Data-Complexity of Operator Learning

Samuel Lanthaler

February, 2024

Brin MRC Workshop
UMD 2024

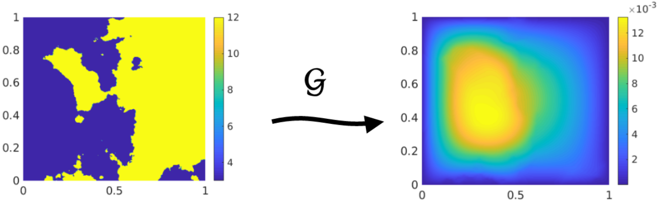
Caltech

Scientific computing

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

$$\mathcal{G} : a \mapsto u$$

$$-\nabla \cdot (a \, \nabla u) = f$$



$$\mathcal{X} \longrightarrow \mathcal{Y}$$

∞-dim. function spaces

Problem setting

$$\left.\begin{array}{ll} \bullet & \text{Function spaces } \mathcal{X}, \mathcal{Y}, \\ \bullet & \text{Operator } \mathcal{G} : \mathcal{X} \to \mathcal{Y},\ u \mapsto \mathcal{G}(u), \\ \bullet & \text{Data } \{u_j, \mathcal{G}(u_j)\}_{j=1}^{N}, \end{array}\right\} \Rightarrow \begin{array}{c} \textbf{Goal:} \\ \text{Find approximation} \\ \Psi(u;\theta) \approx \mathcal{G}(u). \end{array}$$

- Approach: extend neural networks to ∞-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

$$\left.\begin{array}{ll} \bullet & \text{Function spaces } \mathcal{X}, \mathcal{Y}, \\ \bullet & \text{Operator } \mathcal{G} : \mathcal{X} \to \mathcal{Y},\ u \mapsto \mathcal{G}(u), \\ \bullet & \text{Data } \{u_j, \mathcal{G}(u_j)\}_{j=1}^{N}, \end{array}\right\} \Rightarrow \begin{array}{c} \textbf{Goal:} \\ \text{Find approximation} \\ \Psi(u;\theta) \approx \mathcal{G}(u). \end{array}$$

- Approach: extend neural networks to ∞-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

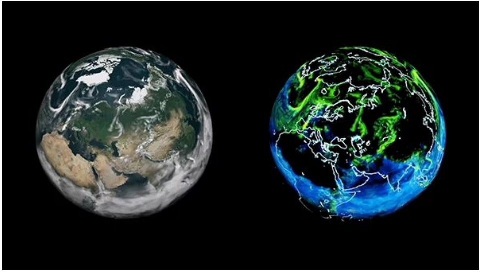


Figure: FourCastNet (NVIDIA)

1. Efforts to apply AI for NWP:
   - Google, Microsoft, NVIDIA, Huawei, ...
2. Promise especially for ensemble forecasting,

$\longrightarrow$ *"45'000x speedup"*.

The Washington Post

How Big Tech AI models nailed forecast for Hurricane Lee a week in advance

Story by Dan Stillman • 3w

Ensemble forecasts from conventional models can miss extreme events, such as excessive rainfall or heat, because they are limited to about 50 simulations due to the time and cost of generating them. AI could enable the generation of much larger ensembles in as little as a few minutes, potentially leading to more useful forecasts and risk assessments for emergency managers, the general public and numerous industries.

"Our hypothesis is we can easily now scale up with AI models to thousands or tens of thousands of ensemble members," Anima Anandkumar, senior director of AI Research at NVIDIA, said in an interview.

Example: Fourier neural operator[1]

- 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 $v(x), \mathcal{L}_\ell(v)(x) \in \mathbb{R}^{d_c}$,

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

[1] Li, Kovachki *et al.*, "Fourier neural operator for parametric partial differential equations", ICLR (2021)

Example: Fourier neural operator[1]

- 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 $v(x), \mathcal{L}_\ell(v)(x) \in \mathbb{R}^{d_c}$,

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

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

[1] Li, Kovachki *et al.*, "Fourier neural operator for parametric partial differential equations", ICLR (2021)

Example: Fourier neural operator[1]

- 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 $v(x), \mathcal{L}_\ell(v)(x) \in \mathbb{R}^{d_c}$,

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

- convolution as Fourier multiplier matrix: $\mathcal{F}^{-1}\big(\underbrace{\mathcal{F}(\kappa)}_{\text{FMM}} \cdot \mathcal{F}(v)\big)$,
- parameter $\theta \in \mathbb{R}^W$ collects components of matrices $(W, \mathcal{F}(\kappa))$ across layers,
- optimize via loss (empirical risk):

$$\theta_{\mathcal{G}} = \operatorname*{argmin}_\theta \frac{1}{N}\sum_{j=1}^N \|\mathcal{G}(u_j) - \Psi(u_j;\theta)\|^2$$

[1] Li, Kovachki *et al.*, "Fourier neural operator for parametric partial differential equations", ICLR (2021)

Approximation Theory

Given

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

Goal

Approximate

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

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

Approximation Theory

Questions:

Parametric complexity

How many parameters $\theta \in \mathbb{R}^W$?

