CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2020-0005

Vol. **2**, No. 3, pp. 484-507
September 2021

Theory of the Frequency Principle for General Deep Neural Networks

Tao Luo[1,*], Zheng Ma[1], Zhi-Qin John Xu[1] and Yaoyu Zhang[1,2]

[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 17 October 2020; Accepted 23 February 2021

Abstract. Along with fruitful applications of Deep Neural Networks (DNNs) to realistic problems, recently, empirical studies reported a universal phenomenon of Frequency Principle (F-Principle), that is, a DNN tends to learn a target function from low to high frequencies during the training. The F-Principle has been very useful in providing both qualitative and quantitative understandings of DNNs. In this paper, we rigorously investigate the F-Principle for the training dynamics of a general DNN at three stages: initial stage, intermediate stage, and final stage. For each stage, a theorem is provided in terms of proper quantities characterizing the F-Principle. Our results are general in the sense that they work for multilayer networks with general activation functions, population densities of data, and a large class of loss functions. Our work lays a theoretical foundation of the F-Principle for a better understanding of the training process of DNNs.

AMS subject classifications: 68Q32, 68T07, 37N40

Key words: Frequency principle, Deep Neural Networks, dynamical system, training process.

1 Introduction

Deep learning has achieved great success as in many fields [15], e.g., speech recognition [1], object recognition [10], natural language processing [35] and computer game control [21]. It has also been adopted into algorithms to solve scientific computing problems [8, 11,12,14]. In principle, the universal approximation theorem states that a commonly-used Deep Neural Network (DNN) of sufficiently large width can approximate any function to

*Corresponding author. *Email addresses:* luotao41@sjtu.edu.cn (T. Luo), zhengma@sjtu.edu.cn (Z. Ma), xuzhiqin@sjtu.edu.cn (Z.-Q. J. Xu), zhyy.sjtu@sjtu.edu.cn (Y. Zhang)

a desired precision [7]. However, it remains a mystery that how a DNN finds a minimum corresponding to such an approximation through the gradient-based training process. To understand the learning behavior of DNNs for the approximation problem, recent works model the gradient flow of parameters in a two-layer ReLU neural networks by a partial differential equation (PDE) in the mean-field limit [19, 25, 26]. However, it is not clear whether this PDE approach, which describes a neural network of one hidden layer of infinite width, can be extended to general DNNs of multiple hidden layers and limited neuron number. For further discussion on the mathematical understanding of DNNs, we refer the readers to a review article [9].

In this work, we take another approach that uses Fourier analysis to study the learning behavior of DNNs based on the phenomenon of *Frequency Principle (F-Principle), i.e., a DNN tends to learn a target function from low to high frequencies during the training* [23, 31, 32, 36]. Empirically, the F-Principle can be widely observed in general DNNs for both benchmark and synthetic data [31, 32]. Conceptually, it provides a qualitative explanation of the success and failure of DNNs [32]. E et al., (2019) [30] propose a continuous viewpoint for studying machine learning and suggest that the F-Principle underlying the gradient flows may be a main reason behind the success of modern machine learning. Empirically, the F-Principle provides us a perspective for quantifying the training process via the convergence of each frequency component [13, 22, 29, 33]. For example, it is used as an important phenomenon to pursue fundamentally different learning trajectories of meta-learning [22] and provides an understanding of why increasing the depth of a neural network may accelerate the training [33]. The F-Principle also provides important theoretical insights to design DNN-based algorithms [2, 3, 5, 16, 17, 20, 27, 28]. For example, Blind et al. [3] designs a loss function with explicit higher priority for high frequencies to significantly accelerate the simulation of fluid dynamics through DNN approach; MscaleDNN [16, 17, 28] is developed to accelerate the fitting of high frequency functions by shifting or rescaling high frequencies to lower ones. These works have signified the importance of the F-Principle. Theoretically, Xu et al. [32] propose a theorem for the characterization of the initial training stage of a two-layer tanh network, which is also adopted in the analysis of DNNs with ReLU activation function [23]. Another series of works [4, 6, 24, 34, 36] attempt to understand the F-Principle in very wide neural networks, which can be well approximated by the first-order expansion with respect to the network parameters (the linear neural tangent kernel (NTK) regime). The studies [6, 24, 34] from the perspective of eigen-decomposition of DNN dynamics in spatial domain require assumptions of very large network width and infinite samples. To study the F-Principle with finite samples, Zhang et al. [36] and Luo et al. [18] study the dynamics in the frequency domain and further obtain an effective model of linear F-Principle dynamics, which accurately predicts the learning results of two-layer ReLU neural networks of large widths, leads to an apriori estimate of the generalization error bound. However, the explanation of DNN's F-Principle beyond the NTK regime (non-linear regime) is still missing.

Following the same direction as in [32], in this work, we propose a theoretical frame-

work of Fourier analysis for the study of the training behavior of *general* DNNs in the following three stages: the initial stage, the intermediate stage, and the final stage. At all stages, we rigorously characterize the F-Principle by estimating some proper quantities. At the initial and final stages with the MSE loss (mean-squared error, also known as L^2 loss), we show that the change of MSE is dominated by low frequencies. Furthermore, in these two stages with general L^p $(2\leq p<\infty)$ loss, we show that the change of the DNN output is dominated by the low-frequency part. A key contribution of this work is on the intermediate stage — with L^p loss, the difference of the MSE over a certain period, in which the MSE is reduced by half, is dominated by the low frequencies. In summary, we verify that the F-Principle is universal in the sense that our results not only work for DNNs of multiple layers with any commonly-used activation function, e.g., ReLU, sigmoid, and tanh, but also work for a general population density of data and for a general class of loss functions. The key insight unraveled by our analysis is that the regularity of DNN converts into the decay rate of a loss function in the frequency domain.

2 Preliminaries

We start with a brief introduction to DNNs and its training dynamics. Under very mild assumptions, we provide some regularity results which are crucial to the proof of the main theorems summarized in the next section.

2.1 Deep Neural Networks

Consider a DNN with $(L-1)$-hidden layers and general activation functions. We regard the d-dimensional input as the 0-th layer and the one-dimensional output as the L-th layer. Let m_l be the number of neurons in the l-th layer. In particular, $m_0=d$ and $m_L=1$. For any $k\in\mathbb{N}$, we denote $[k]:=\{1,2,\cdots,k\}$.

The *hypothesis space* $\mathcal{H}$ is a family of *hypothesis functions* parametrized by the *parameter vector* $\boldsymbol{\theta}\in\mathbb{R}^M$ whose entries are called *parameters* $W_{ij}^{[l]}$'s (also known as *weight*) and $b_i^{[l]}$'s (also known as *bias*). More precisely, we set

$$\boldsymbol{\theta}=\Big(\mathrm{vec}(\boldsymbol{W}^{[1]}),\mathrm{vec}(\boldsymbol{b}^{[1]}),\cdots,\mathrm{vec}(\boldsymbol{W}^{[L]}),\mathrm{vec}(\boldsymbol{b}^{[L]})\Big), \tag{2.1}$$

where for $\boldsymbol{W}^{[l]}\in\mathbb{R}^{m_l\times m_{l-1}}$ and $\boldsymbol{b}^{[l]}\in\mathbb{R}^{m_l}$ for $l\in[L]$. The *size* M of the network is the number of the parameters, i.e.,

$$M=\sum_{l=0}^{L-1}(m_l+1)m_{l+1}. \tag{2.2}$$

To define the hypothesis functions in $\mathcal{H}$, we need some nonlinear functions which are known as *activation functions*:

$$\sigma_i^{[l]}:\mathbb{R}\to\mathbb{R}\quad\text{for}\quad l\in[L-1],\ i\in[m_l]. \tag{2.3}$$

Given $\boldsymbol{\theta}\in\mathbb{R}^M$, the corresponding function $f_{\boldsymbol{\theta}}(\cdot)$ (also denoted by $f(\cdot,\boldsymbol{\theta})$ in this paper) in $\mathcal{H}$ is defined by a series of function compositions. First, we set $\boldsymbol{f}^{[0]}=\mathrm{id}:\mathbb{R}^d\to\mathbb{R}^d$, i.e., $\boldsymbol{f}^{[0]}(\boldsymbol{x})=\boldsymbol{x}$ for all $\boldsymbol{x}\in\mathbb{R}^d$. Then for $l\in[L-1]$, $\boldsymbol{f}^{[l]}$ is defined recursively as

$$\boldsymbol{f}^{[l]}:\mathbb{R}^d\to\mathbb{R}^{m_l}, \tag{2.4}$$

$$(\boldsymbol{f}^{[l]})_i=\sigma_i^{[l]}\left\{(\boldsymbol{W}^{[l]}\boldsymbol{f}^{[l-1]}+\boldsymbol{b}^{[l]})_i\right\},\quad i\in[m_l]. \tag{2.5}$$

Finally, we denote

$$f^{[L]}=\boldsymbol{W}^{[L]}\cdot\boldsymbol{f}^{[L-1]}+b^{[L]}. \tag{2.6}$$

We remark that for the most applications, the activation functions $\sigma_i^{[l]}$ are chosen to be the same, i.e., $\sigma_i^{[l]}=\sigma$, $l\in[L-1]$, $i\in[m_l]$.

Example 2.1. For instance, if a one-hidden layer neural network is used, then $L=2$ and the hypothesis function can be written into the following form:

$$f^{[2]}(\boldsymbol{x},\boldsymbol{\theta})=\sum_{i=1}^{m}w_i^{[2]}\sigma(\boldsymbol{w}_i^{[1]}\cdot\boldsymbol{x}+b_i^{[1]}),\quad w_i^{[2]},b_i^{[1]}\in\mathbb{R},\ \boldsymbol{w}_i^{[1]}\in\mathbb{R}^d. \tag{2.7}$$

Thus the size of the network $M=(d+2)m$ which is consistent with (2.2).

