Data-Complexity of Operator Learning

Samuel Lanthaler

February, 2024

Brin MRC Workshop
UMD 2024

Caltech
Scientific computing

- Neural networks successfully approximate high-dimensional functions.
- In scientific computing, the goal is often to approximate an operator.

\[ G : a \mapsto u \]

\[ -\nabla \cdot (a \nabla u) = f \]
### Problem setting

- Function spaces $\mathcal{X}, \mathcal{Y}$,
- Operator $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$, $u \mapsto \mathcal{G}(u)$,
- Data $\{u_j, \mathcal{G}(u_j)\}_{j=1}^N$,

$$\Rightarrow \quad \text{Goal:} \quad \text{Find approximation}$$

$$\Psi(u; \theta) \approx \mathcal{G}(u).$$

- **Approach:** extend neural networks to $\infty$-dims, e.g.
  - Deep operator networks [Lu, Karniadakis++]
  - Neural operators [Li, Anandkumar, Stuart++]
  - PCA-Net [Bhattacharya, Kovachki, Stuart]
  - Random Feature Model [Nelsen, Stuart]

- **Empirically:** Feasible; potential for model discovery.
Problem setting

- Function spaces $\mathcal{X}, \mathcal{Y}$,
- Operator $G : \mathcal{X} \rightarrow \mathcal{Y}$, $u \mapsto G(u)$,
- Data $\{u_j, G(u_j)\}_{j=1}^N$

Goal:

Find approximation $\Psi(u; \theta) \approx G(u)$.

- Approach: extend neural networks to $\infty$-dims, e.g.
  - Deep operator networks [Lu, Karniadakis++]
  - Neural operators [Li, Anandkumar, Stuart++]
  - PCA-Net [Bhattacharya, Kovachki, Stuart]
  - Random Feature Model [Nelsen, Stuart]

- **Empirically**: Feasible; potential for model discovery.
- **Lack of theory**: When can these methods be effective?
Numerical weather prediction

Figure: FourCastNet (NVIDIA)

1. Efforts to apply AI for NWP:
   - Google, Microsoft, NVIDIA, Huawei, ...

2. Promise especially for ensemble forecasting,
   \[ \rightarrow "45'000x\ speedup". \]
Example: Fourier neural operator\(^1\)

- **composition** \(\Psi(u; \theta) = L_L \circ \cdots \circ L_1(u)\),
- **hidden layers** \(L_\ell : v(x) \mapsto L_\ell(v)(x)\), with vector-valued functions \(v(x), L_\ell(v)(x) \in \mathbb{R}^{d_c}\),

\[
L_\ell(v)(x) = \sigma \left( Wv(x) + \int_D \kappa(x - y)v(y) \, dy \right),
\]

\(^1\)Li, Kovachki et al., “Fourier neural operator for parametric partial differential equations”, ICLR (2021)
Example: Fourier neural operator

- composition $\Psi(u; \theta) = L_L \circ \cdots \circ L_1(u)$,
- hidden layers, $L_\ell : v(x) \mapsto L_\ell(v)(x)$, with vector-valued functions $v(x)$, $L_\ell(v)(x) \in \mathbb{R}^{d_\ell}$,

$$L_\ell(v)(x) = \sigma \left( Wv(x) + \int_D \kappa(x-y)v(y) \, dy \right),$$

- convolution as Fourier multiplier matrix: $\mathcal{F}^{-1}( \mathcal{F}(\kappa) \cdot \mathcal{F}(v) )$, \[ \text{FMM} \]

---

\(^1\text{Li, Kovachki et al., “Fourier neural operator for parametric partial differential equations”, ICLR (2021)} \)
Example: Fourier neural operator

- Composition: $\Psi(u; \theta) = \mathcal{L}_L \circ \cdots \circ \mathcal{L}_1(u)$,

