A Fourier-Based Approach to Generalization and Optimization in Deep Learning

Farzan Farnia, Jesse M. Zhang, and David N. Tse, *Fellow, IEEE*

***Abstract*—The success of deep neural networks stems from their ability to generalize well on real data; however, Zhang *et al.* have observed that neural networks can easily overfit randomly-generated labels. This observation highlights the following question: why do gradient methods succeed in finding generalizable solutions for neural networks while there exist solutions with poor generalization behavior? In this work, we use a Fourier-based approach to study the generalization properties of gradient-based methods over 2-layer neural networks with band-limited activation functions. Our results indicate that in such settings if the underlying distribution of data enjoys nice Fourier properties including bandlimitedness and bounded Fourier norm, then the gradient descent method can converge to local minima with nice generalization behavior. We also establish a Fourier-based generalization error bound for band-limited function spaces, applicable to 2-layer neural networks with general activation functions. This generalization bound motivates a grouped version of path norms for measuring the complexity of 2-layer neural networks with ReLU-type activation functions. We empirically demonstrate that regularization of the group path norms results in neural network solutions that can fit true labels without losing test accuracy while not overfitting random labels.**

***Index Terms*—Deep learning, Fourier analysis, generalization bounds, norm-based regularization.**

I. Introduction

DEEP neural networks (DNNs) have achieved state-of-the-art performance on a wide array of tasks [2]. A given DNN architecture represents a highly expressive space of functions. However, numerous empirical results have shown that a simple stochastic gradient descent (SGD) learner can efficiently search over this complex space to find a solution that achieves high performance on both training and test data. Despite many successful applications of DNNs to practical tasks such as computer vision [3], natural language processing [4], and speech recognition [5], an adequate understanding of the factors driving DNNs' generalization behavior is still lacking.

Addressing generalization for DNNs is hard for two reasons: 1) Empirical risk minimization for neural networks is a non-convex optimization problem with possibly many local minima, and 2) Two different local minima with the same training performance can achieve significantly different performance on test data. For these reasons, the optimization method used for training a DNN plays an important role in the generalizability of the local minima found. Gradient methods have been shown to apply an implicit regularization, resulting in a better generalization performance [1], [6]. Furthermore, the performance of gradient methods can be improved upon by incorporating the geometry of the DNN architecture [7].

However, a gradient-based optimization method is not sufficient for guaranteeing good generalization performance. Zhang *et al.* [1] empirically demonstrate that a neural network trained by SGD can easily overfit random labels on the CIFAR-10 [8] data. Yet, the same neural network fitted by the same optimization algorithm achieves good generalization performance for the original CIFAR-10 labels. This observation challenges the ability of traditional generalization error bounds to explain why SGD learns generalizable hypotheses over a highly-expressive neural network space.

To shed light on this phenomenon, several recent works have developed generalization bounds and complexity measures for neural networks which can distinguish the local minima found for true and random labels. Reference [9] proves a margin-based generalization bound and shows how the classification margin values of a spectrally-normalized DNN help distinguish the DNN solutions fitting true and random labels. Reference [6] explores different complexity scores for DNNs and how they behave differently for true and random labels. The complexity measures investigated in these works help distinguish generalizable from poorly-generalizable local minima. They do not explain, however, why gradient methods converge to generalizable local minima when there exist poorly-generalizable local minima which can also perfectly fit the training examples.

To approach this question, one needs to understand the key properties of CIFAR-10's original labeling which differentiates its real labels from random labels and how those properties are exploited by gradient methods to gain good generalization performance. In this work, we approach this problem in the Fourier domain where a non-random labeling scheme behaves differently from a random labeling. While signals recoverable from few measurements possess nice spectral properties such

Manuscript received October 15, 2019; accepted March 8, 2020. Date of publication March 26, 2020; date of current version June 8, 2020. This work was supported in part by the National Science Foundation (NSF) under Grant CCF-1563098, and in part by the Center for Science of Information, an NSF Science and Technology Center under Grant CCF-0939370. *(Corresponding author: Farzan Farnia.)*

Farzan Farnia was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: farnia@mit.edu).

Jesse M. Zhang was with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA. He is now with Beacons AI, San Francisco, CA 94103 USA (e-mail: jessemzhang@gmail.com).

David N. Tse is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: dntse@stanford.edu).

