Residual-Based Error Correction for Neural Operator Accelerated Infinite-Dimensional Bayesian Inverse Problems

Lianghao Cao

with Thomas O'Leary-Roseberry, Prashant K. Jha, J. Tinsley Oden, Omar Ghattas

The University of Texas at Austin
Oden Institute for Computational Engineering and Sciences

MS17: Recent Advances in Learning

SIAM TX-LA, 11/06/2022

Motivation: Surrogate modeling for complex physical systems

Tasks for decision-making requires many queries of PDE models:

- uncertainty quantification
- optimization
- infinite-dimensional input–output pairs: $(\mathcal{M}, \|\cdot\|_{\mathcal{M}}, \nu)$ and $(\mathcal{U}, \|\cdot\|_{\mathcal{U}})$

An abstract nonlinear variational problem

$$\text{Given } m \in \mathcal{M}, \text{ find } u \in \mathcal{V}_u \subset \mathcal{U} \text{ such that } \underbrace{\mathcal{R}(u,m)}_{\text{PDE residual}} = 0 \in \mathcal{U}_0^*,$$

Full scale PDE solves can be too expensive for these settings:

- high dimensional
- highly nonlinear PDEs

Demand cheap surrogate models of the input–output map $\mathcal{F} : \mathcal{M} \to \mathcal{U}$

$$\mathcal{R}(\mathcal{F}(m), m) = 0 \quad \nu\text{-a.e.}$$

Operator learning using neural network

Operator learning in Bochner spaces

$$\inf_{\boldsymbol{w} \in \mathcal{W} \subset \mathbb{R}^{n_w}} \mathcal{J}(\boldsymbol{w}) := \left\| \mathcal{F} - \widetilde{\mathcal{F}}_{\boldsymbol{w}} \right\|_{\mathcal{T}}$$

$$\mathcal{T} := L^p(\mathcal{M}, \nu; \mathcal{U}) \quad \text{with} \quad \|\mathcal{G}\|_{L^p(\mathcal{M},\nu;\mathcal{U})} = \begin{cases} \left(\mathbb{E}_{M\sim\nu}\left[\|\mathcal{G}(M)\|_{\mathcal{U}}^p\right]\right)^{1/p} & p \in [1,\infty) \\ \operatorname{ess\,sup}_{M\sim\nu} \|\mathcal{G}(M)\|_{\mathcal{U}} & p = \infty \end{cases}$$

- **Neural operator:** neural network (NN) representation of the approximate map $\widetilde{\mathcal{F}}_{\boldsymbol{w}}$

  Key: reduced basis representation of $\widetilde{\mathcal{F}}_{\boldsymbol{w}} : \mathcal{M}^r \to \mathcal{U}^r$

- **NN architectural designs with theoretical guarantees of expressive power:**
  - General multilayer: (Hornik, Stinchcombe, & White, 1989)
  - Fourier neural operator: (Kovachki et al., 2021)
  - POD-NN: (Bhattacharya, Hosseini, Kovachki, & Stuart, 2021)
  - DeepONet: (Lu, Jin, Pang, & Karniadakis, 2021)
  - Derivative-informed neural networks: (O'Leary-Roseberry, 2020)

...but

a significant gap between theory and computation, leading to unreliability

Neural operator training

- Finite dimensional representations: $\mathcal{M}^h \subset \mathcal{M}$, $\mathcal{U}^h \subset \mathcal{U}$, and ν^h
- $\mathcal{F}^h : \mathcal{M}^h \to \mathcal{U}^h$: Full scale numerical PDE solves.

Empirical risk minimization

$$\min_{\boldsymbol{w}\in\mathcal{W}} \widetilde{\mathcal{J}}(\boldsymbol{w}; \{m_j\}_{j=1}^{n_{\text{train}}}) := \frac{1}{n_{\text{train}}} \sum_{j=1}^{n_{\text{train}}} \left\| u_j - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(m_j; \boldsymbol{w}) \right\|_{\mathcal{U}}^p$$

$$\left\{ \left(m_j, u_j = \mathcal{F}^h(m_j) \right) \middle| m_j \sim \nu^h \right\}_{j=1}^{n_{\text{train}}},$$

