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

Takeshi Koshizuka
Department of Computer Science
The University of Tokyo
koshizuka-takeshi938444@g.ecc.u-tokyo.ac.jp

Masahiro Fujisawa
RIKEN AIP
masahiro.fujisawa@riken.jp

Yusuke Tanaka
NTT Communication Science Laboratories
ysk.tanaka@ntt.com

Issei Sato
Department of Computer Science
The University of Tokyo
sato@g.ecc.u-tokyo.ac.jp

Abstract

In this paper, we explores the expressivity and trainability of the Fourier Neural Operator (FNO). We establish a mean-field theory for the FNO, analyzing the behavior of the random FNO from an *edge of chaos* perspective. Our investigation into the expressivity of a random FNO involves examining the ordered-chaos phase transition of the network based on the weight distribution. This phase transition demonstrates characteristics unique to the FNO, induced by mode truncation, while also showcasing similarities to those of densely connected networks. Furthermore, we identify a connection between expressivity and trainability: the ordered and chaotic phases correspond to regions of vanishing and exploding gradients, respectively. This finding provides a practical prerequisite for the stable training of the FNO. Our experimental results corroborate our theoretical findings.

1 Introduction

The recent surge in interest in solving partial differential equations (PDEs) has led to the use of neural network (NN)-based surrogate models. One promising line of work is the neural operator (NO), which learns the solution operator of PDEs, thereby bypassing the need for mesh dependency. Among the variants of NO, the Fourier neural operator (FNO) (Li et al., 2020c) has gained popularity because of its advantageous cost/accuracy trade-off. The FNO can capture long-distance spatial interactions using the Fourier transform, whereas convolutional neural networks (CNNs) (Wen et al., 2019; Jiang et al., 2021b) and message-passing graph neural networks (GNNs) (Li et al., 2020a,b) are limited to operating solely on local variables. From a computational cost perspective, the Fourier transform is performed in quasi-linear time by the fast Fourier transform (FFT), making it significantly faster than the Transformer (Li et al., 2023).

Despite the widespread use of FNO as an architecture, there is a lack of comprehensive theoretical analysis on its expressivity and trainability. The universal approximation property (Kovachki et al., 2021), recognized as the basic expressivity of the FNO, is well-known; however, the exponential expressivity depending on the weight distribution, which are known for the densely connected network (DCN) (Schoenholz et al., 2016), a.k.a. fully connected network, and CNN (Xiao et al., 2018), remains unexplored. Regarding the trainability of FNO, the training instability in deep FNO has been experimentally reported by Tran et al. (2022), but the causes and conditions of the difficulty have not been clarified either theoretically or experimentally.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-0352

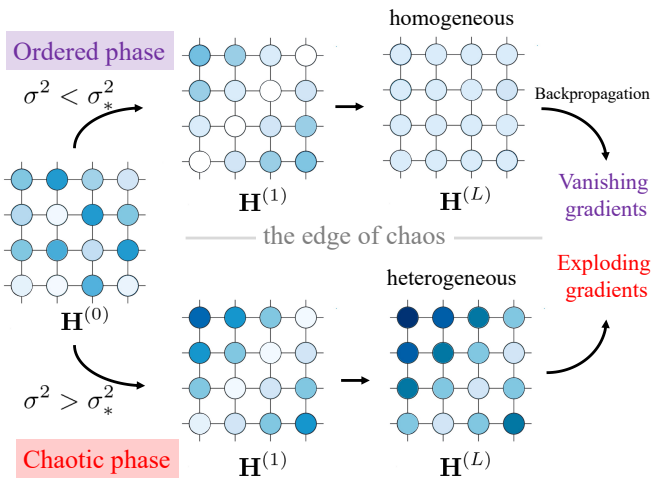


Figure 1: Illustration of ordered-chaos phase transition for the weight initialization parameter σ^2. In the ordered phase, the spatial hidden representations $\mathbf{H}^{(\ell)}$ on the grid converge to a uniform state during forward propagation and the gradient vanishes during backpropagation. In the chaotic phase, the representations either converge to a distinct state or diverge and the gradient explodes.

We analyze the exponential expressivity (how far can two different input vectors be pulled apart) and trainability (how much gradient explosion on average) of the random FNO from the perspective of whether the network is *ordered* or *chaotic*. This viewpoint is grounded in mean-field theory, an analytical framework for NN established by Poole et al. (2016); Schoenholz et al. (2016); Yang & Schoenholz (2017); Xiao et al. (2018). A network is considered ordered when it brings all representations of two different spatial positions closer together, and chaotic when it drives them apart during forward propagation. Furthermore, a network can only be stably trained when initialized close to the *edge of chaos*, which is the transition point between the ordered phase and the chaotic phase. In fact, He initialization (He et al., 2015) is an example of a commonly used edge of chaos initialization for the DCN with ReLU activation (Burkholz & Dubatovka, 2019).

In this study, we establish a mean-field theory to analyze the expressivity and trainability of the FNO. Our investigation reveals the expressivity of random FNO at initialization by examining the transition point between ordered and chaos phases. The phase transition exhibits FNO-specific characteristics induced by mode truncation, as well as similarities with the characteristics of DCN and CNN. We also find a link between expressivity and trainability: the ordered and chaotic phases correspond to regions of vanishing and exploding gradient, respectively. This discovery offers a practical initialization prerequisite for the stable training of the FNO.

2 Background

2.1 Fourier Neural Operators

The FNO (Li et al., 2020c) is one of the well-established methods for solving PDEs across many scientific problems (Yang et al., 2021; Wen et al., 2022b; Hwang et al., 2022; Jiang et al., 2021a; Pathak et al., 2022). An M-dimensional FNO with the number of hidden features D and the spatial size N for learning the operators between scalar-valued functions is defined as follows.

$$\mathbf{X}^{(\ell+1)} = \phi\left(\mathcal{D}^{(\ell)}\left(\mathbf{X}^{(\ell)}\right) + \mathcal{K}^{(\ell)}\left(\mathbf{X}^{(\ell)}\right)\right), \tag{1}$$

where $\mathbf{X}^{(\ell)} \in \mathbb{R}^{\overbrace{N\times\cdots\times N}^{M}\times D}$ is the ℓ-th hidden representation, M is the number of spatial dimensions, and ϕ is activation. The hidden representations $\mathbf{X}^{(0)}$ and $\mathbf{X}^{(L)}$ are the output of the lifting operator and the input of the projection operator, respectively. The architecture of these operators does not affect our analysis as long as the network stays shallow, as implemented in (Li et al., 2020c). The ℓ-th densely connected (DC) module $\mathcal{D}^{(\ell)}$ is an affine point-wise map in the physical space and the ℓ-th

Fourier convolution module $\mathcal{K}^{(\ell)}$ is a parameterized kernel integral operator using Fourier transform. The bias term is considered to be included in either or both modules $\mathcal{K}^{(\ell)}$ and $\mathcal{D}^{(\ell)}$.

Several FNO variants have been developed to address specific challenges, such as geo-FNO (Li et al., 2022) for irregular regions and group equivariant FNO (G-FNO) (Helwig et al., 2023), which maintains equivariance to rotation, reflection, and translation. U-NO (Rahman et al., 2022) and U-FNO (Wen et al., 2022a) integrate FNO with U-Net for multiscale modeling. Additionally, WNO (Tripura & Chakraborty, 2022) utilizes wavelet bases, while CFNO (Brandstetter et al., 2023) enhances the use of geometric relations between different fields and field components through Clifford algebras. Adaptive FNO (Zhao et al., 2022; Guibas et al., 2021) and F-FNO (Tran et al., 2022) have improved computational and memory efficiency through incremental learning and architectural modifications. Other approaches for improving performance include methods with increasing physical inductive bias (Li et al., 2024), data augmentation (Brandstetter et al., 2022), and a variance-preserving weight initialization scheme (Poli et al., 2022).

While numerous new models and learning methods have been proposed, relatively little research has been conducted to understand the intrinsic nature of these methods. Issues such as spectral bias (Zhao et al., 2022) and training instability (Tran et al., 2022) have been reported. Tran et al. (2022) observed that training did not converge even at 24 layers. They successfully addressed the stability and accuracy degradation issues associated with an increase in the number of layers by implementing skip connections behind activation and introducing various training techniques. However, it is still unknown that the theoretical basis for why the original architecture of the FNO has problems with training instability and accuracy degradation.

2.2 Mean-field Theory for Neural Networks

The mean-field theory has been used to provide a mathematical framework for understanding the expressivity and trainability of neural networks (Poole et al., 2016; Schoenholz et al., 2016; Yang & Schoenholz, 2017; Hayou et al., 2018; Xiao et al., 2018). A series of papers (Poole et al., 2016; Schoenholz et al., 2016) delved into the average behavior of infinite-width random deep DCN, with weights and biases initialized by a zero-mean Gaussian distribution. The formulation is given below.

$$\begin{aligned} &\boldsymbol{x}^{(\ell)} = \phi\left(\boldsymbol{h}^{(\ell)}\right),\ \boldsymbol{h}^{(\ell)} = \boldsymbol{W}^{(\ell)}\boldsymbol{x}^{(\ell-1)} + \boldsymbol{b}^{(\ell)}, \\ &W_{i,j}^{(\ell)} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{D}\right),\ b_i^{(\ell)} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_b^2\right), \end{aligned} \tag{2}$$

where $\boldsymbol{x}^{(\ell)} \in \mathbb{R}^D$ is the ℓ-th hidden representation, $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{D\times D}$, $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^D$ are the ℓ-th learnable parameters, and the width is assumed to be sufficiently large $D \gg 1$.

Poole et al. (2016) and Schoenholz et al. (2016) explored the exponential expressivity of random DCN determined by two phases depending on the initial variance parameters σ^2 and σ_b^2, as shown in Fig. 5a. Poole et al. (2016) first examined the forward propagation of a random DCN with Tanh activation. They demonstrated that the covariance $\boldsymbol{\Sigma}^{(\ell)}$ of the ℓ-th pre-activation representations $\boldsymbol{h}^{(\ell)}$ and $\tilde{\boldsymbol{h}}^{(\ell)}$ corresponding to two different inputs $\boldsymbol{x}^{(0)}$ and $\tilde{\boldsymbol{x}}^{(0)}$ are obtained by

$$\forall d \in [D],\ \boldsymbol{\Sigma}^{(\ell)} = \sigma^2 \mathbb{E}\left[\phi\left(h_d^{(\ell-1)}\right)\phi\left(\tilde{h}_d^{(\ell-1)}\right)\right] + \sigma_b^2,$$

where the expectation is taken over the pre-activations $[h_d^{(\ell-1)}, \tilde{h}_d^{(\ell-1)}] \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^{(\ell-1)})$. The covariance converges exponentially to a fixed point $\boldsymbol{\Sigma}^*$ determined by parameters σ^2 and σ_b^2.

A network is considered *ordered* when it brings two distinct representations closer together, which implies a state of small expressivity. Conversely, a network is *chaotic* when it drives them apart during forward propagation, implying a state of large expressivity. Networks with either excessively small or large expressivity can disrupt the structure of the input: the difference between two distinct inputs quickly becomes indistinguishable in networks with small expressivity, while similarities between inputs are no longer recognized in networks with large expressivity. The network is *ordered* if the initial variance of the weights is small. For larger values, and beyond a certain threshold, the phase shifts, and the network behaves chaotically. This phase shift point is termed *the edge of chaos*.

Subsequently, Schoenholz et al. (2016) discovered the connection between expressivity and trainability of DCN through analysis of the backpropagation behavior. In an ordered network, the expected

https://doi.org/10.52202/079017-0352

value of the gradient norm becomes exponentially small during backpropagation, while it becomes exponentially large in a chaotic network. This implies that the gradient vanishes/explodes in ordered or chaotic networks, respectively. These findings suggest that deep DCN can be stably trained only near the edge of chaos. Schoenholz et al. (2016) also provided an estimate of the maximum depth at which a network can be trained when initialized away from the edge of chaos. These insights are not limited to DCN and similar results have been observed for residual networks (Yang & Schoenholz, 2017) and CNN (Xiao et al., 2018).

3 A mean-field theory for FNO

In this section, we establish a mean-field theory of FNO. We demonstrate the exponential expressivity of random FNO by examining the ordered-chaos phase transition during the forward propagation. Furthermore, we identify the connection between expressivity and trainability by concentrating on backward propagation behaviors. Our analysis is an advanced version of the approach developed by Poole et al. (2016); Schoenholz et al. (2016); Yang & Schoenholz (2017); Xiao et al. (2018). In Section 3.1, we outline the problem setup. In Section 3.2, we analyze the forward and backward propagation behavior of random FNO at initialization. In Section 3.3, we discuss the practical prerequisites for initialization to stabilize the training of FNO, leveraging the similarities between FNO and DCN. The proofs for all the lemmas and theorems are provided in Appendices A and B.

3.1 Problem setting

Here, we consider a simplified one-dimensional (1D) FNO. Note that our theory is extensively applicable to the original FNO, as discussed in Section 3.3. The simplified 1D FNO, with a depth of L, is defined by the number of hidden features D, a spatial size $N = 2^m$ (where m is an integer), the number of Fourier modes $K \leq \frac{N}{2} + 1$, two learnable weights $\mathbf{\Theta}^{(\ell,k)} \in \mathbb{R}^{D\times D}$ and $\mathbf{\Xi}^{(\ell,k)} \in \mathbb{R}^{D\times D}$, and a bias $\mathbf{b}^{(\ell)} \in \mathbb{R}^D$. Denote $\phi\colon \mathbb{R} \to \mathbb{R}$ by the non-decreasing activation function. Let $\mathbf{X}^{(\ell)} \in \mathbb{R}^{N\times D}$ and $\mathbf{H}^{(\ell)} \in \mathbb{R}^{N\times D}$ be the post and pre-activation representations defined by

$$\begin{aligned} \mathbf{X}^{(\ell)} = \phi\left(\mathbf{H}^{(\ell)}\right),\ \mathbf{H}^{(\ell)} = \sum_{k=0}^{K-1} \sqrt{\frac{c_k}{2}} \left(\mathbf{H}^{(\ell,k)} + \overline{\mathbf{H}}^{(\ell,k)}\right) + \mathbf{b}^{(\ell)} \mathbf{1}_N^\top, \\ \mathbf{H}^{(\ell,k)} := \boldsymbol{F}^\dagger \boldsymbol{D}^{(k)} \boldsymbol{F} \mathbf{X}^{(\ell-1)} \left(\mathbf{\Theta}^{(\ell,k)} + \sqrt{-1}\mathbf{\Xi}^{(\ell,k)}\right), \end{aligned} \tag{3}$$

where $\delta_{a,b}$ is the Kronecker-delta, $c_k = 2 - \delta_{k,0} - \delta_{k,N/2}$ is a constant, $\mathbf{1}_N$ is all-ones column vector with the size N, $\overline{\mathbf{H}}^{(\ell,k)}$ is the conjugate of $\mathbf{H}^{(\ell,k)}$ corresponding to the $(N-k)$-th frequency components, $\dagger$ is the transpose conjugate, $\boldsymbol{F} \in \mathbb{C}^{N\times N}$ is the Discrete Fourier Transform (DFT) matrix defined by $F_{k,n} = \frac{1}{N}\exp(-\frac{2\pi k}{N}n)$, and $\boldsymbol{D}^{(k)}$ is a diagonal matrix with a 1 at position $D^{(k)}_{k,k}$.

There are two differences from the original FNO proposed by Li et al. (2020a): (1) the DC module is dropped for the simplicity, and (2) $\mathbf{H}^{(\ell,k)}$ is multiplied by $\sqrt{2}$ with respect to $k = 0, \frac{N}{2}$ for appropriate normalization. We assume that the weights of FNO are initialized by i.i.d. samples from Gaussian distribution, *i.e.* $\Theta^{(\ell,k)}_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{\sigma^2}{2D})$, $\Xi^{(\ell,k)}_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0, \frac{\sigma^2}{2D})$, $b^{(\ell)}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_b^2)$. For $k = 0, \frac{N}{2}$, the parameter $\mathbf{\Xi}^{(\ell,k)}$ is set to zero exceptionally. For all $d \in [D] = \{0, \ldots, D-1\}$, the pre-activations $\mathbf{H}^{(\ell)}_{:,d} \in \mathbb{R}^N$ are i.i.d. random variables. When $D \gg 1$, by the central limit theorem, the variables $\mathbf{H}^{(\ell)}_{:,d}$ follow Gaussian distribution with mean 0 and covariance matrix $\Sigma^{(\ell)}_{\alpha,\alpha'} := \mathbb{E}_{\Theta^{1:\ell},\Xi^{1:\ell}}\left[H^{(\ell)}_{\alpha,d} H^{(\ell)}_{\alpha',d}\right]$, where the expectation is taken over all random variables $[\mathbf{\Theta}^{1:\ell}, \mathbf{\Xi}^{1:\ell}] := \{\mathbf{\Theta}^{(\ell',k')}, \mathbf{\Xi}^{(\ell',k')}\}_{\ell'\in[\ell],k'\in[K]}$. Our theory can be easily extended to 2D and 3D FNOs.

3.2 Expressivity and trainability of FNO

Firstly, the forward propagation of a single input signal with spatial features is described as follows.

Lemma 3.1 (Iterated map). *For all $d \in [D]$, the covariance $\mathbf{\Sigma}^{(\ell)} := \mathbb{E}_{\Theta^{1:\ell},\Xi^{1:\ell}}\left[\mathbf{H}_{:,d}^{(\ell)}\,\mathbf{H}_{:,d}^{(\ell)\top}\right]$ is obtained recursively by the iterated map $\mathcal{C}$ defined by*

$$\Sigma_{\alpha,\alpha'}^{(\ell)} = \underbrace{\sigma^2 \sum_{k=0}^{K-1} c_k \mathbb{E}\left[\left|\left[\boldsymbol{F}\phi\left(\mathbf{H}_{:,d}\right)\right]_k\right|^2\right] \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) + \sigma_b^2}_{=:\,\mathcal{C}(\mathbf{\Sigma}^{(\ell-1)})_{\alpha,\alpha'}}, \tag{4}$$

where the expectation is taken over the pre-activations $\mathbf{H}_{:,d} \sim \mathcal{N}(0, \mathbf{\Sigma}^{(\ell-1)})$, $\theta_{\alpha,\alpha'}^{(k)} := \frac{2\pi k}{N}(\alpha - \alpha')$ represents the scaled positional difference.