Digital Object Identifier 10.1109/JSAIT.2020.2983192

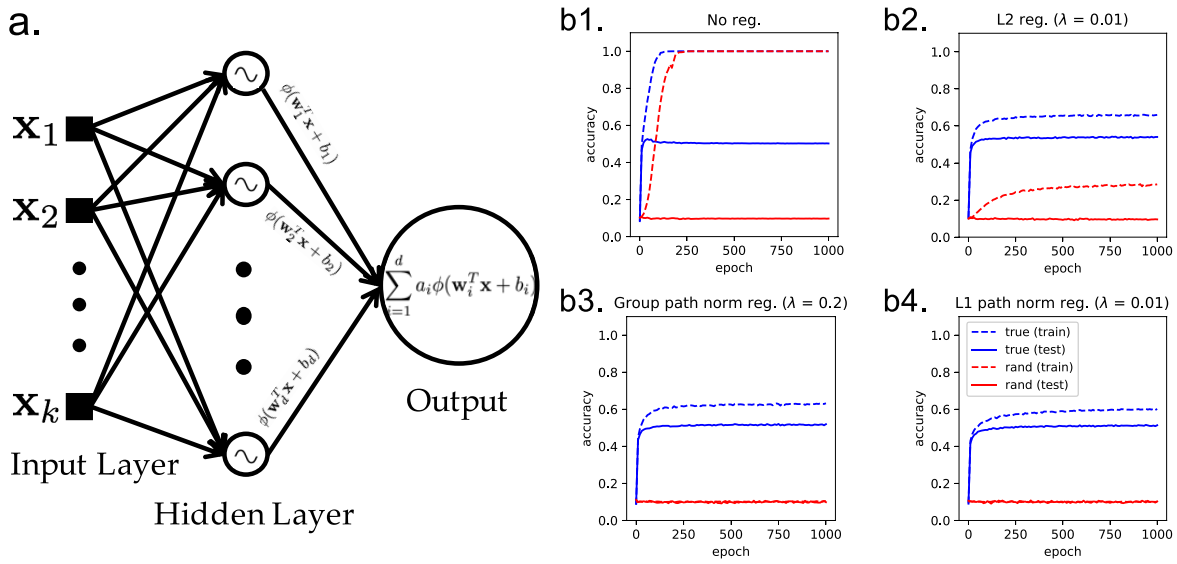


Fig. 1. (a) A 2-layer neural network with activation function ϕ, (b) Training and test accuracy on CIFAR10 with true and random labels on a 2-layer neural network with 512 ReLU hidden units, regularized with an additive penalty: (b1) no penalty, (b2) ℓ_2-norm, (b3) χ_2-group path norm, (b4) ℓ_1-path norm. The χ_2-group path norm and ℓ_1-path norm were successful to close the generalization gap for both true and random labels.

as bandlimitedness, fully random stochastic processes are not band-limited and not recoverable from any finite number of measurements.

Applying Fourier analysis, we focus on characterizing spectral properties of an underlying distribution which can be exploited by gradient-based methods to converge to generalizable local minima. We address this problem for 2-layer neural networks (see Figure 1a) with sinusoidal activation functions, where we show that if the underlying labeling scheme has limited bandwidth and Fourier ℓ_1-norm (i.e., "nice" Fourier properties), a gradient descent optimizer can potentially result in good generalization performance. To show this result, we first develop a Fourier-based generalization bound for 2-layer neural networks using the bandwidth and Fourier ℓ_1-norm of the function space. Next, we prove that the local minima found by the gradient descent method over a 2-layer neural network with sinusoidal activation have bandwidth and Fourier ℓ_1-norm bounded in terms of the Fourier properties of the underlying labeling scheme.

To develop the Fourier-based generalization analysis, we show generalization error bounds applicable to 2-layer neural networks with general activation functions. For band-limited activation functions with bounded Fourier ℓ_1-norm, such as sinusoidal or Gaussian non-linearity, the generalization bound is tighter than the standard generalization bounds based on the activation function's Lipschitz constant. For ReLU-type activation functions, the proposed generalization bound results in a grouped version of path norm functions introduced in [7]. We call the new capacity norm *group path norm* and leverage these norm functions to regularize 2-layer neural networks. Our numerical results suggest that the generalization gap can be effectively tightened by regularizing the group path norm. Figure 1b demonstrates how the same path-norm penalty help close the generalization gap for both true and random labels.