We are only interested in the target function f_{target} in a compact domain Ω, i.e., $\Omega\subset\subset\mathbb{R}^d$. A bump function χ is used to truncate both hypothesis and target functions:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x})=f(\boldsymbol{x},\boldsymbol{\theta})=f^{[L]}(\boldsymbol{x},\boldsymbol{\theta})\chi(\boldsymbol{x}), \tag{2.8}$$

$$f(\boldsymbol{x})=f_{\text{target}}(\boldsymbol{x})\chi(\boldsymbol{x}). \tag{2.9}$$

In the sequel, we will also refer to $f_{\boldsymbol{\theta}}$ and f as the hypothesis and target functions, respectively.

2.2 Loss function and training dynamics

In this work, we investigate the training dynamics of parameters in DNNs with two cases of loss functions:

(i) The MSE loss function with population measure $\mathcal{D}$, i.e.,

$$R_{\mathcal{D}}(\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}(f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x}))^2. \tag{2.10}$$

In this case, the training dynamics of $\boldsymbol{\theta}$ follows the gradient flow:

$$\begin{cases}\dfrac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}=-\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}),\\ \boldsymbol{\theta}(0)=\boldsymbol{\theta}_0.\end{cases} \tag{2.11}$$

(ii) A general loss function with population measure $\mathcal{D}$, i.e.,

$$\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}\ell(f(\boldsymbol{x},\boldsymbol{\theta}) - f(\boldsymbol{x})), \tag{2.12}$$

where the function ℓ satisfies some mild assumptions to be explained later. In this case, the training dynamics of $\boldsymbol{\theta}$ becomes:

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}), \\ \boldsymbol{\theta}(0) = \boldsymbol{\theta}_0. \end{cases} \tag{2.13}$$

In the case of MSE loss function, we have

$$\begin{aligned} R_{\mathcal{D}}(\boldsymbol{\theta}) &= \int_{\mathbb{R}^d} |f_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{\theta}) - f_{\mathcal{D}}(\boldsymbol{x})|^2 \mathrm{d}\boldsymbol{x} \\ &= \int_{\mathbb{R}^d} |\hat{f}_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta}) - \hat{f}_{\mathcal{D}}(\boldsymbol{\xi})|^2 \mathrm{d}\boldsymbol{\xi}, \end{aligned} \tag{2.14}$$

where ρ, satisfying $\mathcal{D}(\mathrm{d}\boldsymbol{x}) = \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$, is called the *population density* and

$$f_{\mathcal{D}}(\cdot,\boldsymbol{\theta}) = f(\cdot,\boldsymbol{\theta})\sqrt{\rho(\cdot)}, \quad f_{\mathcal{D}}(\cdot) = f(\cdot)\sqrt{\rho(\cdot)}. \tag{2.15}$$

The second equality is due to the Plancherel theorem. Here and in the sequel, we use the following conventions for the Fourier transform and its inverse transform on $\mathbb{R}^d$:

$$\mathcal{F}[g](\boldsymbol{\xi}) = \hat{g}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} g(\boldsymbol{x})\mathrm{e}^{-2\pi\mathrm{i}\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x}, \quad g(\boldsymbol{x}) = \int_{\mathbb{R}^d} \hat{g}(\boldsymbol{\xi})\mathrm{e}^{2\pi\mathrm{i}\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi}.$$

For the convenience of proofs, we denote

$$R_{\mathcal{D}}(\boldsymbol{\theta}) = \int_{\mathbb{R}^d} q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi}, \tag{2.16}$$

$$q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta}) = |\hat{f}_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta}) - \hat{f}_{\mathcal{D}}(\boldsymbol{\xi})|^2. \tag{2.17}$$

2.3 Assumptions

The requirements on χ, f, σ, and $\mathcal{D}$ are summarized here.

Assumption 2.1 (regularity)**.** The bump function χ satisfies $\chi(\boldsymbol{x}) = 1$, $\boldsymbol{x}\in\Omega$ and $\chi(\boldsymbol{x}) = 0$, $\boldsymbol{x}\in\mathbb{R}^d\backslash\Omega'$ for domains Ω and Ω' with $\Omega\subset\subset\Omega'\subset\subset\mathbb{R}^d$. There is a positive integer k (can be ∞) such that $f_{\text{target}}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$, $\chi\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;[0,+\infty))$, and $\sigma_i^{[l]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R};\mathbb{R})$ for $l\in[L-1]$, $i\in[m_l]$.

Assumption 2.2 (bounded population density)**.** There exists a function $\rho\in L^\infty\left(\mathbb{R}^d;[0,+\infty)\right)$ satisfying $\mathcal{D}(\mathrm{d}\boldsymbol{x}) = \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$.

Example 2.2. Here we list some commonly-used activation functions:

1. ReLU (Rectified Linear Unit): $\mathrm{ReLU}(x)=\max(0,x)$, $x\in\mathbb{R}$;
2. tanh (hyperbolic tangent): $\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$, $x\in\mathbb{R}$;
3. sigmoid function (also known as logistic function): $S(x)=\frac{1}{1+e^{-x}}$, $x\in\mathbb{R}$.

Remark 2.1. It is also allowed that $k=\infty$ where the functions f and $\sigma_i^{[l]}$ are all C^∞ by Sobolev embedding inequalities. This case includes tanh and sigmoid activation functions.

Remark 2.2. If an activation function is ReLU, then $k=1$.

Remark 2.3. For $\boldsymbol{x}\in\Omega$, we have $f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x})=f^{[L]}(\boldsymbol{x},\boldsymbol{\theta})-f_{\text{target}}(\boldsymbol{x})$.

For the training dynamics (2.11) or (2.13), we suppose the parameters are bounded.

Assumption 2.3 (bounded trajectory)**.** The training dynamics is nontrivial, i.e., $\boldsymbol{\theta}(t)\not\equiv$ const. There exists a constant $C_0>0$ such that $\sup_{t\geq 0}|\boldsymbol{\theta}(t)|\leq C_0$ where the parameter vector $\boldsymbol{\theta}(t)$ is the solution to (2.11) or (2.13).

Remark 2.4. The bound C_0 depends on initial parameter $\boldsymbol{\theta}_0$.

In the case of MSE loss function, we will further take the following assumption.

Assumption 2.4. The density ρ satisfies $\sqrt{\rho}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;[0,+\infty))$.

The general loss function considered in this work satisfies the following assumption.

Assumption 2.5 (general loss function)**.** The function ℓ in the general loss function $\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})$ satisfies $\ell\in C^2(\mathbb{R};[0,+\infty))$ and there exist positive constants C and r_0 such that $C^{-1}[\ell'(z)]^2\leq\ell(z)\leq C|z|^2$ for $|z|\leq r_0$.

Example 2.3. The L^p $(2\leq p<\infty)$ loss function satisfies Assumption 2.5. Here the L^p $(1\leq p<\infty)$ loss functions used in machine learning are defined as $\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}|f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x})|^p$ which is a little bit different from the L^p norm used in mathematics.

2.4 Regularity

Based on the above assumptions, some regularity results can be obtained in terms of the "Japanese bracket" of $\boldsymbol{\xi}$:

$$\langle\boldsymbol{\xi}\rangle:=(1+|\boldsymbol{\xi}|^2)^{1/2}. \tag{2.18}$$

Lemma 2.1. *Suppose that the Assumption* 2.1 *holds. Given any $\boldsymbol{\theta}\in\mathbb{R}^M$, the hypothesis function $f_{\boldsymbol{\theta}}\in W^{k,2}(\mathbb{R}^d;\mathbb{R})$ and its gradient with respect to the parameters $\nabla_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}\in W^{k-1,2}(\mathbb{R}^d;\mathbb{R}^M)$. Also, we have the target function $f\in W^{k,2}(\mathbb{R}^d;\mathbb{R})$.*

Proof. Recall that $f(\boldsymbol{x})=f_{\text{target}}(\boldsymbol{x})\chi(\boldsymbol{x})$ and $f_{\boldsymbol{\theta}}(\boldsymbol{x})=f^{[L]}(\boldsymbol{x},\boldsymbol{\theta})\chi(\boldsymbol{x})$ given $\boldsymbol{\theta}\in\mathbb{R}^M$. By Assumption 2.1, $f_{\text{target}}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$ and $\chi\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;[0,+\infty))$ with a compact support. Thus $f\in W^{k,2}(\mathbb{R}^d;\mathbb{R})$. In order to show $f_{\boldsymbol{\theta}}\in W^{k,2}(\mathbb{R}^d;\mathbb{R})$, it is sufficient to prove that $f^{[L]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$. Indeed, we prove $\boldsymbol{f}^{[l]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^{m_l})$ for $l\in[L]$ by induction. For $l=0$, $\boldsymbol{f}^{[0]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^{m_0})$ because $\boldsymbol{f}^{[0]}(\boldsymbol{x})=\boldsymbol{x}$ and $m_0=d$. Suppose that for l $(0\le l\le H-2)$ we have $\boldsymbol{f}^{[l]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R}^{m_l})$. Now let us consider $\boldsymbol{f}^{[l+1]}$ with $(\boldsymbol{f}^{[l+1]})_i=\sigma_i^{[l+1]}\big\{(\boldsymbol{W}^{[l+1]}\cdot\boldsymbol{f}^{[l]}+\boldsymbol{b}^{[l+1]})_i\big\}$, $i\in[m_{l+1}]$. By the induction assumption, we have $(\boldsymbol{W}^{(l+1)}\cdot\boldsymbol{f}^{[l]}+\boldsymbol{b}^{[l+1]})_i\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$. By Assumption 2.1, $\sigma_i^{[l]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R};\mathbb{R})$. Note that $\sigma_i^{[l]}\in C^{k-1}(\mathbb{R};\mathbb{R})$ by Sobolev embedding. Then $(\boldsymbol{f}^{[l+1]})_i\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$ because of the chain rule and the fact that the composition of continuous functions is still continuous. Finally, for $l=L-1$, we have $f^{[L]}=\boldsymbol{W}^{[L]}\cdot\boldsymbol{f}^{[L-1]}+b^{[L]}\in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$.

The proof for $\nabla_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}$ is similar if we note that $(\sigma_i^{[l]})'\in W^{k-1,\infty}_{\text{loc}}(\mathbb{R};\mathbb{R})$. □

Remark 2.5. The continuity of $\sigma_i^{[l]}$ is necessary because the composition of two Lebesgue measurable functions need not be Lebesgue measurable.