The indices α and α' correspond to different spatial locations as with the mean-field theory for CNN (Xiao et al., 2018). Note that $[\boldsymbol{F}\phi\left(\mathbf{H}_{:,d}\right)]_k$ is the k-th Fourier modes of the post-activation representation. When applying DCN (Poole et al., 2016; Schoenholz et al., 2016) or CNN (Xiao et al., 2018) to the spatial signal, the iterated map depends only on local spatial locations, while in the case of FNO, the iterated map depends on all spatial locations because of the global Fourier convolution. In addition, only periodic spatial correlations with shift-invariant are propagated, and high-frequency components exceeding mode K are eliminated.

Next, we explore the fixed point $\mathbf{\Sigma}^*$ of the iterated map $\mathcal{C}$ satisfying $\mathbf{\Sigma}^* = \mathcal{C}(\mathbf{\Sigma}^*)$. By linearizing the dynamics of signal propagation around this fixed point and analyzing the stability and rate of convergence to the fixed point, we can determine the depth to which each component of the input can propagate. Schoenholz et al. (2016) showed that the iterated map of DCN defined in Eq. (2) has a fixed point of the form:

$$\mathbf{\Sigma}^* = q^* \boldsymbol{I}_N + q^* c^* (\mathbf{1}_N \mathbf{1}_N^\top - \boldsymbol{I}_N), \tag{5}$$

where q^*, c^* are the fixed points of variance and correlation, and $\boldsymbol{I}_N$ is the identity matrix. Meanwhile, Xiao et al. (2018) showed that any fixed point for the iterated map of the DCN is also a fixed point for that of CNN. We show that random FNO has the same fixed points of the form of Eq. (5) with $c^* = 1$ in the following lemma.

Lemma 3.2 (Exsistance of fixed points). *When a random DCN defined by Eq.* (2) *has the fixed point $(q^*, c^* = 1)$ for the initial parameters (σ^2, σ_b^2), then a random simplified FNO defined by Eq.* (3) *has a fixed point $\mathbf{\Sigma}^*$ of the form*

$$\mathbf{\Sigma}^* = q^* \boldsymbol{I}_N + q^* c^* (\mathbf{1}_N \mathbf{1}_N^\top - \boldsymbol{I}_N) = q^* \mathbf{1}_N \mathbf{1}_N^\top.$$

Lemma 3.2 indicates that the fixed point for the iterated map $\mathbf{\Sigma}^*$ of the DCN serves as a fixed point for the iterated map of the simplified FNO (as well as CNN). To analyze the stability and convergence rate, we linearly approximate the C-map around the fixed point $\mathbf{\Sigma}^*$, i.e., $\mathcal{C}(\mathbf{\Sigma}) \approx \mathbf{\Sigma}^* + J_{\mathbf{\Sigma}^*}(\mathbf{\Sigma} - \mathbf{\Sigma}^*)$, where $J_{\mathbf{\Sigma}^*}$ is the Jacobian linear map of the iterated map defined in Eq. (15). We then derive the eigenvalues and eigenvectors for the Jacobian linear map $J_{\mathbf{\Sigma}^*}$ as follows.

Definition 3.3.

$$\chi_{q^*} := \sigma^2 \mathbb{E}\left[\phi'^2(H_{\alpha,d}) + \phi''(H_{\alpha,d})\phi(H_{\alpha,d})\right], \tag{6}$$

$$\chi_{c^*} := \sigma^2 \mathbb{E}[\phi'(H_{\alpha,d})\phi'(H_{\alpha',d})], \tag{7}$$

$$\chi_\kappa := \frac{\sigma^2}{2}\mathbb{E}\left[\phi''(H_{\alpha,d})\phi(H_{\alpha',d}) + \phi(H_{\alpha,d})\phi''(H_{\alpha',d})\right] + \sigma^2 \mathbb{E}\left[c^* \phi'(H_{\alpha,d})\phi'(H_{\alpha',d})\right], \tag{8}$$

where the expectation is taken over the pre-activations $\mathbf{H}_{:,d} \sim \mathcal{N}(0, \mathbf{\Sigma}^*)$, and $\phi',\ \phi''$ are the first- and second-order derivatives of the activation ϕ.

The bases $\boldsymbol{\psi},\ \boldsymbol{\psi}^{(1)},\ \ldots,\ \boldsymbol{\psi}^{(K-1)} \in \mathbb{R}^{N\times N}$ using the above quantities are defined below.

$$\begin{aligned} \psi_{\beta,\beta'} &:= 1 - \frac{1}{N}\left(\frac{\chi_\kappa + \chi_{c^*} - \chi_{q^*}}{\chi_\kappa}\right)\sum_{s=0}^{K-1} c_s \cos\left(\theta_{\beta,\beta'}^{(s)}\right), \\ \psi_{\beta,\beta'}^{(k)} &:= \cos\left(\theta_{\beta,\beta'}^{(k)}\right) - \frac{1}{\sum_{s=0}^{K-1} c_s}\sum_{s=0}^{K-1} c_s \cos\left(\theta_{\beta,\beta'}^{(s)}\right). \end{aligned} \tag{9}$$

https://doi.org/10.52202/079017-0352

From Lemma A.4, $K-1$ matrices in $\{\boldsymbol{\psi}^{(k)}\}_{k\in[K]\setminus\{0\}}$ are eigenbases with the eigenvalue χ_{c^*} of the Jacobian linear map. From Lemma A.5, the matrix $\boldsymbol{\psi}$ is the eigenbases with the eigenvalue χ of the Jacobian linear map. Since the rank of the Jacobian linear map is at most K (Lemma A.3), the deviation from the fixed point $\boldsymbol{\Sigma}^{(\ell)}-\boldsymbol{\Sigma}^*$ is spanned by K-dimensional eigenspace $\operatorname{span}\left(\{\boldsymbol{\psi}^{(k)}\}_{k\in[K]\setminus\{0\}}\cup\{\boldsymbol{\psi}\}\right)$. Then, the fixed point stability and the convergence rate are shown in the following theorem.

Theorem 3.4 (Exponential expressivity)**.** *Let $\boldsymbol{E}^{(\ell)}\coloneqq\boldsymbol{\Sigma}^{(\ell)}-\boldsymbol{\Sigma}^*$ be the deviation from the fixed point at the ℓ-th layer. Suppose that the deviation at the first layer is decomposed as $\boldsymbol{E}^{(0)}=\epsilon\boldsymbol{\psi}+\sum_{k=1}^{K-1}\epsilon_k\boldsymbol{\psi}^{(k)}+\boldsymbol{e}$. The scalars $\epsilon,\ \epsilon_k$ represent the scale of the perturbation for each eigencomponent of the linearly approximated map $\boldsymbol{E}^{(\ell)}\mapsto\boldsymbol{E}^{(\ell+1)}$. The component $\boldsymbol{e}\in\mathbb{R}^{N\times N}$ belongs to the orthogonal complements of the space* $\operatorname{span}\left(\{\boldsymbol{\psi},\boldsymbol{\psi}^{(1)},\ \ldots,\ \boldsymbol{\psi}^{(K-1)}\}\right)$.

Then, the deviation at the ℓ-th layer is obtained by

$$\boldsymbol{E}^{(\ell)}=\chi^{\ell}\epsilon\boldsymbol{\psi}+\sum_{k=1}^{K-1}\chi_{c^*}^{\ell}\epsilon_k\boldsymbol{\psi}^{(k)}, \tag{10}$$

$$\chi\coloneqq\frac{1}{N}\sum_{s=0}^{K-1}c_s\chi_{q^*}+\left(1-\frac{1}{N}\sum_{s=0}^{K-1}c_s\right)(\chi_\kappa+\chi_{c^*}).$$

In particular, when the Fourier mode is $K=\frac{N}{2}+1$, Eqs. (9) *and* (10) *reduce to the following.*

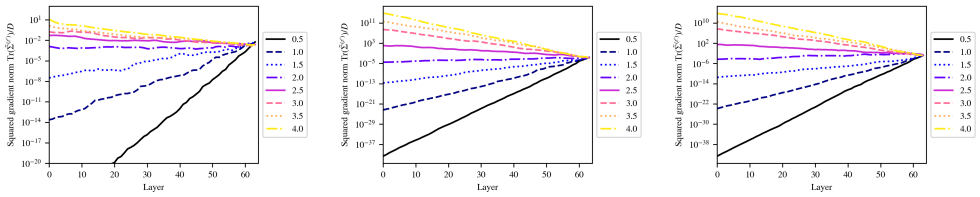
$$\begin{gathered}\boldsymbol{E}^{(\ell)}=\chi_{q^*}^{\ell}\epsilon\boldsymbol{\psi}+\sum_{k=1}^{K-1}\chi_{c^*}^{\ell}\epsilon_k\boldsymbol{\psi}^{(k)},\\ \forall\beta,\beta'\in[N],\ \psi_{\beta,\beta'}=1,\ \psi_{\beta,\beta'}^{(k)}=\cos\left(\theta_{\beta,\beta'}^{(k)}\right)-\delta_{\beta,\beta'}.\end{gathered} \tag{11}$$

Theorem 3.4 shows the expressivity of the FNO, which is characterized by the ordered-chaos phases and varies exponentially with respect to the number of layers. Theorem 3.4 indicates that the asymptotic behavior of the zero-frequency deviation is mostly determined by χ and the periodic deviation is determined by χ_{c^*}. If $\chi<1$ and $\chi_{c^*}<1$, the fixed point is stable as the deviation from the fixed point converges exponentially to zero. When the fixed point remains stable at $c^*=1$, a random network exists in an ordered phase, where all spatial representations are correlated in an asymptotic manner. Conversely, when the fixed point with $c^*=1$ becomes unstable, the network transitions into a chaotic phase, exhibiting behavior dependent on the activation function ϕ. The boundary between these two phases is referred to as *the edge of chaos*.

The convergence rates χ_{q^*} and χ_{c^*} are the same as the convergence rates of the variance and correlation to the fixed point for DCN (Schoenholz et al., 2016) and CNN (Xiao et al., 2018). However, only periodic spatial correlations are propagated in the FNO, resulting in a different eigenspace of the map $\boldsymbol{E}^{(\ell)}\mapsto\boldsymbol{E}^{(\ell+1)}$ compared to the DCN and CNN. In DCN, the deviation belongs to a vector space with dimension $\frac{N(N-1)}{2}$ in DCN, whereas in FNO, the dimension is K, or at most $\frac{N}{2}+1$. CNN possess diagonal eigenspaces associated with eigenvalues χ_{q^*} and non-diagonal eigenspaces associated with eigenvalues χ_{c^*}. In contrast, FNOs without mode truncation exhibit a similarity, possessing eigenspaces χ_{q^*} for zero-frequency and eigenspaces χ_{c^*} for k-frequencies with diagonal components removed. Furthermore, mode truncation increases the convergence rate of zero-frequency deviation from χ_{q^*} to χ and affects all eigenbases as well. For further discussions on the similarities between CNN and FNO, please refer to Appendix C. A visualization of the covariance of the FNO with Tanh and ReLU activations is shown in Appendix F.

Finally, we demonstrate the connection between expressivity and trainability. By examining the covariance of the gradient in each layer during backpropagation, we investigate the conditions under which training is stable without gradient vanishing or exploding.

Theorem 3.5 (Trainability)**.** *Let $\tilde{\boldsymbol{\Sigma}}^{(\ell)}\in\mathbb{R}^{N\times N}$ be the gradient covariance with respect to some loss $\mathcal{L}$,* e.g. *mean squared error, at the ℓ-th layer. Suppose that the gradient covariance at the L-th layer is decomposed as $\tilde{\Sigma}_{\alpha,\alpha'}^{(L)}=\sum_{k=0}^{K-1}\tilde{\epsilon}_k\cos\left(\theta_{\alpha,\alpha'}^{(k)}\right)+\tilde{\boldsymbol{e}}$, where $\tilde{\epsilon}_k$ is the coefficient of each basis and $\tilde{\boldsymbol{e}}$*



(a) Simplified FNO with Tanh (b) Simplified FNO with ReLU (c) Original FNO with ReLU

Figure 2: Average gradient norm $\mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}^{(\ell)})/D$ during the backpropagation of several FNOs plotted as a function of layer ℓ. Each line corresponds to the result of different initial values of σ^2 from 0.5 to 4.0 in increments of 0.5. The x-axis is the layer and the y-axis is the log-scale of the gradient norm. Depending on the value of σ^2, the gradient norm increases or decreases consistently as the gradient propagates to shallower layers.

belongs to the orthogonal complements of $\mathrm{span}(\{\cos\left(\theta^{(k)}_{\alpha,\alpha'}\right)\}_{k=0}^{K-1})$. *Then, the gradient covariance at the ℓ-th layer is obtained by*

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \sum_{k=0}^{K-1} \chi_{c^*}^{L-\ell} \tilde{\epsilon}_k \cos\left(\theta^{(k)}_{\alpha,\alpha'}\right).$$

Theorem 3.5 shows that gradient vanishing occurs when $\chi_{c^*} < 1$ (ordered phase) and gradient explosion occurs when $\chi_{c^*} > 1$ (chaos phase). Thus, stable training of the FNO can be achieved close to *the edge of chaos* by setting the initial parameter σ^2 to satisfy $\chi_{c^*} \approx 1$. We specifically present the initial parameter choices that achieve *the edge of chaos* for several FNOs in Section 3.3.

When $c^* = 1$, there is no change in the dynamics during backpropagation due to mode truncation. When using the full mode $K = \frac{N}{2} + 1$, the condition $\chi_{c^*} = 1$ always achieves *the edge of chaos*, which is consistent with the results for the DCN and CNN. Despite the architectural and iterative map differences among FNO, DCN, and CNN, Theorem 3.4 and Theorem 3.5 demonstrate the similarities in the random behavior of FNO, DCN, and CNN. This allows existing results based on mean-field theory to be applied to the FNO.

3.3 Initialization requirements for stable training

For stable training, Theorem 3.5 suggests the necessity of initializing FNO near *the edge of chaos*, *i.e.*, initializing FNO so that $\chi_{c^*} \approx 1$. In this section, we present the initial parameter choices that achieve *the edge of chaos* for several FNOs, each with slightly different architectures such as activation functions. Furthermore, the behavior of the gradient norm $\mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}^{(\ell)})/D$ as a function of layer ℓ are visualized in Fig. 2 for several variants of random FNO with different initialization parameters σ^2. We used FNO with a width of $D = 32$ and a number of layers $L = 64$, and for simplicity, we computed the absolute value of the output as the loss with respect to the input sampled from the standard normal distribution.

Simplified FNO with Tanh activation.
The behavior of χ_{c^*} for the parameters σ^2 and σ_b^2 of the Tanh network has been extensively studied by Poole et al. (2016); Schoenholz et al. (2016). The phase diagram drawn by Pennington et al. (2017) is shown in Fig. 5a. By using parameters (σ^2, σ_b^2) around the two phase boundaries of ordered and chaotic that achieve $\chi_{c^*} = 1$, the training of the simplified FNO with Tanh activation can be stabilized. Figure 2a depicts the behavior of the gradient backpropagation in the simplified FNO with Tanh activation and the bias parameter being $\sigma_b^2 = 0.1$. Figure 2a shows that when $\sigma^2 \lessapprox 2$, the gradient diminishes exponentially; otherwise, it explodes exponentially.

https://doi.org/10.52202/079017-0352

Simplified FNO with ReLU activation.
The iterated map $\mathcal{C}$ of the DCN with ReLU activation is given by Cho & Saul (2009) as follows.

$$q^{(\ell+1)} = \frac{1}{2}\sigma^2 q^{(\ell)} + \sigma_b^2, \tag{12}$$

$$c^{(\ell+1)} q^{(\ell+1)} = \frac{1}{2}\sigma^2 q^{(\ell)} \mathbb{J}_1\left(\frac{c^{(\ell)} q^{(\ell)}}{q^{(\ell)}}\right) + \sigma_b^2, \tag{13}$$

$$\mathbb{J}_1(c) = \frac{1}{\pi}\left(\sqrt{1-c^2} + (\pi - \arccos(c))c\right),$$

The edge of chaos initialization for the DCN with ReLU activation is known as He initialization (He et al., 2015), which sets the initial variance parameter as $\sigma^2 = 2$ and initial bias as $b_i^{(\ell)} = 0$ for Eq. (2). From the similarity of the DCN and the FNO, we can derive the FNO version of the He initialization that achieves $\chi_{c^*} = 1$ by setting $\sigma^2 = 2,\ b_i^{(\ell)} = 0$ for Eq. (3). The He initialization scheme for the simplified FNO with the activation $\phi = \mathrm{ReLU}$ is derived as follows.

$$\Theta_{i,j}^{(\ell,k)} \overset{i.i.d.}{\sim} \mathcal{N}(0, D^{-1}),\ \Xi_{i,j}^{(\ell,k)} \overset{i.i.d.}{\sim} \mathcal{N}(0, D^{-1}).$$

Figure 2b demonstrates that the choice of $\sigma^2 = 2$ preserves the magnitude of the gradient norm during backpropagation of deep simplified FNO with ReLU activation.

Original FNO.
In the original architecture of the FNO proposed by Li et al. (2020c), the DC module is used together with the Fourier convolution module as shown in Eq. (1). We initialize the weights of both layers consistently as follows. For all $\ell \in [L]$, $k \in [K]$, and $i, j \in [D]$,

$$\Theta_{i,j}^{(\ell,k)} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{4D}\right),\ \Xi_{i,j}^{(\ell,k)} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{4D}\right),$$

$$W_{i,j}^{(\ell)} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \frac{\sigma^2}{2D}\right),\ b_i^{(\ell)} \sim \mathcal{N}\left(0, \sigma_b^2\right).$$

From the similarity of the initial network behavior of the FNO and the DCN, the fixed point with $c^* = 1$ of the simplified FNO is also a fixed point of the original FNO. In the neighborhood of the fixed point $\mathbf{\Sigma}^*$, the eigencomponents spanned by $\{\boldsymbol{\psi}^{(k)}\}_{k=1}^K$ will decay or increase at the rate of χ_{c^*}. Following the derivation of Theorem 3.5, the eigenvalues of the $\cos$ function eigencomponents of the gradient covariance are also χ_{c^*}. These results show that the edge of chaos initialization scheme can be used for the original FNO with each activation function. Figure 2c shows that the original FNO with ReLU activation and the parameter fixed as $b_i^{(\ell)} = 0$ exhibits similar backpropagation behavior as the simplified FNO, *i.e.*, $\sigma^2 \approx 2$ is an appropriate choice.

4 Experiments

In this section, we experimentally demonstrate that deep FNO training requires appropriate initialization settings on a variety of datasets, consistent with the theory discussed in Section 3.

4.1 Datasets

We evaluated three models on commonly used PDEs: the advection equation, Burgers' equation, Darcy Flow equation, and incompressible Navier-Stokes (NS) equation. All datasets were generated by numerical simulations used in (Takamoto et al., 2022; Li et al., 2020c) and are publicly available. A summary of the dataset is provided in Table 1, and more details are given in Appendix D.

