Zongyi Li, Miguel Liu-Schiaffini, Jean Kossaifi, and Anima Anandkumar. Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics*, 6(5):320–328, 2024.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems*, 1996.

Francesca Bartolucci, Emmanuel de Bezenac, Bogdan Raonic, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. In *Advances in Neural Information Processing Systems*, 2023.

Kaushik Bhattacharya, Nikola B Kovachki, Aakila Rajan, Andrew M Stuart, and Margaret Trautner. Learning homogenization for elliptic operators. *SIAM Journal on Numerical Analysis*, 62(4):1844–1873, 2024.

Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2011.

Alex Colagrande, Paul Caillon, Eva Feillet, and Alexandre Allauzen. Limits of resolution equivariance in Fourier neural operators. In *ICLR 2026 Workshop on AI and Partial Differential Equations*, 2026.

Lawrence C Evans. *Partial Differential Equations*, volume 19. American Mathematical Society, 2010.

Wenhan Gao, Ruichen Xu, Yuefan Deng, and Yi Liu. Discretization-invariance? On the discretization mismatch errors in neural operators. In *International Conference on Learning Representations*, pp. 19360–19384, 2025.

Loukas Grafakos. *Classical Fourier Analysis*. Springer, 2014.

Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. Physics-informed machine learning: A survey on problems, methods and applications. *arXiv preprint arXiv:2211.08064*, 2023.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GeLUs). *arXiv preprint arXiv:1606.08415*, 2023.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

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

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

Thorsten Kurth, Shashank Subramanian, Peter Harrington, Jaideep Pathak, Morteza Mardani, David Hall, Andrea Miele, Karthik Kashinath, and Anima Anandkumar. FourCastNet: Accelerating global high-resolution weather forecasting using adaptive Fourier neural operators. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, pp. 1–11, 2023.

Samuel Lanthaler, Andrew M Stuart, and Margaret Trautner. Discretization error of Fourier neural operators. *arXiv preprint arXiv:2405.02221*, 2024.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2020.

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

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.

Chuizheng Meng, Sam Griesemer, Defu Cao, Sungyong Seo, and Yan Liu. When physics meets machine learning: A survey of physics-informed machine learning. *Machine Learning for Computational Science and Engineering*, 1(1):20, 2025.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

Mansi Sakarvadia, Kareem Hegazy, Amin Totounferoush, Kyle Chard, Yaoqing Yang, Ian Foster, and Michael W Mahoney. The false promise of zero-shot super-resolution in machine-learned operators. In *International Conference on Learning Representations*, pp. 69635–69688, 2026.

Tingtao Zhou, Xuan Wan, Daniel Zhengyu Huang, Zongyi Li, Zhiwei Peng, Anima Anandkumar, John F Brady, Paul W Sternberg, and Chiara Daraio. Ai-aided geometric design of anti-infection catheters. *Science Advances*, 10(1):1741, 2024.

A Preliminaries

A.1 Mathematical Preliminaries

The notation used in this section is summarized in Table 4.

Table 4: Notation List for the Mathematical Preliminaries

Notation	Description
Basic notation	
$\mathbb{R}$	The set of real numbers
$\mathbb{C}$	The set of complex numbers
$\mathbb{N}$	The set of natural numbers
$\mathbb{Z}$	The set of integers
A^B	The set of all maps from B to A
$\boldsymbol{v}^\top$	The transpose of a real vector $\boldsymbol{v}$
$\boldsymbol{v}^*$	The conjugate transpose of a complex vector $\boldsymbol{v}$
M^*	The conjugate transpose of a complex matrix M
$\|\cdot\|_2$	The Euclidean norm
$\|\cdot\|_{\mathrm{F}}$	The Frobenius norm for matrices
Function spaces on the torus	
$\mathbb{T}^d$	The d-dimensional torus $\mathbb{R}^d/\mathbb{Z}^d$
π	The canonical quotient map $\pi : \mathbb{R}^d \to \mathbb{T}^d$
$\mathcal{B}(\mathbb{T}^d)$	The Borel σ-algebra
λ	The Haar measure, $\lambda : \mathcal{B}(\mathbb{T}^d) \to [0,1]$
m_d	The Lebesgue measure on $\mathbb{R}^d$
$L^2(\mathbb{T}^d;\mathbb{R}^m)$	The Hilbert space of square-integrable $\mathbb{R}^m$-valued functions
$H^s(\mathbb{T}^d;\mathbb{R}^m)$	The Sobolev space of order s
$\widehat{f}(\boldsymbol{k})$	The $\boldsymbol{k}$-th Fourier coefficient of f
Discretization and sampling	
$\|\cdot\|_{\ell^2(\Lambda_N^d;\mathbb{R}^m)}$	The discretized ℓ^2 norm on the grid Λ_N^d
$\mathsf{DFT}(f)(\boldsymbol{k})$	The multi-dimensional DFT of a grid function f at frequency $\boldsymbol{k}$
$S_{N,m}$	The sampling operator $S_{N,m} : H^s(\mathbb{T}^d;\mathbb{R}^m) \to (\mathbb{R}^m)^{\Lambda_N^d}$