Data complexity

How many samples $\{u_j, \mathcal{G}(u_j)\}_{j=1}^N$?

Given

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

Goal

Approximate

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

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

Prior work – Parametric complexity

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

Results		Required $\mathrm{size}(\Psi(\,\cdot\,;\theta))$	
Universal approximation		$\mathrm{size} \gg 1$	*sufficient*
Lipschitz operators			
	[1]		
	[2]		
	[3]		
Holomorphic operators	[4]		
PDE operators (case-by-case)			

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Lanthaler, Stuart, *"The parametric complexity of operator learning"*, (2023)
[3] Schwab, Stein, and Zech, *"Deep operator network approximation rates for Lipschitz operators"*, (2023)
[4] Hermann, Schwab, and Zech, *"Neural and gpc operator surrogates: Construction and expression rate bounds"* (2022)

Prior work – Parametric complexity

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

Results		Required $\mathrm{size}(\Psi(\,\cdot\,;\theta))$	
Universal approximation		$\mathrm{size} \gg 1$	*sufficient*
Lipschitz operators			
Upper bounds	[1]	$\mathrm{size} \lesssim \exp(c\epsilon^{-\lambda})$	*exponential*
Lower bounds (ReLU)	[2]		
Non-standard architectures	[3]		
Holomorphic operators	[4]		
PDE operators (case-by-case)			

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Lanthaler, Stuart, *"The parametric complexity of operator learning"*, (2023)
[3] Schwab, Stein, and Zech, *"Deep operator network approximation rates for Lipschitz operators"*, (2023)
[4] Hermann, Schwab, and Zech, *"Neural and gpc operator surrogates: Construction and expression rate bounds"* (2022)

Prior work – Parametric complexity

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

Results		Required size$(\Psi(\,\cdot\,;\theta))$	
Universal approximation		size $\gg 1$	*sufficient*
Lipschitz operators			
Upper bounds	[1]	size $\lesssim \exp(c\epsilon^{-\lambda})$	*exponential*
Lower bounds (ReLU)	[2]	size $\gtrsim \exp(c\epsilon^{-\lambda})$	*exponential*
Non-standard architectures	[3]	size $\lesssim \epsilon^{-\gamma}$	*algebraic*
Holomorphic operators	[4]		
PDE operators (case-by-case)			

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Lanthaler, Stuart, *"The parametric complexity of operator learning"*, (2023)
[3] Schwab, Stein, and Zech, *"Deep operator network approximation rates for Lipschitz operators"*, (2023)
[4] Hermann, Schwab, and Zech, *"Neural and gpc operator surrogates: Construction and expression rate bounds"* (2022)

Prior work – Parametric complexity

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

Results		Required $\mathrm{size}(\Psi(\,\cdot\,;\theta))$	
Universal approximation		$\mathrm{size}\gg 1$	*sufficient*
Lipschitz operators			
Upper bounds	[1]	$\mathrm{size}\lesssim\exp(c\epsilon^{-\lambda})$	*exponential*
Lower bounds (ReLU)	[2]	$\mathrm{size}\gtrsim\exp(c\epsilon^{-\lambda})$	*exponential*
Non-standard architectures	[3]	$\mathrm{size}\lesssim\epsilon^{-\gamma}$	*algebraic*
Holomorphic operators	[4]	$\mathrm{size}\lesssim\epsilon^{-\gamma}$	*algebraic*
PDE operators (case-by-case)		$\mathrm{size}\lesssim\epsilon^{-\gamma}$	*algebraic*

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Lanthaler, Stuart, *"The parametric complexity of operator learning"*, (2023)
[3] Schwab, Stein, and Zech, *"Deep operator network approximation rates for Lipschitz operators"*, (2023)
[4] Hermann, Schwab, and Zech, *"Neural and gpc operator surrogates: Construction and expression rate bounds"* (2022)

Prior work – Data complexity

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

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

Results	Required # samples N
Lipschitz operators	
[1]	
Holomorphic operators [3]	

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Mhaskar and Hahm, *"Neural networks for functional approximation and system identification"*, (1997)
[3] Adcock, Dexter, and Moraga, *"Optimal approximation of infinite-dimensional holomorphic functions"*, (2023)