II. Preliminaries

A. Supervised Learning and Generalization Risk

Consider a set of n training samples $(\mathbf{x}_i, y_i)_{i=1}^n$ drawn i.i.d. according to the population distribution $P_{\mathbf{X},Y}$. Here $\mathbf{X}$ denotes the random vector of features and Y denotes the label variable. Using these training samples, the goal in supervised learning is to find a prediction rule f from a function set $\mathcal{F}$ which predicts Y for an unseen test data point $\mathbf{X}$. Therefore, considering a loss function ℓ the supervised learner aims to find $f^* \in \mathcal{F}$ minimizing the population risk averaged under the population distribution of data, i.e.,

$$\mathbb{E}_P\big[\ell(f(\mathbf{X}), Y)\big]. \tag{1}$$

However, the supervised learner's knowledge of the population distribution $P_{\mathbf{X},Y}$ is limited to the observed training samples. Consequently, a standard approach to supervised learning is to minimize the empirical risk, defined as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i), \tag{2}$$

and find f_n^{emp} minimizing the defined emprical risk function. Since the number of observed samples is limited, the empirical risk can be considerably different from the population risk. The generalization risk is defined as the difference between the population risk and empirical risk, i.e.,

$$\mathbb{E}\big[\ell(f(\mathbf{X}), Y)\big] - \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), y_i). \tag{3}$$

Bounding the generalization risk of learning over different function spaces is a subject of central interest in statistical learning theory.

B. Fourier Transform and Band-Limited Functions

Consider a real-valued function $f : \mathbb{R}^k \to \mathbb{R}$. The Fourier transform of this function, which we denote by $\widehat{f}$, is defined as

$$\widehat{f}(\boldsymbol{\xi}) = \int f(\mathbf{x}) \exp\big(-2\pi i \boldsymbol{\xi}^T \mathbf{x}\big)\, \mathrm{d}\mathbf{x}. \tag{4}$$

We will use the following Fourier transform examples several times throughout the paper:

- *Sinusoidal function:* $f(\mathbf{x}) = \exp(2\pi i \boldsymbol{\omega}^T \mathbf{x})$, then $\widehat{f}(\boldsymbol{\xi}) = \boldsymbol{\delta}(\boldsymbol{\xi} - \boldsymbol{\omega})$ where $\boldsymbol{\delta}$ denotes the Dirac delta function, implying
  - $f(\mathbf{x}) = \cos(2\pi \boldsymbol{\omega}^T \mathbf{x})$, then $\widehat{f}(\boldsymbol{\xi}) = 1/2\,[\boldsymbol{\delta}(\boldsymbol{\xi} + \boldsymbol{\omega}) + \boldsymbol{\delta}(\boldsymbol{\xi} - \boldsymbol{\omega})]$.
  - $f(\mathbf{x}) = \sin(2\pi \boldsymbol{\omega}^T \mathbf{x})$, then $\widehat{f}(\boldsymbol{\xi}) = i/2\,[\boldsymbol{\delta}(\boldsymbol{\xi} + \boldsymbol{\omega}) - \boldsymbol{\delta}(\boldsymbol{\xi} - \boldsymbol{\omega})]$.
- *Gaussian function:* $f(\mathbf{x}) = (\sqrt{2\pi}\sigma)^k \exp(-\|\mathbf{x}\|_2^2/2\sigma^2)$, then $\widehat{f}(\boldsymbol{\xi}) = \exp(-\sigma^2 \|\boldsymbol{\xi}\|_2^2/2)$.

We call a function f *B-bandlimited* if $\widehat{f}(\boldsymbol{\xi}) = 0$ for every $\|\boldsymbol{\xi}\|_2 > B$. The smallest B for which this property holds is called f's *bandwidth* which we denote by $\mathcal{B}(f)$. We also use $\|\widehat{f}\|_1$ to denote the ℓ_1-norm of f's Fourier transform, i.e.,

$$\|\widehat{f}\|_1 = \int \big|\widehat{f}(\boldsymbol{\xi})\big|\, \mathrm{d}\boldsymbol{\xi} \tag{5}$$

which we call the *Fourier ℓ_1-norm* of f. Fourier ℓ_1-norm can be interpreted as the absolute volume under f's Fourier transform, and provides an approximate measure of $\widehat{f}$'s sparsity. Note that Fourier ℓ_1-norm is both scale and shift invariant, i.e., if we define $g(\mathbf{x}) = f(\mathbf{W}\mathbf{x} + \mathbf{b})$ for a real-valued f and $\mathbf{W} \in \mathbb{R}^{r \times k}$ and $\mathbf{b} \in \mathbb{R}^r$ with $r \le k$, then $\|\widehat{g}\|_1 = \|\widehat{f}\|_1$.