We denote the set of real numbers by $\mathbb{R}$, the set of complex numbers by $\mathbb{C}$, the set of natural numbers by $\mathbb{N}$, and the set of integers by $\mathbb{Z}$. For sets A and B, we denote by A^B the set of all maps from B to A. For a real vector $\boldsymbol{v} \in \mathbb{R}^d$, we denote its transpose by $\boldsymbol{v}^\top$. For a complex vector $\boldsymbol{v} \in \mathbb{C}^d$ and a matrix $M \in \mathbb{C}^{d\times d}$, we denote their conjugate transposes by $\boldsymbol{v}^*$ and M^*, respectively. Moreover, let $\|\boldsymbol{v}\|_2$ denote the Euclidean norm of $\boldsymbol{v}$, i.e., $\|\boldsymbol{v}\|_2 := \sqrt{\boldsymbol{v}^*\boldsymbol{v}}$. For a complex matrix $M \in \mathbb{C}^{d\times d}$, let $\|M\|_{\mathrm{F}}$ denote the Frobenius norm of M. O stands for Landau's symbol; i.e., $\boldsymbol{y}_T = O(\boldsymbol{x}_T)$ implies that there exist $C > 0$ and $t_0 \in \mathbb{N}$ such that, for all $T \geq t_0$, $\boldsymbol{y}_T \leq C\boldsymbol{x}_T$.

Let $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$ be the d-dimensional torus, More precisely, $\mathbb{T}^d$ is the quotient of $\mathbb{R}^d$ by the equivalence relation,

$$\boldsymbol{x} \sim \boldsymbol{y} \quad \Longleftrightarrow \quad \boldsymbol{x} - \boldsymbol{y} \in \mathbb{Z}^d.$$

We denote by $\pi : \mathbb{R}^d \to \mathbb{T}^d : \boldsymbol{x} \mapsto \boldsymbol{x} + \mathbb{Z}^d$ the canonical quotient map. A function $f : \mathbb{T}^d \to \mathbb{R}^m$ naturally induces a function $\tilde{f} := f \circ \pi : \mathbb{R}^d \to \mathbb{R}^m$. Since $\pi(\boldsymbol{x} + \boldsymbol{k}) = \pi(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{k} \in \mathbb{Z}^d$, it follows that $\tilde{f}(\boldsymbol{x} + \boldsymbol{k}) = \tilde{f}(\boldsymbol{x})$. Thus, $\tilde{f}$ is a 1-periodic function in each coordinate on $\mathbb{R}^d$. Hence, there is a natural one-to-one correspondence between functions on $\mathbb{T}^d$ and 1-periodic functions in each coordinate on $\mathbb{R}^d$. Therefore, throughout this paper, we will freely identify functions on $\mathbb{T}^d$ with their 1-periodic representatives on $\mathbb{R}^d$ (see (Grafakos, 2014, Chapter 3)).

We equip $\mathbb{T}^d$ with the quotient topology induced by π and the corresponding Borel σ-algebra $\mathcal{B}(\mathbb{T}^d)$. Let m_d denote the Lebesgue measure on $\mathbb{R}^d$. The Haar measure $\lambda : \mathcal{B}(\mathbb{T}^d) \to [0,1]$

on $\mathbb{T}^d$ is defined by

$$\lambda(E) := m_d\left(\pi^{-1}(E) \cap [0,1)^d\right),$$

where $E \in \mathcal{B}(\mathbb{T}^d)$. Using the measure λ, the integral of a measurable function $f : \mathbb{T}^d \to \mathbb{R}^m$ is defined by

$$\int_{\mathbb{T}^d} f(\boldsymbol{x})\,\mathrm{d}\lambda(\boldsymbol{x}) = \int_{[0,1)^d} (f \circ \pi)(\boldsymbol{x})\,\mathrm{d}m_d(\boldsymbol{x}),$$

provided the integral on the right-hand side exists. For simplicity, we will often write $\mathrm{d}\boldsymbol{x}$ instead of $\mathrm{d}\lambda(\boldsymbol{x})$ or $\mathrm{d}m_d(\boldsymbol{x})$ in the integrals.

Furthermore, we define $L^2(\mathbb{T}^d;\mathbb{R}^m)$ as the Hilbert space of square-integrable $\mathbb{R}^m$-valued functions on $\mathbb{T}^d$ (see (Brezis, 2011, Chapter 4)). More precisely,

$$L^2(\mathbb{T}^d;\mathbb{R}^m) := \left\{ f\colon \mathbb{T}^d \to \mathbb{R}^m \;\middle|\; f \text{ is measurable and } \int_{\mathbb{T}^d} \|f(\boldsymbol{x})\|_2^2\,\mathrm{d}\boldsymbol{x} < \infty \right\} \Big/ \sim,$$

where $f \sim g$ iff $f = g$ almost everywhere. Through a slight abuse of notation, we write $f \in L^2(\mathbb{T}^d;\mathbb{R}^m)$ for both an equivalence class and a representative of that class. For $f, g \in L^2(\mathbb{T}^d;\mathbb{R}^m)$, we define the inner product by

$$\langle f, g\rangle_{L^2(\mathbb{T}^d;\mathbb{R}^m)} := \int_{\mathbb{T}^d} f(\boldsymbol{x})^\top g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