Lemma 2.2. *Suppose that the Assumptions* 2.1 *and* 2.2 *hold. Then*

(a). For any $0\le r\le k$, we have

$$\langle\cdot\rangle^r|\hat{f}(\cdot,\boldsymbol{\theta})|\in L^2(\mathbb{R}^d;\mathbb{R}),\tag{2.19}$$

$$\langle\cdot\rangle^r|\hat{f}(\cdot)|\in L^2(\mathbb{R}^d;\mathbb{R}).\tag{2.20}$$

(b). For any $0\le r\le k-1$, we have

$$\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\in L^2(\mathbb{R}^d;\mathbb{R}).\tag{2.21}$$

(c). For any $0\le r\le 2k-1$, we have

$$\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}q(\cdot,\boldsymbol{\theta})|\in L^1(\mathbb{R}^d;\mathbb{R}).\tag{2.22}$$

Proof. (a). Let $0\le r\le k$. Given $\boldsymbol{\theta}\in\mathbb{R}^M$, we have $f,f_{\boldsymbol{\theta}}\in W^{k,2}(\mathbb{R}^d;\mathbb{R})$ by Lemma 2.1. It is well known that for any function $g\in W^{k,2}(\mathbb{R}^d)$, for $0\le r\le k$,

$$C\|g\|_{W^{r,2}(\mathbb{R}^d)}\le\|\langle\cdot\rangle^r|\hat{g}|\|_{L^2(\mathbb{R}^d)}\le\tilde{C}\|g\|_{W^{r,2}(\mathbb{R}^d)},\tag{2.23}$$

where the positive constants C and $\tilde{C}$ only depend on d and r. The statements (2.19) and (2.20) follow this.

(b). Let $0\le r\le k-1$. Given $\boldsymbol{\theta}\in\mathbb{R}^M$, we have $\nabla_{\boldsymbol{\theta}}f_{\boldsymbol{\theta}}\in W^{k-1,2}(\mathbb{R}^d;\mathbb{R}^M)$ by Lemma 2.1. Similar to part (a), this leads to (2.21).

(c). Let $r_1 = r - r_2$ and $r_2 = \min\{r,k\}$. Then $0 \le r_1 \le k-1$ and $0 \le r_2 \le k$. Combining the inequalities in parts (a) and (b), we have

$$\begin{aligned}\||\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}} q(\cdot,\boldsymbol{\theta})|\|_{L^1(\mathbb{R}^d)} &= \left\|\left|\langle\cdot\rangle^r\left|\left(\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\right)\overline{\hat{f}(\cdot,\boldsymbol{\theta})-\hat{f}(\cdot)}+\text{c.c.}\right|\right|\right\|_{L^1(\mathbb{R}^d)} \\ &\le 2\left\|\langle\cdot\rangle^{r_1}|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\right\|_{L^2(\mathbb{R}^d)}\left\|\langle\cdot\rangle^{r_2}|\hat{f}(\cdot,\boldsymbol{\theta})-\hat{f}(\cdot)|\right\|_{L^2(\mathbb{R}^d)} < \infty. \quad (2.24)\end{aligned}$$

This completes the proof. □

Lemma 2.3. *Suppose that the Assumptions* 2.1, 2.2, *and* 2.4 *hold. Then*
(a). For any $0 \le r \le k$, *we have*

$$\langle\cdot\rangle^r|\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})| \in L^2(\mathbb{R}^d;\mathbb{R}), \quad (2.25)$$

$$\langle\cdot\rangle^r|\hat{f}_{\mathcal{D}}(\cdot)| \in L^2(\mathbb{R}^d;\mathbb{R}). \quad (2.26)$$

(b). For any $0 \le r \le k-1$, *we have*

$$\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})| \in L^2(\mathbb{R}^d;\mathbb{R}). \quad (2.27)$$

(c). For any $0 \le r \le 2k-1$, *we have*

$$\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}} q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})| \in L^1(\mathbb{R}^d;\mathbb{R}). \quad (2.28)$$

Proof. The proof is similar to the one of Lemma 2.2. The only new ingredient is assumption that $\sqrt{\rho} \in W^{k,\infty}_{\text{loc}}(\mathbb{R}^d;\mathbb{R})$. □

Lemma 2.4. *Suppose that the Assumptions* 2.1 *and* 2.3 *hold. Then*

$$\sup_{|\boldsymbol{\theta}|\le C_0} \||\langle\cdot\rangle^{k-1}|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)} < +\infty. \quad (2.29)$$

If we further suppose that the Assumptions 2.2 and 2.4 hold, then we have

$$\sup_{|\boldsymbol{\theta}|\le C_0} \||\langle\cdot\rangle^{k-1}|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)} < +\infty, \quad (2.30)$$

$$\sup_{|\boldsymbol{\theta}|\le C_0} \||\langle\cdot\rangle^{2k-1}|\nabla_{\boldsymbol{\theta}} q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})|\|_{L^1(\mathbb{R}^d)} < +\infty. \quad (2.31)$$

Proof. We only prove (2.29). The proofs of (2.30) and (2.31) are similar. Here we regard f as a function of both $\boldsymbol{x}$ and $\boldsymbol{\theta}$, i.e., $f: \mathbb{R}^d \times \mathbb{R}^M \to \mathbb{R}$. Then $\nabla_{\boldsymbol{\theta}} f: \mathbb{R}^d \times \mathbb{R}^M \to \mathbb{R}^M$. Note that for any $\boldsymbol{\theta}$ satisfying $|\boldsymbol{\theta}| \le C_0$, we have

$$\||\langle\cdot\rangle^{k-1}|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)} \le C\|\nabla_{\boldsymbol{\theta}} f(\cdot,\boldsymbol{\theta})\|_{W^{k-1,2}(\mathbb{R}^d)}.$$

Let $S(\boldsymbol{\theta}) = \|\nabla_{\boldsymbol{\theta}} f(\cdot,\boldsymbol{\theta})\|_{W^{k-1,2}(\mathbb{R}^d)}$. Similar to the proof of Lemma 2.1, we have $S \in L^\infty_{\text{loc}}(\mathbb{R})$. In other words, $\nabla_{\boldsymbol{\theta}} f \in L^\infty_{\text{loc}}(W^{k-1,2}(\mathbb{R}^d);\mathbb{R}^M)$. Thanks to the compactness of the set $|\boldsymbol{\theta}| \le C_0$, we obtain

$$\sup_{|\boldsymbol{\theta}|\le C_0} \|\nabla_{\boldsymbol{\theta}} f(\cdot,\boldsymbol{\theta})\|_{W^{k-1,2}(\mathbb{R}^d)} \le C.$$

This finishes the proof of (2.29). □

3 Main results

In this section, we first propose several quantitative characterization for the F-Principle. Main results are then summarized with numerical illustrations at the end of this section.

3.1 Characterization of F-Principle

For the MSE loss function, a natural quantity to characterize the F-Principle is the ratio of the loss function decrements caused by low frequencies and the total loss function decrements. To achieve this, we divide the MSE loss function into two parts, contributed by low and high frequencies, respectively, i.e.,

$$R^-_{\mathcal{D},\eta}(\boldsymbol{\theta})=\int_{B_\eta} q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi},\quad R^+_{\mathcal{D},\eta}(\boldsymbol{\theta})=\int_{B^c_\eta} q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi}, \tag{3.1}$$

where B_η and $B^c_\eta=\mathbb{R}^d\backslash B_\eta$ are a ball centered at the origin with radius $\eta>0$ and its complement. Thus $R_{\mathcal{D}}=R^-_{\mathcal{D},\eta}+R^+_{\mathcal{D},\eta}$ for any $\eta>0$. The ratio considered for characterizing the F-Principle is

$$\frac{|\mathrm{d}R^-_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\quad\text{and}\quad\frac{|\mathrm{d}R^+_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}. \tag{3.2}$$

For a general loss function, the training dynamics leads to

$$\frac{\mathrm{d}\tilde{R}_{\mathcal{D}}}{\mathrm{d}t}=-|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}|^2. \tag{3.3}$$

In this case, we study

$$R(\boldsymbol{\theta})=\int_{\mathbb{R}^d}|\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})-\hat{f}(\boldsymbol{\xi})|^2\mathrm{d}\boldsymbol{\xi}. \tag{3.4}$$

We remark that for a given $\boldsymbol{\theta}$, $R(\boldsymbol{\theta})=\int_{\mathbb{R}^d}|f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x})|^2\mathrm{d}\boldsymbol{x}$ has nothing to do with μ. We still take the decomposition $R=R^-_\eta+R^+_\eta$ with

$$R^-_\eta(\boldsymbol{\theta})=\int_{B_\eta} q(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi},\quad R^+_\eta(\boldsymbol{\theta})=\int_{B^c_\eta} q(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi}, \tag{3.5}$$

where

$$q(\boldsymbol{\xi},\boldsymbol{\theta})=|\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})-\hat{f}(\boldsymbol{\xi})|^2. \tag{3.6}$$

One can simply mimic (3.2) and consider

$$\frac{|\mathrm{d}R^-_\eta/\mathrm{d}t|}{|\mathrm{d}R/\mathrm{d}t|}\quad\text{and}\quad\frac{|\mathrm{d}R^+_\eta/\mathrm{d}t|}{|\mathrm{d}R/\mathrm{d}t|}. \tag{3.7}$$

However, there is an issue in this characterization: R may not be monotonically decreasing and the denominator in (3.7) may be zero. To overcome this, a time averaging is

required. Indeed, we investigate the following ratio where integrals are taken for both numerator and denominator in (3.7):

$$\frac{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R_\eta^-}{\mathrm{d}t}\right|\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t} \quad \text{and} \quad \frac{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R_\eta^+}{\mathrm{d}t}\right|\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t}. \tag{3.8}$$

For the general loss function, we also propose another quantity to characterize the F-Principle:

$$\frac{\left\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\right\|_{L^2(B_\eta)}}{\left\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\right\|_{L^2(\mathbb{R}^d)}} \quad \text{and} \quad \frac{\left\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\right\|_{L^2(B_\eta^c)}}{\left\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\right\|_{L^2(\mathbb{R}^d)}}. \tag{3.9}$$

3.2 Main theorems

As we mentioned in the introduction, the training dynamics of a DNN has three stages: initial stage, intermediate stage, and final stage. For each stage, we provide a theorem to characterize the F-Principle.