Some other useful properties of Fourier transform are as follows:

- *Synthesis:* $f(\mathbf{x}) = \int \widehat{f}(\boldsymbol{\xi}) \exp(2\pi i \boldsymbol{\xi}^T \mathbf{x})\, \mathrm{d}\boldsymbol{\xi}$, which also implies $\|\widehat{f}\|_1 = f(0)$ if $\widehat{f}$ is real and non-negative.
- *Shift:* $\widehat{f_{\mathbf{b}}}(\boldsymbol{\xi}) = \exp(2\pi i \mathbf{b}^T \boldsymbol{\xi}) \widehat{f}(\boldsymbol{\xi})$ where $f_{\mathbf{b}}(\mathbf{x}) \coloneqq f(\mathbf{x} - \mathbf{b})$, which implies $\|\widehat{f_{\mathbf{b}}}\|_1 = \|\widehat{f}\|_1$ and $\mathcal{B}(f_{\mathbf{b}}) = \mathcal{B}(f)$.
- *Scale:* $\widehat{f_{\mathbf{W}}}(\boldsymbol{\xi}) = (1/|\det(\mathbf{W})|) \widehat{f}(\mathbf{W}^{-T}\boldsymbol{\xi})$ where $f_{\mathbf{W}}(\mathbf{x}) \coloneqq f(\mathbf{W}\mathbf{x})$, implying $\|\widehat{f_{\mathbf{W}}}\|_1 = \|\widehat{f}\|_1$ and $\mathcal{B}(f_{\mathbf{W}}) \le \|W\|_2 \mathcal{B}(f)$ with $\|W\|_2$ denoting the spectral norm of $\mathbf{W}$.
- *Isometry:* $\int f(\mathbf{x}) \overline{g(\mathbf{x})}\, \mathrm{d}\mathbf{x} = \int \widehat{f}(\boldsymbol{\xi}) \overline{\widehat{g}(\boldsymbol{\xi})}\, \mathrm{d}\boldsymbol{\xi}$ where $\bar{z}$ denotes the complex conjugate of z.
- *Convolution:* $\widehat{fg} = \widehat{f} \star \widehat{g}$ where $\star$ denotes the convolution operator, i.e., $\widehat{f} \star \widehat{g}(\boldsymbol{\xi}) \coloneqq \int \widehat{f}(\boldsymbol{\eta}) \widehat{g}(\boldsymbol{\xi} - \boldsymbol{\eta})\, \mathrm{d}\boldsymbol{\eta}$. Therefore, $\mathcal{B}(fg) \le \mathcal{B}(f) + \mathcal{B}(g)$ and $\|\widehat{fg}\|_1 \le \|\widehat{f}\|_1 \|\widehat{g}\|_1$.

We refer the readers to the Appendix for a summary of the utilized properties of Fourier series.

III. A Fourier-Based Generalization Error Bound

Consider a supervised learning task with n training samples $(\mathbf{x}_i, y_i)_{i=1}^n$ and function space $\mathcal{F}$. We are interested in uniform convergence bounds on the generalization risk. A standard approach to bound the generalization risk is based on the notion of Rademacher complexity. Given samples $(\mathbf{x}_i, y_i)_{i=1}^n$, the empirical Rademacher complexity of $\mathcal{F}$ is defined as

$$\mathcal{R}_n^{\mathrm{emp}}(\mathcal{F}) \coloneqq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \sigma_i f(\mathbf{x}_i)\right] \tag{6}$$

where σ_i's are i.i.d. random variables uniformly distributed over $\{-1, +1\}$. In fact, the Rademacher complexity of $\mathcal{F}$ measures how well $\mathcal{F}$ can fit some random labels over input $\mathbf{x}_i$'s. The following result shows how to bound the generalization risk over $\mathcal{F}$ through its Rademacher complexity.

