CHIN. PHYS. LETT. Vol. 38, No. 3 (2021) 038701

A Linear Frequency Principle Model to Understand the Absence of Overfitting in Neural Networks

Yaoyu Zhang(张耀宇)[1,2], Tao Luo(罗涛)[1], Zheng Ma(马征)[1], and Zhi-Qin John Xu(许志钦)[1*]

[1]*School of Mathematical Sciences, Institute of Natural Sciences, MOE-LSC, and Qing Yuan Research Institute, Shanghai Jiao Tong University, Shanghai 200240, China*
[2]*Shanghai Center for Brain Science and Brain-Inspired Technology, Shanghai 200031, China*

(Received 27 September 2020; accepted 8 January 2021; published online 2 March 2021)

Why heavily parameterized neural networks (NNs) do not overfit the data is an important long standing open question. We propose a phenomenological model of the NN training to explain this non-overfitting puzzle. Our linear frequency principle (LFP) model accounts for a key dynamical feature of NNs: they learn low frequencies first, irrespective of microscopic details. Theory based on our LFP model shows that low frequency dominance of target functions is the key condition for the non-overfitting of NNs and is verified by experiments. Furthermore, through an ideal two-layer NN, we unravel how detailed microscopic NN training dynamics statistically gives rise to an LFP model with quantitative prediction power.

DOI: 10.1088/0256-307X/38/3/038701

Deep learning, a subfield of machine learning achieving huge success in industrial applications, is experiencing a surge in many areas of sciences including physics.[1–7] A typical well-solved problem is supervised learning, where the machine learns a mapping from input $\boldsymbol{x} \in \mathbb{R}^d$ to output $\boldsymbol{y} \in \mathbb{R}^{d_o}$ from a training dataset $S = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$. The machine is realized by a deep neural network (DNN) of proper depth L:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}^{[L]}\sigma\circ[\cdots \boldsymbol{W}^{[2]}\sigma\circ(\boldsymbol{W}^{[1]}\boldsymbol{x}+\boldsymbol{b}^{[1]})+\cdots]+\boldsymbol{b}^{[L]},$$

where $\boldsymbol{\theta} = \{\boldsymbol{W}^{[l]}, \boldsymbol{b}^{[l]}\}_{l=1}^L$, $\boldsymbol{W}^{[l]}$ are weight matrices, $\boldsymbol{b}^{[l]}$ are bias vectors, and $\sigma \circ (\cdots)$ is an element-wise nonlinear activation function. Parameters $\boldsymbol{\theta}$ are updated during the training by minimizing an empirical risk/loss function characterizing the difference between the DNN outputs and the correct outputs, e.g., $R_S(\boldsymbol{\theta}) = \sum_{i=1}^n ||f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - \boldsymbol{y}_i||^2/2n$ for the loss of training dataset S, with gradient-based algorithms. Due to the highly nonlinear nature of the neural network (NN) model, many key theoretical questions raised by Leo Breiman decades ago remain unanswered.[8,9] This work focuses on one of them: why heavily parameterized neural networks do not overfit the data, which is further backed by recent experimental works in large datasets and deep networks.[10] Note that, establishing a good theoretical understanding of this non-overfitting puzzle has become more and more crucial for applications of DNNs because modern DNN architectures with tons of parameters, e.g., $\sim 10^8$ for VGG19,[11] $\sim 10^{11}$ for GPT-3,[12] indeed achieve huge success in practice. However, theoretical understanding to this puzzle is not obvious at all, because it contradicts the doctrine in physics and statistical learning theory implied by von Neumann's famous quote "with four parameters I can fit an elephant".[13] Existing theories based on idealized models of DNNs, e.g., deep linear network,[14–16] committee machine,[17,18] spin glass model,[19] mean-field model,[20–23] neural tangent kernel,[24,25] which emphasize on fully rigorous mathematical proofs, have difficulties in providing a satisfactory explanation.[9]

The training process of NN under gradient flow can be viewed as collective dynamics of a large group of interacting neurons/parameters driven by training data. In analogy to statistical mechanics, there are microscopic levels of NNs caring about the detailed dynamics of each component of $\boldsymbol{\theta}$ and macroscopic level caring about dynamics of statistical quantities of $\boldsymbol{\theta}$, among which $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ as a function-valued quantity is the most important one. At the macroscopic level, it was suggested recently that DNNs learn simple patterns (e.g., certain coarse description or landscape of dataset) first.[26–28] Based on the intuition that low frequency functions, i.e., functions with energy mainly concentrated at low frequencies, are of low complexity, Refs. [29–32] quantify the complexity of $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ by its frequency composition, and demonstrated the general phenomenon of frequency principle (F-principle)—NNs often learn low frequencies first. For example, when a DNN is used to fit data generated from 1-d target function $\sin x + \sin(5x)$, while $\sin x$ and $\sin(5x)$ have the same amplitude, the low frequency $\sin x$ is first captured, and later the target as shown in Fig. 1. This phenomenon can be robustly observed no matter how overparameterized NNs are. The F-principle has initiated a series of subsequent