Prior work – Data complexity

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

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

Results		Required # samples N	
Lipschitz operators			
Upper bounds	[1]	$N \lesssim \exp(c\epsilon^{-\lambda})$	*exponential*
Holomorphic operators	[3]	$N \lesssim \epsilon^{-\gamma}$	*algebraic*

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Mhaskar and Hahm, *"Neural networks for functional approximation and system identification"*, (1997)
[3] Adcock, Dexter, and Moraga, *"Optimal approximation of infinite-dimensional holomorphic functions"*, (2023)

Prior work – Data complexity

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

Results		Required # samples N	
Lipschitz operators			
Upper bounds	[1]	$N \lesssim \exp(c\epsilon^{-\lambda})$	*exponential*
Lower bounds† (sup-norm)	[2]	$N \gtrsim \exp(c\epsilon^{-\lambda})$	*exponential*
Holomorphic operators	[3]	$N \lesssim \epsilon^{-\gamma}$	*algebraic*

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Mhaskar and Hahm, *"Neural networks for functional approximation and system identification"*, (1997)
[3] Adcock, Dexter, and Moraga, *"Optimal approximation of infinite-dimensional holomorphic functions"*, (2023)

Prior work – Data complexity

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

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

Results		Required # samples N	
Lipschitz operators			
Upper bounds	[1]	$N\lesssim\exp(c\epsilon^{-\lambda})$	*exponential*
Lower bounds† (sup-norm)	[2]	$N\gtrsim\exp(c\epsilon^{-\lambda})$	*exponential*
Holomorphic operators	[3]	$N\lesssim\epsilon^{-\gamma}$	*algebraic*

† Specific setting: $\mathcal{G}: H^s(\Omega)\subset L^2(\Omega)\to L^2(\Omega)$, with ϵ-approximation,

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

[1] Liu *et al.*, *"Deep nonparametric estimation of operators between infinite dimensional spaces"*, (2022)
[2] Mhaskar and Hahm, *"Neural networks for functional approximation and system identification"*, (1997)
[3] Adcock, Dexter, and Moraga, *"Optimal approximation of infinite-dimensional holomorphic functions"*, (2023)

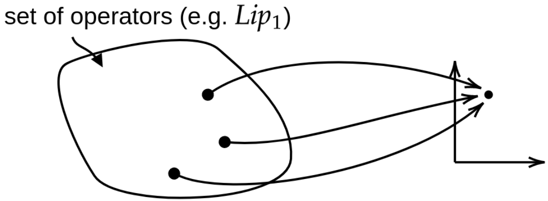
This work – Data complexity

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

	Model complexity	Data complexity	
	sup-norm	sup-norm	L^p-norm
Lipschitz operators	size $\gtrsim \exp(c\epsilon^{-\lambda})$	$N \gtrsim \exp(c\epsilon^{-\lambda})$	???
“Natural” operators	$\underbrace{\mathrm{size}(\Psi(\,\cdot\,;\theta)) \lesssim \epsilon^{-\gamma}}_{\text{by definition}}$		???

Lower bounds via “non-linear widths”

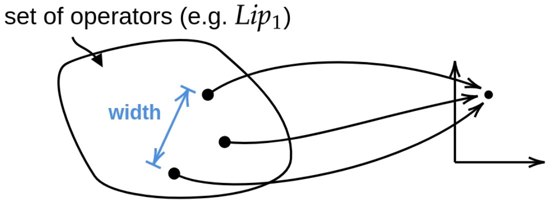
$$\mathcal{G} : \mathcal{X} \to \mathcal{Y} \quad \mapsto \quad (\mathcal{G}(u_1), \ldots, \mathcal{G}(u_N)) \in \mathcal{Y}^N$$



- encoder/decoder point of view,
- many-to-one mapping,

Lower bounds via "non-linear widths"

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



- 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”

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

Lower bounds via "non-linear widths"

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

Sampling N-width

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

- supremum over $\mathcal{G} \in \mathrm{Lip}_1(\mathcal{X}, \mathcal{Y})$, 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

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

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

L^p setting

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

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

Theorem (Kovachki, SL '24)

For any $1 \le p < \infty$, 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 ϵ-accuracy,

$$\sup_{\mathcal{G}\in\mathrm{Lip}_1} \mathbb{E}_{u\sim\mu}[\|\mathcal{G}(u) - \Psi(u;\theta_{\mathcal{G}})\|^p]^{1/p} \le \epsilon,$$

we need exponentially many samples, $N \gtrsim \exp(c\epsilon^{-\lambda})$.

