NEURAL INFORMATION PROCESSING SYSTEMS

Understanding the Expressivity and Trainability of Fourier Neural Operator: A Mean-Field Perspective

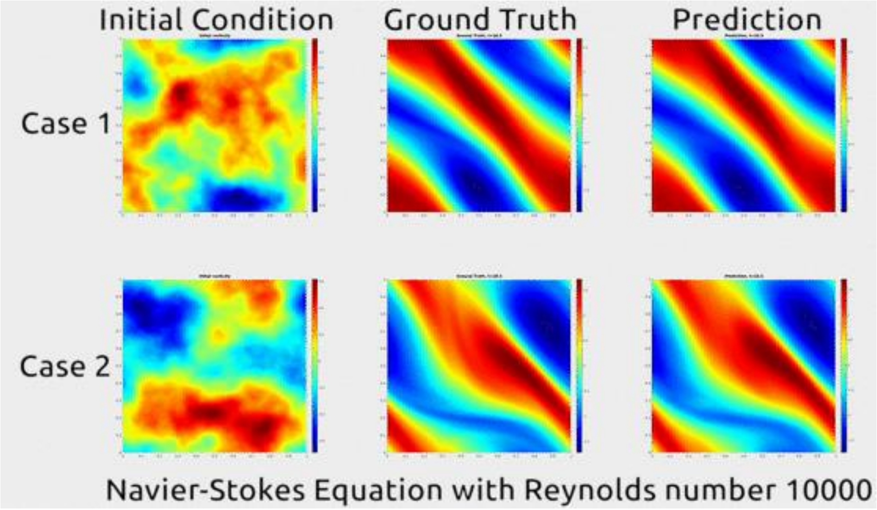
Takeshi Koshizuka (The University of Tokyo)

Masahiro Fujisawa (RIKEN AIP), Yusuke Tanaka (NTT)

Issei Sato (The University of Tokyo)

UTokyo

Background



Machine Learning for PDE

- Accelerate the solution process
- Enhance the precision of solutions using observation
- Enable solutions across a wide range of conditions and parameters

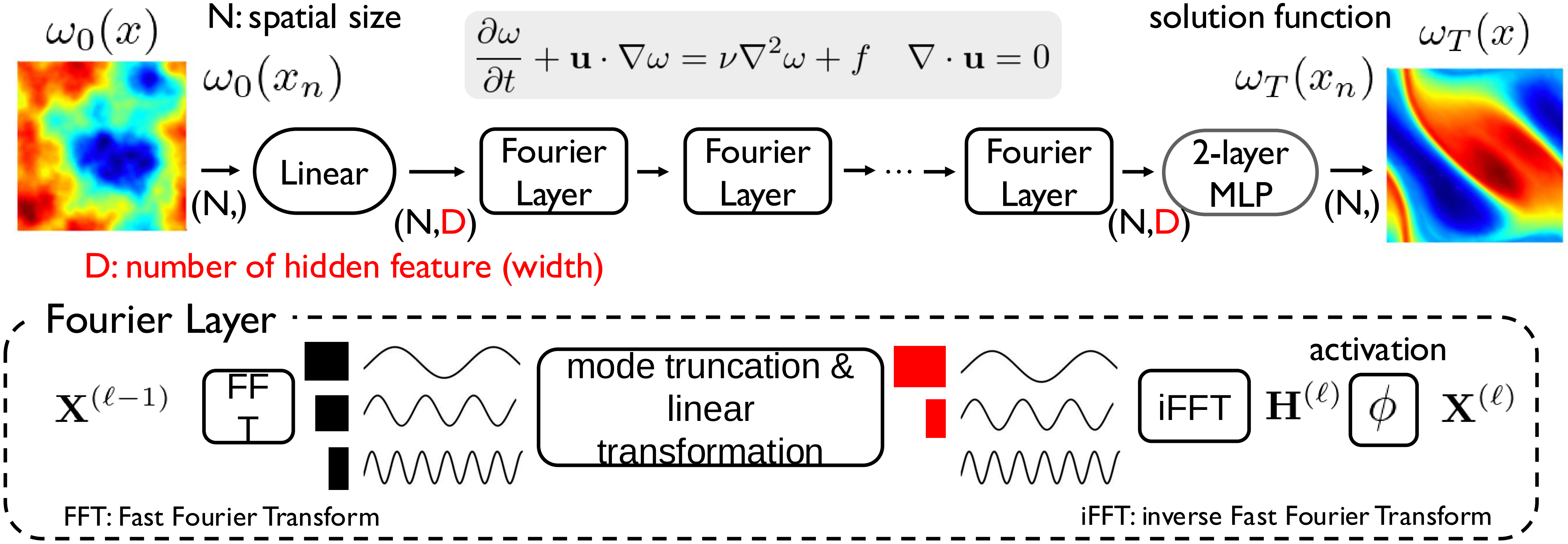
Neural Operators [Kovachiki & Li et al.'21] discretization-invariant