Moreover, the associated norm is then given by

$$\|f\|_{L^2(\mathbb{T}^d;\mathbb{R}^m)} := \left(\int_{\mathbb{T}^d} \|f(\boldsymbol{x})\|_2^2\,\mathrm{d}\boldsymbol{x}\right)^{\frac{1}{2}}.$$

For $s \geq 0$, we define the Sobolev space $H^s(\mathbb{T}^d;\mathbb{R}^m)$ as

$$H^s(\mathbb{T}^d;\mathbb{R}^m) := \left\{ f \in L^2(\mathbb{T}^d;\mathbb{R}^m) \;\middle|\; \sum_{\boldsymbol{k}\in\mathbb{Z}^d} (1 + \|\boldsymbol{k}\|_2^2)^s \left\|\widehat{f}(\boldsymbol{k})\right\|_2^2 < \infty \right\},$$

where $\widehat{f}(\boldsymbol{k}) \in \mathbb{C}^m$ denotes the $\boldsymbol{k}$-th Fourier coefficient of f, defined by

$$\widehat{f}(\boldsymbol{k}) := \int_{\mathbb{T}^d} f(\boldsymbol{x}) \exp(-2\pi i \boldsymbol{k}^\top \boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

See (Kovachki et al., 2021, Appendix B)) for more details. For $f, g \in H^s(\mathbb{T}^d;\mathbb{R}^m)$, we define the Sobolev inner product by

$$\langle f, g\rangle_{H^s(\mathbb{T}^d;\mathbb{R}^m)} := \sum_{\boldsymbol{k}\in\mathbb{Z}^d} (1 + \|\boldsymbol{k}\|_2^2)^s \widehat{g}(\boldsymbol{k})^* \widehat{f}(\boldsymbol{k}).$$

The associated Sobolev norm is then given by

$$\|f\|_{H^s(\mathbb{T}^d;\mathbb{R}^m)} := \left(\sum_{\boldsymbol{k}\in\mathbb{Z}^d} (1 + \|\boldsymbol{k}\|_2^2)^s \left\|\widehat{f}(\boldsymbol{k})\right\|_2^2\right)^{\frac{1}{2}}.$$

Equipped with this inner product, the Sobolev space $H^s(\mathbb{T}^d;\mathbb{R}^m)$ becomes a Hilbert space. For a grid function $f \in (\mathbb{R}^m)^{\Lambda_N^d}$, we define the discretized ℓ^2 norm on the grid Λ_N^d by

$$\|f\|_{\ell^2(\Lambda_N^d;\mathbb{R}^m)} := \left(\sum_{\boldsymbol{x}\in\Lambda_N^d} \|f(\boldsymbol{x})\|_2^2\right)^{\frac{1}{2}}.$$

For a vector-valued grid function $f \in (\mathbb{R}^m)^{\Lambda_N^d}$, the multi-dimensional discretized Fourier Transform (DFT) is defined as:

$$\mathsf{DFT}(f)(\boldsymbol{k}) := \frac{1}{N^d} \sum_{\boldsymbol{x}\in\Lambda_N^d} f(\boldsymbol{x}) \exp\left(-2\pi i \boldsymbol{k}^\top \boldsymbol{x}\right),$$

where $\boldsymbol{k} \in \mathbb{Z}^d$.

We define $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ and $n \in \mathbb{N}_0$. For a multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$, we denote the partial derivative by $\partial^{\boldsymbol{\alpha}} = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d}$. Let $C^n(\mathbb{T}^d)$ be the Banach space of n-times continuously differentiable functions on $\mathbb{T}^d$ equipped with the norm,

$$\|f\|_{C^n(\mathbb{T}^d)} := \sup \left\{ |\partial^{\boldsymbol{\alpha}} f(\boldsymbol{x})| \;\middle|\; \alpha_1 + \cdots + \alpha_d \le n,\, \boldsymbol{x} \in \mathbb{T}^d \right\}.$$

We recall the standard Sobolev embedding theorem Adams & Fournier (2003); Evans (2010); Kovachki et al. (2021) for the Sobolev spaces $H^s(\mathbb{T}^d; \mathbb{R})$, which guarantees that the embedding is continuous, and moreover compact.

Theorem A.1 (Sobolev embedding)**.** *Let $d \in \mathbb{N}$ and $n \in \mathbb{N}_0$. Assume that $s \in \mathbb{R}$ satisfies $s > n + d/2$. Then, there exists a continuous embedding $H^s(\mathbb{T}^d; \mathbb{R}) \hookrightarrow C^n(\mathbb{T}^d)$, which is compact. In particular, there exists a constant $C > 0$ depending only on d, s, and n such that*

$$\|f\|_{C^n(\mathbb{T}^d)} \le C \|f\|_{H^s(\mathbb{T}^d;\mathbb{R})},$$

for all $f \in H^s(\mathbb{T}^d; \mathbb{R})$.