L^∞ setting

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

Also corresponding result in the sup-norm:

Theorem (Kovachki, SL '24)

The sampling N-width decays only logarithmically

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

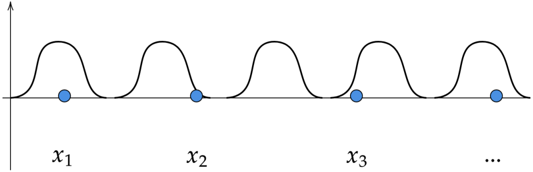
Thus, $N \gtrsim \exp(c\epsilon^{-\gamma})$ samples are required to achieve accuracy ϵ.

Basic idea: it's a counting game

- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate G with error ϵ?
- Equivalently: Given N what error ϵ 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

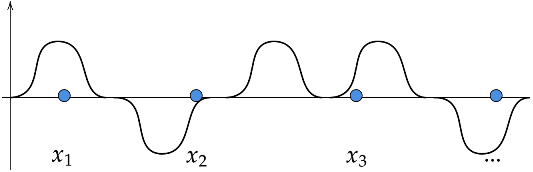
- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate G with error ϵ?
- Equivalently: Given N what error ϵ can be achieved?
- $G : [0, 1] \to \mathbb{R}, \quad \sup_{x\in[0,1]} |G(x)|, |G'(x)| \leq 1,$
- Consider sum of $N + 1$ "bumps",



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

Basic idea: it's a counting game

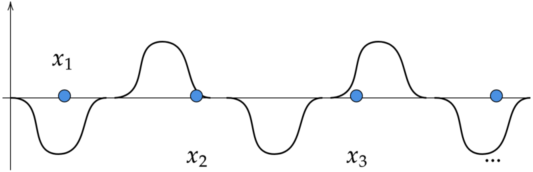
- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate G with error ϵ?
- Equivalently: Given N what error ϵ can be achieved?
- $G : [0, 1] \to \mathbb{R}, \quad \sup_{x\in[0,1]} |G(x)|, |G'(x)| \leq 1,$
- Consider sum of $N + 1$ "bumps",



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

Basic idea: it's a counting game

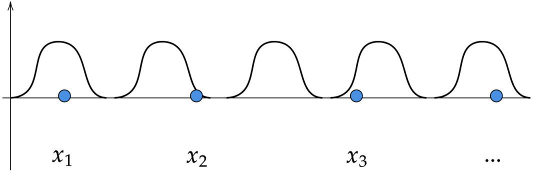
- How many evaluations $G(x_1), \ldots, G(x_N)$ to approximate G with error ϵ?
- Equivalently: Given N what error ϵ can be achieved?
- $G : [0, 1] \to \mathbb{R}, \quad \sup_{x\in[0,1]} |G(x)|, |G'(x)| \leq 1,$
- Consider sum of $N + 1$ "bumps",



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

Basic idea: it's a counting game

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



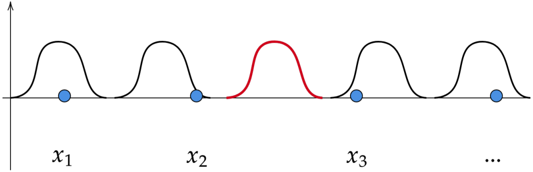
$$G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.$$

- 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 ϵ?
- Equivalently: Given N what error ϵ can be achieved?
- $G : [0, 1] \to \mathbb{R}, \quad \sup_{x\in[0,1]} |G(x)|, |G'(x)| \le 1,$
- Consider sum of $N + 1$ "bumps",



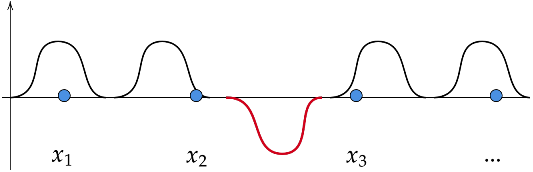
$$G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.$$

- 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 $G(x_1), \dots, G(x_N)$ to approximate G with error ϵ?
- Equivalently: Given N what error ϵ can be achieved?
- $G : [0,1] \to \mathbb{R}, \quad \sup_{x\in[0,1]} |G(x)|, |G'(x)| \le 1,$
- Consider sum of $N+1$ "bumps",



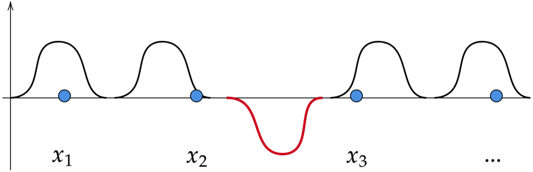
$$G(x) = \sum_{j=1}^{N+1} \sigma_j \phi_j(x), \quad \{\sigma_j = \pm 1\}.$$

- 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 ϵ?
- Equivalently: Given N what error ϵ 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

$$\text{best possible error } \epsilon \gtrsim \text{height of bump} \sim N^{-1} \quad \Rightarrow \quad N \gtrsim \epsilon^{-1}.$$

- Relates achievable accuracy ϵ to number of evaluation points N.
- Theorem generalizes this basic idea to ∞ dimensions.