Initial stage

We start with the F-Principle in the initial stage. Clearly, the constants C in the estimates depend on the initial parameter $\boldsymbol{\theta}_0$ and the time T.

Theorem 3.1 (F-Principle in the initial stage)**.** (L^2 loss function). *Suppose that Assumptions* 2.1, 2.2, 2.3, *and* 2.4 *hold. We consider the training dynamics* (2.11). *Then for any* $1\le r\le 2k-1$ *and any* $T>0$ *satisfying* $|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(T))|>0$ *(if* $k=1$*, we further require that* $\inf_{t\in(0,T]}|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t))|>0$*), there is a constant* $C>0$ *such that*

$$\frac{|\mathrm{d}R_{\mathcal{D},\eta}^+/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\le C\eta^{-r} \quad \text{and} \quad \frac{|\mathrm{d}R_{\mathcal{D},\eta}^-/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\ge 1-C\eta^{-r}, \quad t\in(0,T]. \tag{3.10}$$

(General loss function). *Suppose that Assumptions* 2.1, 2.2, 2.3, *and* 2.5 *hold. We consider the training dynamics* (2.13). *Then for any* $1\le r\le k-1$ *and any* $T>0$ *satisfying* $|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(T))|>0$*, there is a constant* $C>0$ *such that*

$$\frac{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B_\eta^c)}}{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}}\le C\eta^{-r} \quad \text{and} \quad \frac{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B_\eta)}}{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}}\ge 1-C\eta^{-r}, \quad t\in(0,T]. \tag{3.11}$$

Intermediate stage

The theorem of intermediate stage is superior to the other results (initial/final stage) in three aspects. First, for a general loss function considered here, Plancherel theorem

is not helpful. It is even more challenging to show the F-Principle based on the L^2-characterization $R_\eta^-(\boldsymbol{\theta})=\int_{B_\eta}|\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})-\hat{f}(\boldsymbol{\xi})|^2\mathrm{d}\boldsymbol{\xi}$ in the training dynamics which is a gradient flow of a non-L^2 loss function:

$$\frac{\mathrm{d}\tilde{R}_{\mathcal{D}}}{\mathrm{d}t}=-|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}|^2. \tag{3.12}$$

Secondly, although $\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))$ decays as t increases, $R(\boldsymbol{\theta}(t))$ may not be monotonically decreasing. As a result, $\frac{\mathrm{d}R}{\mathrm{d}t}$ might vanish and should not be used in the denominator of the ratio $\frac{\mathrm{d}R_\eta/\mathrm{d}t}{\mathrm{d}R/\mathrm{d}t}$. However, the ratio still makes sense if we replace the infinitesimal change by a finite decrements in both numerator and denominator (see the precise meaning in Eq. (3.13)). The particular choice of a finite decrement is indeed related to the time-scale of the training dynamics. A proper time-scale is the half-life T_2-T_1 satisfying $\frac{1}{2}R(\boldsymbol{\theta}(T_1))=R(\boldsymbol{\theta}(T_2))$. Thirdly, we obtain an upper bound for the dependence of training period T_2-T_1. This bound works for all the situations. If the non-degenerate global minimizer is obtained, the dependence on T_2-T_1 in Eq. (3.13) can also be removed and leads to a consistent result to the results for the final stage.

Theorem 3.2 (F-Principle in the intermediate stage)**.** (General loss function). *Suppose that Assumptions* 2.1, 2.2, 2.3, *and* 2.5 *hold. We consider the training dynamics* (2.13). *Then for any* $1\leq r\leq k-1$, *there is a constant* $C>0$ *such that for any* $0<T_1<T_2$ *satisfying* $\frac{1}{2}R(\boldsymbol{\theta}(T_1))\geq R(\boldsymbol{\theta}(T_2))$, *we have*

$$\frac{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R_\eta^+}{\mathrm{d}t}\right|\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t}\leq C\sqrt{T_2-T_1}\eta^{-r}. \tag{3.13}$$

The gradient in the low frequency part can be very large for each instant but still can oscillate which leads to slower convergence. To avoid this ambiguity and prove the F-Principle in the intermediate stage, we provide the following corollary which works on $|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(T_2))|$ instead of $\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t$.

Corollary 3.1. *Under the same assumptions in Theorem* 3.2, *for any* $1\leq r\leq k-1$, *there is a constant* $C>0$ *such that for any* $0<T_1<T_2$ *satisfying* $\frac{1}{2}R(\boldsymbol{\theta}(T_1))\geq R(\boldsymbol{\theta}(T_2))$ *and* $R(\boldsymbol{\theta}(T_1))\geq R(\boldsymbol{\theta}(t))$ *for all* $t\in[T_1,T_2]$, *we have*

$$\frac{|R_\eta^+(\boldsymbol{\theta}(T_1))-R_\eta^+(\boldsymbol{\theta}(T_2))|}{|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(T_2))|}\leq C\sqrt{T_2-T_1}\eta^{-r}. \tag{3.14}$$

Final stage

If non-degenerate global minimizers are achieved in the training dynamics, we can obtain global-in-time result which characterizing the training dynamics in the final stage. Here we give the definition for non-degenerate minimizers:

Definition 3.1. *A minimizer $\boldsymbol{\theta}^*$ of $R_{\mathcal{D}}$ (or $\tilde{R}_{\mathcal{D}}$, respectively) is global if $R_{\mathcal{D}}(\boldsymbol{\theta}^*)=0$ (or $\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}^*)=0$, respectively). The minimizer is non-degenerate if the Hessian matrix $\nabla^2_{\boldsymbol{\theta}} R_{\mathcal{D}}(\boldsymbol{\theta}^*)$ (or $\nabla^2_{\boldsymbol{\theta}} \tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}^*)$, respectively) exists and is positive definite.*

Theorem 3.3 (F-Principle in the final stage)**.** (L^2 loss function). *Suppose that Assumptions* 2.1, 2.2, 2.3, *and* 2.4 *hold. We consider the training dynamics* (2.11). *If the solution $\boldsymbol{\theta}$ converges to a non-degenerate global minimizer $\boldsymbol{\theta}^*$, then for any $1\le r\le k-1$, there is a constant $C>0$ such that*

$$\frac{|\mathrm{d}R^{+}_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\le C\eta^{-r} \quad \text{and} \quad \frac{|\mathrm{d}R^{-}_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\ge 1-C\eta^{-r}, \quad t\in(0,+\infty). \tag{3.15}$$

(General loss function). *Suppose that Assumptions* 2.1, 2.2, 2.3, *and* 2.5 *hold. We consider the training dynamics* (2.13). *If the solution $\boldsymbol{\theta}$ converges to a non-degenerate global minimizer $\boldsymbol{\theta}^*$, then for any $1\le r\le k-1$, there is a constant $C>0$ such that*

$$\frac{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B^c_\eta)}}{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}}\le C\eta^{-r} \quad \text{and} \quad \frac{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B_\eta)}}{\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}}\ge 1-C\eta^{-r}, \quad t\in(0,+\infty). \tag{3.16}$$

Remark 3.1. Theorems 3.1, 3.2 and 3.3 can be extended to a broad class of models other than DNNs. The key for proving our main results are Lemmas 2.1 and 2.2. Hence for a general parametric model, the similar theorems still hold as long as one can obtain Lemmas 2.1 and 2.2.

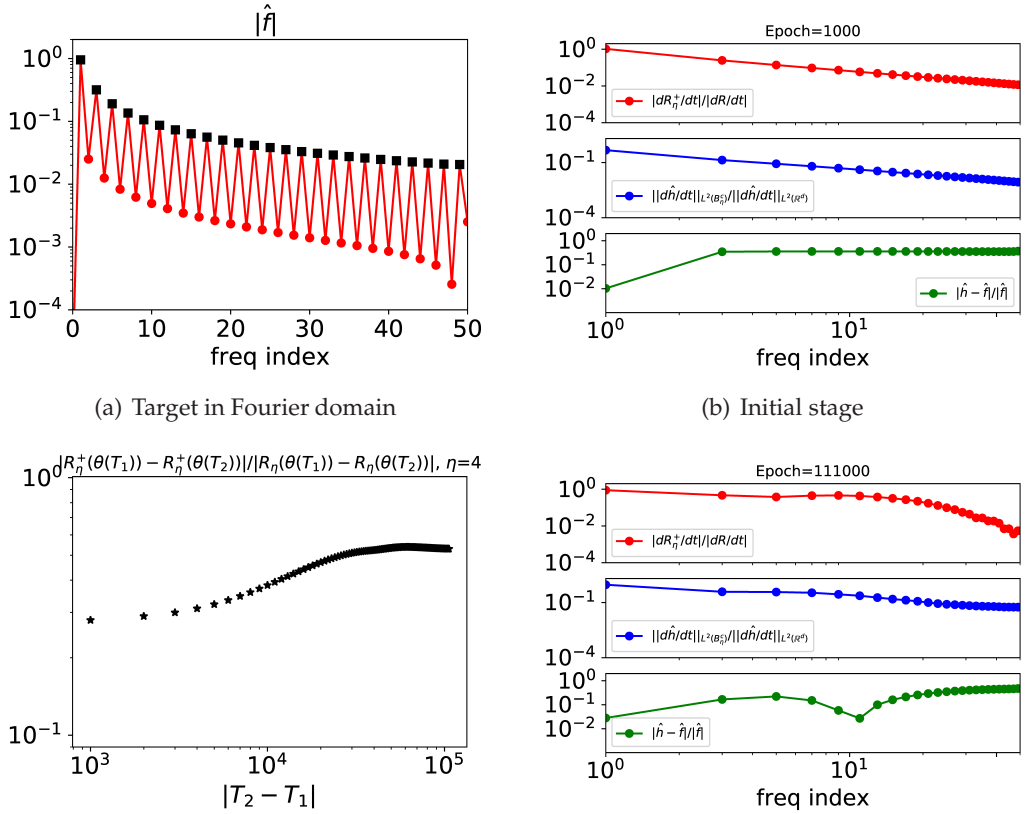
3.3 Discussion and illustrations

To help the readers get some intuitions of the above theorems, we present a numerical example using the following target function

$$f(x)=\sum_{j=1}^{500}\sin(jx/10)/j.$$

The training data are uniformly sampled from $[-3.14,3.14]$ with sample size 300. The discrete Fourier transform of $f(x)$ is shown in Fig. 1(a), in which we focus on the peak frequencies marked by black squares. First, we use the MSE as the training loss function with gradient descent optimizer.