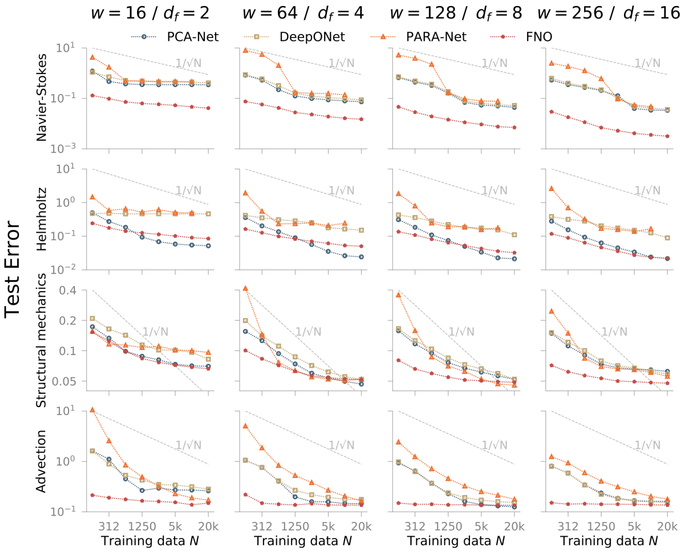
Roughly speaking...

$$\underbrace{\text{Approximation error}}_{>\text{theoretical bound}} = \underbrace{\text{Loss estimation error}}_{\text{Finite samp. and discret.}} + \underbrace{\text{Truncation error}}_{\text{Reduced basis}} + \text{Neural network error}$$

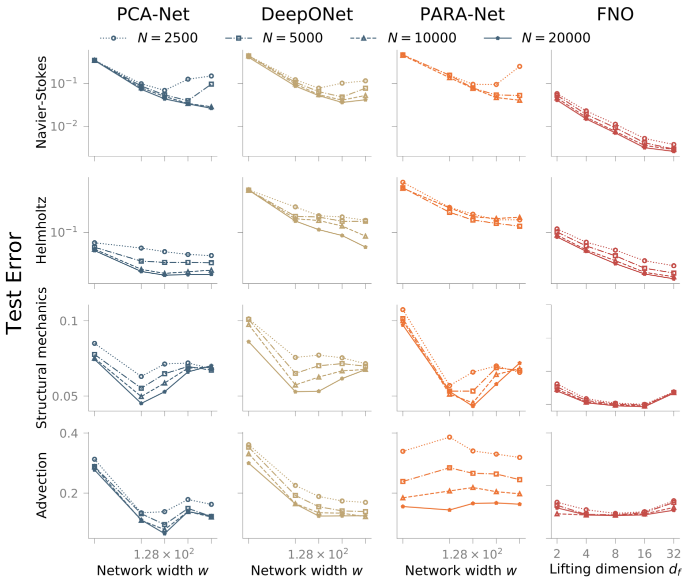
Origin of the neural network error

Interplay between finite samples, nonconvex optimization, and practical architectural designs. Difficult to reduce and control!

Some results from (De Hoop, Huang, Qian, & Stuart, 2022)



Some results from (De Hoop et al., 2022)



Take away: Unreliability in error reduction

More data and more expressive power in theory

$\not\Downarrow$ (does not imply)

Smaller approximation error of trained neural operators.

Improving neural operators in training to meet accuracy requirements of many query applications can be demanding and even intractable.

Outline

- Infinite-dimensional Bayesian inverse problems (BIPs)
- Propagation of approximation error in BIPs
- Residual-based error correction
- Numerical examples

Infinite-dimensional Bayesian inverse problem

- A set of observed data: $\boldsymbol{y} \in \mathbb{R}^{n_y}$
- A discrete observation operator (nonlinear): $\boldsymbol{\mathcal{B}} : \mathcal{U} \to \mathbb{R}^{n_y}$
- Additive noise model: $\boldsymbol{y} = \boldsymbol{\mathcal{B}}(u) + \boldsymbol{n}$, unknown noise corruption.
- $\boldsymbol{N} \sim \nu_N$ (random observation noise), $M \sim \nu_M$ (random function, prior, typically Gaussian)

Data model

$$\boldsymbol{Y} = (\boldsymbol{\mathcal{B}} \circ \mathcal{F})(M) + \boldsymbol{N}$$