The Sobolev embedding theorem plays a fundamental role in the analysis of Fourier neural operators. The continuous embedding ensures that every function in $H^s(\mathbb{T}^d; \mathbb{R})$ with $s > n + d/2$ admits a unique C^n representative and that its C^n norm is controlled by its Sobolev norm. Consequently, pointwise evaluations of functions and their derivatives can be handled in a mathematically rigorous manner. Since pointwise evaluation is well defined, we may introduce the sampling operator $S_{N,m} : H^s(\mathbb{T}^d; \mathbb{R}^m) \to (\mathbb{R}^m)^{\Lambda_N^d}$ by

$$(S_{N,m}(f))(\boldsymbol{x}) := f(\boldsymbol{x}),$$

where $f \in H^s(\mathbb{T}^d; \mathbb{R}^m)$ and $\boldsymbol{x} \in \Lambda_N^d$.

A.2 Extension to the Least-Squares Loss

Let $N_d \in \mathbb{N}$ be the number of training data. For each discretization size $N \in \mathbb{N}$, we define the least-squares loss function $\mathcal{L}_N : \mathbb{R}^{d_{\boldsymbol{\theta}}} \to \mathbb{R}$ by

$$\mathcal{L}_N(\boldsymbol{\theta}) := \frac{1}{N_d} \sum_{h=1}^{N_d} \left\| \Psi_{\boldsymbol{\theta}}^N(a_h) - S_{N,d_u}(G(a_h)) \right\|_{\ell^2(\Lambda_N^d; \mathbb{R}^{d_u})}^2,$$

where $a_h \in \mathcal{A}$, $h = 1, \dots, N_d$, are the training samples. In the following, let $\Theta \subset \mathbb{R}^{d_{\boldsymbol{\theta}}}$ denote the set of all possible values of the trainable parameters. We assume that Θ is nonempty.

B Proof

B.1 Verification of Assumption 3.1 (A2)

Let $\sigma_0 : \mathbb{R} \to \mathbb{R}$ be a continuously differentiable scalar activation function, and define the vector-valued activation function component-wise by $\sigma(\boldsymbol{x}) = (\sigma_0(x_1), \dots, \sigma_0(x_{d_v}))$. If $\sup_{t \in \mathbb{R}} |\sigma_0'(t)| < \infty$, then σ is Lipschitz continuous. Indeed, by the mean value theorem, for each component $j = 1, \dots, d_v$, we have

$$|\sigma_0(x_j) - \sigma_0(y_j)| \le L_\sigma |x_j - y_j|, \quad L_\sigma := \sup_{t \in \mathbb{R}} |\sigma_0'(t)|.$$

Therefore,

$$\|\sigma(\boldsymbol{x}) - \sigma(\boldsymbol{y})\|_2^2 = \sum_{j=1}^{d_v} |\sigma_0(x_j) - \sigma_0(y_j)|^2 \le L_\sigma^2 \sum_{j=1}^{d_v} |x_j - y_j|^2 = L_\sigma^2 \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Hence, σ satisfies Assumption 3.1 (A2) with L_σ.

For the sigmoid activation function $\sigma_0(t) = (1 + e^{-t})^{-1}$, we have $\sigma_0'(t) = \sigma_0(t)(1 - \sigma_0(t))$. Since $0 < \sigma_0(t) < 1$ and $c(1 - c) \leq 1/4$ for $0 < c < 1$, it follows that $|\sigma_0'(t)| \leq 1/4$ for all $t \in \mathbb{R}$. Thus, the sigmoid activation satisfies Assumption 3.1 (A2) with $L_\sigma = 1/4$.

For the GeLU activation function, let $\sigma_0(t) = t\Phi(t)$, where Φ is the cumulative distribution function of the standard normal distribution:

$$\Phi(t) := \int_{-\infty}^{t} \phi(s)\, ds, \quad \phi(t) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right).$$

Here, ϕ is the density function of the standard normal distribution. Since $\Phi'(t) = \phi(t)$, we obtain

$$\sigma_0'(t) = \Phi(t) + t\phi(t).$$

Since $0 \leq \Phi(t) \leq 1$ and

$$|t\phi(t)| = \frac{|t|}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \leq \frac{1}{\sqrt{2\pi e}},$$

we have

$$|\sigma_0'(t)| \leq 1 + \frac{1}{\sqrt{2\pi e}},$$

for all $t \in \mathbb{R}$.

Therefore, GeLU is Lipschitz continuous and satisfies Assumption 3.1 (A2).

B.2 Proof of Lemma 3.3

Proof. By the definition of $\boldsymbol{\theta}$ and the identification of complex-valued Fourier weights with their real and imaginary parts,

$$\|\boldsymbol{\theta}\|_2^2 = \|\boldsymbol{p}\|_2^2 + \|\boldsymbol{q}\|_2^2 + \sum_{t=0}^{T-1} \left(\|W_t\|_{\mathrm{F}}^2 + \|\boldsymbol{b}_t\|_2^2 + \sum_{k \in K} \left\|R_t^k\right\|_{\mathrm{F}}^2 \right).$$