Initial stage in Fig. 1(b). The ratio of the change of the loss function, $|\mathrm{d}R^{+}_{\eta}/\mathrm{d}t|/|\mathrm{d}R/\mathrm{d}t|$ in the upper panel, and the ratio of the change of the DNN output, $\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B^c_\eta)}/\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}$ in the middle panel, both decreases as frequency increases. At such initial stage, only the relative error of the first peak frequency, $|\hat{f}_{\boldsymbol{\theta}}-\hat{f}|/|\hat{f}|$, decreases to a small value.



(a) Target in Fourier domain

(b) Initial stage

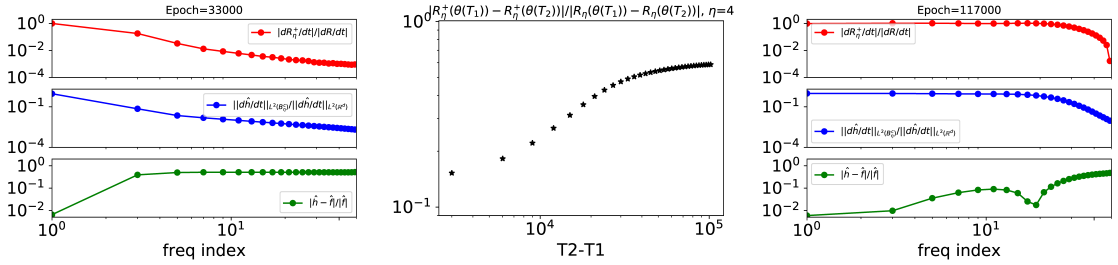
(c) Intermediate stage

(d) Final stage

Figure 1: Numerical understanding of theorems of MSE training loss with gradient descent optimizer. (a) Amplitude of DFT of the training samples against frequency index. Frequencies marked by black squares are analyzed in the second row. (b, d) upper: $|\mathrm{d}R_\eta^+/\mathrm{d}t|/|\mathrm{d}R/\mathrm{d}t|$ vs. frequency index. Middle: $\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B_\eta^c)}/\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}$ vs. frequency index. Lower: Relative error of each selected frequency, $|\hat{f}_{\boldsymbol{\theta}}-\hat{f}|/|\hat{f}|$ vs. frequency index. Each sub-figure is plotted at one training epoch. (c) $|R_\eta^+(\boldsymbol{\theta}(T_1))-R_\eta^+(\boldsymbol{\theta}(T_2))|/|R_\eta(\boldsymbol{\theta}(T_1))-R_\eta(\boldsymbol{\theta}(T_2))|$ vs. $|T_2-T_1|$ with η is selected as the fourth frequency peak. We use a tanh-DNN with widths 1-200-50-1 with full batch training by gradient descent optimizer. The learning rate is 2×10^{-4}.

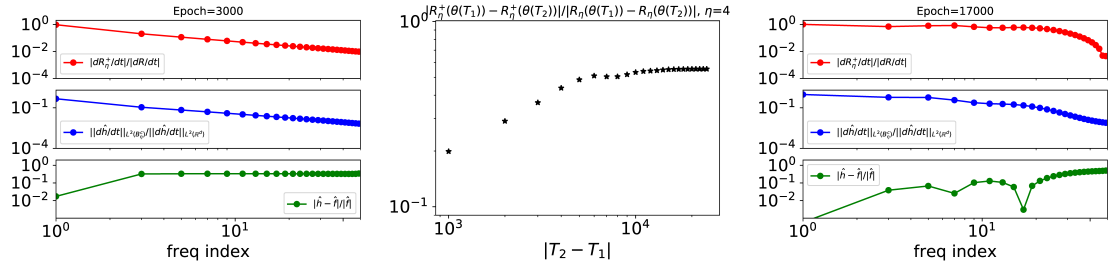
Intermediate stage in Fig. 1(c). The ratio of the change of the loss function in a certain period, $|R_\eta^+(\boldsymbol{\theta}(T_1))-R_\eta^+(\boldsymbol{\theta}(T_2))|/|R_\eta(\boldsymbol{\theta}(T_1))-R_\eta(\boldsymbol{\theta}(T_2))|$, increases with $|T_2-T_1|$ for a fixed η.

Final stage in Fig. 1(d). There exists a frequency η_0 — when $\eta>\eta_0$, the ratio of the change of the loss function, $|\mathrm{d}R_\eta^+/\mathrm{d}t|/|\mathrm{d}R/\mathrm{d}t|$ in the upper panel, and the ratio of the change of the DNN output, $\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(B_\eta^c)}/\|\mathrm{d}\hat{f}_{\boldsymbol{\theta}}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}$ in the middle panel, both decreases as frequency increases. At such final stage, only peak frequencies corresponding



(a) Initial stage (b) Intermediate stage (c) Final stage

Figure 2: Numerical understanding of theorems of MSE training loss with Adam optimizer.The illustrations are same as Fig. 1 (b, c, d), respectively. The learning rate is 2×10^{-5}.



(a) Initial stage (b) Intermediate stage (c) Final stage

Figure 3: Numerical understanding of theorems of L^4 training loss. The illustrations are same as Fig. 1 (b, c, d), respectively. We use a tanh-DNN with widths 1-500-500-500-500-1 with full batch training.

to high frequencies have not converged yet. Note that as shown in Fig. 2, the results for gradient descent optimizer is very similar for Adam optimizer.

Secondly, we use the L^4 training loss $\frac{1}{n}\sum_{i=1}^{n}(f(x_i,\boldsymbol{\theta})-y_i)^4$ as shown in Fig. 3. We obtain similar results.

The different stages may not partition the lifetime of the dynamics into non-intersecting intervals. They may overlap with each other. Theorem 3.1 is regarded as the estimate for the initial stage since the working domain $[0,T]$ starts from the initial time 0. In particular, the constants C in the bounds depend on the initial parameter $\boldsymbol{\theta}_0$ and a given terminal time T. For Theorem 3.2, the time interval $[T_1,T_2]$ is allowed to be more generic. We even provide an explicit dependence on T_2-T_1 in this theorem. In contrast to the local-in-time estimates given by Theorem 3.1 and Theorem 3.2, we prove, under the non-degenerate assumption, the global-in-time Theorem 3.3 where the constant C is uniform in time. These three theorems provide different approaches to characterize the frequency principle for general neural networks, which is lack of an exact mathematical definition in literature.

We further remark that this work focuses on the case of infinite sample where we

have a continuous population distribution density $\rho(x)$. One advantage of such a setting is that the result here directly applies to more general models, not necessarily deep neural networks. For the frequency principle in the over-parameterized case with finite samples, we refer the readers to Ref. [18] where the explicit dependence on the frequency is rigorously derived for the gradient descent dynamics of two-layer neural networks.

4 Proof of theorems

4.1 F-Principle: Initial stage (Theorem 3.1)

In this section, we focus on the initial stage of the training dynamics. The first result shows that the change of loss function concentrates on low frequencies.

In general, C may depend on T. In the next section, we will provide a similar result in some situation where C does not depend on T.

Proof of Theorem 3.1 *(L^2 loss function).* In this proof, we will write $\boldsymbol{\theta}$ for $\boldsymbol{\theta}(t)$. The dynamics for the loss function contributed by high frequency reads as:

$$\frac{\mathrm{d}R^+_{\mathcal{D},\eta}(\boldsymbol{\theta})}{\mathrm{d}t}=\left(\int_{B^c_\eta}\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi}\right)\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}=-\left(\int_{B^c_\eta}\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})\mathrm{d}\boldsymbol{\xi}\right)\cdot\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}). \tag{4.1}$$

The dynamics for the total loss function is

$$\frac{\mathrm{d}R_{\mathcal{D}}(\boldsymbol{\theta})}{\mathrm{d}t}=-|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|^2. \tag{4.2}$$

Therefore

$$\frac{|\mathrm{d}R^+_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\leq\frac{\left(\int_{B^c_\eta}|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})|\mathrm{d}\boldsymbol{\xi}\right)|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|^2}=\frac{\|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})\|_{L^1(B^c_\eta)}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}. \tag{4.3}$$

Note that $\eta\leq\langle\boldsymbol{\xi}\rangle$ for all $0<\eta\leq|\boldsymbol{\xi}|$. Therefore

$$\|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})\|_{L^1(B^c_\eta)}\leq\eta^{-r}\int_{B^c_\eta}\langle\boldsymbol{\xi}\rangle^r|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})|\mathrm{d}\boldsymbol{\xi}\leq\eta^{-r}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})|\|_{L^1(\mathbb{R}^d)}. \tag{4.4}$$

By Assumption 2.3 and Lemma 2.4, $\sup_{t\geq0}|\boldsymbol{\theta}(t)|\leq C_0$ and

$$\sup_{t\in(0,T]}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})|\|_{L^1(\mathbb{R}^d)}<+\infty. \tag{4.5}$$

For $k=1$, we take the assumption that $\inf_{t\in(0,T]}|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t))|>0$. For $k\geq2$, according to Assumption 2.1, we have $\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\cdot)\in W^{1,\infty}_{\mathrm{loc}}(\mathbb{R}^d;\mathbb{R}^M)$, and hence $\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})$ is locally Lipschitz in $\boldsymbol{\theta}$. This together with Assumption 2.3 implies that $\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t))$ is

continuous on $t\in[0,T]$. If $\inf_{t\in(0,T]}|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t))|=0$, then there is a $t_0\in[0,T]$ such that $|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t_0))|=0$. By the uniqueness of ordinary differential equation, we have $|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(T))|=0$ which contradicts with the assumption that $|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t))|>0$. Therefore $\inf_{t\in(0,T]}|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}(t))|>0$. Thus for $k\geq 1$ the following ratio is bounded from above:

$$C:=\sup_{t\in(0,T]}\frac{\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}q_{\mathcal{D}}(\cdot,\boldsymbol{\theta})|\|_{L^1(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}<+\infty. \tag{4.6}$$