Bayes' rule

Given a set of observed data $\boldsymbol{y}^*$ and the data model:

$$\frac{\mathrm{d}\nu_{M|\boldsymbol{Y}}(\cdot|\boldsymbol{y}^*)}{\mathrm{d}\nu_M}(m) = \frac{1}{Z(\boldsymbol{y}^*)} \underbrace{\pi_{\boldsymbol{N}}\left(\boldsymbol{y}^* - (\boldsymbol{\mathcal{B}} \circ \mathcal{F})(m)\right)}_{=:\, \mathcal{L}(m;\boldsymbol{y}^*)} \quad \text{a.s.}$$

Propagation of approximation error

Question: What happens to the BIPs if a learned operator $\widetilde{\mathcal{F}}_{\boldsymbol{w}}$ is used in likelihood evaluations?

$$\boldsymbol{Y} \mapsto \widetilde{\boldsymbol{Y}} = (\boldsymbol{\mathcal{B}} \circ \widetilde{\mathcal{F}}_{\boldsymbol{w}})(M) + \boldsymbol{N} \implies \mathcal{L}(m; \boldsymbol{y}^*) \mapsto \widetilde{\mathcal{L}}(m; \boldsymbol{y}^*) \coloneqq \pi_{\boldsymbol{N}}\left(\boldsymbol{y}^* - (\boldsymbol{\mathcal{B}} \circ \widetilde{\mathcal{F}}_{\boldsymbol{w}})(m)\right)$$

with

$$\mathcal{E}(m) \coloneqq \mathcal{F}(m) - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(m) \quad \nu_M\text{-a.e.} \quad \text{and} \quad \mathcal{E}_{\text{post}} \coloneqq \mathbb{E}_{\nu_{M|\widetilde{\boldsymbol{Y}}}(\cdot|\boldsymbol{y}^*)}\left[\ln\left(\frac{\mathrm{d}\nu_{M|\widetilde{\boldsymbol{Y}}}(\cdot|\boldsymbol{y}^*)}{\mathrm{d}\nu_{M|\boldsymbol{Y}}(\cdot|\boldsymbol{y}^*)}\right)\right]$$

An *a prior* error bound

$$\mathcal{E}_{\text{post}} \leq c_{\text{BIP}} \|\mathcal{E}\|_{L^p(\mathcal{M}, \nu_M; \mathcal{U})}, \quad p \in [2, \infty]$$

Given the typical infinite-dim. BIP setting:

- Gaussian noise $\nu_{\boldsymbol{N}} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{\boldsymbol{N}})$
- well-behaving observations $\boldsymbol{\mathcal{B}}$
- mutual absolute continuity of prior and posteriors

...however, the constant matters a lot for irreducible $\|\mathcal{E}\|_{L^p(\mathcal{M}, \nu_M; \mathcal{U})}$

Propagation of approximation error, cont.

- $\left\| (\boldsymbol{\mathcal{B}} \circ \mathcal{F})(m) - (\boldsymbol{\mathcal{B}} \circ \widetilde{\mathcal{F}}_{\boldsymbol{w}})(m) \right\|_2 \leq c_L \left\| \mathcal{E}(m) \right\|_{\mathcal{U}} \quad \nu_M$-a.e.

Bounding constant

$$\mathcal{E}_{\text{post}} \leq c_{\text{BIP}} \left\| \mathcal{E} \right\|_{L^p(\mathcal{M},\nu_M;\mathcal{U})} \qquad c_{\text{BIP}} = c_1 \left(c_2(1) + c_3(p) \right) c_L$$

$$c_1 = \frac{1}{2} \left\| \boldsymbol{C}_{\boldsymbol{N}}^{-1} \left((\boldsymbol{\mathcal{B}} \circ \mathcal{F})(\cdot) + (\boldsymbol{\mathcal{B}} \circ \widetilde{\mathcal{F}}_{\boldsymbol{w}})(\cdot) - 2\boldsymbol{y}^* \right) \right\|_{L^p(\mathcal{M},\nu_M;\mathbb{R}^{n_y})}$$

$$c_2(p) = \left\| \exp \left(-\frac{1}{2} \left\| \boldsymbol{y}^* - (\boldsymbol{\mathcal{B}} \circ \widetilde{\mathcal{F}}_{\boldsymbol{w}})(\cdot) \right\|^2_{\boldsymbol{C}_{\boldsymbol{N}}^{-1}} \right) \right\|^{-1}_{L^p(\mathcal{M},\nu_M)}$$

$$c_3(p) = \frac{c_2(1)}{c_2(q)} \in [1, c_2(1)] \quad q = \begin{cases} \infty & p = 2 \\ p/(p-2) & p \in (2, \infty) \\ 1 & p = \infty \end{cases}$$