- Hidden layers: $\mathcal{L}_\ell : v(x) \mapsto \mathcal{L}_\ell(v)(x)$, with vector-valued functions $v(x)$, $\mathcal{L}_\ell(v)(x) \in \mathbb{R}^d$,

\[
\mathcal{L}_\ell(v)(x) = \sigma \left( Wv(x) + \int_D \kappa(x - y) v(y) \, dy \right),
\]

- Convolution as Fourier multiplier matrix: $\mathcal{F}^{-1}(\mathcal{F}(\kappa) \cdot \mathcal{F}(v))$,

- Parameter $\theta \in \mathbb{R}^W$ collects components of matrices $(W, \mathcal{F}(\kappa))$ across layers,

- Optimize via loss (empirical risk):

\[
\theta_\mathcal{G} = \arg\min_{\theta} \frac{1}{N} \sum_{j=1}^{N} \| \mathcal{G}(u_j) - \Psi(u_j; \theta) \|^2
\]

---

1Li, Kovachki et al., “Fourier neural operator for parametric partial differential equations”, ICLR (2021)
Approximation Theory

Given

- non-linear operator of interest: \( G : u \mapsto G(u) \)
- distribution of inputs: \( u \sim \mu \) (\( \mu \): probability measure on functions)

Goal

Approximate

\[
\mathbb{E}_{u \sim \mu} [\| G(u) - \Psi(u; \theta) \|^p]^{1/p} \leq \epsilon,
\]

- using parametric model: \( \Psi(u; \theta) \), \( \theta \in \mathbb{R}^W \),
- from sample data: \( (u_1, G(u_1)), \ldots, (u_N, G(u_N)) \).
Approximation Theory

**Questions:**

<table>
<thead>
<tr>
<th>Parametric complexity</th>
<th>Data complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many parameters ( \theta \in \mathbb{R}^W )?</td>
<td>How many samples ( {u_j, G(u_j)}_{j=1}^N )?</td>
</tr>
</tbody>
</table>

**Given**

- non-linear operator of interest: \( G : u \mapsto G(u) \)
- distribution of inputs: \( u \sim \mu \) \( (\mu: \text{probability measure on functions}) \)

**Goal**

Approximate

\[ \mathbb{E}_{u \sim \mu} [\left\| G(u) - \Psi(u; \theta) \right\|^p]^{1/p} \leq \epsilon, \]

- using parametric model: \( \Psi(u; \theta), \theta \in \mathbb{R}^W \),
- from sample data: \( (u_1, G(u_1)), \ldots, (u_N, G(u_N)) \).
Prior work – Parametric complexity

\[ \mathbb{E}_{u \sim \mu} \left[ \| G(u) - \Psi(u; \theta) \|^p \right]^{1/p} \leq \epsilon, \]

How large is \( \text{size}(\Psi(\cdot; \theta)) = \| \theta \|_0 \)?

<table>
<thead>
<tr>
<th>Results</th>
<th>Required size(( \Psi(\cdot; \theta) ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal approximation</td>
<td>size ( \gg 1 ) \textit{sufficient}</td>
</tr>
<tr>
<td>Lipschitz operators</td>
<td>[1]</td>
</tr>
<tr>
<td></td>
<td>[2]</td>
</tr>
<tr>
<td></td>
<td>[3]</td>
</tr>
<tr>
<td>Holomorphic operators</td>
<td>[4]</td>
</tr>
<tr>
<td>PDE operators (case-by-case)</td>
<td></td>
</tr>
</tbody>
</table>

Prior work – Parametric complexity

\[ \mathbb{E}_{u \sim \mu} [\|G(u) - \Psi(u; \theta)\|^p]^{1/p} \leq \epsilon, \quad \text{How large is size}(\Psi(\cdot; \theta)) = \|\theta\|_0? \]

<table>
<thead>
<tr>
<th>Results</th>
<th>Required size((\Psi(\cdot; \theta)))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Universal approximation</strong></td>
<td>size (\gg 1) \quad \text{sufficient}</td>
</tr>
<tr>
<td><strong>Lipschitz operators</strong></td>
<td></td>
</tr>
<tr>
<td>Upper bounds</td>
<td>size (\lesssim \exp(c\epsilon^{-\lambda})) \quad \text{exponential}</td>
</tr>
<tr>
<td>Lower bounds (ReLU)</td>
<td></td>
</tr>
<tr>
<td>Non-standard architectures</td>
<td></td>
</tr>
<tr>
<td><strong>Holomorphobic operators</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>PDE operators</strong> (case-by-case)</td>
<td></td>
</tr>
</tbody>
</table>