Theorem 1 [10]: Consider a ρ-Lipschitz loss function $\ell(f(\mathbf{x}), y)$ bounded as $|\ell(z, y)| \le c$. Then, for any $\delta > 0$, with probability at least $1 - \delta$

$$\forall f \in \mathcal{F}:\ \mathbb{E}\big[\ell(f(\mathbf{X}), Y)\big] - \frac{1}{n}\sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i) \le 2\rho \mathcal{R}_n^{\mathrm{emp}}(\mathcal{F}) + 4c\sqrt{\frac{2\log(4/\delta)}{n}}. \tag{7}$$

Since tight bounds are known for the Rademacher complexity of norm-bounded linear functions [11], Theorem 1 can be applied to bound the generalization risk of learning over norm-bounded linear functions. To extend the application of Theorem 1 to the Fourier space, we provide a Rademacher complexity bound for band-limited functions with bounded Fourier ℓ_1-norm. We apply the following Rademacher complexity bound to bound generalization risk for 2-layer neural networks in Section IV, and subsequently to analyze the performance of gradient-based methods with sinusoidal activation functions in Section V.

Theorem 2: Consider function space $\mathcal{F} = \{f : \mathbb{R}^k \to \mathbb{R} \text{ s.t. } \mathcal{B}(f) \le B,\ \|\widehat{f}\|_1 \le V\}$ of B-band-limited functions with V-bounded Fourier ℓ_1-norm. Then, the empirical Rademacher complexity for samples $(\mathbf{x}_i, y_i)_{i=1}^n$ is bounded as

$$\mathcal{R}_n^{\mathrm{emp}}(\mathcal{F}) \le V\sqrt{\frac{4k\log(64\, nB \max_i \|\mathbf{x}_i\|_2)}{n}}. \tag{8}$$

Proof: We defer the proof to the Appendix. ∎

Corollary 1: Assume that $\|\mathbf{X}\|_2 \le C$ holds with probability 1 and the loss function ℓ is ρ-Lipschitz. Then, for any $\delta > 0$ with probability at least $1 - \delta$ the following generalization bound holds for any B-band-limited function f with V-bounded Fourier ℓ_1-norm:

$$\mathbb{E}\big[\ell(f(\mathbf{X}), Y)\big] - \frac{1}{n}\sum_{i=1}^n \ell(f(\mathbf{x}_i), y_i) \le O\left(\rho V\sqrt{\frac{k\,\log(nBC/\delta)}{n}}\right). \tag{9}$$

Proof: The corollary is a direct result of applying the bound in Theorem 2 to Theorem 1. ∎

The above corollary bounds the generalization risk uniformly over all band-limited f's with $\mathcal{B}(f) \le B$ and $\|\widehat{f}\|_1 \le V$. Next, we apply the above result to 2-layer neural networks. We refer the readers to the Appendix for the application of the above results to shift-invariant kernel spaces.

IV. Application of Theorem 2 to 2-Layer Neural Networks

Consider a 2-layer neural network with d hidden units and activation function ϕ (Figure 1a). The neural network's output can be formulated as

$$f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x}) = \mathbf{a}^T \phi(\mathbf{W}\mathbf{x} + \mathbf{b}). \tag{10}$$

Assuming ϕ's bandwidth and Fourier ℓ_1-norm are bounded, the following corollary applies Theorem 2 to bound the generalization risk over the 2-layer neural network. Note that we use $\|\mathbf{W}\|_{2,\infty}$ to denote $\mathbf{W}$'s $L_{2,\infty}$ group-norm defined as the maximum L_2-norm $\|\mathbf{w}_i\|_2$ among all $\mathbf{W}$'s rows.

Corollary 2: Let $\mathcal{F}_\phi = \{f(\mathbf{x}) = \mathbf{a}^T\phi(\mathbf{W}\mathbf{x}+\mathbf{b}) : \|\mathbf{W}\|_{2,\infty} \leq W,\ \|\mathbf{a}\|_1 \leq A\}$ be the class of 2-layer neural networks where $\mathcal{B}(\phi) = B$ and $\|\widehat{\phi}\|_1 = V$. Then, the empirical Rademacher complexity of $\mathcal{F}_\phi$ for samples $(\mathbf{x}_i, y_i)_{i=1}^n$ is bounded as follows

$$\mathcal{R}_n^{\text{emp}}(\mathcal{F}_\phi) \leq O\left(AV\sqrt{\frac{k\log(nBW\max\|\mathbf{x}_i\|_2)}{n}}\right). \tag{11}$$

Proof: We defer the proof to the Appendix. ∎

Note that for a band-limited activation function the above generalization bound is increasing logarithmically fast in the norm $\|\mathbf{W}\|_{2,\infty}$. For example, this result can be applied to a sinusoidal activation $\phi(x) = \sin(2\pi x)$ with $\|\widehat{\phi}\|_1 = 1$, $\mathcal{B}(\phi) = 1$. On the other hand, the standard Rademacher complexity bounds in the literature use only the Lipschitz constant of the activation function, growing linearly with the norm $\|\mathbf{W}\|_{2,\infty}$ [10]. Therefore, by exploiting ϕ's Fourier properties, Corollary 2 results in a tighter generalization bound than standard bounds based on ϕ's Lipschitz constant.