The conditioning of BIPs is bad in the situations with

uninformative prior, high-dimensional data, small noise corruption, or inadequate models.

Take away: Accuracy requirements in training for BIPs

$$\mathcal{E}_{\text{post}} \leq c_{\text{BIP}} \|\mathcal{E}\|_{L^p(\mathcal{M},\nu_M;\mathcal{U})}$$

- The accuracy requirements should be based on the need of the BIPs, and can be demanding or even intractable for neural operator training.
- Goal-oriented analysis is more useful.

$$\mathcal{E}_{\text{QoI}} \leq c \|\mathcal{E}\|_{L^p(\mathcal{M},\nu_M;\mathcal{U})} \ldots (?)$$

error in training $\Longrightarrow$ error in posterior (inverse) $\Longrightarrow$ error in QoI prediction (forward)

Neural operator with error correction

A well-trained neural operator makes good predictions, but not arbitrarily good

(trained neural operator)	$\mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) \neq 0$
(PDE)	$\mathcal{R}(\mathcal{F}(m), m) = 0$
(error)	$\mathcal{E}(m) := \mathcal{F}(m) - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(m) \quad \nu_M$-a.e.

What about linearizing the PDE residual...

$$\underbrace{\mathcal{R}(\mathcal{F}(m), m)}_{=0} = \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) + \delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m)\underbrace{(\mathcal{F}(m) - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(m))}_{=\mathcal{E}(m)} + \dots$$

$$\delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) \underbrace{\mathcal{F}(m)}_{\text{unknown}} \approx -\mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) + \delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m)\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m)\,.$$

Neural operator with error correction $\widetilde{\mathcal{F}}_C : m \mapsto u_C$

Given $m \in \mathcal{M}$, find $u_C \in \mathcal{V}_u$ such that:

$$\delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) u_C = -\mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) + \delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m)\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m)$$

Error correction as a single Newton step

Neural operator with error correction $\widetilde{\mathcal{F}}_C : m \mapsto u_C$

Given $m \in \mathcal{M}$, find $u_C \in \mathcal{V}_u$ such that:

$$\delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) u_C = -\mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) + \delta_u \mathcal{R}(\widetilde{\mathcal{F}}_{\boldsymbol{w}}(m), m) \widetilde{\mathcal{F}}_{\boldsymbol{w}}(m)$$

The correction step as a Newton step may lead to global (over prior) quadratic error reduction

Corollary to the Newton–Kantorovich theorem in Banach spaces

$$\mathcal{E}(m) \coloneqq \mathcal{F}(m) - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(m) \quad \nu_M\text{-a.e.}\,, \quad \mathcal{E}_C(m) \coloneqq \mathcal{F}(m) - \widetilde{\mathcal{F}}_C(m) \quad \nu_M\text{-a.e.}$$

There exists a global convergence radius $r_C > 0$ such that

$$\|\mathcal{E}\|_{L^\infty(\mathcal{M},\nu_M;\mathcal{U})} < r_C \implies \|\mathcal{E}_C\|_{L^\infty(\mathcal{M},\nu_M;\mathcal{U})} \leq c_R \|\mathcal{E}\|^2_{L^\infty(\mathcal{M},\nu_M;\mathcal{U})}$$

Given regularity, continuity, and boundedness conditions on the PDE residual and its derivative.