Prior work – Parametric complexity

$$\mathbb{E}_{u \sim \mu} [\|G(u) - \Psi(u; \theta)\|^p]^{1/p} \leq \epsilon,$$

How large is size($\Psi(\cdot; \theta)$) = $\|\theta\|_0$?

<table>
<thead>
<tr>
<th>Results</th>
<th>Required size($\Psi(\cdot; \theta)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Universal approximation</strong></td>
<td>$\text{size} \gg 1$ <strong>sufficient</strong></td>
</tr>
<tr>
<td><strong>Lipschitz operators</strong></td>
<td>$\text{size} \lesssim \exp(c \epsilon^{-\lambda})$ <strong>exponential</strong></td>
</tr>
<tr>
<td>Upper bounds</td>
<td>[1]</td>
</tr>
<tr>
<td>Lower bounds (ReLU)</td>
<td>[2]</td>
</tr>
<tr>
<td>Non-standard architectures</td>
<td>[3]</td>
</tr>
<tr>
<td><strong>Holomorphic operators</strong></td>
<td>[4]</td>
</tr>
<tr>
<td><strong>PDE operators</strong> (case-by-case)</td>
<td></td>
</tr>
</tbody>
</table>

Prior work – Parametric complexity

\[ \mathbb{E}_{u \sim \mu} \left[ \| G(u) - \Psi(u; \theta) \|_p \right]^{1/p} \leq \epsilon, \quad \text{How large is size}(\Psi(\cdot; \theta)) = \| \theta \|_0? \]

<table>
<thead>
<tr>
<th>Results</th>
<th>Required size(\Psi(\cdot; \theta))</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Universal approximation</strong></td>
<td>size ( \gg 1 )</td>
</tr>
<tr>
<td><strong>Lipschitz operators</strong></td>
<td></td>
</tr>
<tr>
<td>Upper bounds</td>
<td>size ( \lesssim \exp(c\epsilon^{-\lambda}) )</td>
</tr>
<tr>
<td>Lower bounds (ReLU)</td>
<td>size ( \gtrsim \exp(c\epsilon^{-\lambda}) )</td>
</tr>
<tr>
<td>Non-standard architectures</td>
<td>size ( \lesssim \epsilon^{-\gamma} )</td>
</tr>
<tr>
<td><strong>Holomorphic operators</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>size ( \lesssim \epsilon^{-\gamma} )</td>
</tr>
<tr>
<td><strong>PDE operators (case-by-case)</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>size ( \lesssim \epsilon^{-\gamma} )</td>
</tr>
</tbody>
</table>

Prior work – Data complexity

\[ \mathbb{E}_{u \sim \mu} \left[ \| G(u) - \Psi(u; \theta) \|^p \right]^{1/p} \leq \epsilon, \quad \text{How many samples } (u_1, G(u_1)), \ldots, (u_N, G(u_N))? \]

<table>
<thead>
<tr>
<th>Results</th>
<th>Required # samples N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lipschitz operators</td>
<td>[1]</td>
</tr>
<tr>
<td>Holomorphic operators</td>
<td>[3]</td>
</tr>
</tbody>
</table>

Prior work – Data complexity

\[ \mathbb{E}_{u \sim \mu} \left[ \| G(u) - \Psi(u; \theta) \|^p \right]^{1/p} \leq \epsilon, \]

How many \( (u_1, G(u_1)), \ldots, (u_N, G(u_N)) \)?

