Frequency Principle in deep learning: an overview

Zhi-Qin John Xu
zhiqinxu@nyu.edu

August 7, 2019

Abstract

This report provides an overview the research of the Frequency Principle (F-Principle) in deep learning that DNNs (Deep neural networks) often fit target functions from low to high frequencies during the training. Xu et al. (2018), Rahaman et al. (2018), Xu et al. (2019) empirically verify the F-Principle on synthetic and real datasets. Xu et al. (2019) also points out a key mechanism underlying the F-Principle. Luo et al. (2019) then gives a rigorous characterization of the F-Principle during the DNN training. Zhang et al. (2019b) proposes an effective linear F-Principle (LFP) model that accurately predict two-layer DNN outputs when neuron number is large. Zhang et al. (2019b) also utilizes an kernel-norm minimization framework to study the LFP model, which gives an a prior generalization error bound of two-layer DNNs. The F-Principle has been adopted in the empirical study, theoretical study and application of deep learning. The series work of the F-Principle lays a new direction for quantitatively understanding deep learning. This overview would be updated frequently as the development of the F-Principle.

1 The development trajectory of the F-Principle

The Frequency Principle (F-Principle) is stated as follows (Xu et al. 2018, 2019):

DNNs often fit target functions from low to high frequencies during the training.

With numerical simulations, the F-Principle is first proposed in Xu et al. (2018) through regression problems on synthetic data and real data with a projection method. Xu (2018a) subsequently propose a theoretical analysis framework for one hidden layer neural network with 1-d input, which illustrates the key mechanism underlying the F-Principle—the activation function (including tanh and Relu) in the Fourier domain decays as frequency increases. Note that (Rahaman et al. 2018) also found a similar phenomenon as the F-Principle. Xu (2018b) examined the F-Principle through classification problems on benchmark datasets with cross-entropy loss with a projection method and solving a Poisson equation with a variation loss function. Xu (2018b) also proposes that DNN can be adopted to accelerate the convergence of low frequencies for scientific computing problems, in which most of the conventional methods (e.g., Jacobi method) exhibit the opposite convergence behavior—faster convergence for higher frequencies. Xu et al. (2019) summarizes some works in Xu (2018a,b). In addition, Xu et al. (2019) proposes a Gaussian filtering method which can directly verify the F-Principle in high dimensional datasets for both regression and classification problems. Xu et al. (2019) also utilizes the F-Principle to understand both the success and failure of DNNs in different types of problems. Therefore, Xu et al. (2019) is a foundation work for the study of the F-Principle.

Then, we rigorously investigate the F-Principle for the training dynamics of a general DNN at three stages: initial stage, intermediate stage, and final stage (Luo et al. 2019). For each stage, a theorem is provided in terms of proper quantities characterizing the F-Principle. Our results are general in the

Preprint. Work in progress.
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.

To advance the quantitative understanding of DNNs by the F-Principle, we develop an analysis framework in the kernel regime, in which the training dynamics of a DNN is found to be well approximated by the gradient flow of a linearized model of the DNN resembling kernel methods. In Zhang et al. (2019b), we develop a kernel-norm minimization framework in the kernel regime and find that for any loss in a general class of functions, the DNN finds the same global minima—the one that is nearest to the initial value in the parameter space, or equivalently, the one that is closest to the initial DNN output in the corresponding reproducing kernel Hilbert space. With this framework, we prove that a non-zero initial output increases the generalization error of DNN. We further propose an antisymmetrical initialization (ASI) trick that eliminates this type of error and accelerates the training, even for DNNs in the non-kernel regime.

With the kernel-norm minimization framework, in Zhang et al. (2019a), we propose an effective model of linear F-Principle (LFP) dynamics which accurately predicts the learning results of two-layer ReLU neural networks (NNs) of large widths. Importantly, the long-time limit solution of this LFP dynamics is equivalent to the solution of a constrained optimization problem explicitly minimizing an FP-norm, in which higher frequencies of feasible solutions are more heavily penalized. Using this optimization formulation, an a priori estimate of the generalization error bound is provided, revealing that a higher FP-norm of the target function increases the generalization error. Overall, by explicitizing the implicit bias of the F-Principle as an explicit penalty for two-layer NNs, our work makes a step towards a quantitative understanding of the learning and generalization of general DNNs.

2 The influence of the F-Principle


**Empirical study.** The F-Principle is used as an important phenomenon to pursue fundamentally different learning trajectories of meta-learning (Rabinowitz 2019). The F-Principle is also used as a tool to observe the performance of adaptive activation function (Jagtap & Karniadakis 2019).

**Theoretical study.** The theoretical framework in Xu (2018a), Xu et al. (2019) of analyzing the F-Principle is used to analyze a nonlinear collaborative scheme for deep network training (Zhen et al. 2018) and the DNN with ReLU activation function (Rahaman et al. 2018). Basri et al. (2019) analyze the convergence rate of one-hidden layer neural networks for learned functions of different frequencies in the kernel regime through eigenvalue analysis. Note that our study in Zhang et al. (2019a) propose an accurate model that explicitly express the dynamics of one-hidden layer neural networks for learned functions of different frequencies in the kernel regime. **Application.** Based on the F-Principle, a fast algorithm by shifting high frequencies to lower ones is developed for fitting high frequency functions (Cai et al. 2019). These subsequent works show the importance of the F-Principle.

Note that some papers (Arora et al. 2019, Bukweon et al. 2019, Nakkiran et al. 2019, Michoski et al. 2019, Zajc et al. 2019, Fort et al. 2019) have cited the phenomenon of the F-Principle, but they only cite Rahaman et al. (2018) and miss other studies of Xu et al. (2018, 2019), which independently found this phenomenon at the same time.

3 The future of the F-Principle

Communities, including computer science, mathematics, statistics, optimization etc., have realized the increasing importance of the understanding of deep learning provided by the F-Principle, which is exhibited in the empirical study, the theoretical study and the application of deep learning. In each aspect, the F-Principle reserves many cutting-edge research problems in the coming future. Several examples are followed. **Empirical study.** F-Principle can be used a fundamental phenomenon to study the training trajectory and performance of deep learning. **Theoretical study.** Linear F-Principle
(LFP) in the kernel regime provides an opportunity that scientists from mathematics, statistics and optimization can use various tools to study the differential equation model and the optimization model of the LFP. Also, how to use the F-Principle to provide similar quantitative understanding of the deep learning in non-kernel regime is also an important topic. **Application.** The deep learning in applications often heavily depends on the experiences and chances, which suffers the criticism of “black box”. The F-Principle provides a deep understanding of the low-frequency priority of the deep learning. How to utilize this understanding in applications would significantly enhance the credibility of deep learning.

In addition, F-Principle as a training principle should be one of characteristics of deep learning. F-Principle provides an example of how to search, quantify and utilize other characteristics of deep learning.

Finally, if you have found any research or idea related to the F-Principle, please contact me.

**References**


URL: http://arxiv.org/abs/1905.10264

URL: http://arxiv.org/abs/1905.07777