Numerical example: Deformation of hyperelastic materials

Strong form PDE model:

$$
\begin{aligned}
\nabla \cdot (\boldsymbol{F}(\boldsymbol{X})\boldsymbol{S}(\boldsymbol{X}, m(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X}))) &= \boldsymbol{0} & \boldsymbol{X} &\in \Omega_0 \\
\boldsymbol{u}(\boldsymbol{X}) &= \boldsymbol{0} & \boldsymbol{X} &\in \Gamma_l \\
\boldsymbol{F}(\boldsymbol{X})\boldsymbol{S}(\boldsymbol{X}, m(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) \cdot \boldsymbol{n} &= \boldsymbol{0} & \boldsymbol{X} &\in \Gamma_t \cup \Gamma_b \\
\boldsymbol{F}(\boldsymbol{X})\boldsymbol{S}(\boldsymbol{X}, m(\boldsymbol{X}), \boldsymbol{C}(\boldsymbol{X})) \cdot \boldsymbol{n} &= \boldsymbol{t}(\boldsymbol{X}) & \boldsymbol{X} &\in \Gamma_r
\end{aligned}
$$

$$
\boldsymbol{t}(\boldsymbol{X}) = a \exp\left(-\frac{|X_2 - 0.5|^2}{b}\right)\boldsymbol{e}_1 + c\left(1 + \frac{X_2}{d}\right)\boldsymbol{e}_2
$$

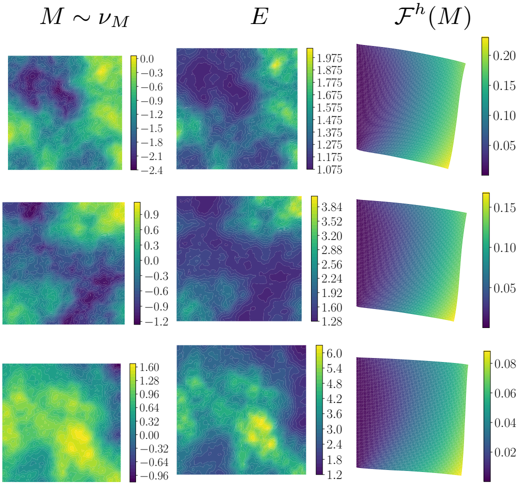
Discretization via FEniCS (Alnæs et al., 2015):

- $\mathcal{M}^h \subset L^2(\Omega_0)$: linear triangular FE, 4225 DoFs
- $\mathcal{U}^h \subset H^1(\Omega_0; \mathbb{R}^2)$: linear triangular FE, 8450 DoFs

Gauss. prior via hIPPYlib (Villa, Petra, & Ghattas, 2018):

$$
E = \exp(M) + 1 \quad M \sim \nu_M := \mathcal{N}(0.37, \mathcal{C}_{\text{pr}})
$$

$$
\mathcal{C}_{\text{pr}} = (\alpha I + \beta \Delta)^{-2}
$$



Neural operator training and accuracy

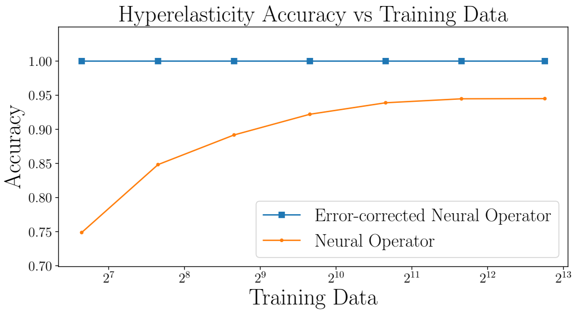
Neural network architecture: (O'Leary-Roseberry, 2020; O'Leary-Roseberry et al., 2021)