Contradiction (?)

In Theory

Learning $\mathcal{G} \in \mathrm{Lip}(\mathcal{X}; \mathcal{Y})$ 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 $\mathcal{G}$, μ, $\Psi(\,\cdot\,;\theta)$.
- Assuming only $\mathcal{G} \in \mathrm{Lip}$ leads to very pessimistic bounds.
- Intuition: Lip is too large; does not capture "operators of interest".

Operators of interest

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

[2] Kovachki, Lanthaler, Mishra, *"On Universal Approximation and Error Bounds for Fourier Neural Operators"*, (2021)

Operators of interest

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

Fourier neural operator (FNO) approximation space

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

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

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

[2] Kovachki, Lanthaler, Mishra, *"On Universal Approximation and Error Bounds for Fourier Neural Operators"*, (2021)

Operators of interest

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

Fourier neural operator (FNO) approximation space

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

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

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

[2] Kovachki, Lanthaler, Mishra, *"On Universal Approximation and Error Bounds for Fourier Neural Operators"*, (2021)

Operators of interest

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

Fourier neural operator (FNO) approximation space

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

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

- 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 $\mathcal{G} \in \mathcal{A}^\gamma(\mathit{FNO})$ be efficiently approximated in terms of sample complexity?

[2] Kovachki, Lanthaler, Mishra, *"On Universal Approximation and Error Bounds for Fourier Neural Operators"*, (2021)

- Consider unit ball $\mathcal{B}^\gamma \subset \mathcal{A}^\gamma(FNO)$, $\mathcal{G} \in \mathcal{B}^\gamma$.
- **Empirical risk minimizer:** Fix $\Psi(\,\cdot\,;\theta)$ Fourier neural operator architecture,

$$\mathcal{G} \approx \Psi(\,\cdot\,;\theta_{\mathcal{G}}), \qquad \theta_{\mathcal{G}} := \operatorname*{argmin}_{\theta} \frac{1}{N}\sum_{j=1}^{N} \|\mathcal{G}(u_j) - \Psi(u_j;\theta)\|^2.$$

Theorem (Kovachki, SL '24)

For any N, there exist sample points $u_1, \dots, u_N$, 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

$$\underbrace{\sup_{\mathcal{G}\in\mathcal{B}^\gamma} \mathbb{E}_{u\sim\mu}\left[\|\mathcal{G}(u) - \Psi(u;\theta_{\mathcal{G}})\|^2\right]^{1/2}}_{\textit{worst-case error of ERM}} \lesssim N^{-\frac{1}{2}\frac{\gamma}{\gamma+8}}$$

- Consider unit ball $\mathcal{B}^\gamma \subset \mathcal{A}^\gamma(FNO)$, $\mathcal{G} \in \mathcal{B}^\gamma$.
- **Empirical risk minimizer:** Fix $\Psi(\,\cdot\,;\theta)$ Fourier neural operator architecture,

$$\mathcal{G} \approx \Psi(\,\cdot\,;\theta_{\mathcal{G}}), \qquad \theta_{\mathcal{G}} := \operatorname*{argmin}_{\theta} \frac{1}{N}\sum_{j=1}^{N} \|\mathcal{G}(u_j) - \Psi(u_j;\theta)\|^2.$$

Theorem (Kovachki, SL '24)

For any N, there exist sample points $u_1, \dots, u_N$, 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

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

Summary

Data complexity

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

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

	Model complexity	**Data complexity**	
	sup-norm	sup-norm	L^p-norm
Lipschitz operators	size $\gtrsim \exp(c\epsilon^{-\lambda})$	$N \gtrsim \exp(c\epsilon^{-\lambda})$	$N \gtrsim \exp(c\epsilon^{-\lambda})$
(Fréchet-)C^k operators	size $\gtrsim \exp(c\epsilon^{-\lambda})$		$N \gtrsim \exp(c\epsilon^{-\lambda})$
"Natural" operators	$\underbrace{\text{size}(\Psi(\,\cdot\,;\theta)) \lesssim \epsilon^{-\gamma}}_{\text{by definition}}$	(exponential?)	$N \lesssim \epsilon^{-\gamma^*}$