Advection equation and Burgers' equation. The linear advection equation for the function $u(x,t)$ is given by

$$\partial_t u(x,t) + \beta \partial_x \left(u(x,t)/2\right) = 0,\ u(x,0) = u_0(x),$$

where u_0 is the initial condition and $\beta = 2.0$ is an advection speed. The non-linear Burgers' equation for the function $u(x,t)$ is given by

$$\partial_t u(x,t) + \partial_x \left(u^2(x,t)/2\right) = \nu \partial_{xx} u(x,t),\ u(x,0) = u_0(x),$$

Table 1: Overview of dataset with number of spatial dimensions M, time dependence, spatial resolution $N_s = N \times \cdots \times N$, temporal resolution N_t, and number of samples for training, validation, and testing.

PDE	M	time	N_s	N_t	Number of samples (Train / Val / Test)
advection	1	-	64	-	8000 / 1000 / 1000
Burgers'	1	-	64	-	8000 / 1000 / 1000
Darcy flow	2	-	64×64	-	900 / 100 / 100
Navier-Stokes ($\nu = 1e$-3)	2	✓	64×64	25	1000 / 100 / 100
Navier-Stokes ($\nu = 1e$-4)	2	✓	64×64	20	8000 / 1000 / 1000
Navier-Stokes ($\nu = 1e$-5)	2	✓	64×64	20	1000 / 100 / 100

where u_0 is the initial condition and $\nu = 4.0$ is the diffusion coefficient.

Darcy Flow equation. The Darcy Flow equation for the function $u(x)$ with a Dirichlet boundary is given by

$$-\nabla \cdot (a(x)\nabla u(x)) = f(x) \quad (x \in (0,1)^2),\ u(x) = 0 \quad (x \in \partial(0,1)^2),$$

where a is the diffusion coefficient and $f(x) = 1$ is the forcing function.

Incompressible Navier-Stokes equation. The 2D NS equation on the unit torus is defined by

$$\partial_t \omega(x,t) + u(x,t) \cdot \nabla \omega(x,t) = \nu \nabla^2 \omega(x,t) + f(x),\ \nabla \cdot u(x,t) = 0,\ \omega(x,0) = \omega_0(x),$$

where $\omega(x,t)$ is the vorticity, ω_0 is the initial vorticity, $u(x,t)$ is the velocity field for any $r > 0$, $\nu \in \mathbb{R}_+$ is the viscosity, and f is the external forcing function defined by $f(x) = 0.1\,(\sin(2\pi(x_1 + x_2) + \cos(2\pi(x_1 + x_2))$. We experimented with the viscosities $\nu = 1e{-}3, 1e{-}4, 1e{-}5$. For the data with $\nu = 1e{-}4$, the time resolution was also downsampled by half.

4.2 Experimental Settings

In the experiments on the 1D advection and Burgers' equation, we compared the results of simplified FNOs defined in Eq. (3) with Tanh and ReLU activations for varying number of layer L and initial parameters σ^2. In the experiments on the 2D Darcy Flow and NS equations, we compared the results of the original FNO (Li et al., 2020c) as shown in Eq. (1) for varying number of layer L and initial parameters σ^2. All models have a width of $D = 32$ and the Fourier mode of $K = 12$. When using Tanh as the activation, we fixed the initial bias parameter $\sigma_b^2 = 0.1$, and when using ReLU activation, we fixed the initial bias $b_i^{(\ell)} = 0$. We used the AdamW optimizer and cosine annealing scheduler. Training was stopped early at the epoch of minimal normalized mean squared error (nMSE) on the validation data. Details of the experimental setup are given in Appendix D. In the experiments on the NS equation, we trained the original FNO with the autoregressive scheme using a teacher-forcing technique, input normalization, and regularization that adds small-Gaussian noise to the input (Tran et al., 2022). During the evaluation phase, only the initial 10 steps are provided as input, and the rollout results of all subsequent steps are evaluated. For all tasks, the mean squared error (MSE) is used as the training loss and nMSE as the validation and testing (Li et al., 2020a; Tran et al., 2022).

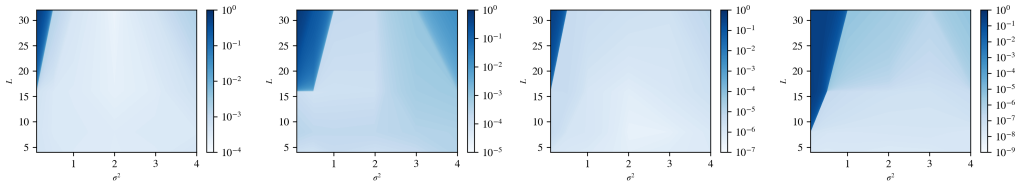
4.3 Results

Heatmaps of the training loss measured by MSE at the last epoch and the test performance measured by nMSE on six different datasets, for different depths L and initial parameters σ^2, are shown in Figs. 3 and 4, respectively. Despite the differences in architectures and datasets, Figure 3 shows the same trend supporting the theory in all experiments. As the number of layers L increases, the range of acceptable initial σ^2 value settings becomes narrower, and initialization near *the edge of chaos* ($\sigma^2 \approx 2$) is required for stable training of deep FNO. Detailed analyses on training loss and test performance are presented in Appendix E.1 and Appendix E.2, respectively.

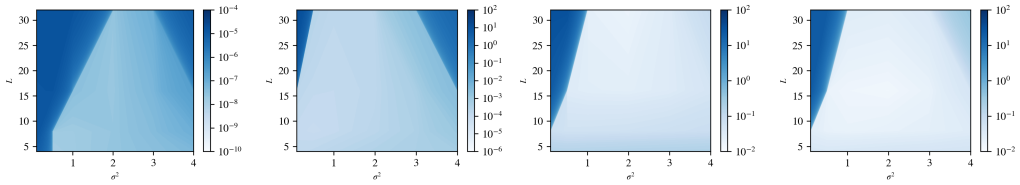
5 Conclusion

In this study, we developed a mean-field theory for FNO. We showed the expressivity and trainability of the FNO, which is characterized by the ordered-chaos phases. Furthermore, we observed both

https://doi.org/10.52202/079017-0352

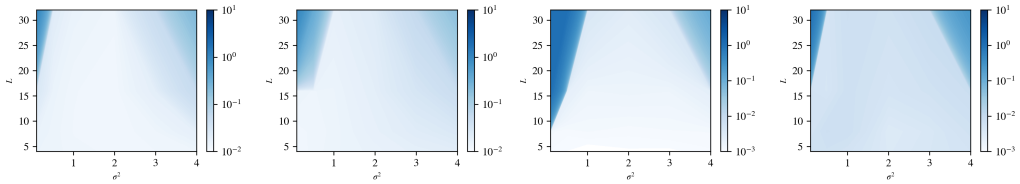


(a) Tanh on advection eq. (b) ReLU on advection eq. (c) Tanh on Burgers' eq. (d) ReLU on Burgers' eq.

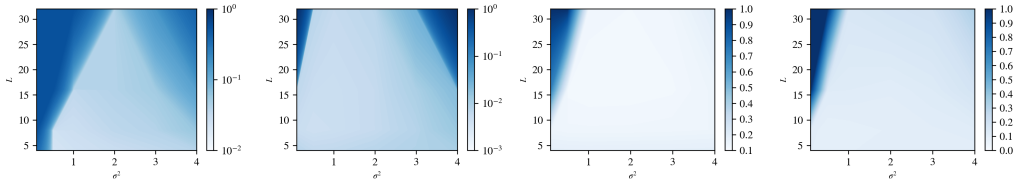


(e) ReLU on Darcy Flow (f) ReLU on NS eq. (1e-3) (g) ReLU on NS eq. (1e-4) (h) ReLU on NS eq. (1e-5)

Figure 3: Training loss of FNOs at last epoch for four distinct PDEs. **(a, b):** the advection equation, **(c, d):** the Burgers' equation, **(e):** Darcy Flow, **(f-h):** the NS equation. The heatmaps represents the training loss values for varying depth $L \in \{4, 8, 16, 32\}$ and initial weight parameter $\sigma^2 \in \{0.1, 0.5, 1.0, 2.0, 3.0, 4.0\}$, with lighter colors signifying lower training loss. The presented results are the mean training loss at the last epoch over three different seeds.



(a) Tanh on advection eq. (b) ReLU on advection eq. (c) Tanh on Burgers' eq. (d) ReLU on Burgers' eq.



(e) ReLU on Darcy Flow (f) ReLU on NS eq. (1e-3) (g) ReLU on NS eq. (1e-4) (h) ReLU on NS eq. (1e-5)

Figure 4: nMSE of FNOs on test datasets for four distinct PDEs. **(a, b):** the advection equation **(c, d):** the Burgers' equation, **(e):** Darcy Flow, **(f-h):** the NS equation. The heatmaps for each nMSE correspond to the results of each heatmap of training loss in Fig. 3. The lighter colors, the better the test performance. The presented results are the mean nMSE calculated over three different seeds.

unique FNO-specific behaviors caused by mode truncation, as well as common behaviors akin to those of DCN. With our analysis as a basis, we identified the necessity of initializing FNO near *the edge of chaos* for stable training of the FNO. Experimental results supported our theoretical results.

A limitation of our analysis is that it is limited to the network at initialization and does not address the stability of the entire optimization process. While we do not provide sufficient conditions for stable training, we do offer one necessary condition for achieving stable training. Future work may consider a mean-field analysis of the FNO when using skip-connection (Tran et al., 2022), Dropout and batch normalization, as well as initialization methods that ensure the input-output Jacobian of the FNO satisfies dynamical isometry (Pennington et al., 2017, 2018).

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 24H00709 Japan, JST ASPIRE Grant Number JPMJAP2329, RIKEN Special Postdoctoral Researcher Program, JST ACT-X Grant Number JPMJAX210K, and JST ACT-X Grant Number JPMJAX210D.

References

Brandstetter, J., Welling, M., and Worrall, D. E. Lie point symmetry data augmentation for neural pde solvers. In *International Conference on Machine Learning*, pp. 2241–2256. PMLR, 2022.

Brandstetter, J., van den Berg, R., Welling, M., and Gupta, J. K. Clifford neural layers for pde modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Burkholz, R. and Dubatovka, A. Initialization of relus for dynamical isometry. *Advances in Neural Information Processing Systems*, 32, 2019.

Cho, Y. and Saul, L. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

Guibas, J., Mardani, M., Li, Z., Tao, A., Anandkumar, A., and Catanzaro, B. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*, 2021.

Hayou, S., Doucet, A., and Rousseau, J. On the selection of initialization and activation function for deep neural networks. *arXiv preprint arXiv:1805.08266*, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

Helwig, J., Zhang, X., Fu, C., Kurtin, J., Wojtowytsch, S., and Ji, S. Group equivariant fourier neural operators for partial differential equations. *arXiv preprint arXiv:2306.05697*, 2023.

Hwang, R., Lee, J. Y., Shin, J. Y., and Hwang, H. J. Solving pde-constrained control problems using operator learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, No. 4, pp. 4504–4512, 2022.

Jiang, P., Meinert, N., Jordão, H., Weisser, C., Holgate, S., Lavin, A., Lutjens, B., Newman, D., Wainright, H., Walker, C., et al. Digital twin earth-coasts: Developing a fast and physics-informed surrogate model for coastal floods via neural operators. In *Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021)*, 2021a.

Jiang, Z., Tahmasebi, P., and Mao, Z. Deep residual u-net convolution neural networks with autoregressive strategy for fluid flow predictions in large-scale geosystems. *Advances in Water Resources*, 150:103878, 2021b.

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

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., and Anandkumar, A. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020a.

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Stuart, A., Bhattacharya, K., and Anandkumar, A. Multipole graph neural operator for parametric partial differential equations. *Advances in Neural Information Processing Systems*, 33:6755–6766, 2020b.

Li, Z., Kovachki, N. B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A., et al. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2020c.

 https://doi.org/10.52202/079017-0352

Li, Z., Huang, D. Z., Liu, B., and Anandkumar, A. Fourier neural operator with learned deformations for pdes on general geometries. *arXiv preprint arXiv:2207.05209*, 2022.

Li, Z., Meidani, K., and Farimani, A. B. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023.

Li, Z., Zheng, H., Kovachki, N., Jin, D., Chen, H., Liu, B., Azizzadenesheli, K., and Anandkumar, A. Physics-informed neural operator for learning partial differential equations. *ACM/JMS Journal of Data Science*, 1(3):1–27, 2024.

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., Kurth, T., Hall, D., Li, Z., Azizzadenesheli, K., et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

Pennington, J., Schoenholz, S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *Advances in neural information processing systems*, 30, 2017.

Pennington, J., Schoenholz, S., and Ganguli, S. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1924–1932. PMLR, 2018.

Poli, M., Massaroli, S., Berto, F., Park, J., Dao, T., Ré, C., and Ermon, S. Transform once: Efficient operator learning in frequency domain. *Advances in Neural Information Processing Systems*, 35: 7947–7959, 2022.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

Rahman, M. A., Ross, Z. E., and Azizzadenesheli, K. U-no: U-shaped neural operators. *arXiv preprint arXiv:2204.11127*, 2022.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. In *International Conference on Learning Representations*, 2016.

Takamoto, M., Praditia, T., Leiteritz, R., MacKinlay, D., Alesiani, F., Pflüger, D., and Niepert, M. Pdebench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Tran, A., Mathews, A., Xie, L., and Ong, C. S. Factorized fourier neural operators. In *The Eleventh International Conference on Learning Representations*, 2022.

Tripura, T. and Chakraborty, S. Wavelet neural operator: a neural operator for parametric partial differential equations. *arXiv preprint arXiv:2205.02191*, 2022.

Wen, G., Tang, M., and Benson, S. M. Multiphase flow prediction with deep neural networks. *arXiv preprint arXiv:1910.09657*, 2019.

Wen, G., Li, Z., Azizzadenesheli, K., Anandkumar, A., and Benson, S. M. U-fno—an enhanced fourier neural operator-based deep-learning model for multiphase flow. *Advances in Water Resources*, 163:104180, 2022a.

Wen, G., Li, Z., Long, Q., Azizzadenesheli, K., Anandkumar, A., and Benson, S. M. Accelerating carbon capture and storage modeling using fourier neural operators. *arXiv preprint arXiv:2210.17051*, 3, 2022b.

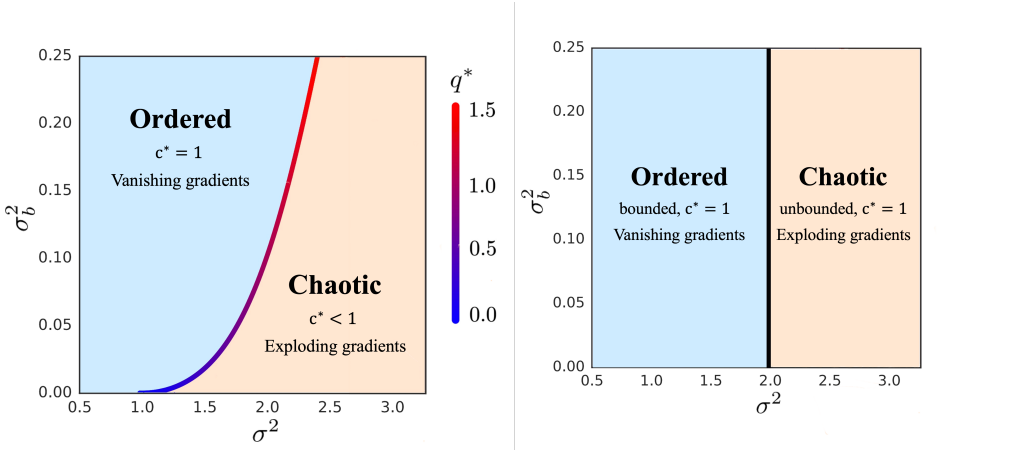
Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pp. 5393–5402. PMLR, 2018.

Yang, G. and Schoenholz, S. Mean field residual networks: On the edge of chaos. *Advances in neural information processing systems*, 30, 2017.

Yang, Y., Gao, A. F., Castellanos, J. C., Ross, Z. E., Azizzadenesheli, K., and Clayton, R. W. Seismic wave propagation and inversion with neural operators. *The Seismic Record*, 1(3):126–134, 2021.

Zhao, J., George, R. J., Zhang, Y., Li, Z., and Anandkumar, A. Incremental fourier neural operator. *arXiv preprint arXiv:2211.15188*, 2022.

 https://doi.org/10.52202/079017-0352



(a) Tanh activation (Pennington et al., 2017) (b) ReLU activation

Figure 5: Ordered-chaos phase transition diagram for the DCN

A Proof of Theorem 3.4

Theorem 3.4 (Exponential expressivity)**.** *Let $\boldsymbol{E}^{(\ell)} := \boldsymbol{\Sigma}^{(\ell)} - \boldsymbol{\Sigma}^*$ be the deviation from the fixed point at the ℓ-th layer. Suppose that the deviation at the first layer is decomposed as $\boldsymbol{E}^{(0)} = \epsilon\boldsymbol{\psi} + \sum_{k=1}^{K-1} \epsilon_k \boldsymbol{\psi}^{(k)} + \boldsymbol{e}$. The scalars $\epsilon,\ \epsilon_k$ represent the scale of the perturbation for each eigencomponent of the linearly approximated map $\boldsymbol{E}^{(\ell)} \mapsto \boldsymbol{E}^{(\ell+1)}$. The component $\boldsymbol{e} \in \mathbb{R}^{N\times N}$ belongs to the orthogonal complements of the space* $\mathrm{span}\left(\{\boldsymbol{\psi}, \boldsymbol{\psi}^{(1)}, \ldots, \boldsymbol{\psi}^{(K-1)}\}\right)$.