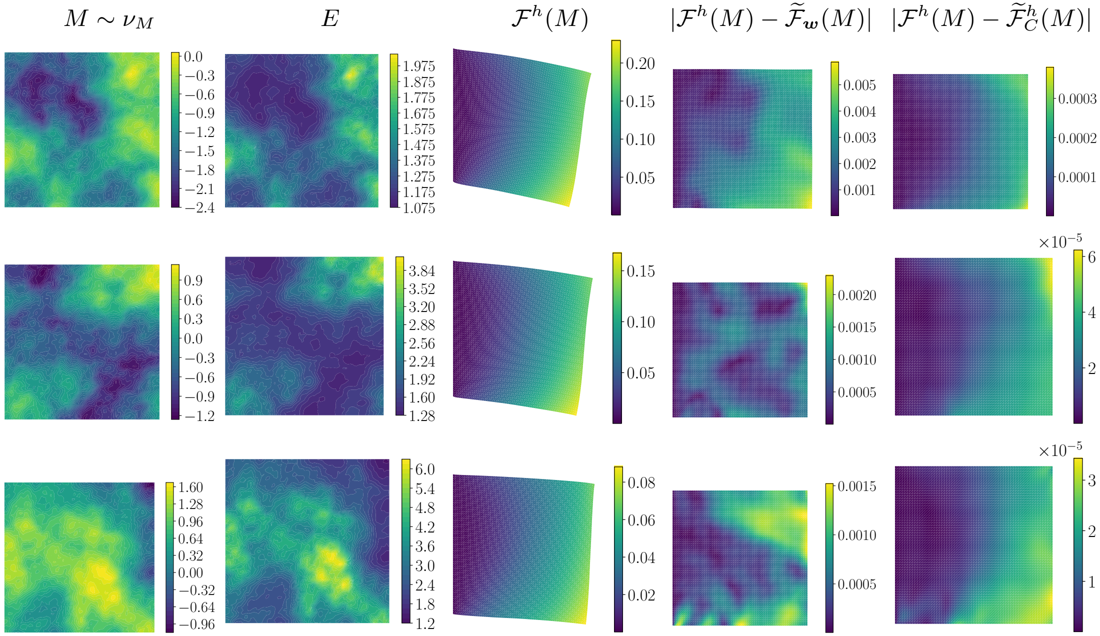
- 50-dim. derivative-informed reduced basis of $\mathcal{M}^h$
- 50-dim. POD reduced basis of $\mathcal{U}^h$
- 10 nonlinear ResNet layers, each with a rank of 10

Adaptive neural network training: (O'Leary-Roseberry et al., 2021)

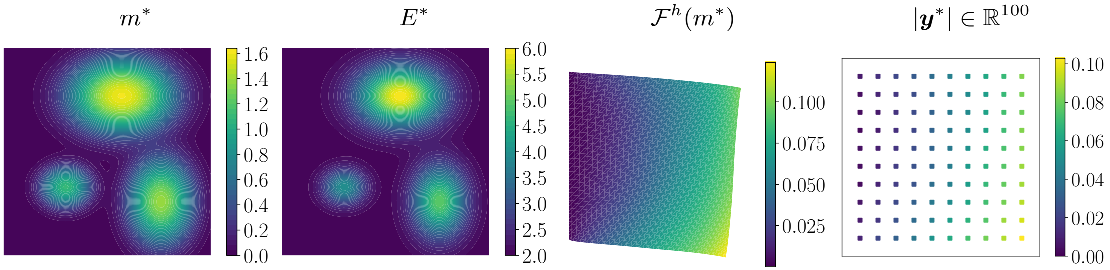
- $p = 2$ in the Bochner norm.
- varying training data size: $n_{\text{train}} = 100, 201, 403, 806, 1612, 3225, 6912.$
- 256 samples for accuracy estimate

$$\text{Accuracy} = 100\left(1 - \mathbb{E}_{M\sim\nu_M}\left[\frac{\left\|\mathcal{F}(M) - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(M)\right\|_{\mathcal{U}}}{\|\mathcal{F}(M)\|_{\mathcal{U}}}\right]\right)$$




$M \sim \nu_M$
E
$\mathcal{F}^h(M)$
$|\mathcal{F}^h(M) - \widetilde{\mathcal{F}}_{\boldsymbol{w}}(M)|$
$|\mathcal{F}^h(M) - \widetilde{\mathcal{F}}^h_C(M)|$
0.0 −0.3 −0.6 −0.9 −1.2 −1.5 −1.8 −2.1 −2.4
1.975 1.875 1.775 1.675 1.575 1.475 1.375 1.275 1.175 1.075
0.20 0.15 0.10 0.05
0.005 0.004 0.003 0.002 0.001
0.0003 0.0002 0.0001
0.9 0.6 0.3 0.0 −0.3 −0.6 −0.9 −1.2
3.84 3.52 3.20 2.88 2.56 2.24 1.92 1.60 1.28
0.15 0.10 0.05
0.0020 0.0015 0.0010 0.0005
$\times 10^{-5}$
6 4 2
1.60 1.28 0.96 0.64 0.32 0.00 −0.32 −0.64 −0.96
6.0 5.4 4.8 4.2 3.6 3.0 2.4 1.8 1.2
0.08 0.06 0.04 0.02
0.0015 0.0010 0.0005
$\times 10^{-5}$
3 2 1