<table>
<thead>
<tr>
<th>Results</th>
<th>Required # samples ( N )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lipschitz operators</strong></td>
<td></td>
</tr>
<tr>
<td>Upper bounds</td>
<td>( N \lesssim \exp(c \epsilon^{-\lambda}) )</td>
</tr>
<tr>
<td>( \text{exponential} )</td>
<td></td>
</tr>
<tr>
<td><strong>Holomorphic operators</strong></td>
<td></td>
</tr>
<tr>
<td>( \text{[3]} )</td>
<td>( N \lesssim \epsilon^{-\gamma} )</td>
</tr>
<tr>
<td>( \text{algebraic} )</td>
<td></td>
</tr>
</tbody>
</table>

---

Prior work – Data complexity

\[ \mathbb{E}_{u \sim \mu} [ \| \mathcal{G}(u) - \Psi(u; \theta) \|^p ]^{1/p} \leq \epsilon, \]

How many samples \((u_1, \mathcal{G}(u_1)), \ldots, (u_N, \mathcal{G}(u_N))\)?

<table>
<thead>
<tr>
<th>Results</th>
<th>Required # samples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lipschitz operators</strong></td>
<td></td>
</tr>
<tr>
<td>Upper bounds</td>
<td>(N \lesssim \exp(c\epsilon^{-\lambda}))</td>
</tr>
<tr>
<td>Lower bounds† (sup-norm)</td>
<td>(N \gtrsim \exp(c\epsilon^{-\lambda}))</td>
</tr>
<tr>
<td><strong>Holomorphic operators</strong></td>
<td>(N \lesssim \epsilon^{-\gamma})</td>
</tr>
</tbody>
</table>

Prior work – Data complexity

\[ \mathbb{E}_{u \sim \mu} \left[ \| G(u) - \Psi(u; \theta) \|^p \right]^{1/p} \leq \epsilon, \]

How many samples \((u_1, G(u_1)), \ldots, (u_N, G(u_N))\)?

<table>
<thead>
<tr>
<th>Results</th>
<th>Required # samples (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lipschitz operators</strong></td>
<td></td>
</tr>
<tr>
<td>Upper bounds</td>
<td>([1]) (N \lesssim \exp(c \epsilon^{-\lambda})) exponential</td>
</tr>
<tr>
<td>Lower bounds(\dagger) (sup-norm)</td>
<td>([2]) (N \gtrsim \exp(c \epsilon^{-\lambda})) exponential</td>
</tr>
<tr>
<td><strong>Holomorphic operators</strong></td>
<td>([3]) (N \lesssim \epsilon^{-\gamma}) algebraic</td>
</tr>
</tbody>
</table>

\(\dagger\) Specific setting: \(G : H^s(\Omega) \subset L^2(\Omega) \rightarrow L^2(\Omega)\), with \(\epsilon\)-approximation,

\[
\sup_{\|u\|_{H^s} \leq 1} \| G(u) - \Psi(u; \theta) \|_{L^2} \leq \epsilon.
\]

This work – Data complexity

\[ \mathbb{E}_{u \sim \mu} \left[ \| G(u) - \psi(u; \theta) \|^p \right]^{1/p} \leq \epsilon, \]

How many samples \((u_1, G(u_1)), \ldots, (u_N, G(u_N))\)?

<table>
<thead>
<tr>
<th>Model complexity</th>
<th>Data complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>sup-norm</td>
<td>sup-norm</td>
</tr>
<tr>
<td>Lipschitz operators</td>
<td>size (\gtrsim \exp(ce^{-\lambda}))</td>
</tr>
<tr>
<td>“Natural” operators</td>
<td>size (\psi(\cdot; \theta)) \lesssim \epsilon^{-\gamma})</td>
</tr>
</tbody>
</table>
Lower bounds via “non-linear widths”

\[ G : \mathcal{X} \to \mathcal{Y} \mapsto (G(u_1), \ldots, G(u_N)) \in \mathcal{Y}^N \]

- encoder/decoder point of view,
- many-to-one mapping,
Lower bounds via “non-linear widths”

\[ G : \mathcal{X} \to \mathcal{Y} \iff (G(u_1), \ldots, G(u_N)) \in \mathcal{Y}^N \]