Then, the deviation at the ℓ-th layer is obtained by

$$\boldsymbol{E}^{(\ell)} = \chi^{\ell}\epsilon\boldsymbol{\psi} + \sum_{k=1}^{K-1} \chi_{c^*}^{\ell}\epsilon_k\boldsymbol{\psi}^{(k)}, \tag{10}$$

$$\chi := \frac{1}{N}\sum_{s=0}^{K-1} c_s\chi_{q^*} + \left(1 - \frac{1}{N}\sum_{s=0}^{K-1} c_s\right)(\chi_\kappa + \chi_{c^*}).$$

In particular, when the Fourier mode is $K = \frac{N}{2} + 1$, Eqs. (9) *and* (10) *reduce to the following.*

$$\begin{aligned} \boldsymbol{E}^{(\ell)} &= \chi_{q^*}^{\ell}\epsilon\boldsymbol{\psi} + \sum_{k=1}^{K-1} \chi_{c^*}^{\ell}\epsilon_k\boldsymbol{\psi}^{(k)}, \\ \forall\beta,\beta' \in [N],\ \psi_{\beta,\beta'} &= 1,\ \psi_{\beta,\beta'}^{(k)} = \cos\left(\theta_{\beta,\beta'}^{(k)}\right) - \delta_{\beta,\beta'}. \end{aligned} \tag{11}$$

Proof. The theorem is obtained by the eigenvalue analysis on the the first-order Taylor approximation of the iterated map $\mathcal{C}$ at the fixed point $\boldsymbol{\Sigma}^*$. The Jacobian matrix $J(\boldsymbol{\Sigma}^*) \in \mathbb{R}^{N^2\times N^2}$ of the iterated map at the fixed point and the Jacobian linear map $J_{\boldsymbol{\Sigma}^*}(\cdot)\colon \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$ are defined as follows.

$$[J(\boldsymbol{\Sigma}^*)]_{(\alpha,\alpha'),(\beta,\beta')} := \left.\frac{\partial[\mathcal{C}(\boldsymbol{\Sigma})]_{\alpha,\alpha'}}{\partial\Sigma_{\beta,\beta'}}\right|_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^*}, \tag{14}$$

$$\forall\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N},\ J_{\boldsymbol{\Sigma}^*}(\boldsymbol{\Sigma}) := \mathrm{mat}\left(J(\boldsymbol{\Sigma}^*)\,\mathrm{vec}(\boldsymbol{\Sigma})\right), \tag{15}$$

where vec performs the vectorization, *i.e.* transforming an $N \times N$ matrix into a vector of size N^2, and mat performs the inverse operation of vec.

From Lemma A.3, $K-1$ matrices in $\{\mathrm{vec}(\boldsymbol{\psi}^{(k)})\}_{k\in[K]\setminus\{0\}}$ are eigenbases with the eigenvalue χ_{c^*} of the Jacobian linear map. From Lemma A.4, the matrix $\boldsymbol{\psi}$ is the eigenbases with the eigenvalue χ of the Jacobian linear map. Since the sets of K matrices in $\{\boldsymbol{\psi}^{(k)}\}_{k\in[K]\setminus\{0\}} \cup \{\boldsymbol{\psi}\}$ are linearly

independent (yet non-orthogonal) in $\mathbb{R}^{N\times N}$, the subspace $\operatorname{span}\left(\{\boldsymbol{\psi}^{(k)}\}_{k\in[K]\setminus\{0\}}\cup\{\boldsymbol{\psi}\}\right)$ is the K-dimensional eigenspace of the Jacobian linear map. From Lemma A.2, the rank of the Jacobian matrix $J(\boldsymbol{\Sigma}^*)$ at the fixed point $\boldsymbol{\Sigma}^*$ is at most K, thereby the rank of the Jacobian linear map in Eq. (15) is at most K. Therefore, we have

$$\begin{aligned}
&\forall \boldsymbol{e}\in \operatorname{span}\left(\{\boldsymbol{\psi}^{(k)}\}_{k\in[K]\setminus\{0\}}\cup\{\boldsymbol{\psi}\}\right)^{\perp},\quad J_{\boldsymbol{\Sigma}^*}(\boldsymbol{e})=\boldsymbol{O}.\\
\mathcal{C}(\boldsymbol{\Sigma}^*+\boldsymbol{E}^{(\ell-1)})&\approx \boldsymbol{\Sigma}^*+J_{\boldsymbol{\Sigma}^*}(\boldsymbol{E}^{(\ell-1)})\\
&=\boldsymbol{\Sigma}^*+\left(\chi^{\ell-1}\epsilon J_{\boldsymbol{\Sigma}^*}(\boldsymbol{\psi})+\sum_{k=1}^{K-1}\chi_{c^*}^{\ell-1}\epsilon_k J_{\boldsymbol{\Sigma}^*}(\boldsymbol{\psi}^{(k)})\right)\\
&=\boldsymbol{\Sigma}^*+\underbrace{\chi^{\ell}\epsilon\boldsymbol{\psi}+\sum_{k=1}^{K-1}\chi_{c^*}^{\ell}\epsilon_k\boldsymbol{\psi}^{(k)}}_{=\boldsymbol{E}^{(\ell)}}.
\end{aligned}$$

□

Lemma 3.1 (Iterated map)**.** *For all $d\in[D]$, the covariance $\boldsymbol{\Sigma}^{(\ell)}:=\mathbb{E}_{\Theta^{1:\ell},\Xi^{1:\ell}}\left[\mathbf{H}_{:,d}^{(\ell)}\,\mathbf{H}_{:,d}^{(\ell)\top}\right]$ is obtained recursively by the iterated map $\mathcal{C}$ defined by*

$$\Sigma_{\alpha,\alpha'}^{(\ell)}=\underbrace{\sigma^2\sum_{k=0}^{K-1}c_k\mathbb{E}\left[\left|\left[\boldsymbol{F}\phi\left(\mathbf{H}_{:,d}\right)\right]_k\right|^2\right]\cos\left(\theta_{\alpha,\alpha'}^{(k)}\right)+\sigma_b^2}_{=:\,\mathcal{C}(\boldsymbol{\Sigma}^{(\ell-1)})_{\alpha,\alpha'}},\tag{4}$$

where the expectation is taken over the pre-activations $\mathbf{H}_{:,d}\sim\mathcal{N}(0,\boldsymbol{\Sigma}^{(\ell-1)})$, $\theta_{\alpha,\alpha'}^{(k)}:=\frac{2\pi k}{N}(\alpha-\alpha')$ represents the scaled positional difference.

Proof. For simplicity, we introduce $Y_{\alpha,k,\beta,i}$ as follows.

$$H_{\alpha,d}^{(\ell,k)}=\sum_{i=1}^{D}\sum_{\beta=0}^{N-1}\underbrace{F_{\alpha,k}^{\dagger}F_{k,\beta}X_{\beta,i}^{(\ell-1)}}_{=:\,Y_{\alpha,k,\beta,i}}\left(\Theta_{i,d}^{(\ell,k)}+\left(1-\left(\delta_{k,0}+\delta_{k,N/2}\right)\right)\sqrt{-1}\Xi_{i,d}^{(\ell,k)}\right).$$

Since the weights are sampled independently, for different $k\neq k'$,

$$\mathbb{E}\left[H_{\alpha,d}^{(\ell,k)}H_{\alpha,d}^{(\ell,k')}\right]=\mathbb{E}\left[H_{\alpha,d}^{(\ell,k)}\overline{H}_{\alpha,d}^{(\ell,k')}\right]=\mathbb{E}\left[\overline{H}_{\alpha,d}^{(\ell,k)}\overline{H}_{\alpha,d}^{(\ell,k')}\right]=0.$$

From Eq. (3), we have

$$\mathbb{E}_{\Theta^{(\ell)},\Xi^{(\ell)}}[H_{\alpha,d}^{(\ell)}H_{\alpha',d}^{(\ell)}]=\sum_{k=0}^{K-1}\frac{c_k}{2}\mathbb{E}\left[\left(H_{\alpha,d}^{(\ell,k)}+\overline{H}_{\alpha,d}^{(\ell,k)}\right)\left(H_{\alpha',d}^{(\ell,k)}+\overline{H}_{\alpha',d}^{(\ell,k)}\right)\right]+\mathbb{E}\left[(b_d^{(\ell)})^2\right].$$

https://doi.org/10.52202/079017-0352

First, we calculate the term $k \neq 0, \frac{N}{2}$, where $c_k = 2$.

$$
\begin{aligned}
&\mathbb{E}\left[\left(H_{\alpha,d}^{(\ell,k)} + \overline{H}_{\alpha,d}^{(\ell,k)}\right)\left(H_{\alpha',d}^{(\ell,k)} + \overline{H}_{\alpha',d}^{(\ell,k)}\right)\right] \\
&= \sum_{i,i'=1}^{D} \sum_{\beta,\beta'=0}^{N-1} Y_{\alpha,k,\beta,i} Y_{\alpha',k,\beta',i'} \delta_{i,i'} \left(\sigma^2 - \sigma^2\right)/2D \\
&\quad + \sum_{i,i'=1}^{D} \sum_{\beta,\beta'=1}^{N} \overline{Y}_{\alpha,k,\beta,i} \overline{Y}_{\alpha',k,\beta',i'} \delta_{i,i'} \left(\sigma^2 - \sigma^2\right)/2D \\
&\quad + \sum_{i,i'=1}^{D} \sum_{\beta,\beta'=0}^{N-1} Y_{\alpha,k,\beta,i} \overline{Y}_{\alpha',k,\beta',i'} \delta_{i,i'} \left(\sigma^2 + \sigma^2\right)/2D \\
&\quad + \sum_{i,i'=1}^{D} \sum_{\beta,\beta'=0}^{N-1} \overline{Y}_{\alpha,k,\beta,i} Y_{\alpha',k,\beta',i'} \delta_{i,i'} \left(\sigma^2 + \sigma^2\right)/2D \\
&= \frac{\sigma^2}{D} \sum_{i=1}^{D} \sum_{\beta,\beta'=0}^{N-1} Y_{\alpha,k,\beta,i} \overline{Y}_{\alpha',k,\beta',i} + \overline{Y}_{\alpha,k,\beta,i} Y_{\alpha',k,\beta',i} \\
&= \frac{\sigma^2}{D} \sum_{i=1}^{D} \underbrace{\left(\sum_{\beta=0}^{N-1} F_{k,\beta} X_{\beta,i}^{(\ell-1)}\right)}_{=\hat{X}_{k,i}^{(\ell-1)}} \overline{\left(\sum_{\beta'=0}^{N-1} F_{k,\beta'} X_{\beta',i}^{(\ell-1)}\right)} \left(F_{\alpha,k}^{\dagger} F_{k,\alpha'} + \overline{F_{\alpha,k}^{\dagger} F_{k,\alpha'}}\right) \\
&= \frac{2\sigma^2}{D} \sum_{i=1}^{D} \left|\hat{X}_{k,i}^{(\ell-1)}\right|^2 \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right),
\end{aligned}
$$

where $\hat{X}_{k,i}^{(\ell-1)}$ is the k-th Fourier mode of the representaion $\mathbf{X}_{:,i}^{(\ell-1)}$.

Second, we calculate the terms $k = 0, \frac{N}{2}$, where $c_k = 1$ and $H_{\alpha,d}^{(\ell,k)} = \overline{H}_{\alpha,d}^{(\ell,k)}$.

$$
\begin{aligned}
\mathbb{E}\left[\frac{c_k}{2}\left(H_{\alpha,d}^{(\ell,k)} + \overline{H}_{\alpha,d}^{(\ell,k)}\right)\left(H_{\alpha',d}^{(\ell,k)} + \overline{H}_{\alpha',d}^{(\ell,k)}\right)\right] &= 2 \sum_{i,i'=1}^{D} \sum_{\beta,\beta'=0}^{N-1} \left(Y_{\alpha,k,\beta,i} Y_{\alpha',k,\beta',i'}\right) \delta_{i,i'} \frac{\sigma^2}{2D} \\
&= \frac{\sigma^2}{D} \sum_{i=1}^{D} \left|\hat{X}_{k,i}^{(\ell-1)}\right|^2 \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right).
\end{aligned}
$$

Since the Fourier modes $|\hat{X}_{k,i}^{(\ell-1)}|^2$ are i.i.d. for each hidden dimension $i \in [D]$, we have

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\Theta}^{0:\ell-1},\boldsymbol{\Xi}^{0:\ell-1},\boldsymbol{b}^{0:\ell-1}} \mathbb{E}_{\boldsymbol{\Theta}^{(\ell)},\boldsymbol{\Xi}^{(\ell)},\boldsymbol{b}^{(\ell)}} \left[H_{\alpha,d}^{(\ell)} H_{\alpha',d}^{(\ell)}\right] &= \sigma^2 \sum_{k=0}^{K-1} c_k \mathbb{E}_{\boldsymbol{\Theta}^{0:\ell-1},\boldsymbol{\Xi}^{0:\ell-1}} \left[\left|\hat{X}_{k,i}^{(\ell-1)}\right|^2\right] \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) + \sigma_b^2 \\
&= \sigma^2 \sum_{k=0}^{K-1} c_k \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}\left(0,\boldsymbol{\Sigma}^{(\ell-1)}\right)} \left[\left|\hat{X}_{k,i}^{(\ell-1)}\right|^2\right] \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) + \sigma_b^2.
\end{aligned} \tag{16}
$$

To obtain a more tractable expression in the subsequent proofs of theorems, we express the iterated map without using Fourier modes as follows.

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\Theta}^{0:\ell},\boldsymbol{\Xi}^{0:\ell},\boldsymbol{b}^{0:\ell}} \left[H_{\alpha,d}^{(\ell)} H_{\alpha',d}^{(\ell)}\right] \\
&= \frac{\sigma^2}{N^2} \sum_{k=0}^{K-1} c_k \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) \cos\left(\theta_{\beta,\beta'}^{(k)}\right) \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}\left(0,\boldsymbol{\Sigma}^{(\ell-1)}\right)} \left[\phi(H_{\beta,d}) \phi(H_{\beta',d})\right] + \sigma_b^2.
\end{aligned} \tag{17}
$$

□

Lemma 3.2 (Exsistance of fixed points)**.** *When a random DCN defined by Eq.* (2) *has the fixed point* $(q^*, c^* = 1)$ *for the initial parameters* (σ^2, σ_b^2)*, then a random simplified FNO defined by Eq.* (3) *has a fixed point* $\mathbf{\Sigma}^*$ *of the form*

$$\mathbf{\Sigma}^* = q^* \boldsymbol{I}_N + q^* c^* (\mathbf{1}_N \mathbf{1}_N^\top - \boldsymbol{I}_N) = q^* \mathbf{1}_N \mathbf{1}_N^\top.$$

Proof. Using the properties of cosine functions, the following holds.

$$\sigma_b^2 = \frac{1}{N^2} \sum_{k=0}^{K-1} c_k \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) \underbrace{\sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta_{\beta,\beta'}^{(k)}\right)}_{=N^2\delta_{k,0}} \sigma_b^2. \tag{18}$$

Then, the following holds for all $\alpha, \alpha' \in [N]$.

$$[\mathcal{C}(\mathbf{\Sigma}^*)]_{\alpha,\alpha'} = \frac{1}{N^2} \sum_{k=0}^{K-1} c_k \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta_{\beta,\beta'}^{(k)}\right) \left(\sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma}^*)} \left[\phi(H_{\beta,d})\phi(H_{\beta',d})\right] + \sigma_b^2\right).$$

When the DCN is applied pointwise to each spatial representation $\boldsymbol{H}_{\beta,:},\ \beta \in [N]$, the iterated map of the random DCN (Poole et al., 2016) is given by $\mathcal{C}_{\mathrm{DCN}}(\mathbf{\Sigma}) \coloneqq \sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma})} \left[\phi(H_{\beta,d})\phi(H_{\beta',d})\right] + \sigma_b^2$. Since the covariance $\mathbf{\Sigma}^*$ is a fixed point with respect to the iterative map of the DCN, *i.e.* $\mathcal{C}_{\mathrm{DCN}}(\mathbf{\Sigma}^*) = \mathbf{\Sigma}^*$, the following holds.

$$[\mathcal{C}(\mathbf{\Sigma}^*)]_{\alpha,\alpha'} = \frac{1}{N^2} \sum_{k=0}^{K-1} c_k \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) \underbrace{\sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta_{\beta,\beta'}^{(k)}\right)}_{=N^2\delta_{k,0}} \underbrace{q^*(\delta_{\beta,\beta'} + (1-\delta_{\beta,\beta'})c^*)}_{=q^*} = q^*.$$

Thus, we confirm that the covariance $\mathbf{\Sigma}^*$ satisfies the definition of a fixed point in the map $\mathcal{C}$, *i.e.* $\mathbf{\Sigma}^* = \mathcal{C}(\mathbf{\Sigma}^*)$. This means that the fixed point for the iterated map $\mathbf{\Sigma}^*$ of the DCN also serves as a fixed point for the iterated map of the simplified FNO. □

Lemma A.1. *Let* $\mathbf{\Sigma}^*$ *be the fixed point of the form in Eq.* (5)*. Suppose that the symmetric perturbation* $\boldsymbol{E} \in \mathbb{R}^{N \times N}$*, where* $E_{\beta,\beta} = E_{\beta',\beta'}$ *and* $E_{\beta,\beta'}$ *are non-zero for some* $\beta, \beta' \in [N],\ \beta \neq \beta$*, and all other elements are zero. Then, we have*

$$\sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma}^* + \boldsymbol{E})} \left[\phi(H_{\beta,d})^2\right] + \sigma_b^2 = q^* + E_{\beta,\beta}\chi_{q^*} + \mathcal{O}\left(|E_{\beta,\beta}|^2\right), \tag{19}$$

$$\sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma}^* + \boldsymbol{E})} \left[\phi(H_{\beta,d})\phi(H_{\beta',d})\right] + \sigma_b^2 = q^* c^* + E_{\beta,\beta}\chi_\kappa + E_{\beta,\beta'}\chi_{c^*} + \mathcal{O}\left(|E_{\beta,\beta'}|^2\right), \tag{20}$$

where $\chi_{q^*},\ \chi_{c^*}$*, and* χ_κ *are the constants defined by*

$$\chi_{q^*} \coloneqq \sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma}^*)} \left[\phi'^2(H_{\beta,d}) + \phi''(H_{\beta,d})\phi(H_{\beta,d})\right],$$

$$\chi_{c^*} \coloneqq \sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma}^*)} \left[\phi'(H_{\beta,d})\phi'(H_{\beta',d})\right],$$

$$\chi_\kappa \coloneqq \frac{\sigma^2}{2} \mathbb{E}\left[\phi''(H_{\alpha,d})\phi(H_{\alpha',d}) + \phi(H_{\alpha,d})\phi''(H_{\alpha',d}) + 2c^*\phi'(H_{\alpha,d})\phi'(H_{\alpha',d})\right].$$

Proof. Equation (19) is obviously shown by the result of Section 2 in (Poole et al., 2016) and Section 3 in (Schoenholz et al., 2016). We prove Eq. (20) with reference to the results of (Schoenholz et al., 2016).

$$\sigma^2 \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(0,\mathbf{\Sigma}^* + \boldsymbol{E})} \left[\phi(H_{\beta,d})\phi(H_{\beta',d})\right] + \sigma_b^2 = \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2,$$

where $\int \mathcal{D}\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int \mathrm{d}z e^{-\frac{1}{2}z^2}$ is the measure for a standard Gaussian distribution, $u_1 = \sqrt{q} z_1$, $u_2 = \sqrt{q}\left(c_{\beta,\beta'} z_1 + \sqrt{1 - c_{\beta,\beta'}^2} z_2\right)$, $q = q^* + E_{\beta,\beta}$ and $c_{\beta,\beta'} = c^* + \frac{E_{\beta,\beta'}}{q}$.

https://doi.org/10.52202/079017-0352

We consider the case where $c^* < 1$ and $c^* = 1$ separately. Later, we will show that the two results agree with each other. First, we consider the case where $c^* < 1$. Using Taylor expansion, we can approximate $\phi(u_1)$ and $\phi(u_2)$ as follows. For the simplicity, we assume $\mathcal{O}(|E_{\beta,\beta}|) = \mathcal{O}(|E_{\beta,\beta'}|)$.

$$\phi(u_1) = \phi(u_1^*) + \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}} z_1 \phi'(u_1^*) + \mathcal{O}(|E_{\beta,\beta}|^2),$$

$$\phi(u_2) = \phi\left(u_2^*\right) + \frac{E_{\beta,\beta'}}{\sqrt{q^*}}\left(z_1 - \frac{c^*}{\sqrt{1-(c^*)^2}} z_2\right)\phi'(u_2^*) + \frac{E_{\beta,\beta}}{2\sqrt{q^*}}(c^* z_1 + \sqrt{1-(c^*)^2} z_2)\phi'(u_2^*) + \mathcal{O}(|E_{\beta,\beta}|^2),$$

where $u_1^* = \sqrt{q^*} z_1$ and $u_2^* = \sqrt{q^*}(c^* z_1 + \sqrt{1-(c^*)^2} z_2)$.