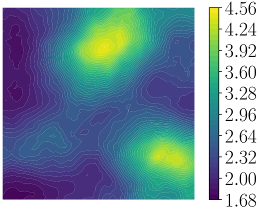
Bayesian inverse problems: Setting



- Posterior sampling using preconditioned Crank–Nicolson (MCMC) (Cotter, Roberts, Stuart, & White, 2013)
- Additive white noise with std of 1% max value of $\boldsymbol{y}^*$

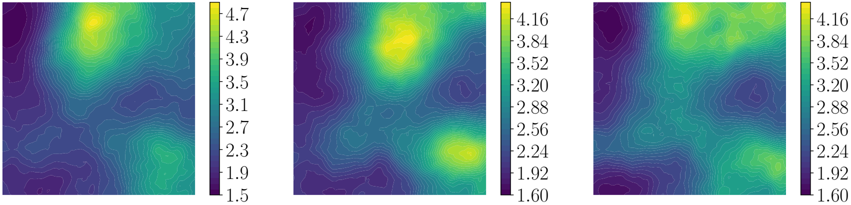
Bayesian inverse problems: Posterior mean Young's modulus

Model $\mathcal{F}^h$



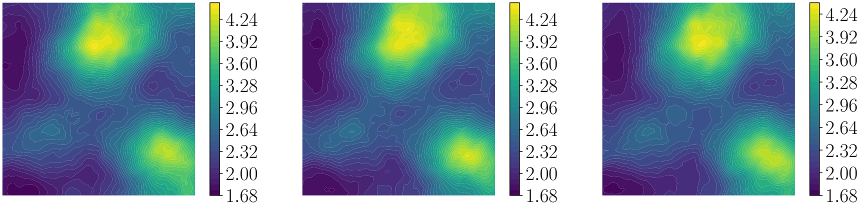
Neural operator $\widetilde{\mathcal{F}}_w$

$n_{\text{train}} = 1612$ | $n_{\text{train}} = 3225$ | $n_{\text{train}} = 6912$

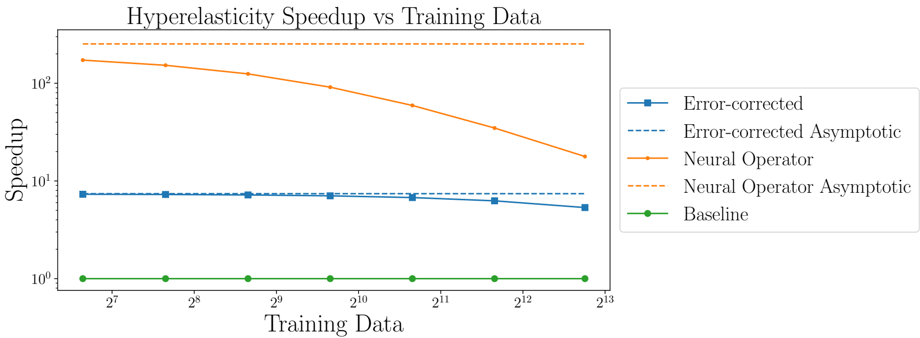


Neural operator with error correction $\widetilde{\mathcal{F}}_C$

$n_{\text{train}} = 1612$ | $n_{\text{train}} = 3225$ | $n_{\text{train}} = 6912$



Comparison of computational speedups



Conclusion

We are interested in...

- using neural operators to accelerate consequential many-query applications of PDEs

The challenges to address are...

- difficulty in reducing and controlling its approximation error in training
- understanding the implication of the approximation error in decision-making

What we did (or tried)...

- establish theoretical understanding of the error propagation in BIPs
- using residual-based error correction to drastically (quadratically) reduce the approximation error

Link to the pre-print paper:

Contact

Lianghao Cao
lianghao@oden.utexas.edu