Supported by the National Key R&D Program of China (Grant No. 2019YFA0709503), the Shanghai Sailing Program, the Natural Science Foundation of Shanghai (Grant No. 20ZR1429000), the National Natural Science Foundation of China (Grant No. 62002221), Shanghai Municipal of Science and Technology Project (Grant No. 20JC1419500), and the HPC of School of Mathematical Sciences at Shanghai Jiao Tong University
*Corresponding author. Email: xuzhiqin@sjtu.edu.cn

works,[33–36] and inspired the design of DNN-based algorithms.[37–42]

In this Letter, starting from this key macroscopic dynamical feature of F-principle, we establish a theory for the non-overfitting puzzle. We propose a linear frequency principle (LFP) model for the phenomenological characterization of the F-principle. Based on the LFP model, we establish a theory which explains the non-overfitting puzzle, and experimentally test its qualitative predictions about failures of NNs. Furthermore, through an ideal example of two-layer NN in the infinite proper width limit, we unravel how microscopic reality of NN training dynamics statistically gives rise to an LFP model. Finally, we demonstrate the quantitative prediction power of the LFP model through experiments.

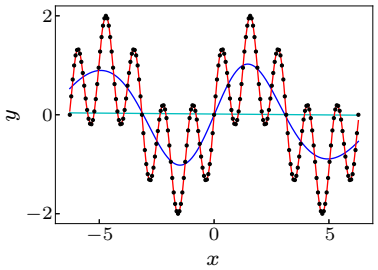

Fig. 1. Illustration of the training process of a DNN. Black dots are training data sampled from target function $\sin x + \sin(5x)$. Cyan, blue and red curves indicates $f_{\boldsymbol{\theta}(t)}(\boldsymbol{x})$ at training epochs $t = 0$, 2000, 17000, respectively.

The LFP Model. F-principle is "opposite" to common physical processes with diffusion, in which high frequency modes dissipate faster than low frequency ones. To phenomenologically model such a process, we consider a dynamics in frequency domain, in which each frequency mode evolves to certain target determined by training data $\{\boldsymbol{x}_i \in \mathbb{R}^d, f^*(\boldsymbol{x}_i) \in \mathbb{R}\}_{i=1}^n$ with a positive rate $\gamma(\boldsymbol{\xi})$ decaying as frequency $\boldsymbol{\xi} \in \mathbb{R}^d$ increases. In particular, in this study we show that for a wide two-layer NN, explicit form of $\gamma(\boldsymbol{\xi})$, which is a linear combination of $\frac{1}{||\boldsymbol{\xi}||^{d+1}}$ and $\frac{1}{||\boldsymbol{\xi}||^{d+3}}$, can be derived to accurately predict the NN outputs after training. Before that, we begin with proposing the following general model for F-principle,

$$\partial_t \hat{h}(\boldsymbol{\xi}, t) = -\gamma(\boldsymbol{\xi})\Big(\hat{h_\rho}(\boldsymbol{\xi}, t) - \hat{f}^*_\rho(\boldsymbol{\xi})\Big), \tag{1}$$

where $h(\boldsymbol{x}, t)$ models $f_{\boldsymbol{\theta}(t)}(\boldsymbol{x})$ with microscopic details neglected, $\hat{(\cdots)}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} (\cdots)(\boldsymbol{x}) e^{-I\boldsymbol{x}\cdot\boldsymbol{\xi}} d\boldsymbol{x}$ is the Fourier transform. The initial condition is set to $h(\boldsymbol{x}, 0) = h_{\text{ini}}(\boldsymbol{x})$, $(\cdots)_\rho(\boldsymbol{x}) = (\cdots)(\boldsymbol{x})\rho(\boldsymbol{x})$; $\rho(\boldsymbol{x})$ is the data distribution, which can be a continuous function or a probability function for discrete training data points, that is, $\rho(\boldsymbol{x}) = \sum_{i=1}^n \delta(\boldsymbol{x} - \boldsymbol{x}_i)/n$ with $\delta(\cdots)$ being the dirac delta function, an uncommon part of this dynamics. Since the steady state requires the model prediction equal to the target function only at the empirical training data points, at the steady state, no explicit constraint is imposed on the unseen data points, therefore $h(\boldsymbol{x}, \infty)$ can drastically deviate from the target $f^*(\boldsymbol{x})$ at unseen data points. We call model (1) *Linear frequency principle (LFP) model*, in which "linear" refers to the fact that model (1) is a linear differential equation in h. For simplicity, we set $u(\boldsymbol{x}, t) = h(\boldsymbol{x}, t) - f^*(\boldsymbol{x})$ with the LFP model simplified to $\partial_t \hat{u}(\boldsymbol{\xi}, t) = -\gamma(\boldsymbol{\xi})\hat{u_\rho}(\boldsymbol{\xi}, t)$.