- encoder/decoder point of view,
- many-to-one mapping,
- best reconstruction limited by the width of the pre-image,
- different notions of widths
  - continuous \( n \)-width: arbitrary continuous encoder, arbitrary decoder
  - sampling \( n \)-width: encoder by point-evaluation, arbitrary decoder
Lower bounds via “non-linear widths”

- $\Psi = D_N(G(u_1), \ldots, G(u_N))$ reconstruction from samples,
  - $\{u_1, \ldots, u_N\}$ chosen sampling points,
  - $D_N : \mathcal{Y}^N \rightarrow \text{Lip}(\mathcal{X}, \mathcal{Y})$ chosen decoder/reconstruction algorithm.
Lower bounds via “non-linear widths”

- $\Psi = \mathcal{D}_N(G(u_1), \ldots, G(u_N))$ reconstruction from samples,
  - $\{u_1, \ldots, u_N\}$ chosen sampling points,
  - $\mathcal{D}_N : \mathcal{Y}^N \rightarrow \text{Lip}(\mathcal{X}, \mathcal{Y})$ chosen decoder/reconstruction algorithm.

### Sampling $N$-width

\[
\text{sampling } N\text{-width} = \inf_{\{u_j\}_{j=1}^N} \sup_{\mathcal{D}_N \mathcal{G}} \mathbb{E}_{u \sim \mu} \left[ \|G(u) - \Psi(u)\|_Y^p \right]^{1/p}
\]

- supremum over $\mathcal{G} \in \text{Lip}_1(\mathcal{X}, \mathcal{Y})$, i.e. 1-Lipschitz operators.

This measures:
- **Worst-case** reconstruction-error ...
- ... of the **best-possible choice** of sampling points and the best reconstruction.
$L^p$ setting

- $G \in \text{Lip}_1(\mathcal{X}; \mathcal{Y})$ 1-Lipschitz operator,
- Input functions drawn from $\mu = \text{Gaussian random field}$,

\[ u = \sum_{j=1}^{\infty} \lambda_j Z_j e_j, \quad Z_j \sim \mathcal{N}(0, 1), \quad \lambda_j \sim j^{-\alpha}. \]
**$L^p$ setting**

- $\mathcal{G} \in \text{Lip}_1(\mathcal{X}; \mathcal{Y})$ 1-Lipschitz operator,
- Input functions drawn from $\mu = \text{Gaussian random field},$

\[ u = \sum_{j=1}^{\infty} \lambda_j Z_j e_j, \quad Z_j \sim \mathcal{N}(0, 1), \quad \lambda_j \sim j^{-\alpha}. \]

**Theorem (Kovachki, SL ’24)**

For any $1 \leq p < \infty$, we have

\[ \inf_{\{u_j\}_{j=1}^N, \mathcal{D}_N} \sup_{\mathcal{G}} \mathbb{E}_{u \sim \mu} \left[ \| \mathcal{G}(u) - \Psi(u) \|_{\mathcal{Y}}^p \right]^{1/p} \gtrsim \log(N)^{-\alpha+3}. \]

Thus, with any neural operator architecture, to achieve $\epsilon$-accuracy,

\[ \sup_{\mathcal{G} \in \text{Lip}_1} \mathbb{E}_{u \sim \mu} \left[ \| \mathcal{G}(u) - \Psi(u; \theta_{\mathcal{G}}) \|_{\mathcal{Y}}^p \right]^{1/p} \leq \epsilon, \]

we need exponentially many samples, $N \gtrsim \exp(ce^{-\lambda}).$
$L^\infty$ setting

$$\mathbb{E}_{u \sim \mu} \left[ \|G(u) - \Psi(u)\|^p \right]^{1/p} \xrightarrow{(p \to \infty)} \sup_u \|G(u) - \Psi(u)\|$$

Also corresponding result in the sup-norm:

**Theorem (Kovachki, SL ’24)**

The sampling $N$-width decays only logarithmically

$$s_N(\text{error in sup-norm}) \gtrsim \log(N)^{-\alpha}.$$  

Thus, $N \gtrsim \exp(c\epsilon^{-\gamma})$ samples are required to achieve accuracy $\epsilon$. 
Basic idea: it’s a counting game

- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate $G$ with error $\epsilon$?
- Equivalently: Given $N$ what error $\epsilon$ can be achieved?
- $G : [0, 1] \to \mathbb{R}, \quad \sup_{x \in [0, 1]} |G(x)|, |G'(x)| \leq 1,$
Basic idea: it’s a counting game

- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate $G$ with error $\epsilon$?
- Equivalently: Given $N$ what error $\epsilon$ can be achieved?
- $G : [0, 1] \to \mathbb{R}$, $\sup_{x \in [0,1]} |G(x)|, |G'(x)| \leq 1$,
- Consider sum of $N + 1$ “bumps”,

$$G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.$$
Basic idea: it’s a counting game

• How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate $G$ with error $\epsilon$?
• Equivalently: Given $N$ what error $\epsilon$ can be achieved?
• $G : [0, 1] \to \mathbb{R}$, $\sup_{x \in [0, 1]} |G(x)|, |G'(x)| \leq 1$,
• Consider sum of $N + 1$ “bumps”,

$$G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.$$
Basic idea: it’s a counting game

- How many evaluations \( G(x_1), \ldots, G(x_N) \) to approximate \( G \) with error \( \epsilon \)?
- Equivalently: Given \( N \) what error \( \epsilon \) can be achieved?
- \( G : [0, 1] \rightarrow \mathbb{R}, \quad \sup_{x \in [0,1]} |G(x)|, |G'(x)| \leq 1 \),
- Consider sum of \( N + 1 \) “bumps”,

\[
G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.
\]
Basic idea: it's a counting game

- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate $G$ with error $\epsilon$?
- Equivalently: Given $N$ what error $\epsilon$ can be achieved?
- $G : [0, 1] \to \mathbb{R}, \quad \sup_{x \in [0, 1]} |G(x)|, |G'(x)| \leq 1$,
- Consider sum of $N + 1$ “bumps”,

$$G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.$$ 

- Can reconstruct sign of at most $N$ bumps, so reconstruction error

$$\text{best possible error } \epsilon \gtrsim \text{height of bump } \sim N^{-1} \quad \Rightarrow \quad N \gtrsim \epsilon^{-1}.$$
Basic idea: it’s a counting game

• How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate $G$ with error $\epsilon$?
• Equivalently: Given $N$ what error $\epsilon$ can be achieved?
• $G : [0, 1] \rightarrow \mathbb{R}, \quad \sup_{x \in [0,1]} |G(x)|, |G'(x)| \leq 1$,
• Consider sum of $N + 1$ “bumps”,

\[ G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}. \]

• Can reconstruct sign of at most $N$ bumps, so reconstruction error

\[ \text{best possible error } \epsilon \gtrsim \text{height of bump } \sim N^{-1} \quad \Rightarrow \quad N \gtrsim \epsilon^{-1}. \]
Basic idea: it’s a counting game

- How many evaluations \( G(x_1), \ldots, G(x_N) \) to approximate \( G \) with error \( \epsilon \)?
- Equivalently: Given \( N \) what error \( \epsilon \) can be achieved?
- \( G : [0, 1] \to \mathbb{R}, \quad \sup_{x \in [0,1]} |G(x)|, |G'(x)| \leq 1, \)
- Consider sum of \( N + 1 \) “bumps”,

\[
G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.
\]

- Can reconstruct sign of at most \( N \) bumps, so reconstruction error

\[
\text{best possible error } \epsilon \gtrsim \text{height of bump } \sim N^{-1} \quad \Rightarrow \quad N \gtrsim \epsilon^{-1}.
\]
Basic idea: it’s a counting game

- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate $G$ with error $\epsilon$?
- Equivalently: Given $N$ what error $\epsilon$ can be achieved?
- $G : [0, 1] \rightarrow \mathbb{R}$, $\sup_{x \in [0,1]} |G(x)|, |G'(x)| \leq 1$,
- Consider sum of $N + 1$ “bumps”,
Contradiction (?)