By Assumption 3.1 (A3), $\|\boldsymbol{\theta}\|_2 \leq D$. Since all terms on the right-hand side are nonnegative, each individual term is bounded by D^2. Hence,

$$\|\boldsymbol{p}\|_2 \leq D, \quad \|\boldsymbol{q}\|_2 \leq D,$$

and

$$\|W_t\|_{\mathrm{F}} \leq D, \quad \|\boldsymbol{b}_t\|_2 \leq D, \quad \left\|R_t^k\right\|_{\mathrm{F}} \leq D,$$

for all $t = 0, \ldots, T-1$ and $k \in K$. Since $K = \{-k_{\max}, \ldots, k_{\max}\}$ contains $2k_{\max} + 1$ Fourier modes, it follows that

$$\sum_{k \in K} \left\|R_t^k\right\|_{\mathrm{F}} \leq (2k_{\max} + 1)D.$$

□

B.3 Proof of Theorem 3.4

Proof. By the definition of $\boldsymbol{\theta}$ and the identification of complex-valued Fourier weights with their real and imaginary parts,

$$\|\boldsymbol{\theta}\|_2^2 = \|\boldsymbol{p}\|_2^2 + \|\boldsymbol{q}\|_2^2 + \sum_{t=0}^{T-1} \left(\|W_t\|_{\mathrm{F}}^2 + \|\boldsymbol{b}_t\|_2^2 + \sum_{k \in K} \left\|R_t^k\right\|_{\mathrm{F}}^2 \right).$$

By Assumption 3.1 (A3), $\|\boldsymbol{\theta}\|_2 \leq D$. Since all terms on the right-hand side are nonnegative, each individual term is bounded by D^2. Hence,

$$\|\boldsymbol{p}\|_2 \leq D, \quad \|\boldsymbol{q}\|_2 \leq D,$$

and

$$\|W_t\|_{\mathrm{F}} \leq D, \quad \|\boldsymbol{b}_t\|_2 \leq D, \quad \left\|R_t^k\right\|_{\mathrm{F}} \leq D,$$

for all $t = 0, \ldots, T-1$ and $k \in K$. Since $K = \{-k_{\max}, \ldots, k_{\max}\}$ contains $2k_{\max} + 1$ Fourier modes, it follows that

$$\sum_{k \in K} \left\|R_t^k\right\|_{\mathrm{F}} \leq (2k_{\max} + 1)D.$$

□

B.4 Proof of Theorem 3.5

Proof. By the definition of the projection layer, we have

$$\begin{aligned}\left|\left(\Psi_{\boldsymbol{\theta}}^{M}(a)\right)(x)-\left(\Psi_{\boldsymbol{\theta}}^{N}(a)\right)(x)\right| &= \left|\boldsymbol{q}^{\top}\boldsymbol{v}_1^{M}(x)-\boldsymbol{q}^{\top}\boldsymbol{v}_1^{N}(x)\right| \\ &= \left|\boldsymbol{q}^{\top}\left(\boldsymbol{v}_1^{M}(x)-\boldsymbol{v}_1^{N}(x)\right)\right|.\end{aligned}$$

From the Cauchy-Schwarz inequality, this implies

$$\left|\left(\Psi_{\boldsymbol{\theta}}^{M}(a)\right)(x)-\left(\Psi_{\boldsymbol{\theta}}^{N}(a)\right)(x)\right| \leq \|\boldsymbol{q}\|_2 \left\|\boldsymbol{v}_1^{M}(x)-\boldsymbol{v}_1^{N}(x)\right\|_2. \tag{2}$$

From the definition of the Fourier layer,

$$\boldsymbol{v}_1^{M}(x)-\boldsymbol{v}_1^{N}(x)=\sigma\left(W\boldsymbol{v}_0^{M}(x)+\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)+b_0\right)-\sigma\left(W\boldsymbol{v}_0^{N}(x)+\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)+b_0\right).$$

By Assumption 3.1 (A2), the activation function σ is the Lipschitz continuous. Therefore,

$$\begin{aligned}\left\|\boldsymbol{v}_1^{M}(x)-\boldsymbol{v}_1^{N}(x)\right\|_2 &\leq L_\sigma \left\|\left(W\boldsymbol{v}_0^{M}(x)+\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)+b_0\right)-\left(W\boldsymbol{v}_0^{N}(x)+\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)+b_0\right)\right\|_2 \\ &= L_\sigma \left\|W\left(\boldsymbol{v}_0^{M}(x)-\boldsymbol{v}_0^{N}(x)\right)+\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)-\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)\right\|_2.\end{aligned}$$

Since $x \in \Lambda_N^1 \subset \Lambda_M^1$, the lifting layer gives

$$\boldsymbol{v}_0^{M}(x)=a(x)\boldsymbol{p}=\boldsymbol{v}_0^{N}(x).$$

Thus,

$$W\left(\boldsymbol{v}_0^{M}(x)-\boldsymbol{v}_0^{N}(x)\right)=0.$$

Hence,

$$\left\|\boldsymbol{v}_1^{M}(x)-\boldsymbol{v}_1^{N}(x)\right\|_2 \leq L_\sigma \left\|\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)-\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)\right\|_2. \tag{3}$$

We now estimate the right-hand-side term of equation 3. From the definition of $\mathcal{K}_t^N : \mathbb{R}^{\Lambda_N^1} \to \mathbb{R}^{\Lambda_N^1}$, we have

$$\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)-\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)=\sum_{k\in K} R_0^k\left(\mathsf{DFT}(\boldsymbol{v}_0^{M})(k)-\mathsf{DFT}(\boldsymbol{v}_0^{N})(k)\right)\exp(2\pi ikx).$$