We relate dynamics of model (1) with the dynamics of least square loss, that is, model (1) is a dissipative process with a decreasing loss (in analogy to energy),

$$R_S = \frac{1}{2}\int u^2 \rho d\boldsymbol{x} = \frac{1}{2}\int \hat{u}\hat{u_\rho}^* d\boldsymbol{\xi}, \tag{2}$$

governed by $\frac{d}{dt}R_S = -\int \gamma |\hat{u_\rho}|^2 d\boldsymbol{\xi} < 0$. The dissipation of loss at each frequency is governed by $\partial_t(\hat{u}\hat{u_\rho}^*/2) = -\gamma|\hat{u_\rho}|^2$. More importantly, because $\gamma(\boldsymbol{\xi})$ is a decaying function by F-principle, e.g., $\gamma(\boldsymbol{\xi}) = \frac{1}{||\boldsymbol{\xi}||^{d+1}}$, loss decreases faster over lower frequencies. This behavior is essential for overcoming the singularity in u_ρ as a summation of delta functions. Otherwise, if $\gamma(\boldsymbol{\xi}) = ||\boldsymbol{\xi}||^2$, model (1) becomes a heat-diffusion-type equation $\partial_t u = -\Delta u_\rho$ in spatial domain and is not well-posed for non-differentiable u_ρ. As the study of waves by mode decomposition, the coefficient $\gamma(\boldsymbol{\xi})$ as a function of frequency $\boldsymbol{\xi}$ plays an important role in governing the macroscopic phenomenon of training dynamics. For example, for a $\gamma(\boldsymbol{\xi})$ decaying with $\boldsymbol{\xi}$, model (1) first learns the landscape or a simple pattern of the training data, followed by more details or complex patterns, exemplified by the case in Fig. 1. However, for a $\gamma(\boldsymbol{\xi})$ increasing with $\boldsymbol{\xi}$, the learning behavior is opposite. Note that, there are infinite feasible decay functions of $\gamma(\boldsymbol{\xi})$ obeying the F-principle. In general, power-law decay is relevant to an activation of singularity in derivatives, e.g., ReLU, whereas exponential decay is relevant to a smooth activation, e.g., tanh.

In the following, we further analyze our proposed LFP model (1) in two folds. First, we show that the long-time solution of model (1) is equivalent to the solution of an optimization problem, which reveals the low-frequency bias of the LFP model. Based on the optimization problem, we obtain a generalization error estimate for understanding the non-overfitting puzzle. Second, as an example, we exactly compute the LFP model for two-layer wide ReLU networks.

Theory for the Non-Overfitting Puzzle. Our LFP model searches for the fitting of n points of f^* in an infinite dimensional function space. Clearly, it possesses infinite steady states (minimizers of E) that satisfy $h(\boldsymbol{x}_i) = f^*(\boldsymbol{x}_i)$ for $i = 1, \ldots, n$. If we arbitrarily pick one steady state of h, it is likely to generalize poorly, i.e., h deviates drastically from target f^* on unobserved positions, resulting in overfitting as commonly expected from an overparameterized model. However, given proper $h_{\text{ini}}(\boldsymbol{x})$ and $\gamma(\boldsymbol{\xi})$, we obtain a unique

steady state $h(\boldsymbol{x},\infty)$, denoted by $h_\infty(\boldsymbol{x})$ for simplicity. Exploiting the linearity of the LFP model, we derive that $h_\infty(\boldsymbol{x})$ satisfies the following constrained minimization problem:

$$\begin{aligned} &\min_h \int \gamma(\boldsymbol{\xi})^{-1}|\hat{h}(\boldsymbol{\xi})-\hat{h}_{\rm ini}(\boldsymbol{\xi})|^2 d\boldsymbol{\xi}, \\ &\text{s.t.}\quad h(\boldsymbol{x}_i)=f^*(\boldsymbol{x}_i), \quad i=1,\ldots,n. \end{aligned} \tag{3}$$

Note that solution to this problem may generalize poorly if $h_{\rm ini}$ attains any function. For example, if $h_{\rm ini}$ attains a "bad" steady state, then solution of the problem $h=h_{\rm ini}$ will also be "bad". However, in practice, common initialization of NNs yields small output. Without loss of generality, we consider in the following an unbiased initial function $h_{\rm ini}=0$, which can be achieved in NNs by applying the AntiSymmetrical Initialization (ASI) trick.[43]

This static minimization problem defines an FP-energy $E_\gamma(h)=\int \gamma^{-1}|\hat{h}|^2 d\boldsymbol{\xi}$ that quantifies the preference of the LFP model among all its steady states. Because $\gamma(\boldsymbol{\xi})^{-1}$ is an increasing function, say $\gamma(\boldsymbol{\xi})^{-1}=||\boldsymbol{\xi}||^{d+1}$, the FP-energy $\int ||\boldsymbol{\xi}||^{d+1}|\hat{h}|^2 d\boldsymbol{\xi}$ amplifies the high frequencies while diminishing low frequencies. By minimizing $E_\gamma(h)$, problem (3) gives rise to a low frequency fitting, instead of an arbitrary one, of training data. By intuition, if target f^* is indeed low frequency dominant, then h_∞ will likely well approximate f^* at unobserved positions.