In Theory
Learning $\mathcal{G} \in \text{Lip}(\mathcal{X}; \mathcal{Y})$ requires
- **exponential** amounts of data,
- *(and exponential model size).*

In Practice
Learning operators of interest with
- **moderate** amounts of data,
- *(and moderate model size).*
Sample bounds for “operators of interest”?

**Given**

- non-linear operator of interest: \( \mathcal{G} : u \mapsto \mathcal{G}(u) \),
- distribution of inputs: \( u \sim \mu \),
- parametric model: \( \Psi(u; \theta) \).

**Goal**

Approximate from sample data, \((u_1, \mathcal{G}(u_1)), \ldots, (u_N, \mathcal{G}(u_N))\),

\[
\mathbb{E}_{u \sim \mu} \left[ \| \mathcal{G}(u) - \Psi(u; \theta) \|^2 \right]^{1/2} \leq \epsilon,
\]

**Question**

How many samples are sufficient?

- Answer depends on \( \mathcal{G}, \mu, \Psi(\cdot; \theta) \).
- Assuming only \( \mathcal{G} \in \text{Lip} \) leads to very pessimistic bounds.
- Intuition: \( \text{Lip} \) is too large; does not capture “operators of interest”.
Operators of interest

• Difficult to characterize “operators of interest”
  • $\text{Lip}(\mathcal{X}; \mathcal{Y})$ too broad,
  • Holomorphic operators too narrow (?)
Operators of interest

• Difficult to characterize “operators of interest”
  • Lip($\mathcal{X}; \mathcal{Y}$) too broad,
  • Holomorphic operators too narrow (?)

Fourier neural operator (FNO) approximation space

Given $\mu$ supported on compact set $\mathcal{K} \subset \mathcal{X}$, parameter $\gamma > 0$:

$$A^\gamma(FNO) := \left\{ G : \mathcal{K} \subset \mathcal{X} \to \mathcal{Y} \Bigg| \inf_{\text{size}(\Psi) \leq W} \| G - \Psi(\cdot; \theta) \|_{C(\mathcal{K})} \lesssim W^{-\gamma} \right\}$$

• efficiently approximated by Fourier neural operator, in terms of model size,

---

Operators of interest

- Difficult to characterize “operators of interest”
  - $\text{Lip}(\mathcal{X}; \mathcal{Y})$ too broad,
  - Holomorphic operators too narrow (?)

Fourier neural operator (FNO) approximation space

Given $\mu$ supported on compact set $\mathcal{K} \subset \mathcal{X}$, parameter $\gamma > 0$:

$$
\mathcal{A}^\gamma(\text{FNO}) := \left\{ G : \mathcal{K} \subset \mathcal{X} \to \mathcal{Y} \left| \inf_{\text{size}(\Psi) \leq W} \| G - \Psi(\cdot; \theta) \|_{C(\mathcal{K})} \lesssim W^{-\gamma} \right. \right\}
$$

- efficiently approximated by Fourier neural operator, in terms of model size,
  - examples\(^2\): Navier-Stokes in 2D, coeff-to-sol map of elliptic PDE $-\nabla \cdot (a \nabla u) = f$

---

Operators of interest

- Difficult to characterize “operators of interest”
  - Lip$(\mathcal{X}; \mathcal{Y})$ too broad,
  - Holomorphic operators too narrow (?)

Fourier neural operator (FNO) approximation space

Given $\mu$ supported on compact set $\mathcal{K} \subset \mathcal{X}$, parameter $\gamma > 0$:

$$\mathcal{A}^{\gamma}(\text{FNO}) := \left\{ \mathcal{G} : \mathcal{K} \subset \mathcal{X} \rightarrow \mathcal{Y} \bigg| \inf_{\text{size}(\Psi) \leq W} \| \mathcal{G} - \Psi(\cdot ; \theta) \|_{C(\mathcal{K})} \lesssim W^{-\gamma} \right\}$$

- efficiently approximated by Fourier neural operator, in terms of model size,
- examples$^2$: Navier-Stokes in 2D, coeff-to-sol map of elliptic PDE $-\nabla \cdot (a \nabla u) = f$

**Question:** Can $\mathcal{G} \in \mathcal{A}^{\gamma}(\text{FNO})$ be efficiently approximated in terms of sample complexity?