Therefore

$$\frac{|\mathrm{d}R^+_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|}\leq C\eta^{-r},\quad t\in(0,T]. \tag{4.7}$$

This completes the proof. □

Corollary 4.1 (dissipation)**.** *In the situation of Theorem* 3.1 *for L^2 loss function, we have that for sufficiently large η*

$$\frac{\mathrm{d}R^-_{\mathcal{D},\eta}}{\mathrm{d}t}\leq-(1-C\eta^{-r})|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}|^2\leq 0. \tag{4.8}$$

Proof. For sufficiently large η, the dynamics of $R^-_{\mathcal{D},\eta}$ is dissipative because

$$\frac{\mathrm{d}R^-_{\mathcal{D},\eta}(\boldsymbol{\theta})}{\mathrm{d}t}=\frac{\mathrm{d}R_{\mathcal{D}}(\boldsymbol{\theta})}{\mathrm{d}t}-\frac{\mathrm{d}R^+_{\mathcal{D},\eta}(\boldsymbol{\theta})}{\mathrm{d}t}\leq-(1-C\eta^{-r})|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|^2\leq 0.$$

This completes the proof. □

Next we prove the case of general loss function.

Proof of Theorem 3.1 *(general loss function).* On the one hand, we estimate the numerator by studying the dynamics for $\hat{f}_{\boldsymbol{\theta}}$:

$$\frac{\mathrm{d}\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})}{\mathrm{d}t}=\nabla_{\boldsymbol{\theta}}\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}=-\nabla_{\boldsymbol{\theta}}\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}). \tag{4.9}$$

Taking square and integrating both sides on B^c_η leads to the upper bound on the numerator

$$\left\|\frac{\mathrm{d}\hat{f}(\cdot,\boldsymbol{\theta})}{\mathrm{d}t}\right\|_{L^2(B^c_\eta)}\leq|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\|_{L^2(B^c_\eta)}. \tag{4.10}$$

On the other hand, note the dynamics for the hypothesis function

$$\frac{\mathrm{d}f(\boldsymbol{x},\boldsymbol{\theta})}{\mathrm{d}t}=\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x},\boldsymbol{\theta})\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}=-\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}f(\boldsymbol{x},\boldsymbol{\theta}) \tag{4.11}$$

and the dynamics for the total loss function

$$\frac{\mathrm{d}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})}{\mathrm{d}t}=-|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|^2. \tag{4.12}$$

Thus we have

$$\begin{aligned}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|^2&=\left|\frac{\mathrm{d}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})}{\mathrm{d}t}\right|=\left|\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}\ell(f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x}))\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\\&=\left|\int_{\mathbb{R}^d}\frac{\mathrm{d}f(\boldsymbol{x},\boldsymbol{\theta})}{\mathrm{d}t}\ell'(f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x}))\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right|\\&\leq\|\sqrt{\rho}\|_{L^\infty}\left\|\frac{\mathrm{d}f(\cdot,\boldsymbol{\theta})}{\mathrm{d}t}\right\|_{L^2(\mathbb{R}^d)}\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)},\end{aligned}\tag{4.13}$$

where we used the Cauchy–Schwarz inequality in the last step. Combining Eqs. (4.10) and (4.13), we obtain

$$\begin{aligned}\frac{\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\|_{L^2(B_\eta^c)}}{\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\|_{L^2(\mathbb{R}^d)}}&\leq\frac{\|\sqrt{\rho}\|_{L^\infty}\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\|_{L^2(B_\eta^c)}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|^2}\\&\leq\|\sqrt{\rho}\|_{L^\infty}\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\|_{L^2(B_\eta^c)}\frac{\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|}.\end{aligned}\tag{4.14}$$

Similar to the case of L^2 loss function,

$$\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\|_{L^2(B_\eta^c)}\leq\eta^{-r}\left(\int_{B_\eta^c}\langle\boldsymbol{\xi}\rangle^{2r}|\nabla_{\boldsymbol{\theta}}\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta})|^2\mathrm{d}\boldsymbol{\xi}\right)^{1/2}\leq\eta^{-r}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)}.\tag{4.15}$$

Again, by Assumption 2.3 and Lemma 2.4, $\sup_{t\geq0}|\boldsymbol{\theta}(t)|\leq C_0$ and

$$\sup_{t\in(0,T]}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)}<+\infty.\tag{4.16}$$

For $k\geq2$, the same argument of the L^2 loss case leads to $\inf_{t\in(0,T]}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|>0$. The proof is completed by the following bound

$$\sup_{t\in(0,T]}\frac{\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|}\leq\sup_{t\in(0,T]}\frac{\left(C\int_{\mathbb{R}^d}\ell(f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x}))\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/2}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|}<+\infty,$$

where we used Assumption 2.5. □

4.2 F-Principle: Intermediate stage (Theorem 3.2)

In this section, we prove the key theorem for the intermediate stage. This theorem then implies several useful corollaries.

Proof of Theorem 3.2. The numerator can be controlled as follows

$$\begin{aligned}
&\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R_\eta^+(\boldsymbol{\theta})}{\mathrm{d}t}\right|\mathrm{d}t\\
=&\int_{T_1}^{T_2}\left|\left(\int_{B_\eta^c}\nabla_{\boldsymbol{\theta}}q(\boldsymbol{\xi},\boldsymbol{\theta}(t))\mathrm{d}\boldsymbol{\xi}\right)\cdot\frac{\mathrm{d}\boldsymbol{\theta}(t)}{\mathrm{d}t}\right|\mathrm{d}t\\
\le&\int_{T_1}^{T_2}\left(\int_{B_\eta^c}|\nabla_{\boldsymbol{\theta}}q(\boldsymbol{\xi},\boldsymbol{\theta}(t))|\mathrm{d}\boldsymbol{\xi}\right)|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|\mathrm{d}t\\
=&\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|\int_{B_\eta^c}\left|\nabla_{\boldsymbol{\theta}}\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta}(t))\overline{\hat{f}(\boldsymbol{\xi},\boldsymbol{\theta}(t))-\hat{f}(\boldsymbol{\xi})}+\text{c.c.}\right|\mathrm{d}\boldsymbol{\xi}\mathrm{d}t\\
\le&2\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta}(t))\|_{L^2(B_\eta^c)}\|h(\cdot,\boldsymbol{\theta}(t))-f(\cdot)\|_{L^2(\mathbb{R}^d)}\mathrm{d}t\\
\le&2\eta^{-r}\left(\sup_{t\in[T_1,T_2]}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta}(t))|\|_{L^2(\mathbb{R}^d)}\right)\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|R(\boldsymbol{\theta}(t))^{1/2}\mathrm{d}t,
\end{aligned}\tag{4.17}$$

where in the second-to-last step we used the Cauchy–Schwarz inequality and the Plancherel theorem, and in the last step we used the following

$$\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta}(t))\|_{L^2(B_\eta^c)}\le\eta^{-r}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta}(t))|\|_{L^2(\mathbb{R}^d)}.\tag{4.18}$$

By Assumption 2.3 and Lemma 2.4, $\sup_{t\ge0}|\boldsymbol{\theta}(t)|\le C_0$ and

$$C_1:=\sup_{t\ge0}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta}(t))|\|_{L^2(\mathbb{R}^d)}<+\infty.\tag{4.19}$$

By the assumption that $\frac{1}{2}R(\boldsymbol{\theta}(T_1))\ge R(\boldsymbol{\theta}(T_2))$, we have

$$\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t\ge|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(T_2))|\ge\frac{1}{2}R(\boldsymbol{\theta}(T_1)).\tag{4.20}$$

Therefore,

$$\begin{aligned}
\frac{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R_\eta^+}{\mathrm{d}t}\right|\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t}&\le\frac{2C_1\eta^{-r}\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|R(\boldsymbol{\theta}(t))^{1/2}\mathrm{d}t}{|\frac{1}{2}R(\boldsymbol{\theta}(T_1))|^{1/2}(\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R(\boldsymbol{\theta}(t))}{\mathrm{d}t}\right|\mathrm{d}t)^{1/2}}\\
&\le\frac{2\sqrt{2}C_1\eta^{-r}(\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|^2\mathrm{d}t)^{1/2}}{|R(\boldsymbol{\theta}(T_1))|^{1/2}}\times\frac{(\int_{T_1}^{T_2}R(\boldsymbol{\theta}(t))\mathrm{d}t)^{1/2}}{(\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R(\boldsymbol{\theta}(t))}{\mathrm{d}t}\right|\mathrm{d}t)^{1/2}}.
\end{aligned}\tag{4.21}$$

Recall the training dynamics

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}=-\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}),\tag{4.22}$$

where for $k\geq 2$, $\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\cdot)\in W^{1,\infty}_{\rm loc}(\mathbb{R}^M)\subset C^{0,1}(\mathbb{R}^M)$. Hence $\boldsymbol{\theta}(\cdot)\in C^{1,1}([0,+\infty))$. Taking further time derivative of $\boldsymbol{\theta}$, we obtain

$$\frac{\mathrm{d}^2\boldsymbol{\theta}}{\mathrm{d}t^2}=-\nabla^2_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}=\nabla^2_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}). \tag{4.23}$$

Since $\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\cdot),\nabla^2_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\cdot)\in L^{\infty}_{\rm loc}(\mathbb{R}^M)$ and $\boldsymbol{\theta}(\cdot)$ is continuous, we have $\frac{\mathrm{d}^2\boldsymbol{\theta}(\cdot)}{\mathrm{d}t^2}\in L^{\infty}_{\rm loc}([0,+\infty))$. Taking time derivatives of the L^2 loss function R, we obtain

$$\frac{\mathrm{d}R}{\mathrm{d}t}=\nabla_{\boldsymbol{\theta}}R\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}, \tag{4.24}$$

$$\frac{\mathrm{d}^2R}{\mathrm{d}t^2}=(\nabla^2_{\boldsymbol{\theta}}R\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t})\cdot\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}+\nabla_{\boldsymbol{\theta}}R\cdot\frac{\mathrm{d}^2\boldsymbol{\theta}}{\mathrm{d}t^2}. \tag{4.25}$$