To theoretically demonstrate above intuition, we derive in the following an estimate of the generalization error of h_∞ using *a priori* error estimate technique.[44] Because $h(\boldsymbol{x})=f^*(\boldsymbol{x})$ is a viable steady state, $E_\gamma(h_\infty)\le E_\gamma(f^*)$ by the minimization problem. Using this constraint on h_∞, we obtain that, with probability of at least $1-\delta$,

$$\mathbb{E}_{\boldsymbol{x}}[h_\infty(\boldsymbol{x})-f^*(\boldsymbol{x})]^2 \le \frac{E_\gamma(f^*)}{\sqrt{n}} C_\gamma\left(2+4\sqrt{2\log(4/\delta)}\right), \tag{4}$$

where C_γ is a constant depending on γ. Error reduces with more training data as expected with a decay rate $1/\sqrt{n}$ similar to the Monte Carlo method. Importantly, because $E_\gamma(f^*)$ strongly amplifies high frequencies of f^*, the more high-frequency components the target function f^* possesses, the worse h_∞ may generalize.

The above theory explains the non-overfiting puzzle of NNs as follows: regardless of the number of parameters of NNs, the F-principle dynamics finds for an overparameterized NN with a low frequency fitting of training data, which unlikely overfits a low frequency target function (since the FP-norm is small for low-frequency function). Specifically, it predicts the following qualitative behaviors of NNs. (i) *Preference*: NNs preferentially learn low frequency fittings of training data. (ii) *Success*: NNs often generalize for low frequency dominant target functions. (iii) *Failure*: NNs likely overfit a high frequency target function.

In the following, we test whether these predictions well hold for NNs in experiments. In the first experiment, we use a DNN to fit high dimensional high frequency dominant data sampled from a parity function $f(\boldsymbol{x})=\prod_{j=1}^d x_j$ defined on $\Omega=\{-1,1\}^d$, whose Fourier transform $(-I)^d\prod_{j=1}^d \sin 2\pi k_j$ for $\boldsymbol{k}\in[-\frac{1}{4},\frac{1}{4}]^d$ peaks at its highest frequencies $\boldsymbol{k}\in\{-\frac{1}{4},\frac{1}{4}\}^d$. The difficulty of learning the parity function with NNs is well-known.[45,46] We provide a frequency perspective to understand this learning difficulty. For high-dimensional function, we perform a non-uniform discrete Fourier transform on the first principle direction of a training data set. As demonstrated in Fig. 2(a), the well-trained DNN indeed preferentially learns more low frequencies and less high frequencies compared to the target. Furthermore, as predicted by the model, the DNN generalizes badly with a low test accuracy 38% no more than chance-level 50% (while training accuracy is 100%!). In the second experiment, we use the widely considered image classification dataset of CIFAR10 as an example, on which a well-trained DNN achieves a test accuracy 68% much higher than chance-level 10%, and compute its frequency composition by non-uniform discrete Fourier transform. As shown in Fig. 2(b), the target is indeed dominated by low frequencies. Actually, this low frequency dominance property for most real high dimensional image data can be intuitively understood based on the common sense that a small perturbation in input image mostly does not change the its category as output. Furthermore, as is predicted, DNN preferentially learns the low frequencies better than the high ones, leading to a good generalization.

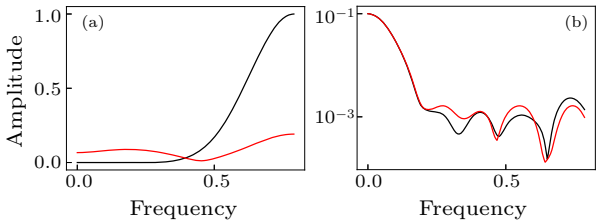

Fig. 2. Frequency composition (amplitude vs frequency) of target (black) and the well-trained DNN (red) along the first principle component direction of inputs of training data. (a) Target: 10-d parity function; NN: three-layer fully connected net. (b) Target: CIFAR10; NN: two convolutional layers with a fully connected layer.

LFP Model Derived from a Two-Layer NN. Analysis of the training process of a multi-layer ($L\ge 2$) NN is well known to be difficult.[9] Recently, based on a dynamical regime of neural tangent kernel (NTK), where the gradient flow of overparameterized NNs can be effectively linearized around initialization, fruitful mathematical theorems were proved at an abstract level about the behavior of NNs.[24,47,48] Still, deriving explicitly the linearized dynamics of even a two-layer NN, which already possesses similar nontrivial training and generalization behavior as deeper NNs,

for quantitative analysis is a challenging task. In this part, we present such a derivation in frequency domain, which yields an LFP model with a specific $\gamma(\boldsymbol{\xi})$ depending on detailed setups of the target NN, such as smoothness of σ and statistics of $\boldsymbol{\theta}(0)$. Note that, since the F-principle generally exists in deeper NNs and in both NTK and non-NTK regimes, the mechanism unraveled by the above theoretical analysis, i.e., low frequency first learning dynamics leads to a low frequency fitting of data, applies to general DNNs where the NTK theory can drastically fail.