Using $|\exp(2\pi ikx)|=1$, we obtain

$$\left\|\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)-\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)\right\|_2 \leq \sum_{k\in K}\left\|R_0^k\right\|_{\mathrm{F}}\left\|\mathsf{DFT}(\boldsymbol{v}_0^{M})(k)-\mathsf{DFT}(\boldsymbol{v}_0^{N})(k)\right\|_2.$$

From Lemma 3.4, we have

$$\left\|\mathsf{DFT}(\boldsymbol{v}_0^{M})(k)-\mathsf{DFT}(\boldsymbol{v}_0^{N})(k)\right\|_2 \leq \frac{C_f(a)D(2^r-1)}{2\cdot 2^r}\cdot\frac{1}{N},$$

for all $k \in K$. Therefore, we have

$$\begin{aligned}\left\|\mathcal{K}^{M}\boldsymbol{v}_0^{M}(x)-\mathcal{K}^{N}\boldsymbol{v}_0^{N}(x)\right\|_2 &\leq \sum_{k\in K}\left\|R_0^k\right\|_{\mathrm{F}}\frac{C_f(a)D(2^r-1)}{2\cdot 2^r}\cdot\frac{1}{N} \\ &= \frac{C_f(a)D(2^r-1)}{2\cdot 2^r}\cdot\frac{1}{N}\sum_{k\in K}\left\|R_0^k\right\|_{\mathrm{F}}.\end{aligned}$$

Combining this estimate with equation 2 and equation 3, and using Lemma 3.3 and $N=2^n$, we obtain

$$\left|\left(\Psi_{\boldsymbol{\theta}}^{M}(a)\right)(x)-\left(\Psi_{\boldsymbol{\theta}}^{N}(a)\right)(x)\right| \leq \frac{L_\sigma C_f(a)D^3(2k_{max}+1)}{2}\cdot\left(1-\frac{1}{2^r}\right)\frac{1}{2^n}.$$

This completes the proof. □

B.5 Verification of the Sobolev Regularity of a_1

The input function $a_1 : \mathbb{T} \to \mathbb{R}$ is given by

$$a_1(x) = \sin(2\pi x) + 2\cos(4\pi x) \quad (x \in \mathbb{R}),$$

where we identify a_1 with its 1-periodic representative on $\mathbb{R}$. Expanding the trigonometric functions into complex exponentials, we obtain

$$\hat{a}_1(k) = \begin{cases} -\dfrac{i}{2}, & k = 1, \\ \dfrac{i}{2}, & k = -1, \\ 1, & k = \pm 2, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, a_1 is a trigonometric polynomial whose Fourier coefficients vanish outside $\{\pm 1, \pm 2\}$. Consequently,

$$\|a_1\|^2_{H^s(\mathbb{T})} = \sum_{k\in\mathbb{Z}} \left(1 + |k|^2\right)^s |\hat{a}_1(k)|^2 = 2(1+1)^s \cdot \frac{1}{4} + 2(1+4)^s < \infty$$

for every $s \geq 0$. Hence,

$$a_1 \in H^s(\mathbb{T}) \quad (s \geq 0).$$

In particular, a_1 satisfies Assumption 3.1 (A1) for all $s > 3/2$.

C Loss discrepancy across discretizations

Theorem 3.5 gives a pointwise estimate for the discrepancy between the FNO outputs under the two discretizations. To translate this estimate into a bound on the difference between the corresponding loss functions, we first establish a uniform bound on the pointwise residual between the FNO output and the target operator.

Lemma C.1. *Suppose that Assumptions 3.1 (A1), (A2) and (A3) hold. There exists a constant $B > 0$ such that*

$$\left|\left(\Psi^N_{\boldsymbol{\theta}}(a_h)\right)(x) - (G(a_h))(x)\right| \leq B,$$

for all $\boldsymbol{\theta} \in \Theta$, $N \in \mathbb{N}$, $x \in \Lambda^1_N$ and $h = 1, \ldots, N_d$.

Proof. From the triangle inequality, we have

$$\left|\left(\Psi^N_{\boldsymbol{\theta}}(a_h)\right)(x) - (G(a_h))(x)\right| \leq \left|\left(\Psi^N_{\boldsymbol{\theta}}(a_h)\right)(x)\right| + \left|(G(a_h))(x)\right|, \tag{4}$$

for all $h = 1, \ldots, N_d$. We first estimate the second term on the right-hand side of equation 4. Since $G(a_h) \in H^s(\mathbb{T}; \mathbb{R})$, it follows from Assumption 3.1 (A1) and the Sobolev embedding theorem (Theorem A.1) that $(G(a_h))(x)$ is finite for every $x \in [0, 1]$. Hence, defining

$$B_0 := \max_{x\in[0,1]} |(G(a_h))(x)| < \infty,$$

we obtain

$$|(G(a_h))(x)| \leq B_0,$$

for all $x \in [0, 1]$. Next, we estimate the first term on the right-hand side of equation 4. From the Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left|\left(\Psi^N_{\boldsymbol{\theta}}(a_h)\right)(x)\right| &= \left|\boldsymbol{q}^\top u(x)\right| \\ &\leq \|\boldsymbol{q}\|_2 \left\|\sigma\left(a_h(x) W_0 \boldsymbol{p} + (\mathcal{K}^N_0 \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right)\right\|_2. \end{aligned}$$