Integral Operator σ_ℓ : non-linear activation

$$v_{t+1}(x) = \sigma\left(Wv_t(x) + \int_D G_\theta(x, y, a(x), a(y))v_t(y)\mathrm{d}\nu_x(y)\right)$$

Model the integral kernel (Green's function) G_θ with a neural network and stack multiple layers

Fourier Neural Operator (FNO) [Zongyi et al '20]



The deep FNO has poor performance.

e.g. instability of initial training

[Lu et al. '20, Tran et al. '23]

Analyze **expressivity** & **trainability** from a mean-field perspective

Infinite-width FNO at initialization

<u>Fourier Layer</u> $\mathbf{X}^{(\ell-1)} \mapsto \mathbf{X}^{(\ell)}$

concatenation

learnable parameters

$$\mathbf{X}^{(\ell)} = \phi(\mathbf{H}^{(\ell)}),\ \mathbf{H}^{(\ell)} = \mathrm{iFFT}\left(\Big\Vert_{k=0}^{N-1} \mathrm{FFT}(\mathbf{X}^{(\ell-1)})_{k,:} \left(\boldsymbol{\Theta}^{(\ell,k)} + \sqrt{-1}\,\boldsymbol{\Xi}^{(\ell,k)} \right) \right) + \mathbf{1}_N \mathbf{b}^{(\ell)}$$

linear transformation

<u>Initialization</u>

tunable hyperparameters

$$\Theta_{i,j}^{(\ell,k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{2D}\right),\ \Xi_{i,j}^{(\ell,k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{2D}\right)\ (0 \le k \le K-1),\ b_i^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma_b^2\right)$$

D: number of hidden feature (width)

<u>Infinite width</u> central limit theorem

signals

$$\mathbf{H}_{:,1}^{(\ell)}, \mathbf{H}_{:,2}^{(\ell)}, \ldots, \mathbf{H}_{:,D}^{(\ell)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{(\ell)}\right)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{H}_{:,1}^{(\ell)}}, \frac{\partial \mathcal{L}}{\partial \mathbf{H}_{:,2}^{(\ell)}}, \ldots, \frac{\partial \mathcal{L}}{\partial \mathbf{H}_{:,D}^{(\ell)}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}^{(\ell)}\right)$$

<u>Mode truncation</u>

$$\boldsymbol{\Theta}^{(\ell,k)} = \boldsymbol{\Xi}_{i,j}^{(\ell,k)} = \mathbf{0} \quad \text{for } K \le k \le N-1-K$$