Considering the following two-layer neural network

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{m}}\sum_{j=1}^{m} a_j \sigma(\boldsymbol{w}_j^{\mathbf{T}}\boldsymbol{x} + r_j c_j), \tag{5}$$

where $r_j := |\boldsymbol{w}_j|$ and $\sigma(x) = \max(0, x)$, i.e., the widely used ReLU (rectified linear unit) activation. Note that our following derivation applies similarly to other σ such as sigmoid or tanh activation. Denote $\boldsymbol{p}_j = (a_j, \boldsymbol{w}_j, c_j)^{\mathbf{T}} \in \mathbb{R}^{d+2}$, $\boldsymbol{\theta} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m)$. During the learning process, i.e., fitting training data $\{[\boldsymbol{x}_i, f^*(\boldsymbol{x}_i)]\}_{i=1}^n$ generated from a target function $f^*(\boldsymbol{x})$ by model (5), $\boldsymbol{\theta}$ evolves by the gradient descent with dynamics at continuous limit

$$\dot{\boldsymbol{\theta}} = -\nabla R_S(\boldsymbol{\theta}), \tag{6}$$

with mean-squared error (MSE) loss $R_S(\boldsymbol{\theta})$ of the empirical sample distribution in Eq. (2).

At initialization, a_j, $\boldsymbol{w}_j$, and c_j for $j = 1, \ldots, m$ are sampled independently from random distributions under mild assumptions that (i) distribution of $\boldsymbol{w}_j/|\boldsymbol{w}_j|$ is uniform on the unit sphere; (ii) variance of c_j, denoted by σ_c^2, is sufficiently large.

In general, dynamics (6) is difficult to be analyzed due to its high-dimensional and highly nonlinear nature similar to particle systems in statistical mechanics.[21] In the following, we show how an LFP macroscopic statistical description of above dynamics can be derived at the infinite neuron limit $m \to \infty$, which has been considered in Refs. [20,21,23–25], in analogy to the thermodynamic limit. This limit with the scaling factor of $1/\sqrt{m}$ in NN (5) makes its linearization around initialization

$$f^{\mathrm{lin}}[\boldsymbol{x};\boldsymbol{\theta}(t)] = f[\boldsymbol{x};\boldsymbol{\theta}(0)] + \nabla_{\boldsymbol{\theta}} f[\boldsymbol{x};\boldsymbol{\theta}(0)][\boldsymbol{\theta}(t) - \boldsymbol{\theta}(0)] \tag{7}$$

an effective approximation of $f[\boldsymbol{x};\boldsymbol{\theta}(t)]$, i.e., $f^{\mathrm{lin}}[\boldsymbol{x};\boldsymbol{\theta}(t)] \approx f[\boldsymbol{x};\boldsymbol{\theta}(t)]$ for any t, as demonstrated by both theoretical and empirical studies of neural tangent kernels (NTK).[24,25] Note that, $f^{\mathrm{lin}}[\boldsymbol{x};\boldsymbol{\theta}(t)]$, linear in $\boldsymbol{\theta}$ and nonlinear in $\boldsymbol{x}$, reserves the universal approximation power of $f[\boldsymbol{x};\boldsymbol{\theta}(t)]$ at $m \to \infty$. In the following, we do not distinguish $f[\boldsymbol{x};\boldsymbol{\theta}(t)]$ from $f^{\mathrm{lin}}[\boldsymbol{x};\boldsymbol{\theta}(t)]$.

Again, analogous to statistical mechanics, while dynamics (6) acts at a microscopic level on parameters of each neuron, function $f[\boldsymbol{x},\boldsymbol{\theta}(t)]$ for the fitting problem is macroscopic. For simplicity, we denote $f[\boldsymbol{x},\boldsymbol{\theta}(t)]$ by $f(\boldsymbol{x},t)$. The evolution of $f(\boldsymbol{x},t)$ in the NTK regime follows gradient flow, i.e.,

$$\begin{aligned}\partial_t f(\boldsymbol{x},t) &= \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x},t)\cdot\partial_t\boldsymbol{\theta} \\ &= -\int_{\mathbb{R}} u_\rho(\boldsymbol{x}') K_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{x}')d\boldsymbol{x}',\end{aligned}$$

where $K_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{x}') = \nabla_{\boldsymbol{\theta}} f[\boldsymbol{x}',\boldsymbol{\theta}(0)]\nabla_{\boldsymbol{\theta}} f[\boldsymbol{x},\boldsymbol{\theta}(0)]$, $u(\boldsymbol{x}) = f(\boldsymbol{x},t) - f^*(\boldsymbol{x})$, $u_\rho(\boldsymbol{x}) = u(\boldsymbol{x})\rho(\boldsymbol{x})$. This gradient flow applies for deep neural networks with arbitrary hidden layers in the NTK regime. However, to derive an explicit form of the kernel $K_{\boldsymbol{\theta}}$, we limit our analysis to the two-layer ReLU neural network. By applying Fourier transform to both sides of the above equation, with approximation we obtain

$$\partial_t \hat{u}(\boldsymbol{\xi},t) = -\left[\frac{\langle a^2+r^2\rangle_{a,r}}{||\boldsymbol{\xi}||^{d+3}} + \frac{\langle a^2 r^2\rangle_{a,r}}{||\boldsymbol{\xi}||^{d+1}}\right]\hat{u_\rho}(\boldsymbol{\xi},t), \tag{8}$$