Thus, we have

$$\begin{aligned}
&\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2 \\
&= \underbrace{\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1^*)\phi(u_2^*) + \sigma_b^2}_{=q^* c^*} + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{E_{\beta,\beta'}}{\sqrt{q^*}}\left(z_1 - \frac{c^*}{\sqrt{1-(c^*)^2}} z_2\right)\phi(u_1^*)\phi'(u_2^*) \\
&\quad + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}}(c^* z_1 + \sqrt{1-(c^*)^2} z_2)\phi(u_1^*)\phi'(u_2^*) + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}} z_1 \phi'(u_1^*)\phi(u_2^*) + \mathcal{O}(|E_{\beta,\beta}^2|).
\end{aligned}$$

The results of the second term are obtained from the transformation of equations 36 to 39 in Appendix 7.2 of (Schoenholz et al., 2016).

$$\sigma^2 \frac{E_{\beta,\beta'}}{\sqrt{q^*}} \int \mathcal{D}z_1 \mathcal{D}z_2 \left(z_1 - \frac{c^*}{\sqrt{1-(c^*)^2}} z_2\right)\phi(u_1^*)\phi'(u_2^*) = E_{\beta,\beta'} \underbrace{\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi'(u_1^*)\phi'(u_2^*)}_{=\chi_{c^*}}.$$

Utilizing the identity, $\int \mathcal{D}z z f(z) = \int \mathcal{D}z f'(z)$, we obtain the third term as follows.

$$\begin{aligned}
&\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}}(c^* z_1 + \sqrt{1-(c^*)^2} z_2)\phi(u_1^*)\phi'(u_2^*) \\
&= \sigma^2 \frac{1}{2} E_{\beta,\beta} \int \mathcal{D}z_1 \mathcal{D}z_2 \left(c^*\left(\phi'(u_1^*)\phi'(u_2^*) + c^*\phi(u_1^*)\phi''(u_2^*)\right) + (1-(c^*)^2)\phi(u_1^*)\phi''(u_2^*)\right) \\
&= \sigma^2 \frac{1}{2} E_{\beta,\beta} \int \mathcal{D}z_1 \mathcal{D}z_2 \left(c^*\phi'(u_1^*)\phi'(u_2^*) + \phi(u_1^*)\phi''(u_2^*)\right).
\end{aligned}$$

The last term is calculated as follows.

$$\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}} z_1 \phi'(u_1^*)\phi(u_2^*) = \sigma^2 \frac{1}{2} E_{\beta,\beta} \int \mathcal{D}z_1 \mathcal{D}z_2 (\phi''(u_1^*)\phi(u_2^*) + c^*\phi'(u_1^*)\phi'(u_2^*)).$$

Summing the last two terms, we obtain the term $E_{\beta,\beta}\chi_\kappa$.

Next, we consider the case where $c^* = 1$. As with the discussion of (Schoenholz et al., 2016), the perturbed correlation is defined by $c_{\beta,\beta'} = c^* - \frac{E_{\beta,\beta'}}{q}$ where $E_{\beta,\beta'} > 0$ and then $\phi(u_2)$ is expanded as follows.

$$\phi(u_2) = \phi(u_2^*) + \left(\sqrt{2\frac{E_{\beta,\beta'}}{q^*}} z_2 - \frac{E_{\beta,\beta'}}{\sqrt{q^*}} z_1\right)\phi'(u_2^*) + E_{\beta,\beta'} z_2^2 \phi''(u_2^*) + \frac{E_{\beta,\beta}}{2\sqrt{q^*}} z_1 \phi'(u_2^*) + \mathcal{O}(|E_{\beta,\beta'}|^{3/2}).$$

Thus, we have

$$
\begin{aligned}
&\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2 \\
&= \underbrace{\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1^*)\phi(u_2^*) + \sigma_b^2}_{=q^* c^*} + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \sqrt{2\frac{E_{\beta,\beta'}}{q^*}} z_2 \phi(u_1^*)\phi'(u_2^*) \\
&\quad - \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{E_{\beta,\beta'}}{\sqrt{q^*}} z_1 \phi(u_1^*)\phi'(u_2^*) + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 E_{\beta,\beta'} z_2^2 \phi(u_1^*)\phi''(u_2^*) \\
&\quad + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{E_{\beta,\beta}}{2\sqrt{q^*}} z_1 \phi(u_1^*)\phi'(u_2^*) + \sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}} z_1 \phi'(u_1^*)\phi(u_2^*) + \mathcal{O}(|E_{\beta,\beta}^{3/2}|).
\end{aligned}
\tag{21}
$$

Using the fact that $u_2^* = u_1^*$ and u_2^* is independent of z_2,

$$
\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \sqrt{2\frac{E_{\beta,\beta'}}{q^*}} z_2 \phi(u_1^*)\phi'(u_2^*) = \sigma^2 \int \mathcal{D}z_1 \sqrt{2\frac{E_{\beta,\beta'}}{q^*}} \phi(u_1^*)\phi'(u_2^*) \underbrace{\left(\int \mathcal{D}z_2 z_2\right)}_{=0}.
$$

$$
\begin{aligned}
\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{E_{\beta,\beta'}}{\sqrt{q^*}} z_1 \phi(u_1^*)\phi'(u_2^*) &= \sigma^2 \int \mathcal{D}z_1 \frac{E_{\beta,\beta'}}{\sqrt{q^*}} z_1 \phi(u_1^*)\phi'(u_1^*) \\
&= \sigma^2 E_{\beta,\beta'} \int \mathcal{D}z_1 (\phi'(u_1^*)^2 + \phi(u_1^*)\phi''(u_1^*)).
\end{aligned}
$$

$$
\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 E_{\beta,\beta'} z_2^2 \phi(u_1^*)\phi''(u_2^*) = \sigma^2 E_{\beta,\beta'} \int \mathcal{D}z_1 \phi(u_1^*)\phi''(u_1^*) \underbrace{\left(\int \mathcal{D}z_2 z_2^2\right)}_{=1}.
$$

$$
\begin{aligned}
\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{E_{\beta,\beta}}{2\sqrt{q^*}} z_1 \phi(u_1^*)\phi'(u_2^*) &= \sigma^2 \int \mathcal{D}z_1 \frac{E_{\beta,\beta}}{2\sqrt{q^*}} z_1 \phi(u_1^*)\phi'(u_1^*) \\
&= \sigma^2 \frac{E_{\beta,\beta}}{2} \int \mathcal{D}z_1 (\phi'(u_1^*)^2 + \phi(u_1^*)\phi''(u_1^*)).
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}} z_1 \phi'(u_1^*)\phi(u_2^*) &= \sigma^2 \int \mathcal{D}z_1 \frac{1}{2}\frac{E_{\beta,\beta}}{\sqrt{q^*}} z_1 \phi'(u_1^*)\phi(u_1^*) \\
&= \sigma^2 \frac{E_{\beta,\beta}}{2} \int \mathcal{D}z_1 (\phi'(u_1^*)^2 + \phi(u_1^*)\phi''(u_1^*)).
\end{aligned}
$$

Substituting these facts into Eq. (21), we obtain

$$
\begin{aligned}
&\sigma^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1)\phi(u_2) + \sigma_b^2 \\
&\approx q^* c^* - \sigma^2 E_{\beta,\beta'} \int \mathcal{D}z_1 \phi'(u_1^*)^2 + \sigma^2 E_{\beta,\beta} \int \mathcal{D}z_1 (\phi'(u_1^*)^2 + \phi(u_1^*)\phi''(u_1^*))
\end{aligned}
$$

The above result agrees with that obtained by substituting $c^* = 1$ for the result obtained when the case $c^* < 1$.

□

Lemma A.2. *The Jacobian matrix $J(\mathbf{\Sigma}^*)$ of the iterated map $\mathcal{C}$ defined in Eq.* (14) *is obtained as follows.*

$$
[J(\mathbf{\Sigma}^*)]_{(\alpha,\alpha'),(\beta,\beta')} = \frac{1}{N^2} \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \cos\left(\theta_{\beta,\beta'}^{(k')}\right) \left(\delta_{\beta,\beta'}(\chi_{q^*} - \chi_\kappa + \delta_{k',0} N \chi_\kappa) + (1 - \delta_{\beta,\beta'})\chi_{c_*}\right).
$$

Furthermore, the rank of the Jacobian matrix $J(\mathbf{\Sigma}^*) \in \mathbb{R}^{N^2 \times N^2}$ is at most K.

https://doi.org/10.52202/079017-0352

Proof. Let some semi-positive definite matrix $\boldsymbol{E} \in \mathbb{R}^{N\times N}$ be a deviation from the fixed point $\boldsymbol{\Sigma}^*$. From Lemma 3.1 and Lemma A.1, we have

$$
\begin{aligned}
&[\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E})]_{\alpha,\alpha'} \\
&= \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) \left(\sigma^2 \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(0,\boldsymbol{\Sigma}^*+\boldsymbol{E})}\left[\phi(H_{\gamma,d})\phi(H_{\gamma',d})\right] + \sigma_b^2\right) \\
&\approx \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) \left(\delta_{\beta,\beta'}(q^* + E_{\beta,\beta}\chi_{q^*}) + (1-\delta_{\beta,\beta'})(q^* c^* + E_{\beta,\beta}\chi_\kappa + E_{\beta,\beta'}\chi_{c_*})\right).
\end{aligned}
\tag{22}
$$

Note that Eq. (22) is obtained by neglecting higher order terms in Eqs. (19) and (20).

By the definition of the fixed point,

$$
\Sigma^*_{\alpha,\alpha'} = \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) q^*(\delta_{\beta,\beta'} + (1-\delta_{\beta,\beta'})c^*).
$$

By substituting the fact into Eq. (22), we obtain

$$
\begin{aligned}
&[\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}) - \boldsymbol{\Sigma}^*]_{\alpha,\alpha'} \\
&\approx \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) \left(\delta_{\beta,\beta'} E_{\beta,\beta}\chi_{q^*}) + (1-\delta_{\beta,\beta'})(E_{\beta,\beta}\chi_\kappa + E_{\beta,\beta'}\chi_{c_*})\right) \\
&= \sum_{\beta,\beta'=0}^{N-1} \underbrace{\left[\frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \cos\left(\theta^{(k')}_{\beta,\beta'}\right) \left(\delta_{\beta,\beta'}(\chi_{q^*} - \chi_\kappa + \delta_{k',0}N\chi_\kappa) + (1-\delta_{\beta,\beta'})\chi_{c_*}\right)\right]}_{=[J(\boldsymbol{\Sigma}^*)]_{(\alpha,\alpha'),(\beta,\beta')}} E_{\beta,\beta'}.
\end{aligned}
\tag{23}
$$

The last equation can be rewritten using the matrix calculation as follows.

$$
\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}) - \boldsymbol{\Sigma}^* \approx \mathrm{mat}(J(\boldsymbol{\Sigma}^*)\,\mathrm{vec}(\boldsymbol{E})).
$$

Since $J(\boldsymbol{\Sigma}^*)$ is the first-order coefficient to the deviaction $\boldsymbol{E}$, $J(\boldsymbol{\Sigma}^*)$ is exactly the Jacobian matrix of the iterated map $\mathcal{C}$.

Furthermore, the Jacobian matrix can be decomposed to two matricies $\boldsymbol{A} \in \mathbb{R}^{N^2\times K}$ and $\boldsymbol{B} \in \mathbb{R}^{K\times N^2}$ as follows.

$$
\begin{aligned}
[J(\boldsymbol{\Sigma}^*)]_{(\alpha,\alpha'),(\beta,\beta')} &= \sum_{k'=0}^{K-1} A_{(\alpha,\alpha'),k'} B_{k',(\beta,\beta')},\ A_{(\alpha,\alpha'),k'} \coloneqq \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right), \\
B_{k',(\beta,\beta')} &\coloneqq \frac{1}{N^2} c_{k'} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) \left(\delta_{\beta,\beta'}(\chi_{q^*} - \chi_\kappa + \delta_{k',0}N\chi_\kappa) + (1-\delta_{\beta,\beta'})\chi_{c_*}\right).
\end{aligned}
$$

Therefore, the rank of the Jacobian matrix is at most K. □

Lemma A.3. *Let $\boldsymbol{\Sigma}^*$ be the fixed point of the form in Eq.* (5) *and $\boldsymbol{E}^{(k)}$ be the perturbation expressed as*

$$
E^{(k)}_{\beta,\beta'} = \epsilon_k \psi^{(k)}_{\beta,\beta'}, \quad \psi^{(k)}_{\beta,\beta'} \coloneqq \left[\cos\left(\theta^{(k)}_{\beta,\beta'}\right) - \left(\sum_{s=0}^{K-1} c_s\right)^{-1} \sum_{s=0}^{K-1} c_s \cos\left(\theta^{(s)}_{\beta,\beta'}\right)\right],
$$

where ϵ_k denotes the scale of the perturbation, assumed to be sufficiently small.

Then, we have, for all $k \in [K]\backslash\{0\}$,

$$
\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}^{(k)}) \approx \boldsymbol{\Sigma}^* + \chi_{c^*}\boldsymbol{E}^{(k)}.
$$

Proof. From Eq. (23), we have

$$[\mathcal{C}(\mathbf{\Sigma}^* + \boldsymbol{E}^{(k)})]_{\alpha,\alpha'} \approx \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta=0}^{N-1} (q^* + E_{\beta,\beta}^{(k)}\chi_{q^*})$$
$$+ \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta=0}^{N-1}\sum_{\beta\neq\beta'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) (q^*c^* + E_{\beta,\beta}^{(k)}\chi_\kappa + E_{\beta,\beta'}^{(k)}\chi_{c^*}).$$

From the definition of the fixed point, we obtain

$$\Sigma_{\alpha,\alpha'}^* = \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta,\beta'=0}^{N-1} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) q^*(\delta_{\beta,\beta'} + (1-\delta_{\beta,\beta'})c^*).$$

By combining the above two results, we have

$$[\mathcal{C}(\mathbf{\Sigma}^* + \boldsymbol{E}^{(k)}) - \mathbf{\Sigma}^*]_{\alpha,\alpha'} \approx \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta=0}^{N-1} E_{\beta,\beta}^{(k)}\chi_{q^*}$$
$$+ \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta=0}^{N-1}\sum_{\beta\neq\beta'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) (E_{\beta,\beta}^{(k)}\chi_\kappa + E_{\beta,\beta'}^{(k)}\chi_{c^*}).$$

Since $\cos(\theta_{\beta,\beta}^{(k)}) = 1$ for all $k \in [K]$, we have $\psi_{\beta,\beta}^{(k)} = 0$ and $E_{\beta,\beta}^{(k)} = 0$. Thus, only the term with $E_{\beta,\beta'}^{(k)}$ remains.

$$[\mathcal{C}(\mathbf{\Sigma}^* + \boldsymbol{E}^{(k)}) - \mathbf{\Sigma}^*]_{\alpha,\alpha'} \approx \epsilon_k\chi_{c^*}\frac{1}{N^2}\left(\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta\neq\beta'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) \cos\left(\theta_{\beta,\beta'}^{(k)}\right) \right. \quad (24)$$
$$\left. - \left(\sum_{s=0}^{K-1} c_s\right)^{-1} \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta\neq\beta'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) \sum_{s=0}^{K-1} c_s \cos\left(\theta_{\beta,\beta'}^{(s)}\right)\right). \quad (25)$$

Given the orthogonality of the cosine functions, we have

$$\sum_{\substack{\beta,\beta'=0\\ \beta\neq\beta'}}^{N-1} c_{k'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) \cos\left(\theta_{\beta,\beta'}^{(k)}\right) = \begin{cases} N^2 - c_{k'}N & (k' = k) \\ -c_{k'}N & (\text{otherwise}). \end{cases} \quad (26)$$

Utilizing Eq. (26) leads to the following two facts:

$$\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta\neq\beta'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) \cos\left(\theta_{\beta,\beta'}^{(k)}\right) = N^2 \cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) - N\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right).$$

$$\sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) \sum_{\beta\neq\beta'} \cos\left(\theta_{\beta,\beta'}^{(k')}\right) \sum_{s=0}^{K-1} c_s \cos\left(\theta_{\beta,\beta'}^{(s)}\right) = N^2 \sum_{k'=0}^{K-1} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) - N\left(\sum_{s=0}^{K-1} c_s\right) \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta_{\alpha,\alpha'}^{(k')}\right).$$

https://doi.org/10.52202/079017-0352

By substituting these facts into Eq. (24), we obtain

$$
\begin{aligned}
&[\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}^{(k)}) - \boldsymbol{\Sigma}^*]_{\alpha,\alpha'} \\
&\approx \chi_{c^*}\epsilon_k \frac{1}{N^2}\left(N^2 \cos\left(\theta^{(k)}_{\alpha,\alpha'}\right) - \cancel{N \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right)} \right. \\
&\quad \left. - \left(\sum_{s=0}^{K-1} c_s\right)^{-1} N^2 \sum_{k'=0}^{K-1} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) + \cancel{N \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right)} \right) \\
&= \chi_{c^*}\epsilon_k \underbrace{\left[\cos\left(\theta^{(k)}_{\alpha,\alpha'}\right) - \left(\sum_{s=0}^{K-1} c_s\right)^{-1} \sum_{k'=0}^{K-1} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \right]}_{=\psi^{(k)}_{\alpha,\alpha'}}.
\end{aligned}
$$

This completes the proof. □

Lemma A.4. *Let $\boldsymbol{\Sigma}^*$ be the fixed point of the form in Eq.* (5) *and $\boldsymbol{E}$ be the perturbation expressed as*

$$
E_{\beta,\beta'} = \epsilon\psi_{\beta,\beta'},\ \psi_{\beta,\beta'} := \left[1 - \frac{1}{N}\left(\frac{\chi_\kappa + \chi_{c^*} - \chi_{q^*}}{\chi_\kappa}\right) \sum_{s=0}^{K-1} c_s \cos\left(\theta^{(s)}_{\beta,\beta'}\right) \right],
$$

where ϵ denotes the scale of the perturbation, assumed to be sufficiently small, $\mathbf{1}_N\mathbf{1}_N^\top$ is all-one matrix with size $N \times N$.