From Assumption 3.1 (A2), we have

$$\begin{aligned}
&\left\|\sigma\left(a_h(x)W_0\boldsymbol{p} + (\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right)\right\|_2 \\
&= \left\|\sigma\left(a_h(x)W_0\boldsymbol{p} + (\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right) + \sigma(0) - \sigma(0)\right\|_2 \\
&= \left\|\sigma\left(a_h(x)W_0\boldsymbol{p} + (\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right) - \sigma(0)\right\|_2 + \|\sigma(0)\|_2 \\
&\leq L_\sigma \left\|a_h(x)W_0\boldsymbol{p} + (\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right\|_2 + \|\sigma(0)\|_2 \\
&\leq L_\sigma \left\|a_h(x)W_0\boldsymbol{p}\right\|_2 + L_\sigma \left\|(\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x)\right\|_2 + L_\sigma \left\|\boldsymbol{b}_0\right\|_2 + \|\sigma(0)\|_2 .
\end{aligned}$$

From Lemma 3.3, we have

$$\begin{aligned}
\left\|\sigma\left(a_h(x)W_0\boldsymbol{p} + (\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right)\right\|_2 &\leq \underbrace{L_\sigma \tilde{a} D^2 + L_\sigma D + \|\sigma(0)\|_2}_{B_1} + L_\sigma \left\|(\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x)\right\|_2 \\
&= B_1 + L_\sigma \left\|(\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x)\right\|_2 ,
\end{aligned}$$

where $\tilde{a} \geq 0$ is defined by

$$\tilde{a} := \max_{h=1,\ldots,N_d} \max_{x\in[0,1]} |a_h(x)| .$$

Here, from the definition of $\mathcal{K}_0^N : (\mathbb{R}^{d_v})^{\Lambda_N^1} \to (\mathbb{R}^{d_v})^{\Lambda_N^1}$, we have

$$\begin{aligned}
\left\|(\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x)\right\|_2 &= \left\| \sum_{k\in K} R_0^k \mathsf{DFT}(\mathcal{P}^N(a_h))(k) \exp\left(2\pi i k x\right) \right\|_2 \\
&\leq \sum_{k\in K} \left\| R_0^k \mathsf{DFT}(\mathcal{P}^N(a_h))(k) \exp\left(2\pi i k x\right) \right\|_2 \\
&\leq \sum_{k\in K} \left\| R_0^k \right\|_{\mathrm{F}} \left\| \mathsf{DFT}(\mathcal{P}^N(a_h))(k) \right\|_2 .
\end{aligned}$$

From the definition of DFT, we have

$$\begin{aligned}
\left\| \mathsf{DFT}(\mathcal{P}^N(a_h))(k) \right\|_2 &= \left\| \frac{1}{N} \sum_{y\in\Lambda_N^1} a_h(y)\boldsymbol{p} \exp\left(-2\pi i k y\right) \right\|_2 \\
&\leq \frac{1}{N} \sum_{y\in\Lambda_N^1} |a_h(y)| \left|\exp\left(-2\pi i k y\right)\right| \|\boldsymbol{p}\|_2 \\
&\leq \frac{\tilde{a} D}{N} \sum_{y\in\Lambda_N^1} 1 \\
&= \tilde{a} D.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\left|\left(\Psi_{\boldsymbol{\theta}}^N(a_h)\right)(x)\right| &\leq \|\boldsymbol{q}\|_2 \left\|\sigma\left(a_h(x)W_0\boldsymbol{p} + (\mathcal{K}_0^N \circ \mathcal{P}^N(a_h))(x) + \boldsymbol{b}_0\right)\right\|_2 \\
&\leq D(B_1 + L_\sigma \tilde{a} D^2 (2k_{\max} + 1)).
\end{aligned}$$

This implies that

$$\left|\left(\Psi_{\boldsymbol{\theta}}^N(a_h)\right)(x) - (G(a_h))(x)\right| \leq \underbrace{D(B_1 + L_\sigma \tilde{a} D^2 (2k_{\max} + 1)) + B_0}_{B} .$$

This completes the proof. □

We now extend the pointwise output estimate to the corresponding loss functions. Theorem 3.5 and Lemma C.1 then allow us to bound the difference between the two losses.

Theorem C.2. *Suppose that Assumptions 3.1 (A1), (A2) and (A3) hold. Let $n, r \in \mathbb{N}$, and set $N = 2^n$ and $M = 2^{n+r}$. We define $\tilde{\mathcal{L}}_M : \mathbb{R}^{d_{\boldsymbol{\theta}}} \to \mathbb{R}$ by*

$$\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) := \frac{1}{N_d} \sum_{h=1}^{N_d} \left\| \Psi_{\boldsymbol{\theta}}^M(a_h) - g(a_h) \right\|_{\ell^2(\Lambda_N^1;\mathbb{R})}^2 ,$$

where $\boldsymbol{\theta} \in \mathbb{R}^{d_{\boldsymbol{\theta}}}$. There exists a constant $C \geq 0$ such that

$$\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| \leq C\left(1 - \frac{1}{2^r}\right),$$

for all $\boldsymbol{\theta} \in \Theta$.