where $\langle\cdots\rangle_{a,r}$ is the expectation with respect to the initial distribution of a and r. Clearly, it is an LFP model that prioritizes the learning of low frequencies quantified by mixed power law decay. This power law decay results from the decay of the spectrum of σ depending on its smoothness. For a sigmoid or tanh activation, an exponential decay will be obtained. This model signifies the analogy between NN and statistical mechanics that the learning process of NN with a large number of neurons is effectively captured by several statistics, e.g., $\langle a^2+r^2\rangle_{a,r}$ and $\langle a^2 r^2\rangle_{a,r}$, with microscopic details neglected. Remark that, to derive model (8), we ignore an additional term arising from the rotation of $\boldsymbol{w}$'s for $d \geq 2$, that is, $\frac{\langle r^2\rangle_r}{\|\boldsymbol{\xi}\|_2^{d+1}}\Delta_{\boldsymbol{\xi}^\perp}\hat{u_\rho}(\boldsymbol{\xi})$, where $\Delta_{\boldsymbol{\xi}^\perp}$ indicates a Laplacian at the subspace orthogonal to $\boldsymbol{\xi}$. While F-principle always holds due to the power-law decay, there are mild extra effective results from this term in practice. Details about such an effect remains a problem for future study. This suggests a wider class of *generalized LFP models*, in which $\gamma(\boldsymbol{\xi})$ can be a general linear operator in frequency domain. Detailed properties about the generalized model remains a problem for future study.

To analyze model (8), we resort to its equivalent optimization problem as discussed before. Based on the equivalent optimization problem in Eq. (3) and the error estimate in Eq. (4), the non-overfitting puzzle for two-layer wide ReLU NNs can be explained. Next, we analyze each decaying term for 1-d problems ($d = 1$). When $1/\xi^2$ term dominates, the corresponding minimization problem Eq. (3) rewritten in spatial domain yields

$$\begin{aligned}&\min_h \int |h'(x) - h'_{\mathrm{ini}}(x)|^2 dx, \\ &\text{s.t.} h(\boldsymbol{x}_i) = f^*(\boldsymbol{x}_i), \quad i = 1, \ldots, n,\end{aligned} \tag{9}$$

where primes represent differentiations. For $h_{\mathrm{ini}}(x) = 0$, Eq. (9) indicates a linear spline interpolation. Similarly, when $1/\xi^4$ dominates, $\int |h''(x) - h''_{\mathrm{ini}}(x)|^2 dx$ is minimized, indicating a cubic spline. In general, above

two power law decays coexist, giving rise to a specific mixture of linear and cubic splines. For high dimensional problems, the model prediction is difficult to interpret because the order of differentiation depends on d and can be fractal.

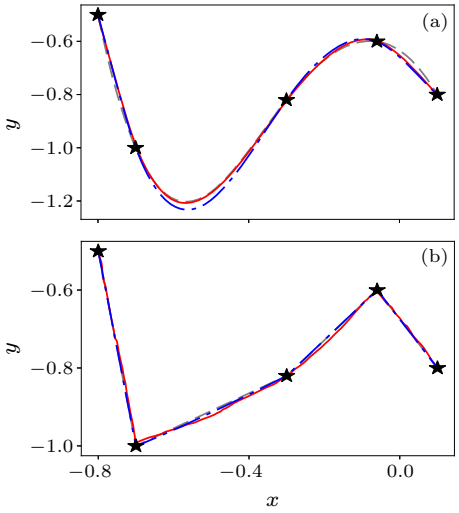

Fig. 3. Characteristics of $f_{\rm NN}$ (red solid) vs $f_{\rm LFP}$ (blue dashed dot) and splines [grey dashed, cubic spline in (a) and linear spline in (b)] for a 1-d problem. All curves nearly overlap with one other. Two-layer NN Eq. (5) of 40000 hidden neurons is initialized with (a) $\langle a^2+r^2\rangle_{a,r} \gg \langle a^2r^2\rangle_{a,r}$ and (b) $\langle a^2+r^2\rangle_{a,r} \ll \langle a^2r^2\rangle_{a,r}$. Black stars indicates training data.

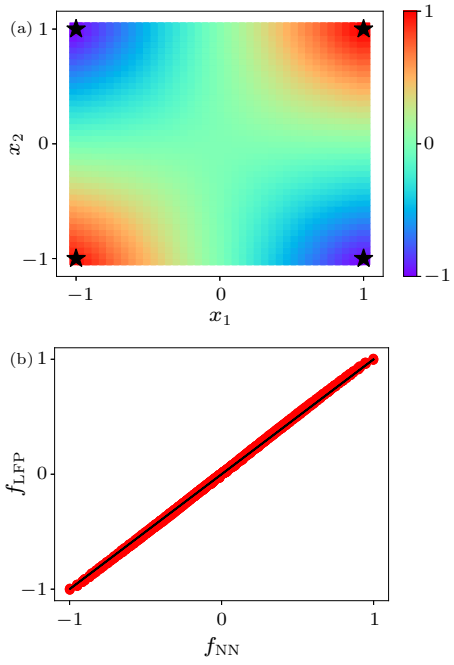

Fig. 4. The 2-d XOR problem with four training data indicated by black stars learned by a two-layer NN Eq. (5) of 160000 hidden neurons: (a) $f_{\rm NN}$ illustrated in color scale, (b) $f_{\rm LFP}$ (ordinate) vs $f_{\rm NN}$ (abscissa) represented by red dots evaluated over whole input domain $[-1,1]^2$. The black line indicates the identity function.