Then, we have

$$
\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}) \approx \boldsymbol{\Sigma}^* + \underbrace{\left(\frac{\sum_{s=0}^{K-1} c_s}{N}\chi_{q^*} + \left(1 - \frac{\sum_{s=0}^{K-1} c_s}{N}\right)(\chi_\kappa + \chi_{c^*}) \right)}_{:=\chi} \boldsymbol{E}.
$$

Proof. For simplicity, we introduce $x := \frac{1}{N}\left(\frac{\chi_\kappa + \chi_{c^*} - \chi_{q^*}}{\chi_\kappa}\right)\left(\sum_{s=0}^{K-1} c_s\right)$ as follows.

$$
\begin{aligned}
\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}) &\approx \boldsymbol{\Sigma}^* + \left(\frac{\sum_{s=0}^{K-1} c_s}{N}\chi_{q^*} + \left(1 - \frac{\sum_{s=0}^{K-1} c_s}{N}\right)(\chi_\kappa + \chi_{c^*}) \right) \boldsymbol{E} \\
&= \boldsymbol{\Sigma}^* + \left[\left(1 - \underbrace{\frac{1}{N}\left(\frac{\chi_\kappa + \chi_{c^*} - \chi_{q^*}}{\chi_\kappa}\right)\left(\sum_{s=0}^{K-1} c_s\right)}_{=:x} \right) \chi_\kappa + \chi_{c^*} \right] \boldsymbol{E}. \qquad (27)
\end{aligned}
$$

From Eq. (23), we have

$$
\begin{aligned}
&[\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}) - \boldsymbol{\Sigma}^*]_{\alpha,\alpha'} \\
&\approx \frac{1}{N^2} \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\alpha,\alpha'}\right) \sum_{\beta=0}^{N-1} \left(E_{\beta,\beta}\chi_{q^*} + \sum_{\beta\neq\beta'} \cos\left(\theta^{(k')}_{\beta,\beta'}\right)(E_{\beta,\beta}\chi_\kappa + E_{\beta,\beta'}\chi_{c^*}) \right). \qquad (28)
\end{aligned}
$$

Our goal is to derive Eq. (27) from Eq. (28). The following results are useful for the computation of each term of Eq. (28).

$$
\forall \beta,\beta' \in [N],\ \beta \neq \beta', \quad E_{\beta,\beta} = \epsilon(1-x), \quad E_{\beta,\beta'} = \epsilon\left(1 - x\left(\sum_{s=0}^{K-1} c_s\right)^{-1} \sum_{s=0}^{K-1} c_s \cos\left(\theta^{(s)}_{\beta,\beta'}\right)\right).
$$

$$
\sum_{\beta=0}^{N-1} \sum_{\beta\neq\beta'} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) = \begin{cases} N^2 - N & (k' = 0) \\ -N & (\text{otherwise}) \end{cases}.
$$

$$
\sum_{\substack{\beta,\beta'=0 \\ \beta\neq\beta'}}^{N-1} c_{k'} \cos\left(\theta^{(k')}_{\beta,\beta'}\right) \cos\left(\theta^{(k)}_{\beta,\beta'}\right) = \begin{cases} N^2 - c_{k'}N & (k' = k) \\ -c_{k'}N & (\text{otherwise}) \end{cases}.
$$

Using the above results, the first and second terms of Eq. (28) are calculated as follows.

$$\frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\sum_{\beta=0}^{N-1} E_{\beta,\beta}\chi_{q^*} = \epsilon(1-x)\chi_{q^*}\frac{1}{N}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right).$$

$$\frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\sum_{\beta=0}^{N-1}\sum_{\beta\neq\beta'}\cos\left(\theta_{\beta,\beta'}^{(k')}\right)E_{\beta,\beta}\chi_{\kappa} = \epsilon(1-x)\chi_{\kappa} - \epsilon(1-x)\chi_{\kappa}\frac{1}{N}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right).$$

The third term of Eq. (28) is obtained as follows.

$$\begin{aligned}
&\frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\sum_{\beta=0}^{N-1}\sum_{\beta\neq\beta'}\cos\left(\theta_{\beta,\beta'}^{(k')}\right)E_{\beta,\beta'}\chi_{c^*}\\
&= \frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\sum_{\beta=0}^{N-1}\sum_{\beta\neq\beta'}\cos\left(\theta_{\beta,\beta'}^{(k')}\right)\epsilon\chi_{c^*}\\
&\quad - x\left(\sum_{s=0}^{K-1} c_s\right)^{-1}\frac{1}{N^2}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\sum_{s=0}^{K-1}\sum_{\beta=0}^{N-1}\sum_{\beta\neq\beta'} c_s\cos\left(\theta_{\beta,\beta'}^{(s)}\right)\cos\left(\theta_{\beta,\beta'}^{(k')}\right)\epsilon\chi_{c^*}\\
&= \epsilon\chi_{c^*} - \epsilon\chi_{c^*}\frac{1}{N}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\\
&\quad - \epsilon\chi_{c^*}x\left(\sum_{s=0}^{K-1} c_s\right)^{-1}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) + \epsilon\chi_{c^*}x\frac{1}{N}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\\
&= \epsilon\chi_{c^*} - \epsilon\chi_{c^*}x\left(\sum_{s=0}^{K-1} c_s\right)^{-1}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right) - \epsilon(1-x)\chi_{c^*}\frac{1}{N}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right).
\end{aligned}$$

By substituting these facts into Eq. (28), we have

$$\begin{aligned}
&[\mathcal{C}(\boldsymbol{\Sigma}^* + \boldsymbol{E}) - \boldsymbol{\Sigma}^*]_{\alpha,\alpha'}\\
&\approx \epsilon((1-x)\chi_{\kappa} + \chi_{c^*}) - \epsilon\chi_{c^*}x\left(\sum_{s=0}^{K-1} c_s\right)^{-1}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\\
&\quad - \epsilon(1-x)\underbrace{\frac{1}{N}(\chi_{\kappa} + \chi_{c^*} - \chi_{q^*})}_{=\chi_{\kappa}x\left(\sum_{s=0}^{K-1} c_s\right)^{-1}}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\\
&= ((1-x)\chi_{\kappa} + \chi_{c^*})\,\underbrace{\epsilon\left(1 - x\left(\sum_{s=0}^{K-1} c_s\right)^{-1}\sum_{k'=0}^{K-1} c_{k'}\cos\left(\theta_{\alpha,\alpha'}^{(k')}\right)\right)}_{=E_{\alpha,\alpha'}}.
\end{aligned}$$

Finally, we obtain Eq. (27). □

B Proof of Theorem 3.5

Theorem 3.5 (Trainability)**.** *Let $\tilde{\boldsymbol{\Sigma}}^{(\ell)} \in \mathbb{R}^{N\times N}$ be the gradient covariance with respect to some loss $\mathcal{L}$,* e.g. *mean squared error, at the ℓ-th layer. Suppose that the gradient covariance at the L-th layer is decomposed as $\tilde{\Sigma}_{\alpha,\alpha'}^{(L)} = \sum_{k=0}^{K-1}\tilde{\epsilon}_k\cos\left(\theta_{\alpha,\alpha'}^{(k)}\right) + \tilde{\boldsymbol{e}}$, where $\tilde{\epsilon}_k$ is the coefficient of each basis and $\tilde{\boldsymbol{e}}$ belongs to the orthogonal complements of* $\mathrm{span}(\{\cos\left(\theta_{\alpha,\alpha'}^{(k)}\right)\}_{k=0}^{K-1})$*. Then, the gradient covariance*

https://doi.org/10.52202/079017-0352

at the ℓ-th layer is obtained by

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \sum_{k=0}^{K-1} \chi_{c^*}^{L-\ell} \tilde{\epsilon}_k \cos\left(\theta^{(k)}_{\alpha,\alpha'}\right).$$

Proof. Recall the definition of the gradient covariance. We first demonstrate the iterated map of the gradient covariance, starting from this definition.

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} := \mathbb{E}_{\mathbf{\Theta}^{\ell:L},\mathbf{\Xi}^{\ell:L}}\left[g^{(\ell)}_{\alpha,j} g^{(\ell)}_{\alpha',j}\right],\ g^{(\ell)}_{\alpha,j} := \frac{\partial \mathcal{L}}{\partial H^{(\ell)}_{\alpha,j}} = \sum_{i=1}^{D}\sum_{\beta=0}^{N-1} \frac{\partial \mathcal{L}}{\partial H^{(\ell+1)}_{\beta,i}} \frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}}.$$

The Jacobian $\frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}}$ is calculated as follows.

$$\begin{aligned}
&\frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}} \\
&= \sum_{k=0}^{K-1} \sqrt{\frac{c_k}{2}} \left(F^{\dagger}_{\beta,k} F_{k,\alpha}\left(\Theta^{(\ell+1,k)}_{j,i} + \sqrt{-1}\Xi^{(\ell+1,k)}_{j,i}\right) + \overline{F^{\dagger}_{\beta,k} F_{k,\alpha}}\left(\Theta^{(\ell+1,k)}_{j,i} - \sqrt{-1}\Xi^{(\ell+1,k)}_{j,i}\right)\right) \phi'\left(H^{(\ell)}_{\alpha,j}\right) \\
&= \frac{1}{N}\sum_{k=1}^{K-1} 2\sqrt{\frac{c_k}{2}} \left(\Theta^{(\ell+1,k)}_{j,i} \cos\left(\theta^{(k)}_{\beta,\alpha}\right) + \Xi^{(\ell+1,k)}_{j,i} \sin\left(\theta^{(k)}_{\beta,\alpha}\right)\right) \phi'(H^{(\ell)}_{\alpha,j}).
\end{aligned}$$

The covariance of Jacobian $\frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}}$ is as follows.

$$\begin{aligned}
&\mathbb{E}_{\mathbf{\Theta}^{\ell:L},\mathbf{\Xi}^{\ell:L}}\left[\frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}} \frac{\partial H^{(\ell+1)}_{\beta',i}}{\partial H^{(\ell)}_{\alpha',j}}\right] \\
&= \frac{\sigma^2}{N^2 D} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(0,\mathbf{\Sigma}^{(\ell)})}\left[\phi'(H^{(\ell)}_{\alpha,j})\phi'(H^{(\ell)}_{\alpha,j})\right] \\
&\quad \times \left(\sum_{k=0}^{K-1} c_k \left(\cos\left(\theta^{(k)}_{\beta,\alpha}\right)\cos\left(\theta^{(k)}_{\beta',\alpha'}\right) + \sin\left(\theta^{(k)}_{\beta,\alpha}\right)\sin\left(\theta^{(k)}_{\beta',\alpha'}\right)\right)\right) \\
&= \frac{\sigma^2}{N^2 D} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(0,\mathbf{\Sigma}^{(\ell)})}\left[\phi'(H^{(\ell)}_{\alpha,j})\phi'(H^{(\ell)}_{\alpha,j})\right] \left(\sum_{k=0}^{K-1} c_k \cos\left(\theta^{(k)}_{\beta-\beta',\alpha-\alpha'}\right)\right).
\end{aligned}$$

Then, the covariance of the gradient is given by the following recurrence relation for all $\alpha, \alpha' \in [N]$,

$$\begin{aligned}
\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} &:= \mathbb{E}_{\mathbf{\Theta}^{\ell:L},\mathbf{\Xi}^{\ell:L}}\left[g^{(\ell)}_{\alpha,j} g^{(\ell)}_{\alpha',j}\right] \\
&= \sum_{i,i'=1}^{D}\sum_{\beta,\beta'=0}^{N-1} \mathbb{E}_{\mathbf{\Theta}^{\ell:L},\mathbf{\Xi}^{\ell:L}}\left[g^{(\ell+1)}_{\beta,i} g^{(\ell+1)}_{\beta',i'} \frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}} \frac{\partial H^{(\ell+1)}_{\beta',i'}}{\partial H^{(\ell)}_{\alpha',j}}\right] \\
&= \sum_{\beta,\beta'=0}^{N-1}\sum_{i,i'=1}^{D} \mathbb{E}_{\mathbf{\Theta}^{\ell+1:L},\mathbf{\Xi}^{\ell+1:L}}\left[g^{(\ell+1)}_{\beta,i} g^{(\ell+1)}_{\beta',i'}\right] \mathbb{E}_{\mathbf{\Theta}^{\ell:L},\mathbf{\Xi}^{\ell:L}}\left[\frac{\partial H^{(\ell+1)}_{\beta,i}}{\partial H^{(\ell)}_{\alpha,j}} \frac{\partial H^{(\ell+1)}_{\beta',i'}}{\partial H^{(\ell)}_{\alpha',j}}\right] \delta_{i,i'} \\
&= \frac{\sigma^2}{N^2}\sum_{\beta,\beta'=0}^{N-1} \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(0,\mathbf{\Sigma}^{(\ell)})}\left[\phi'(H^{(\ell)}_{\alpha,j})\phi'(H^{(\ell)}_{\alpha',j})\right] \left(\sum_{k=0}^{K-1} c_k \cos\left(\theta^{(k)}_{\beta-\beta',\alpha-\alpha'}\right)\right).
\end{aligned}$$

As with (Schoenholz et al., 2016), we approximate $\mathbf{\Sigma}^{(\ell+1)} \approx \mathbf{\Sigma}^*$ since the number of layer ℓ is assumed to be sufficiently large. Then, the linear iterated map of the gradient covariance $\mathbf{\Sigma}^{(\ell+1)} \mapsto \mathbf{\Sigma}^{(\ell)}$ is given as follows.

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \frac{1}{N^2}\sum_{\beta,\beta'=0}^{N-1} \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'} \chi_{c^*}\left(\sum_{k=0}^{K-1} c_k \cos\left(\theta^{(k)}_{\beta-\beta',\alpha-\alpha'}\right)\right). \tag{29}$$

The rank of the linear iterated map $\mathbf{\Sigma}^{(\ell+1)} \mapsto \mathbf{\Sigma}^{(\ell)}$ is less than K since the matrix representation of the linear map can be decomposed into two matrices $\tilde{\boldsymbol{A}} \in \mathbb{R}^{N^2 \times K}$ and $\tilde{\boldsymbol{B}} \in \mathbb{R}^{K \times N^2}$ as follows.

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \frac{1}{N^2} \sum_{k=0}^{K-1} c_k \exp\left(-\sqrt{-1}\theta^{(k)}_{\beta,\beta'}\right) \exp\left(\sqrt{-1}\theta^{(k)}_{\alpha',\alpha}\right) \chi_{c^*} \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'} = \left(\sum_{k=0}^{K-1} \tilde{A}_{(\alpha,\alpha'),k} \tilde{B}_{k,(\beta,\beta')}\right) \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'},$$

$$\tilde{A}_{(\alpha,\alpha'),k} \coloneqq \frac{1}{N^2} c_k \exp\left(\sqrt{-1}\theta^{(k)}_{\alpha',\alpha}\right) \chi_{c^*},\ \tilde{B}_{k,(\beta,\beta')} \coloneqq \exp\left(-\sqrt{-1}\theta^{(k)}_{\beta,\beta'}\right).$$

Next, we show that the subspace $\mathrm{span}\left(\{\cos\left(\theta^{(k)}_{\alpha,\alpha'}\right)\}_{k=0}^{K-1}\right) \subset \mathbb{R}^{N \times N}$ is the K-dimensional eigenspace with eigenvalue χ_{c^*} of the linear iterated map $\mathbf{\Sigma}^{(\ell+1)} \mapsto \mathbf{\Sigma}^{(\ell)}$.

By substituting $\tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'} = \sum_{k=0}^{K-1} \chi_{c^*}^{L-(\ell+1)} \tilde{\epsilon}_k \cos\left(\theta^{(k)}_{\beta,\beta'}\right)$ into Eq. (29), we obtain

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \frac{1}{N^2} \sum_{\beta,\beta'=0}^{N-1} \chi_{c^*} \left(\sum_{k=0}^{K-1} \tilde{\epsilon}_k \cos\left(\theta^{(k)}_{\beta,\beta'}\right)\right) \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\beta-\beta',\alpha-\alpha'}\right).$$

From the orthogonality of the cosine and sine function, we obtain

$$\begin{aligned}
&\frac{1}{N^2} \sum_{\beta,\beta'=0}^{N-1} \sum_{k'=0}^{K-1} c_{k'} \cos\left(\theta^{(k')}_{\beta-\beta',\alpha-\alpha'}\right) \cos\left(\theta^{(k)}_{\beta,\beta'}\right) \\
&= \frac{1}{N^2} \sum_{\beta,\beta'=0}^{N-1} \sum_{k'=0}^{K-1} c_{k'} \left(\cos\left(\theta^{(k')}_{\beta,\beta'}\right) \cos\left(\theta^{(k')}_{\alpha',\alpha}\right) \cos\left(\theta^{(k)}_{\beta,\beta'}\right)\right. \\
&\qquad \left. + \sin\left(\theta^{(k')}_{\beta,\beta'}\right) \sin\left(\theta^{(k')}_{\alpha',\alpha}\right) \cos\left(\theta^{(k)}_{\beta,\beta'}\right)\right) \\
&= \cos\left(\theta^{(k)}_{\alpha,\alpha'}\right).
\end{aligned} \tag{30}$$

Hence, we have

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \sum_{k=0}^{K-1} \chi_{c^*}^{L-\ell} \tilde{\epsilon}_k \cos\left(\theta^{(k)}_{\alpha,\alpha'}\right).$$

This completes the proof. □

C Discussions on the similarity of DCN and CNN

While CNNs perform local convolutions in the spatial domain, FNOs execute convolutions in the frequency domain, thereby achieving global convolutions in the spatial domain. Our theory for the FNO and the mean-field theory for CNNs (Xiao et al., 2018) share a common focus on the correlation dynamics $\mathbf{\Sigma}^{(0)}, \mathbf{\Sigma}^{(1)}, \ldots, \mathbf{\Sigma}^{(L)}$ of the spatial representations $\mathbf{H}^{(0)}, \mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(L)} \in \mathbb{R}^{N \times D}$. The iterated maps for CNNs and FNOs are obtained as follows.

$$\Sigma^{(\ell+1)}_{\alpha,\alpha'} = \frac{\sigma^2}{2r+1} \sum_{\beta \in \ker} \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(\ell)})} \left[\phi(H_{\alpha+\beta,d}) \phi(H_{\alpha'+\beta,d})\right] + \sigma_b^2 =: \mathcal{C}_{\mathrm{CNN}}(\mathbf{\Sigma}^{(\ell)}),$$

$$\Sigma^{(\ell)}_{\alpha,\alpha'} = \sigma^2 \sum_{k=0}^{K-1} c_k \mathbb{E}_{\mathbf{H}_{:,d} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{(\ell)})} \left[\left|\left[\boldsymbol{F}\phi\left(\mathbf{H}_{:,d}\right)\right]_k\right|^2\right] \cos\left(\theta^{(k)}_{\alpha,\alpha'}\right) + \sigma_b^2 =: \mathcal{C}_{\mathrm{FNO}}(\mathbf{\Sigma}^{(\ell)}),$$

where $2r+1$ is the number of filter width and $\ker = \{\beta \in \mathbb{Z} | |\beta| \leq r\}$ is set of indices referring to the elements of the filter.