The facts that $\nabla^2_{\boldsymbol{\theta}}R(\cdot),\nabla_{\boldsymbol{\theta}}R(\cdot)\in L^{\infty}_{\rm loc}(\mathbb{R}^M)$ and that $\boldsymbol{\theta}(\cdot)$ is continuous lead to $\nabla^2_{\boldsymbol{\theta}}R(\boldsymbol{\theta}(\cdot))$, $\nabla_{\boldsymbol{\theta}}R(\boldsymbol{\theta}(\cdot))\in L^{\infty}_{\rm loc}([0,+\infty))$. This with $\frac{\mathrm{d}\boldsymbol{\theta}(\cdot)}{\mathrm{d}t},\frac{\mathrm{d}^2\boldsymbol{\theta}(\cdot)}{\mathrm{d}t^2}\in L^{\infty}_{\rm loc}([0,+\infty))$ implies that $\frac{\mathrm{d}^2R(\boldsymbol{\theta}(\cdot))}{\mathrm{d}t^2}\in L^{\infty}_{\rm loc}([0,+\infty))$. Therefore $\frac{\mathrm{d}R(\boldsymbol{\theta}(\cdot))}{\mathrm{d}t}\in W^{1,\infty}_{\rm loc}([0,+\infty))\subset C^{0,1}([0,+\infty))$. Thus $Q:=\max_{t\in[T_1,T_2]}R(\boldsymbol{\theta}(t))$ is finite. If $Q\leq 2R(\boldsymbol{\theta}(T_1))$, we have

$$\frac{\int_{T_1}^{T_2}R(\boldsymbol{\theta})\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R(\boldsymbol{\theta})}{\mathrm{d}t}\right|\mathrm{d}t}\leq\frac{(T_2-T_1)Q}{|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(T_2))|}\leq\frac{(T_2-T_1)2R(\boldsymbol{\theta}(T_1))}{\frac{1}{2}R(\boldsymbol{\theta}(T_1))}=4(T_2-T_1). \tag{4.26}$$

If $Q>2R(\boldsymbol{\theta}(T_1))$, then we choose $t_Q\in[T_1,T_2]$ such that $R(\boldsymbol{\theta}(t_Q))=Q$. We have

$$\begin{aligned}\frac{\int_{T_1}^{T_2}R(\boldsymbol{\theta})\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R(\boldsymbol{\theta})}{\mathrm{d}t}\right|\mathrm{d}t}&\leq\frac{(T_2-T_1)Q}{|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(t_Q))|+|R(\boldsymbol{\theta}(t_Q))-R(\boldsymbol{\theta}(T_2))|}\\&=\frac{(T_2-T_1)Q}{Q-R(\boldsymbol{\theta}(T_1))+Q-R(\boldsymbol{\theta}(T_2))}\\&\leq\frac{(T_2-T_1)Q}{2Q-\frac{3}{2}R(\boldsymbol{\theta}(T_1))}\leq\frac{4}{5}(T_2-T_1).\end{aligned} \tag{4.27}$$

Combining Eqs. (4.26) and (4.27), we have

$$\frac{\int_{T_1}^{T_2}R(\boldsymbol{\theta})\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R(\boldsymbol{\theta})}{\mathrm{d}t}\right|\mathrm{d}t}\leq 4(T_2-T_1). \tag{4.28}$$

Therefore

$$\begin{aligned}\frac{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R^+_\eta}{\mathrm{d}t}\right|\mathrm{d}t}{\int_{T_1}^{T_2}\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t}&\leq\frac{2\sqrt{8}C_1\eta^{-r}\sqrt{T_2-T_1}|\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(T_1))-\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(T_2))|^{1/2}}{|R(\boldsymbol{\theta}(T_1))|^{1/2}}\\&\leq 4\sqrt{2}C_1\eta^{-r}\sqrt{T_2-T_1}\frac{(\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(T_1)))^{1/2}}{(R(\boldsymbol{\theta}(T_1)))^{1/2}}=C\sqrt{T_2-T_1}\eta^{-r},\end{aligned} \tag{4.29}$$

where $C=4\sqrt{2}C_1C_2^{1/2}$ and $C_2:=\sup_{t\geq 0}\frac{\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))}{R(\boldsymbol{\theta}(t))}$. Now it is sufficient to show that $C_2<+\infty$. In fact, there is a constant C_3 such that $\sup_{|\boldsymbol{\theta}|\leq C_0}\sup_{\boldsymbol{x}\in\mathbb{R}^d}|f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x})|\leq C_3$. This with Assumption 2.5 implies that $\ell(z)\leq C_4|z|^2$ for $|z|\leq C_3$. Therefore

$$\begin{aligned} C_2 &\leq \sup_{t\geq 0}\frac{\int_{\mathbb{R}^d}\ell(f(\boldsymbol{x},\boldsymbol{\theta}(t))-f(\boldsymbol{x},\boldsymbol{\theta}^*))\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^d}(f(\boldsymbol{x},\boldsymbol{\theta}(t))-f(\boldsymbol{x},\boldsymbol{\theta}^*))^2\mathrm{d}\boldsymbol{x}} \\ &\leq \|\rho\|_{L^\infty}\sup_{t\geq 0}\frac{\int_{\mathbb{R}^d}C_4(f(\boldsymbol{x},\boldsymbol{\theta}(t))-f(\boldsymbol{x},\boldsymbol{\theta}^*))^2\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^d}(f(\boldsymbol{x},\boldsymbol{\theta}(t))-f(\boldsymbol{x},\boldsymbol{\theta}^*))^2\mathrm{d}\boldsymbol{x}}<+\infty. \end{aligned} \tag{4.30}$$

This completes the proof. □

Remark 4.1. If the condition $\frac{1}{2}R(\boldsymbol{\theta}(T_1))\geq R(\boldsymbol{\theta}(T_2))$ is replaced by $\delta R(\boldsymbol{\theta}(T_1))\geq R(\boldsymbol{\theta}(T_2))$ for any $\delta\in(0,1)$, the estimates in Theorem 3.2 and the following corollaries still hold.

Proof of Corollary 3.1. Similar to the proof of Theorem 3.2, we have the upper bound for the numerator

$$|R_\eta^+(\boldsymbol{\theta}(T_1))-R_\eta^+(\boldsymbol{\theta}(T_2))|\leq 2\eta^{-r}C_1\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|R(\boldsymbol{\theta}(t))^{1/2}\mathrm{d}t \tag{4.31}$$

and lower bound for the denominator $|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(T_2))|\geq R(\boldsymbol{\theta}(T_1))/2$ with $C_1:=\sup_{t\in[T_1,T_2]}\|\langle\cdot\rangle^r\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta}(t))\|_{L^2(\mathbb{R}^d)}<+\infty$. Therefore, these bounds with the assumption that $R(\boldsymbol{\theta}(T_1))\geq R(\boldsymbol{\theta}(t))$ for all $t\in[T_1,T_2]$ leads to

$$\begin{aligned} \frac{|R_\eta^+(\boldsymbol{\theta}(T_1))-R_\eta^+(\boldsymbol{\theta}(T_2))|}{|R(\boldsymbol{\theta}(T_1))-R(\boldsymbol{\theta}(T_2))|} &\leq \frac{2C_1\eta^{-r}\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|R(\boldsymbol{\theta}(t))^{1/2}\mathrm{d}t}{\frac{1}{2}R(\boldsymbol{\theta}(T_1))} \\ &\leq \frac{4C_1\eta^{-r}\int_{T_1}^{T_2}|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}(t))|\mathrm{d}t}{(R(\boldsymbol{\theta}(T_1)))^{1/2}}\leq C\sqrt{T_2-T_1}\eta^{-r}, \end{aligned} \tag{4.32}$$

where the last inequality is due to the same reason as Theorem 3.2. □

Corollary 4.2. *Under the same assumptions in Theorem* 3.2, *if the solution* $\boldsymbol{\theta}$ *converges to a non-degenerate global minimizer* $\boldsymbol{\theta}^*$, *then for any* $1\leq r\leq k-1$, *the above upper bound can be improved to the following: there is a constant* $C>0$ *such that for any* $T>0$, *we have*

$$\frac{\int_0^T\left|\frac{\mathrm{d}R_\eta^+}{\mathrm{d}t}\right|\mathrm{d}t}{\int_0^T\left|\frac{\mathrm{d}R}{\mathrm{d}t}\right|\mathrm{d}t}\leq C\eta^{-r} \tag{4.33}$$

and

$$\frac{|R_\eta^+(\boldsymbol{\theta}(0))-R_\eta^+(\boldsymbol{\theta}(T))|}{|R(\boldsymbol{\theta}(0))-R(\boldsymbol{\theta}(T))|}\leq C\eta^{-r}. \tag{4.34}$$

We skip the proof since this corollary can be obtained directly from Theorem 3.3.

4.3 F-Principle: Final stage (Theorem 3.3)

In this section, we prove the F-Principle in final stage of the training dynamics.