In the following, we examine experimentally the quantitative prediction power of the LFP model Eq. (8). For convenience of notation, solution predicted by the LFP model (8) is denoted as $f_{\rm LFP}(\boldsymbol{x}) = f(\boldsymbol{x},\infty)$. The function learned by NN is denoted by $f_{\rm NN}(\boldsymbol{x})$. As shown in Fig. 3, for a 1-d problem, $f_{\rm LFP}(\boldsymbol{x})$ accurately predicts $f_{\rm NN}(\boldsymbol{x})$ over two different initializations. As predicted by above analysis, a wide NN initialized with $\langle a^2+r^2\rangle_{a,r} \gg \langle a^2r^2\rangle_{a,r}$ learns approximately a cubic spline, whereas $\langle a^2+r^2\rangle_{a,r} \ll \langle a^2r^2\rangle_{a,r}$ a linear spline. For $d=2$, we consider the famous XOR problem, which cannot be solved by one-layer neural networks.[45] The training samples consist of four points represented by black stars in Fig. 4(a). As shown in Fig. 4(b), our LFP model predicts accurately outputs of the well-trained NN over the input domain $[-1,1]^2$.

Discussion. In this study, we propose the phenomenological LFP model that explains the absence of overfitting in NNs by its low frequency preference. Our theory informs that NNs are no panacea to all difficult problems and are bad in general for fitting high frequency target functions. As an example, it has been demonstrated that a standard DNN fails drastically for ground state fitting of a frustrated quantum magnet with a rapidly oscillating ground-state characteristic function.[49] Therefore, to solve a broad spectrum of problems with practical success, it is important to take into account the low frequency preference of DNNs in the algorithm design. Our work on F-principle is only a starting point to a more comprehensive understanding of NNs. In the future, the role of depth, width, optimization methods and other hyperparameters in fine tuning the F-principle dynamics will be studied in detail. Importantly, more preferences (inductive biases) of NNs, which are keys to open the "black box", need to be unraveled. Specifically, the physics approach from phenomenological study based on carefully designed experiments to theoretical study based on effective models can play an important role as demonstrated by the series of works on F-principle.

We thank Hugues Chate for critical reading and suggestions on the manuscript. We also thank David W. MacLaughlin, Haijun Zhou, Leihan Tang, Hepeng Zhang, and Yongfeng Zhao for helpful comments on the manuscript.

References

[1] Aurisano A, Radovic A, Rocco D, Himmel A, Messier M D, Niner E, Pawloski G, Psihas F, Sousa A and Vahle P 2016 *J. Instrum.* **11** P09001
[2] Zhang L, Han J, Wang H, Car R and E W 2018 *Phys. Rev. Lett.* **120** 143001
[3] Guest D, Cranmer K and Whiteson D 2018 *Annu. Rev. Nucl. Part. Sci.* **68** 161
[4] Radovic A, Williams M, Rousseau D, Kagan M, Bonacorsi D, Himmel A, Aurisano A, Terao K and Wongjirad T 2018 *Nature* **560** 41
[5] Levine Y, Sharir O, Cohen N and Shashua A 2019 *Phys. Rev. Lett.* **122** 065301
[6] Carleo G, Cirac I, Cranmer K, Daudet L, Schuld M, Tishby N, Vogt-Maranto L and Zdeborová L 2019 *Rev. Mod. Phys.* **91** 045002