When we consider the FNO without mode truncation ($K = N/2+1$), the propagation of the diagonal components (α, α) for any $\alpha \in [N]$ is equivalent to a CNN with filter size N performing global

https://doi.org/10.52202/079017-0352

convolution.

$$\Sigma^{(\ell)}_{\alpha,\alpha} = \sigma^2 \sum_{k=0}^{\frac{N}{2}} c_k \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}^{(\ell)})} \left[\left| [\boldsymbol{F}\phi\left(\mathbf{H}_{:,d}\right)]_k \right|^2 \right] \underbrace{\cos\left(\theta^{(k)}_{\alpha,\alpha}\right)}_{=1} + \sigma_b^2$$

$$= \frac{\sigma^2}{N} \sum_{\beta=0}^{N-1} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}^{(\ell)})} \left[|\phi(H_{\beta,d})|^2 \right] + \sigma_b^2$$

$$= \frac{\sigma^2}{N} \sum_{\beta=0}^{N-1} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(\mathbf{0},\mathbf{\Sigma}^{(\ell)})} \left[\phi(H_{\alpha+\beta,d})\phi(H_{\alpha+\beta,d}) \right] + \sigma_b^2.$$

The first equality follows from Perseval's equality and the second from the periodic boundary condition. In contrast, the propagation of the off-diagonal components differs between CNN and FNO, and the iterated map is different even in the presence of mode truncation.

Regarding the fixed point, Xiao et al. (2018) demonstrated that any fixed point for the iterated map of the DCN is also a fixed point for that of the CNN. Consequently, Lemma A.1 indicates that the fixed points for CNN and FNO are consistent.

The behaviour of the iterated map around fixed points reflects the nature of each architecture. CNN possess diagonal eigenspaces associated with eigenvalues χ_{q^*} and non-diagonal eigenspaces associated with eigenvalues χ_{c^*}. FNOs without mode truncation exhibit a similarity, possessing eigenspaces χ_{q^*} for zero-frequency and eigenspaces χ_{c^*} for k-frequencies with diagonal components removed.

Finally, the iterated map during the backpropagation for CNN and FNO are given by

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \sigma^2 \sum_{\beta\in\ker} v_\beta \tilde{\Sigma}^{(\ell+1)}_{\alpha-\beta,\alpha'-\beta} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(0,\mathbf{\Sigma}^{(\ell)})} \left[\phi'(H^{(\ell)}_{\alpha,j})\phi'(H^{(\ell)}_{\alpha',j}) \right]$$

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha'} = \frac{\sigma^2}{N^2} \sum_{\beta,\beta'=0}^{N-1} \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'} \mathbb{E}_{\mathbf{H}_{:,d}\sim\mathcal{N}(0,\mathbf{\Sigma}^{(\ell)})} \left[\phi'(H^{(\ell)}_{\alpha,j})\phi'(H^{(\ell)}_{\alpha',j}) \right] \left(\sum_{k=0}^{K-1} c_k \cos\left(\theta^{(k)}_{\beta-\beta',\alpha-\alpha'}\right) \right),$$

where v_β is the variance weight parameter dependent of the filter position β, i.e., $w_{i,j}(\beta) \sim \mathcal{N}(0, \sigma^2 v_\beta/D)$, $\sum_{\beta\in\ker} v_\beta = 1$.

Using the approximation $\mathbf{\Sigma}^{(\ell+1)} \approx \mathbf{\Sigma}^*$ as with (Schoenholz et al., 2016), the backpropagation of the diagonal components (α,α) of the FNO without mode truncation is equivalent to that of the global CNN (see Eq. (2.16) in (Xiao et al., 2018)) with $v_\beta = \frac{1}{N}$, as shown below:

$$\tilde{\Sigma}^{(\ell)}_{\alpha,\alpha} \approx \frac{1}{N^2} \sum_{\beta,\beta'=0}^{N-1} \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta'} \chi_{c^*} \underbrace{\left(\sum_{k=0}^{N/2} c_k \cos\left(\theta^{(k)}_{\beta-\beta',\alpha-\alpha}\right) \right)}_{=N\delta_{\beta,\beta'}} = \chi_{c^*} \sum_{\beta=0}^{N-1} \frac{1}{N} \tilde{\Sigma}^{(\ell+1)}_{\beta,\beta}.$$

This equivalence suggests that the edge of chaos initialization (e.g., He initialization) is also valid for the FNO since the problems of gradient vanishing and explosion are determined by the diagonal components of $\tilde{\mathbf{\Sigma}}$.

D Details of Experimental Setup

In this section, we summarize the detailed setup of the all experiments, including the experiments in Section 4.

D.1 Datasets

D.1.1 Advection equation

We used the advection equation data published by Takamoto et al. (2022). The advection equation for the function $u(x,t) \in L^2((0,1)\times(0,2];\mathbb{R})$ is given by

$$\partial_t u(x,t) + \beta\partial_x \left(u(x,t)/2\right) = 0,\ u(x,0) = u_0(x),$$

where $u_0 \in L^2((0,1);\mathbb{R})$ is the initial condition and $\beta \in \mathbb{R}$ is an advection speed set to 2.0. The exact solution is given as $u(x,t) = u_0(x - \beta t)$ for any initial condition u_0.

Only periodic boundary conditions were used in this dataset. The initial conditions are the superposition of sinusoidal wave given by

$$u_0(x) = \sum_{i=1}^{k_{\max}} A_i \sin(k_i x + \phi_i), \tag{31}$$

where $k_i = 2\pi \sum_{j=1}^{N} n_{i,j}/L_x$ are wave numbers whose $n_{i,j}$ are integer numbers randomly chosen in $[1, k_{\max}]$, $L_x = 1$ is the calculation domain size, $N = 2$ is the number of wave to be added, and $k_{\max} = 8$ is the maximum wave number. The amplitude A_i is uniformly chosen in $[0,1]$, and the phase ϕ_i is the randomly chosen in $(0, 2\pi)$. The 2nd-order temporal and spatial upwind finite difference scheme was used for generating the data. Settings are described in Appendix D of (Takamoto et al., 2022).

D.1.2 Burgers' equation

We used the Burgers' equation data published by Takamoto et al. (2022). The Burgers' equation for the function $u(x,t) \in L^2((0,1) \times (0,2];\mathbb{R})$ is given by

$$\begin{aligned} \partial_t u(x,t) + \partial_x \left(u^2(x,t)/2\right) &= \nu \partial_{xx} u(x,t), \\ u(x,0) &= u_0(x), \end{aligned}$$

where $u_0 \in L^2((0,1);\mathbb{R})$ is the initial condition and ν is the diffusion coefficient set to 4.0. The periodic boundary conditions and Equation (31) are used as the initial conditions. The 2nd-order temporal and spatial upwind finite difference scheme is used for generating the data. Settings are described in Appendix D of (Takamoto et al., 2022).

D.1.3 Darcy Flow equation

We used the data of 2D Darcy Flow equation on a regular grid published by Li et al. (2020a). The Darcy Flow equation for the function $u \in H_0^1((0,1)^2;\mathbb{R}_+)$ with a Dirichlet boundary is given by

$$\begin{aligned} -\nabla \cdot (a(x)\nabla u(x)) &= f(x), & x &\in (0,1)^2, \\ u(x) &= 0, & x &\in \partial(0,1)^2, \end{aligned}$$

where $a \in L^\infty((0,1)^2;\mathbb{R}_+)$ is the diffusion coefficient and $f \in L^2((0,1)^2;\mathbb{R})$ is the forcing function. The coefficients a was generated by measure $\mu = \psi_\sharp \mathcal{N}(0, (-\Delta + 9I)^{-2})$ using the Laplacian with zero Neumann boundary and the binary point-wise mapping $\psi(x) = 12\ (x \geq 0),\ 3\ (x < 0)$. The forcing function is fixed $f(x) = 1$. Our aim is to predict the operator mapping the diffusion coefficient to the solution $a \to u$. The solution function u was generated by using the second-order finite difference scheme on a 421×421 grid. Settings are described in Appendix A.3.2 of (Li et al., 2020c).

D.1.4 Incompressible Navier-Stokes equation

We used the 2D NS equation on the unit torus defined by

$$\begin{aligned} \partial_t \omega(x,t) + u(x,t) \cdot \nabla \omega(x,t) &= \nu \nabla^2 \omega(x,t) + f(x), \\ \nabla \cdot u(x,t) = 0,\ \omega(x,0) &= \omega_0(x), \end{aligned}$$

where $\omega(x,t) \in C([0,T]; H^r((0,1)^2;\mathbb{R}^2))$ is the vorticity, $\omega_0 \in H^r((0,1)^2;\mathbb{R}^2)$ is the initial vorticity, $u(x,t) \in C([0,T]; H^r((0,1)^2;\mathbb{R}^2))$ is the velocity field for any $r > 0$, $\nu \in \mathbb{R}_+$ is the viscosity, and $f \in L^2((0,1)^2;\mathbb{R})$ is the external forcing function defined by $f(x) = 0.1\,(\sin(2\pi(x_1 + x_2) + \cos(2\pi(x_1 + x_2))$. The initial vorticity ω_0 was generated by $\omega_0 \sim \mu$ where $\mu = \mathcal{N}(0, 7^{\frac{3}{2}}(-\Delta + 49I)^{-2.5})$ with periodic boundary conditions. The viscosity was set to $1e{-}3$, $1e{-}4$, or $1e{-}5$. Our aim is to predict the operator that maps a solution u up to time 10 to a solution up to some later time $T > 10$. The data was generated by the pseudo-spectral Crak-Nicholson second-order method on 64×64 grid. For the data with $\nu = 1e{-}4$, the time resolution was also downsampled by half. Settings are described in Section 5.3 of (Li et al., 2020c).

https://doi.org/10.52202/079017-0352

Table 2: Training settings

PDE	architecture	batch size	initial lr	max_epochs
advection	Simplified FNO with Tanh	40	5.0×10^{-5}	200
advection	Simplified FNO with ReLU	40	5.0×10^{-5}	200
Burgers'	Simplified FNO with Tanh	40	5.0×10^{-5}	200
Burgers'	Simplified FNO with ReLU	40	1.0×10^{-3}	200
Darcy flow	2D FNO with ReLU	20	1.0×10^{-3}	500
Navier-Stokes ($\nu = 1e$-3)	2D FNO with ReLU	20	1.0×10^{-4}	500
Navier-Stokes ($\nu = 1e$-4)	2D FNO with ReLU	50	2.5×10^{-3}	400
Navier-Stokes ($\nu = 1e$-5)	2D FNO with ReLU	20	2.5×10^{-3}	500

D.2 Training settings

Our detailed training settings of the experiments in Section 4 are provided in Table 2. Our experimental environment consists of an Intel Xeon Plantinum 8360Y (36-core) CPU and a single NVIDIA A100 GPU. Most of our code for experiments are based on the code of PDEBench (`https://github.com/pdebench/PDEBench`) (Takamoto et al., 2022). The only modifications to the model are to multiply the outputs (variable `out_ft` in code of class FNO1d and FNO2d) corresponding to mode $k = 0, N/2$ by $\sqrt{2}$ and to initialize the weights by Gaussian distribution, as described in Section 3.1.

E Detailed Experimental Analysis

E.1 Analysis of Training Loss

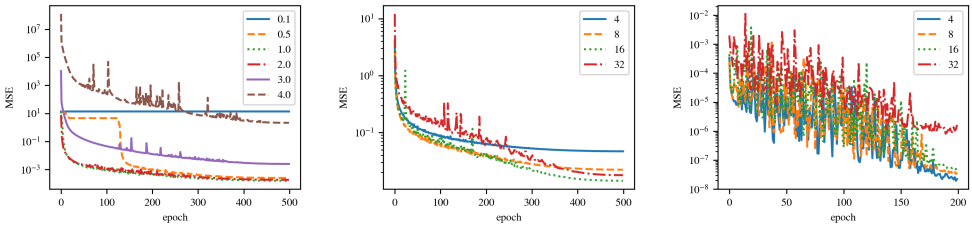
Figure 6a shows the training loss for each epoch of the 32-layer FNOs with parameters $\sigma^2 \in \{0.1, 0.5, 1.0, 2.0, 3.0, 4.0\}$ on the NS dataset with $\nu = 1e{-}3$. When the initial parameter σ^2 is too small, the training loss is not well reduced due to gradient vanishing. On the other hand, when the initial parameter σ^2 is too large, the initial training loss blows up due to gradient exploding. The proposed edge of chaos initialization smoothly reduces the training loss in the initial epoch and enables stable training.

E.2 Analysis of Test Performance

The nMSE of the FNOs on test datasets for six distinct PDEs is presented in Tables 3 and 4. Results are shown only for the FNOs with initial parameters where training was successful in many cases. For the NS equation with viscosity values of $\nu = 1e{-}3, 1e{-}4$, where sufficient data is available, Table 4 shows that best performance is achieved with 8 or 16 layers. This suggests that while shallow FNOs are currently prevalent, deep FNOs could be advantageous in certain tasks, underscoring the significance of our analysis of the bias in deep FNOs. Conversely, for other equations, the 4-layer FNO performs best and deeper FNOs result in a drop in performance even with the edge of chaos initialization. We will discuss this test performance deterioration in detail.

The over-fitting phenomenon is observed in the Darcy Flow and NS equation datasets with $\nu = 1e{-}5$, where only limited training data is available. The training loss for each epoch on the NS equation is depicted in Fig. 6b. As demonstrated in Fig. 6b, the 16 and 32-layer FNOs yield a lower training loss than the 4-layer FNO, but exhibit poorer performance on the test dataset as shown in Table 4. These results suggest over-fitting to the training data, necessitating either abundant training data or appropriate regularization.

Conversely, the under-fitting phenomenon is apparent in the 1D advection and Burgers' equation datasets in Table 3. The training loss of the FNO with ReLU activation for each epoch on the Burgers' equation is presented in Fig. 6c. Figure 6c indicates that the larger the number of layers, the higher the training loss in the final epoch, and the worse the test performance. This under-fitting to the training data could be attributed to the escalating complexity of the loss landscape as the layer count increases, a known issue for DCN and CNN (Li et al., 2018). This may be due to the emergence of local minima corresponding to operators that generate too complex functions, preventing the



(a) NS eq. (1e-3) (b) NS eq. (1e-5) (c) Burgers' eq.

Figure 6: Training Loss Curve. (a): training loss curve of the 32-layer original FNOs with varying initial parameters $\sigma^2 \in \{0.1, 0.5, 1.0, 2.0, 3.0, 4.0\}$, on the NS equation with $\nu = 1e{-}3$. (b): training loss curve of the original FNOs with an initial parameter $\sigma^2 = 2.0$ with a varying number of layers $L \in \{4, 8, 16, 32\}$ on the NS equation with $\nu = 1e{-}5$. (c): training loss curve of the simplified FNOs with ReLU activation and the initial parameter $\sigma^2 = 2.0$ with varying number of layers $L \in \{4, 8, 16, 32\}$ on the Burgers' equation.

Table 3: Test performance measured by nMSE of 1D simplified FNO on 1D PDEs

nMSE	Layers	Advection		Burgers'	
		Tanh	ReLU	Tanh	ReLU
$\sigma^2 = 1.0$	4	0.013	0.015	0.0055	0.00088
	8	0.013	0.015	0.0069	0.0012
	16	0.013	0.015	0.0068	0.0016
	32	0.016	0.018	0.0071	0.0041
$\sigma^2 = 2.0$	4	0.013	0.017	0.0036	0.00098
	8	0.012	0.018	0.0034	0.0011
	16	0.012	0.020	0.0050	0.0016
	32	0.014	0.024	0.0062	0.0027
$\sigma^2 = 3.0$	4	0.013	0.019	0.0044	0.00096
	8	0.014	0.022	0.0042	0.0012
	16	0.020	0.032	0.0060	0.0016
	32	0.053	0.059	0.0093	0.0045

attainment of parameters that achieve global minima. This issue could be mitigated by introducing more suitable regularization, an appropriate optimizer, or a skip connection (Tran et al., 2022).

Our theory and experiments suggest that the training of deep FNOs has suffered from problems including gradient vanishing and exploding due to improper initialization, over-fitting caused by insufficient training data, and under-fitting caused by loss landscapes with strong non-convexity. While our edge of chaos initialization prevents the gradient vanishing and exploding, techniques to solve over-fitting and under-fitting problems are still needed in practice.