Proof. To simplify the notation, we write $g := S_{N,1}(G) : \Lambda_N^1 \to \mathbb{R}$. From the definition of $\tilde{\mathcal{L}}_M$ and $\mathcal{L}_N$, we have

$$\begin{aligned}\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| &= \left|\frac{1}{N_d}\sum_{h=1}^{N_d}\left\|\Psi_{\boldsymbol{\theta}}^M(a_h) - g(a_h)\right\|_{\ell^2(\Lambda_N^1;\mathbb{R})}^2 - \frac{1}{N_d}\sum_{h=1}^{N_d}\left\|\Psi_{\boldsymbol{\theta}}^N(a_h) - g(a_h)\right\|_{\ell^2(\Lambda_N^1;\mathbb{R})}^2\right| \\ &= \left|\frac{1}{N_d}\sum_{h=1}^{N_d}\left\{\left\|\Psi_{\boldsymbol{\theta}}^M(a_h) - g(a_h)\right\|_{\ell^2(\Lambda_N^1;\mathbb{R})}^2 - \left\|\Psi_{\boldsymbol{\theta}}^N(a_h) - g(a_h)\right\|_{\ell^2(\Lambda_N^1;\mathbb{R})}^2\right\}\right|,\end{aligned}$$

for all $\boldsymbol{\theta} \in \Theta$. From the definition of the discretized ℓ^2 norm on the gird Λ_N^1, we have

$$\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| = \left|\frac{1}{N_d}\sum_{x\in\Lambda_N^1}\sum_{h=1}^{N_d}\left\{\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right\}\right|,$$

for all $\boldsymbol{\theta} \in \Theta$. From the triangle inequality, this implies that

$$\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| \leq \sum_{x\in\Lambda_N^1}\left|\frac{1}{N_d}\sum_{h=1}^{N_d}\left\{\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right\}\right|,$$

for all $\boldsymbol{\theta} \in \Theta$. From the Jensen's inequality, we have

$$\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| \leq \frac{1}{N_d}\sum_{x\in\Lambda_N^1}\sum_{h=1}^{N_d}\left\{\left|\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right|\right\}, \tag{5}$$

for all $\boldsymbol{\theta} \in \Theta$. Here, we estimate the quantity inside the braces as follows:

$$\begin{aligned}&\left|\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right| \\ &= \left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (\Psi_{\boldsymbol{\theta}}^N(a_h))(x)\right|\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) + (\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - 2(g(a_h))(x))\right|,\end{aligned}$$

for all $h = 1, \ldots, N_d$ and $x \in \Lambda_N^1$. From the triangle inequality, we have

$$\begin{aligned}&\left|\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right| \\ &\leq \left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (\Psi_{\boldsymbol{\theta}}^N(a_h))(x)\right|\left(\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x))\right| + \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x))\right|\right),\end{aligned}$$

for all $h = 1, \ldots, N_d$ and $x \in \Lambda_N^1$. From Lemma C.1, we have

$$\left|\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right| \leq 2B\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (\Psi_{\boldsymbol{\theta}}^N(a_h))(x)\right|,$$

for all $h = 1, \ldots, N_d$ and $x \in \Lambda_N^1$. From Theorem 3.5, we have

$$\left|\left|(\Psi_{\boldsymbol{\theta}}^M(a_h))(x) - (g(a_h))(x)\right|^2 - \left|(\Psi_{\boldsymbol{\theta}}^N(a_h))(x) - (g(a_h))(x)\right|^2\right| \leq L_\sigma C_f(a_h) B D^3(2k_{max}+1)\cdot\left(1 - \frac{1}{2^r}\right)\frac{1}{N},$$

for all $h = 1, \ldots, N_d$ and $x \in \Lambda_N^1$. Combining this estimate with equation 5, and using $N = 2^n$, we obtain

$$\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| \leq \frac{1}{N_d}\sum_{x\in\Lambda_N^1}\sum_{h=1}^{N_d}\left\{L_\sigma C_f(a_h) B D^3(2k_{max}+1)\cdot\left(1 - \frac{1}{2^r}\right)\frac{1}{N}\right\}.$$

Here, by defining a constant $\overline{C_f} \geq 0$ independent of $h = 1, \ldots, N_d$ by

$$\overline{C_f} := \max_{h=1,\ldots,N_d} C_f(a_h).$$

This implies that

$$\begin{aligned}\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| &\leq L_\sigma C_f(a_h) B D^3 (2k_{max}+1) \cdot \left(1 - \frac{1}{2^r}\right) \frac{1}{N} \sum_{x \in \Lambda_N^1} \frac{1}{N_d} \cdot \sum_{h=1}^{N_d} 1 \\ &= L_\sigma \overline{C_f} B D^3 (2k_{max}+1) \cdot \left(1 - \frac{1}{2^r}\right),\end{aligned}$$

for all $\boldsymbol{\theta} \in \Theta$. From the Assumption 3.1 (A3), we have

$$\left|\tilde{\mathcal{L}}_M(\boldsymbol{\theta}) - \mathcal{L}_N(\boldsymbol{\theta})\right| \leq \underbrace{L_\sigma \overline{C_f} B D^3 (2k_{\max}+1)}_{C} \left(1 - \frac{1}{2^r}\right),$$

for all $\boldsymbol{\theta} \in \Theta$. This completes the proof. □