[7] Mehta P, Bukov M, Wang C H, Day A G R, Richardson C, Fisher C K and Schwab D J 2019 *Phys. Rep.* **810** 1
[8] Breiman L 1995 *The Mathematics of Generalization* (Addison Wesley Reading, MA) XX 11
[9] Zdeborová L 2020 *Nat. Phys.* **16** 1
[10] Zhang C, Bengio S, Hardt M, Recht B and Vinyals O 2017 *The International Conference on Learning Representations* (Toulon, France 24–26 April 2017)
[11] Simonyan K and Zisserman A 2014 arXiv:1409.1556 [cs.CV]
[12] Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A *et al.* 2020 arXiv:2005.14165 [cs.CL]
[13] Dyson F 2004 *Nature* **427** 297
[14] Saxe A M, McClelland J L and Ganguli S 2014 *The International Conference on Learning Representations* (Banff, Canada 14–16 April 2014)
[15] Saxe A M, Bansal Y, Dapello J, Advani M, Kolchinsky A, Tracey B D and Cox D D 2019 *J. Stat. Mech.: Theory Exp.* **2019** 124020
[16] Lampinen A K and Ganguli S 2019 *The International Conference on Learning Representations* (New Orleans, United States 6–9 May 2019)
[17] Engel A and Broeck C V D 2001 *Statistical Mechanics ofLearning* (Cambridge: Cambridge University Press)
[18] Aubin B, Maillard A, barbier J, Krzakala F, Macris N and Zdeborová L 2018 *Advances in Neural Information Processing Systems* (NeurIPS 2018) (Publisher: Curran Associates, Inc.) vol 31 p 3223
[19] Choromanska A, Henaff M, Mathieu M, Arous G B and LeCun Y 2015 *Artificial Intelligence and Statistics* (Publisher: Curran Associates, Inc.) p 192
[20] Mei S, Montanari A and Nguyen P M 2018 *Proc. Natl. Acad. Sci. USA* **115** E7665–E7671
[21] Rotskoff G and Vanden-Eijnden E 2018 *Advances in Neural Information Processing Systems* (NeurIPS 2018) (Publisher: Curran Associates, Inc.) vol 31 p 7146
[22] Chizat L and Bach F 2018*Advances in Neural Information Processing Systems* (NeurIPS 2018) (Publisher: Curran Associates, Inc.) vol 31 p 3036
[23] Sirignano J and Spiliopoulos K 2020 *Stochastic Processes and Their Applications* **130** 1820
[24] Jacot A, Gabriel F and Hongler C 2018 *Advances in Neural Information Processing Systems* (NeurIPS 2018) (Publisher: Curran Associates, Inc.) vol 31 p 8571
[25] Lee J, Xiao L, Schoenholz S, Bahri Y, Novak R, Sohl-Dickstein J and Pennington J 2019 *Advances in Neural Information Processing Systems* (NIPS 2019) (Publisher: Curran Associates, Inc.) vol 32 p 8572
[26] Arpit D, Jastrzbski S, Ballas N, Krueger D, Bengio E, Kanwal M S, Maharaj T, Fischer A, Courville A, Bengio Y *et al.* 2017 *Proceedings of the 34th International Conference on Machine Learning* PMLR 70 p 233
[27] Kalimeris D, Kaplun G, Nakkiran P, Edelman B, Yang T, Barak B and Zhang H 2019 *Advances in Neural Information Processing Systems* (NIPS 2019) (Publisher: Curran Associates, Inc.) vol 32 p 3496
[28] Valle-Perez G, Camargo C Q and Louis A A 2019 *The International Conference on Learning Representations* (New Orleans, United States 6–9 May 2019)
[29] Xu Z Q J, Zhang Y and Xiao Y 2019 *Neural Information Processing* in *Lecture Notes in Computer Science* p 264
[30] Xu Z Q J, Zhang Y, Luo T, Xiao Y and Ma Z 2020 *Commun. Comput. Phys.* **28** 1746
[31] Rahaman N, Baratin A, Arpit D, Draxler F, Lin M, Hamprecht F, Bengio Y and Courville A 2019 *International Conference on Machine Learning* PMLR 97 p 5301
[32] Ronen B, Jacobs D, Kasten Y and Kritchman S 2019 *Advances in Neural Information Processing Systems* (NIPS 2019) (Publisher: Curran Associates, Inc.) vol 32 p 4763
[33] Rabinowitz N C 2019 arXiv:1905.01320[cs.LG]
[34] Jagtap A D, Kawaguchi K and Karniadakis G E 2020 *J. Comput. Phys.* **404** 109136
[35] Yang G and Salman H 2019 arXiv:1907.10599 [cs.LG]
[36] Cao Y, Fang Z, Wu Y, Zhou D X and Gu Q 2019 arXiv:1912.01198 [cs.LG]
[37] Cai W, Li X and Liu L 2019 arXiv:1909.11759 [cs.LG]
[38] Biland S, Azevedo V C, Kim B and Solenthaler B 2019 arXiv:1912.08776 [cs.LG]
[39] Biland S, Azevedo V C, Kim B and Solenthaler B 2020 *Eurographics Conferences* (Publisher: The Eurographics Association)
[40] Liu Z, Cai W and Xu Z Q J 2020 *Commun. Comput. Phys.* **28** 1970
[41] Li X A, Xu Z Q J and Zhang L 2020 *Commun. Comput. Phys.* **28** 1886
[42] Wang B, Zhang W and Cai W 2020 *Commun. Comput. Phys.* **28** 2139
[43] Zhang Y, Xu Z Q J, Luo T and Ma Z 2019 arXiv:1905.07777 [cs.LG]
[44] Weinan E, Ma C and Wu L 2019 *Commun. Math. Sci.* **17** 1407
[45] Minsky M and Papert S A 2017 *Perceptrons: An introduction to Computational Geometry* (Massachusetts: MIT Press)
[46] Allender E 1996 *International Conference on Foundations of Software Technology and Theoretical Computer Science* (Berlin: Springer) p 1
[47] Arora S, Du S, Hu W, Li Z and Wang R 2019 *International Conference on Machine Learning* PMLR 97 p 322
[48] Weinan E, Ma C and Wu L 2020 *Sci. Chin. Math.* **63** 1235
[49] Cai Z and Liu J 2018 *Phys. Rev. B* **97** 035116