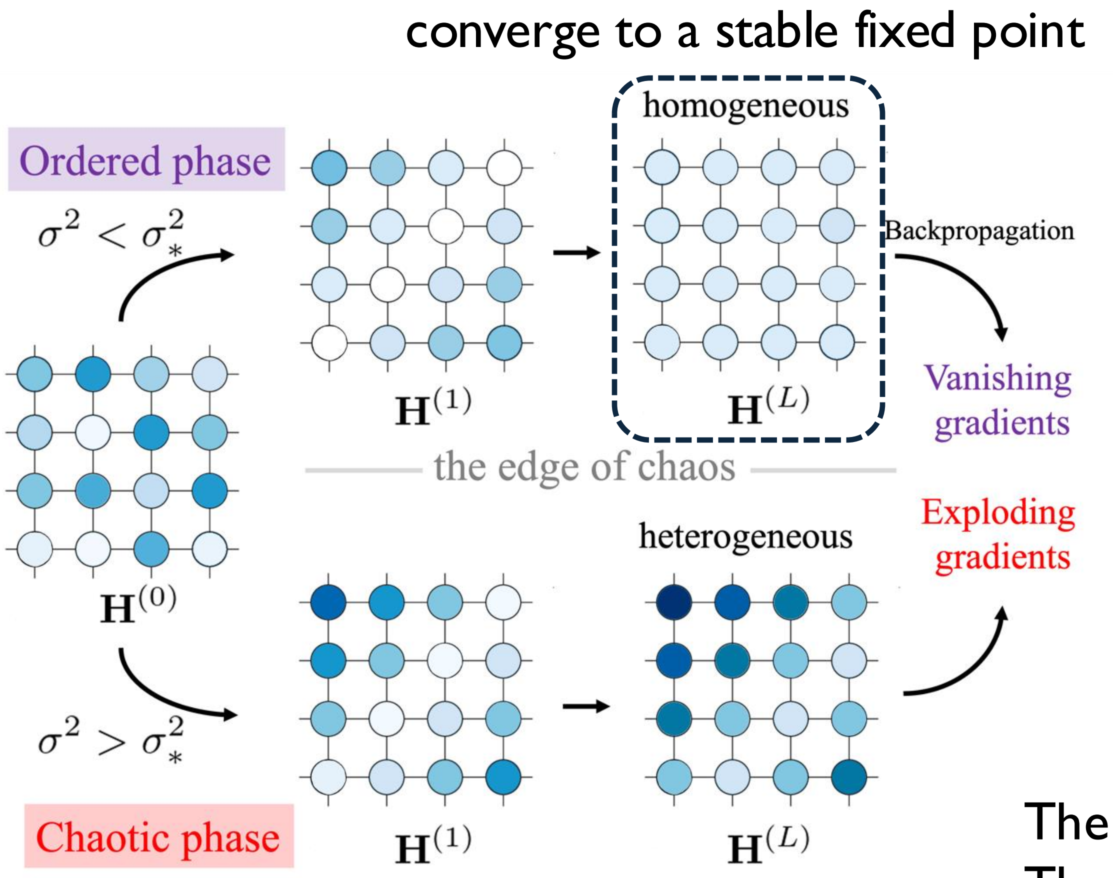
Analyze the dynamics

$$\boldsymbol{\Sigma}^{(1)} \xrightarrow{\mathcal{C}} \ldots \xrightarrow{\mathcal{C}} \boldsymbol{\Sigma}^{(\ell-1)} \xrightarrow{\mathcal{C}} \boldsymbol{\Sigma}^{(\ell)}$$

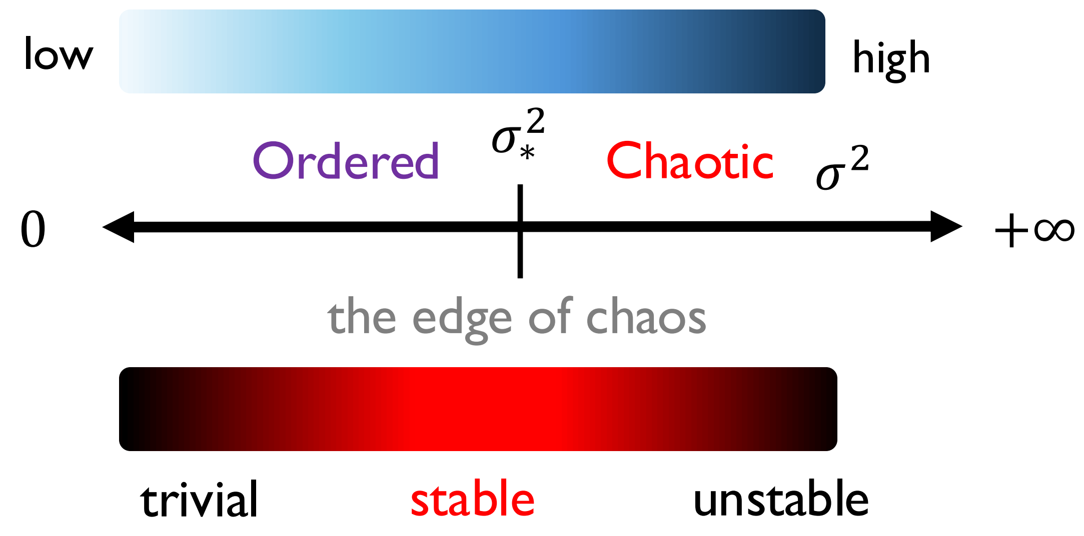
$$\tilde{\boldsymbol{\Sigma}}^{(\ell)} \xleftarrow{\tilde{\mathcal{C}}} \tilde{\boldsymbol{\Sigma}}^{(\ell+1)} \xleftarrow{\tilde{\mathcal{C}}} \ldots \xleftarrow{\tilde{\mathcal{C}}} \tilde{\boldsymbol{\Sigma}}^{(L)}$$

Main Results

Ordered-Chaos phase transition for the weight initialization parameter σ^2



Expressivity $\operatorname{diag} \boldsymbol{\Sigma}^{(\ell)} \sim e^{\ell \log C\sigma^2}, \;\; C : \text{const}$



Trainability $\operatorname{diag} \tilde{\boldsymbol{\Sigma}}^{(\ell)} \sim e^{(L-\ell) \log C\sigma^2}$

The constant C is determined by the activation function.
The transition point σ_*^2 is obtained by $\sigma_*^2 = 1/C$.