---

• Consider unit ball $B^\gamma \subset A^\gamma (FNO), G \in B^\gamma$.

• **Empirical risk minimizer:** Fix $\Psi(\cdot; \theta)$ Fourier neural operator architecture, 

$$G \approx \Psi(\cdot; \theta_G), \quad \theta_G := \arg\min_{\theta} \frac{1}{N} \sum_{j=1}^{N} \|G(u_j) - \Psi(u_j; \theta)\|^2.$$

**Theorem (Kovachki, SL ’24)**

For any $N$, there exist sample points $u_1, \ldots, u_N$, and FNO architecture $\Psi(\cdot; \theta)$ of size $W = W(N)$ depending on $N$, such that empirical risk minimizers $\Psi(\cdot; \theta_G)$ satisfy

$$\sup_{G \in B^\gamma} \mathbb{E}_{u \sim \mu} \left[\|G(u) - \Psi(u; \theta_G)\|^2\right]^{1/2} \lesssim N^{-\frac{1}{2} \frac{\gamma}{\gamma+8}}$$

**worst-case error of ERM**
• Consider unit ball $B^\gamma \subset A^\gamma (FNO)$, $\mathcal{G} \in B^\gamma$.

• **Empirical risk minimizer:** Fix $\Psi(\cdot ; \theta)$ Fourier neural operator architecture,

$$
\mathcal{G} \approx \Psi(\cdot ; \theta_\mathcal{G}), \quad \theta_\mathcal{G} := \arg\min_\theta \frac{1}{N} \sum_{j=1}^{N} \| \mathcal{G}(u_j) - \Psi(u_j; \theta) \|^2.
$$

**Theorem (Kovachki, SL ’24)**

For any $N$, there exist sample points $u_1, \ldots, u_N$, and FNO architecture $\Psi(\cdot ; \theta)$ of size $W = W(N)$ depending on $N$, such that empirical risk minimizers $\Psi(\cdot ; \theta_\mathcal{G})$ satisfy

$$
\sup_{\mathcal{G} \in B^\gamma} \mathbb{E}_{u \sim \mu} \left[ \| \mathcal{G}(u) - \Psi(u; \theta_\mathcal{G}) \|^2 \right]^{1/2} \lesssim N^{-\frac{1}{2} \frac{\gamma}{\gamma+8}} = \underbrace{N^{-1/2}}_{\text{Monte-Carlo}} \cdot \underbrace{\text{(correction)}}_{\text{complexity of } B^\gamma}
$$

\[\text{worst-case error of ERM}\]
### Data complexity

How many samples \( \{u_j, \mathcal{G}(u_j)\}_{j=1}^N \) are needed to approximate operator,

\[
\mathbb{E}_{u \sim \mu} \left[ \|\mathcal{G}(u) - \Psi(u; \theta)\|^p \right]^{1/p} \leq \epsilon?
\]

<table>
<thead>
<tr>
<th>Model complexity</th>
<th>Data complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>sup-norm</td>
<td>sup-norm</td>
</tr>
<tr>
<td>Lipschitz operators</td>
<td>size ( \gtrsim \exp(c\epsilon^{-\lambda}) )</td>
</tr>
<tr>
<td>(Fréchet-)( C^k ) operators</td>
<td>size ( \gtrsim \exp(c\epsilon^{-\lambda}) )</td>
</tr>
<tr>
<td>“Natural” operators</td>
<td>size ( \Psi(\cdot; \theta) ) ( \lesssim \epsilon^{-\gamma} ) (by definition)</td>
</tr>
</tbody>
</table>