F Visualization of Forward Propagation

We visualized the behavior of the simplified FNO's covariance matrix $\mathbf{\Sigma}^{(\ell)}$ with varying initialization parameters $\sigma^2 \in \{0.1, 1.0, 2.0, 4.0\}$ and activation functions. The FNO, with a width of $D = 1024$, was used and the input was sampled from the standard normal distribution with a spatial size of $N = 32$. The results of the FNO with Tanh activation, both with and without mode truncation, are shown in Figs. 7 and 8 and Figs. 9 and 10 respectively. Similarly, the results of the FNO with ReLU activation, both with and without mode truncation, are displayed in Figs. 11 and 12 and Figs. 13 and 14 respectively. In the ordered phase, all figures illustrate convergence to the fixed point $\mathbf{\Sigma}^*$ where $c^* = 1$, with the rate of convergence increasing as the parameter σ^2 decreases. In the chaotic phase, the activation function dictates the covariance behavior. Without mode truncation, the covariance behavior of the FNO is identical to those of the DCN; otherwise non-uniform, FNO-specific periodic covariance is exhibited.

https://doi.org/10.52202/079017-0352

Table 4: Test performance measured by nMSE of 2D original FNO with ReLU activation on Darcy Flow and NS equation.

nMSE	Layers	Darcy Flow	Navier-Stokes $\nu = 1e{-}3$	$\nu = 1e{-}4$	$\nu = 1e{-}5$
$\sigma^2 = 1.0$	4	0.025	0.0063	0.18	0.10
	8	0.028	0.0047	0.14	0.094
	16	0.035	0.0048	0.12	0.12
	32	0.56	0.0057	0.13	0.16
$\sigma^2 = 2.0$	4	0.029	0.0075	0.18	0.10
	8	0.036	0.0057	0.14	0.11
	16	0.041	0.0057	0.12	0.11
	32	0.042	0.0072	0.13	0.18
$\sigma^2 = 3.0$	4	0.033	0.0089	0.18	0.10
	8	0.040	0.0080	0.14	0.11
	16	0.052	0.0098	0.13	0.13
	32	0.16	0.028	0.14	0.19

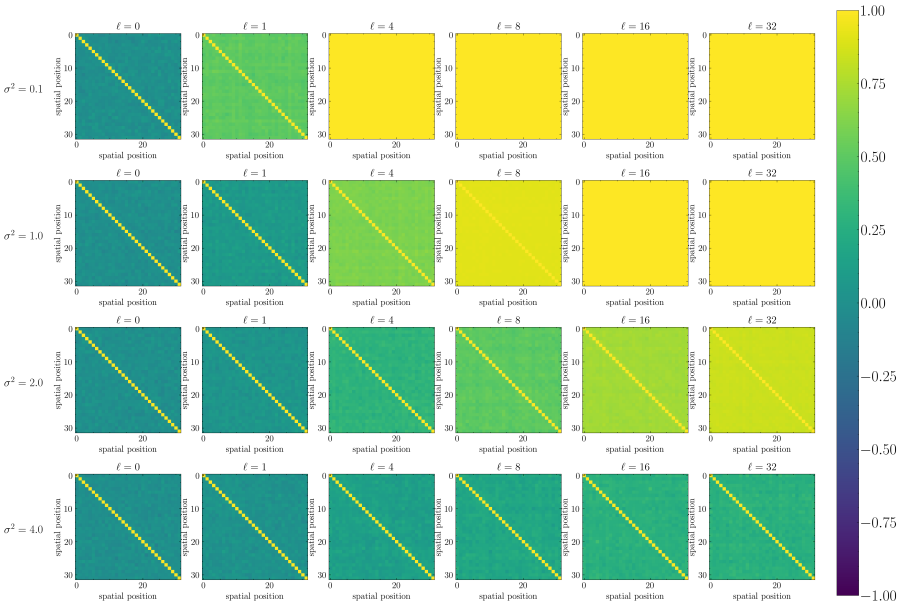


Figure 7: Visualization of the correlation $\Sigma^{(\ell)}_{\beta,\beta'} / \sqrt{\Sigma^{(\ell)}_{\beta,\beta} \Sigma^{(\ell)}_{\beta',\beta'}}$ for the simplified FNO with Tanh activation and no mode truncation.

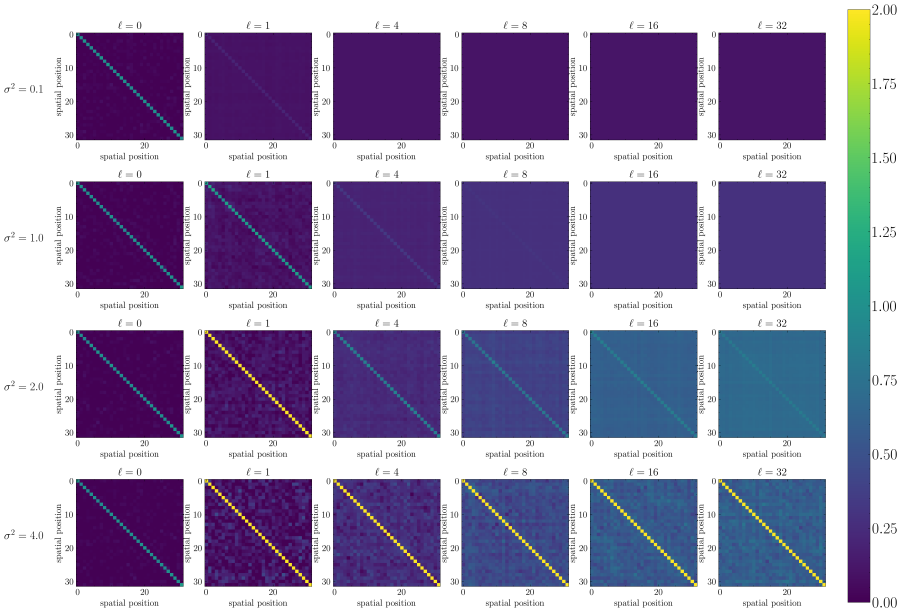


Figure 8: Visualization of the covariance $\mathbf{\Sigma}^{(\ell)}$ for the simplified FNO with Tanh activation and no mode truncation.

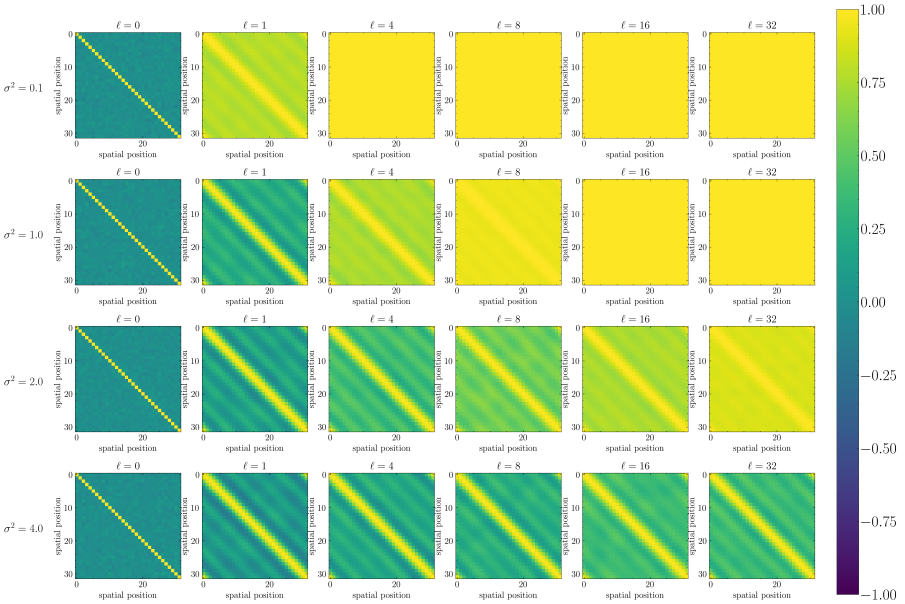


Figure 9: Visualization of the correlation $\Sigma^{(\ell)}_{\beta,\beta'} / \sqrt{\Sigma^{(\ell)}_{\beta,\beta}\Sigma^{(\ell)}_{\beta',\beta'}}$ for the simplified FNO with Tanh activation and the Fourier mode $K = 5$.

https://doi.org/10.52202/079017-0352

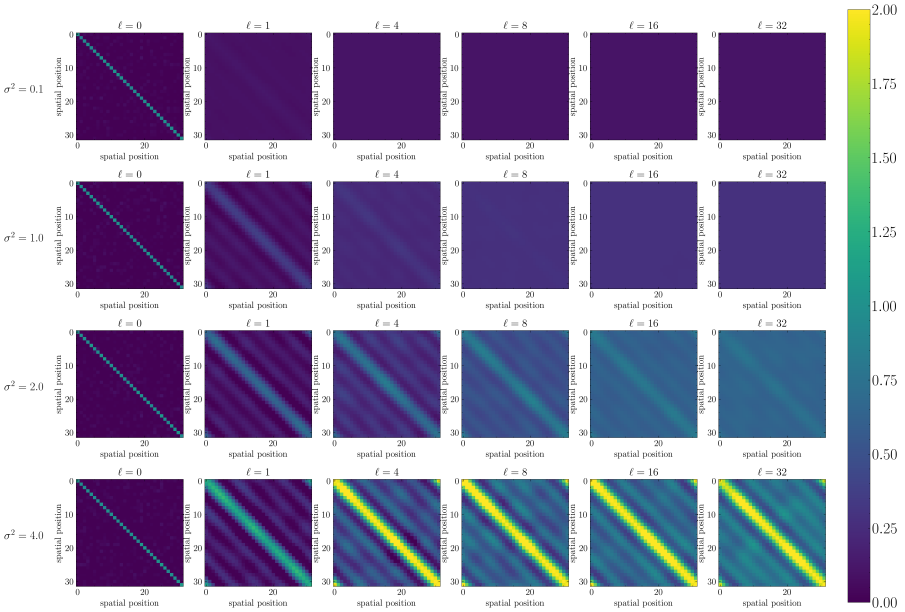


Figure 10: Visualization of the covariance $\mathbf{\Sigma}^{(\ell)}$ for the simplified FNO with Tanh activation and the Fourier mode $K = 5$.

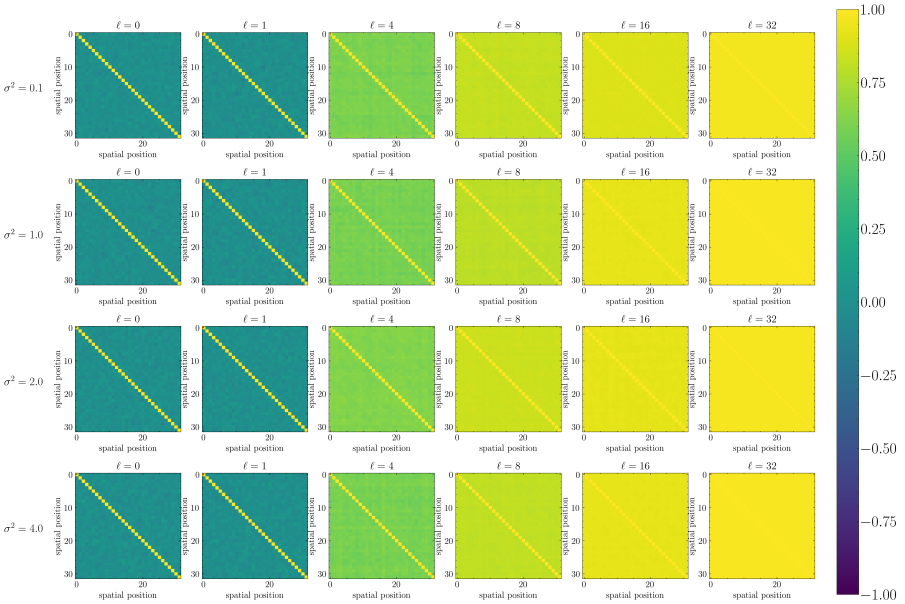
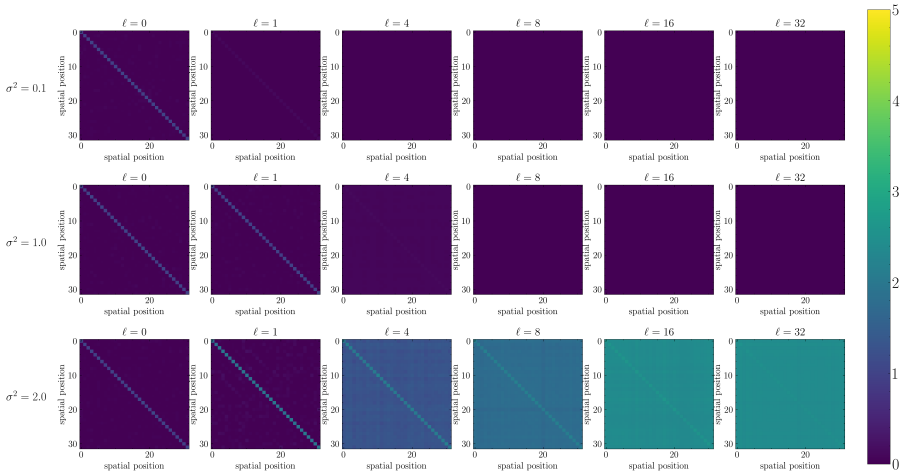
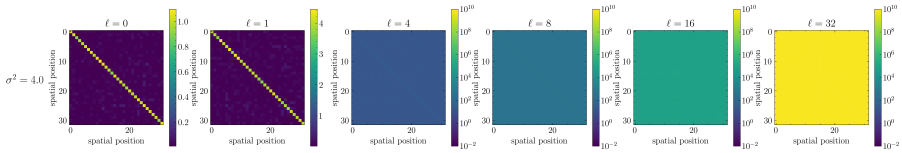


Figure 11: Visualization of the correlation $\Sigma^{(\ell)}_{\beta,\beta'} / \sqrt{\Sigma^{(\ell)}_{\beta,\beta} \Sigma^{(\ell)}_{\beta',\beta'}}$ for the simplified FNO with ReLU activation and no mode truncation.



(a) Initial parameters $\sigma^2 \in \{0.1, 1.0, 2.0\}$.



(b) Initial parameter $\sigma^2 = 4.0$

Figure 12: Visualization of the covariance $\mathbf{\Sigma}^{(\ell)}$ for the simplified FNO with ReLU activation and no mode truncation.

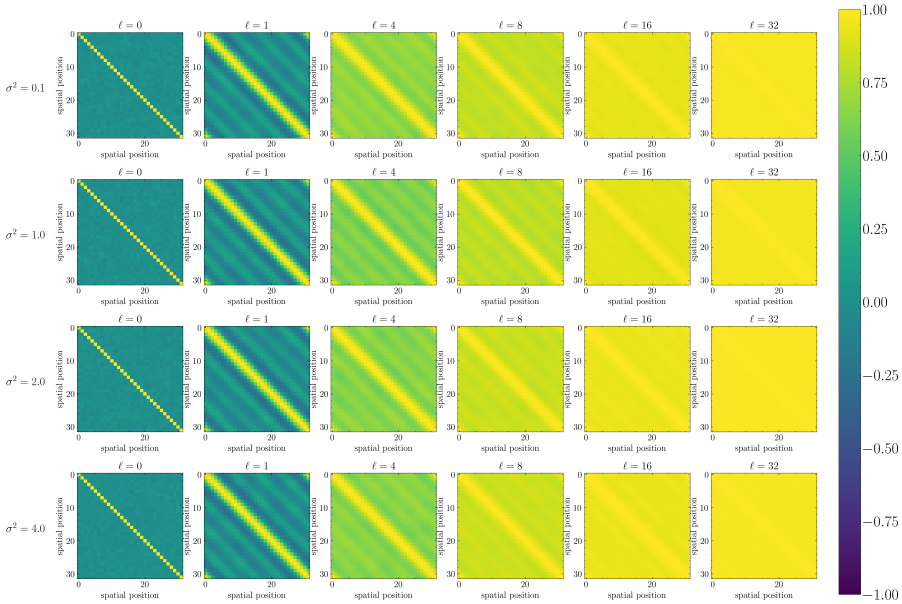
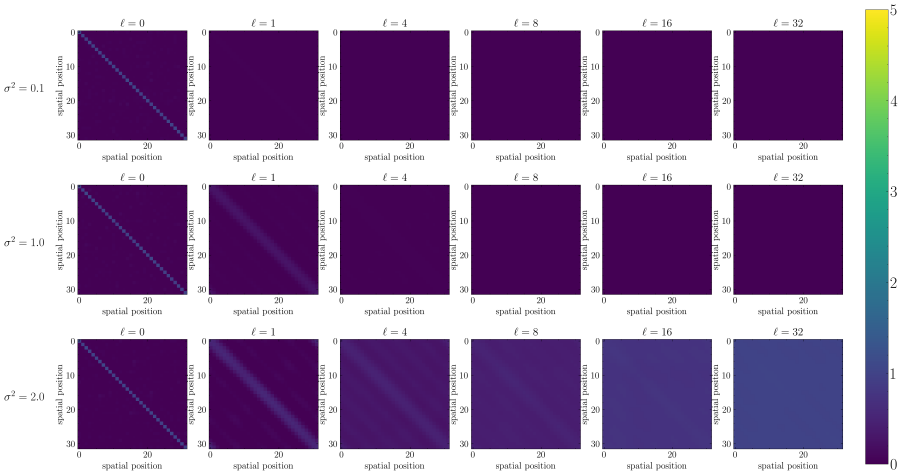
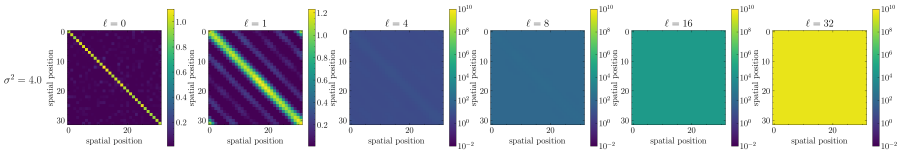


Figure 13: Visualization of the correlation $\Sigma^{(\ell)}_{\beta,\beta'} / \sqrt{\Sigma^{(\ell)}_{\beta,\beta} \Sigma^{(\ell)}_{\beta',\beta'}}$ for the simplified FNO with ReLU activation and the Fourier mode $K = 5$.

https://doi.org/10.52202/079017-0352



(a) Initial parameters $\sigma^2 \in \{0.1, 1.0, 2.0\}$.



(b) Initial parameter $\sigma^2 = 4.0$

Figure 14: Visualization of the covariance $\mathbf{\Sigma}^{(\ell)}$ for the simplified FNO with ReLU activation and the Fourier mode $K = 5$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contribution and scope of the paper are described in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention the limitation of this paper in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-0352

Justification: The full set of assumptions and a complete proof are provided in Section 3 and Appendices A and B.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the information needed to reproduce the main experimental results are provided in Section 4 and Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We use existing public datasets and cite them appropriately.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: All training and test details are provided in Section 4 and Appendix D.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: In the experimental results depicted in Figs. 3 and 4, we determined that the variation in the reported values has minimal impact. In fact, the inclusion of error bars seems to obstruct intuitive understanding through visualization.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

https://doi.org/10.52202/079017-0352

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All sufficient information on the computer resources are provided in Section 4 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our paper is foundational research and not tied to particular applications.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have stated the license and copyright of the data used.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

https://doi.org/10.52202/079017-0352

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The content of our public code is well described.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.