Proof of Theorem 3.3 *(L^2 loss function).* Following (4.3), we have

$$\begin{aligned}\frac{|\mathrm{d}R^+_{\mathcal{D},\eta}/\mathrm{d}t|}{|\mathrm{d}R_{\mathcal{D}}/\mathrm{d}t|} &\le \frac{\int_{B^c_\eta}\left|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})\overline{\hat{f}_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})-\hat{f}_{\mathcal{D}}(\boldsymbol{\xi})}+\text{c.c.}\right|\mathrm{d}\boldsymbol{\xi}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}\\ &\le \frac{2\|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})\|_{L^2(B^c_\eta)}\|\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})-\hat{f}_{\mathcal{D}}(\cdot)\|_{L^2(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}\\ &= 2\|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})\|_{L^2(B^c_\eta)}\frac{|R_{\mathcal{D}}(\boldsymbol{\theta})|^{1/2}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|},\end{aligned} \tag{4.35}$$

where we used $\|\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})-\hat{f}_{\mathcal{D}}(\cdot)\|^2_{L^2(\mathbb{R}^d)}=R_{\mathcal{D}}(\boldsymbol{\theta})$ in the last inequality. Similar to the local-in-time situation,

$$\begin{aligned}\|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})\|_{L^2(B^c_\eta)} &\le \eta^{-r}\left(\int_{B^c_\eta}\langle\boldsymbol{\xi}\rangle^{2r}|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\boldsymbol{\xi},\boldsymbol{\theta})|^2\mathrm{d}\boldsymbol{\xi}\right)^{1/2}\\ &\le \eta^{-r}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)}.\end{aligned} \tag{4.36}$$

By Assumption 2.3 and Lemma 2.4, $\sup_{t\ge0}|\boldsymbol{\theta}(t)|\le C_0$ and

$$\sup_{t\in[0,+\infty)}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}_{\mathcal{D}}(\cdot,\boldsymbol{\theta}(t))|\|_{L^2(\mathbb{R}^d)}<+\infty. \tag{4.37}$$

Now it is sufficient to prove that

$$C:=\lim_{t\to+\infty}\frac{|R_{\mathcal{D}}(\boldsymbol{\theta})|^{1/2}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}<+\infty. \tag{4.38}$$

This is true because

$$C=\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\frac{|R_{\mathcal{D}}(\boldsymbol{\theta})|^{1/2}}{|\nabla_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta})|}=\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\frac{|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\Lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+o(|\boldsymbol{\theta}-\boldsymbol{\theta}^*|^2)|^{1/2}}{|2\Lambda(\boldsymbol{\theta}-\boldsymbol{\theta}^*)|+o(|\boldsymbol{\theta}-\boldsymbol{\theta}^*|)}<+\infty, \tag{4.39}$$

where we used the assumption that the minimizer is non-degenerate with the Hessian $\Lambda=\nabla^2_{\boldsymbol{\theta}}R_{\mathcal{D}}(\boldsymbol{\theta}^*)$. □

Now we finish the proof for general loss function.

Proof of Theorem 3.3 *(general loss function).* By the proof of Theorem 3.1, we have

$$\frac{\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\|_{L^2(B^c_\eta)}}{\|\frac{\mathrm{d}\hat{f}_{\boldsymbol{\theta}}}{\mathrm{d}t}\|_{L^2(\mathbb{R}^d)}}\le 2\|\sqrt{\rho}\|_{L^\infty}\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\|_{L^2(B^c_\eta)}\frac{\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|} \tag{4.40}$$

and

$$\|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})\|_{L^2(B_\eta^c)} \leq \eta^{-r}\|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)}. \tag{4.41}$$

By Assumption 2.3 and Lemma 2.4, $\sup_{t\geq 0}|\boldsymbol{\theta}(t)| \leq C_0$ and

$$\sup_{t\in[0,+\infty)} \|\langle\cdot\rangle^r|\nabla_{\boldsymbol{\theta}}\hat{f}(\cdot,\boldsymbol{\theta})|\|_{L^2(\mathbb{R}^d)} < +\infty. \tag{4.42}$$

Now it is sufficient to prove that

$$\sup_{t\in(0,\infty]} \frac{\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|} < +\infty. \tag{4.43}$$

This is true because

$$\begin{aligned}
\lim_{t\to+\infty} \frac{\|\ell'(f(\cdot,\boldsymbol{\theta})-f(\cdot))\sqrt{\rho(\cdot)}\|_{L^2(\mathbb{R}^d)}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|} &= \lim_{t\to+\infty} \frac{\left(\int_{\mathbb{R}^d}[\ell'(f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x}))]^2\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/2}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|} \\
&\leq \lim_{t\to+\infty} \frac{\left(C\int_{\mathbb{R}^d}\ell(f(\boldsymbol{x},\boldsymbol{\theta})-f(\boldsymbol{x}))\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)^{1/2}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|} \\
&= \lim_{t\to+\infty} \frac{C^{1/2}|\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|^{1/2}}{|\nabla_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta})|} \\
&= \lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} \frac{|(\boldsymbol{\theta}-\boldsymbol{\theta}^*)^T\tilde{\Lambda}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+o(|\boldsymbol{\theta}-\boldsymbol{\theta}^*|^2)|^{1/2}}{|2\tilde{\Lambda}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)|+o(|\boldsymbol{\theta}-\boldsymbol{\theta}^*|)} \\
&< +\infty,
\end{aligned} \tag{4.44}$$

where we used Assumption 2.5 and the assumption that the minimizer is non-degenerate with the Hessian $\tilde{\Lambda} = \nabla^2_{\boldsymbol{\theta}}\tilde{R}_{\mathcal{D}}(\boldsymbol{\theta}^*)$. □

Acknowledgments

This work is sponsored by the National Key R&D Program of China Grant No. 2019YFA0709503 (Z. X.), the Shanghai Sailing Program, the Natural Science Foundation of Shanghai Grant No. 20ZR1429000 (Z. X.), the National Natural Science Foundation of China Grant No. 62002221 (Z. X.), Shanghai Municipal of Science and Technology Project Grant No. 20JC1419500 (Y. Z.), Shanghai Municipal of Science and Technology Major Project NO. 2021SHZDZX0102, and the HPC of School of Mathematical Sciences and the Student Innovation Center at Shanghai Jiao Tong University. The authors would like to thank Prof. Weinan E for the helpful discussion.

References

[1] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International Conference on Machine Learning*, pages 173–182, 2016.

[2] S. Bi, Z. Xu, P. Srinivasan, B. Mildenhall, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020.

[3] S. Biland, V. C. Azevedo, B. Kim, and B. Solenthaler. Frequency-aware reconstruction of fluid simulations with generative networks. *arXiv preprint arXiv:1912.08776*, 2019.

[4] B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034, 2020.

[5] W. Cai, X. Li, and L. Liu. A phase shift deep neural network for high frequency wave equations in inhomogeneous media. *SIAM Journal on Scientific Computing*, 42(5):A3285–A3312, 2020.

[6] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu. Towards understanding the spectral bias of deep learning. *arXiv preprint arXiv:1912.01198*, 2019.

[7] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[8] W. E, J. Han, and A. Jentzen. Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics*, 5(4):349–380, 2017.

[9] W. E, C. Ma, L. Wu, and S. Wojtowytsch. Towards a mathematical understanding of neural network-based machine learning: What we know and what we don't. *CSIAM Trans. Appl. Math.*, 1(4):561–615, 2020.

[10] A. Eitel, J. T. Springenberg, L. Spinello, Ma. Riedmiller, and W. Burgard. Multimodal deep learning for robust RGB-D object recognition. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 681–687. IEEE, 2015.

[11] Y. Fan, L. Lin, L. Ying, and L. Zepeda-Núnez. A multiscale neural network based on hierarchical matrices. *Multiscale Modeling & Simulation*, 17(4):1189–1213, 2019.

[12] J. He, L. Li, J. Xu, and C. Zheng. ReLU deep neural networks and linear finite elements. *Journal of Computational Mathematics*, 2019.

[13] A. D. Jagtap, K. Kawaguchi, and G. E. Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020.

[14] Y. Khoo and L. Ying. SwitchNet: A neural network model for forward and inverse scattering problems. *SIAM Journal on Scientific Computing*, 41(5):A3182–A3201, 2019.

[15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[16] X.-A. Li, Z.-Q. J. Xu, and L. Zhang. A multi-scale DNN algorithm for nonlinear elliptic equations with multiple scales. *Communications in Computational Physics*, 28(5):1886–1906, 2020.

[17] Z. Liu, W. Cai, and Z.-Q. J. Xu. Multi-scale deep neural network (MscaleDNN) for solving Poisson-Boltzmann equation in complex domains. *Communications in Computational Physics*, 28(5):1970–2001, 2020.

[18] T. Luo, Z. Ma, Z.-Q. J. Xu, and Y. Zhang. On the exact computation of linear frequency principle dynamics and its generalization. *arXiv preprint arXiv:2010.08153*, 2020.

[19] S. Mei, A. Montanari, and P.-M. Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.
[20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020.
[21] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.
[22] N. C. Rabinowitz. Meta-learners' learning dynamics are unlike learners'. *arXiv preprint arXiv:1905.01320*, 2019.
[23] N. Rahaman, D. Arpit, A. Baratin, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of deep neural networks. *International Conference on Machine Learning*, 2019.
[24] B. Ronen, D. Jacobs, Y. Kasten, and S. Kritchman. The convergence rate of neural networks for learned functions of different frequencies. In *Advances in Neural Information Processing Systems*, pages 4763–4772, 2019.
[25] G. Rotskoff and E. Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, pages 7146–7155, 2018.
[26] J. Sirignano and K. Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.
[27] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.
[28] F. Wang, A. Eljarrat, J. Müller, T. R. Henninen, R. Erni, and C. T. Koch. Multi-resolution convolutional neural networks for inverse problems. *Scientific Reports*, 10(1):1–11, 2020.
[29] J. Wang, Z.-Q. J. Xu, J. Zhang, and Y. Zhang. Implicit bias with Ritz-Galerkin method in understanding deep learning for solving pdes. *arXiv preprint arXiv:2002.07989*, 2020.
[30] W. E, C. Ma, and L. Wu. Machine learning from a continuous viewpoint, I. *Science China Mathematics*, 63(11):2233–2266, 2020.
[31] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao. Training behavior of deep neural network in frequency domain. *International Conference on Neural Information Processing*, pages 264–274, 2019.
[32] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.
[33] Z.-Q. J. Xu and H. Zhou. Deep frequency principle towards understanding why deeper learning is faster. *arXiv preprint arXiv:2007.14313, AAAI-21*, 2020.
[34] G. Yang and H. Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.
[35] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018.
[36] Y. Zhang, Z.-Q. J. Xu, T. Luo, and Z. Ma. Explicitizing an implicit bias of the frequency principle in two-layer neural networks. *arXiv preprint arXiv:1905.10264*, 2019.