However, an unbounded function such as the popular ReLU activation $\phi(x) = \max(x, 0)$ has an unbounded Fourier ℓ_1-norm. Therefore, Corollary 2 is not directly applicable to the unbounded functions. The following result leverages the assumption of having a bounded input $\mathbf{X}$ to extend Theorem 2 to ReLU-type activation functions. While the resulted generalization bound is growing faster than logarithmically with $\mathbf{W}$'s norm, it results in new capacity norms for 2-layer ReLU-based neural networks.

Theorem 3: Suppose that $\phi_\alpha(x) = \max\{x, \alpha x\}$ where $\alpha \in [0, 1]$ is an arbitrary constant. Consider a dual norm pair $(\|\cdot\|_p, \|\cdot\|_q)$ where $1 \leq p, q \leq \infty$ satisfy $1/p+1/q = 1$. Suppose that $\|\mathbf{x}_i\|_p \leq C$ holds for each $\mathbf{x}_i$. Then, for $\mathcal{F}_{\phi_\alpha} = \{f_{\mathbf{a},\mathbf{W}}(\mathbf{x}) = \mathbf{a}^T\phi_\alpha(\mathbf{W}\mathbf{x}) : \sum_{i=1}^d |a_i|\|\mathbf{w}_i\|_q \leq V\}$

$$\mathcal{R}_n^{\text{emp}}(\mathcal{F}_{\phi_\alpha}) \leq O\left(VC\sqrt{\frac{k\log(nkC)}{n}}\right). \tag{12}$$

Proof: We defer the proof to the Appendix. ∎

The above bound is based on the complexity score $\sum_{i=1}^d |a_i|\|\mathbf{w}_i\|_q$ for function $f_{\mathbf{a},\mathbf{W}}(\mathbf{x}) = \mathbf{a}^T\phi_\alpha(\mathbf{W}\mathbf{x})$. This complexity score can be interpreted as the $\ell_{1,q}$-group norm on the product of weights on each path from the input nodes to the output node of the 2-layer neural network,

$$\chi_q\big(f_{\mathbf{a},\mathbf{W}}\big) = \sum_{i=1}^d\left(\sum_{j=1}^k\big(|a_i||w_{i,j}|\big)^q\right)^{1/q}. \tag{13}$$

Note that $w_{i,j}$ denotes the weight between the jth node in the input layer to the ith node in the hidden layer. Similar to the path-norm function defined in [7], we call $\chi_q(f_{\mathbf{a},\mathbf{W}})$ the *group path norm*. For $q = 1$, χ_1-group path norm revisits the ℓ_1-path norm for 2-layer neural networks. We can further apply group path norms to regularize learning over 2-layer neural networks. In our numerical experiments, we test the performance of regularizing χ_2-group path norm and ℓ_1-path norm for controlling the generalization risk of learning over 2-layer neural networks.

V. Fourier Analysis of Gradient-Based Methods for 2-Layer Neural Networks With Sine Activation

In this section, we apply Fourier analysis to analyze the generalization performance of gradient methods learning over a 2-layer neural network with a sinusoidal activation. We aim to study the connection between the generalization behavior of local minima found by gradient-based methods and Fourier properties of the population distribution $P_{\mathbf{X},Y}$. In our discussion, we assume that label variable Y is a deterministic function $Y(\mathbf{x})$ of input $\mathbf{X}$, which we call the *labeling scheme*. In our analysis, we consider the squared-error loss function $\ell(y, y') = (y - y')^2$.

To analyze the generalization performance of gradient methods, we follow a similar approach to [12]'s by establishing generalization bounds for both the empirical risk functions and its first-order derivative. First, we show that the bandwidth and Fourier ℓ_1-norm for the local minima of the population risk can be bounded in terms of the bandwidth and Fourier ℓ_1-norm of $Y(\mathbf{x})$ and $P_{\mathbf{X}}(\mathbf{x})$. Next, we establish a generalization result for the gradient of the empirical risk, proving that the gradient of empirical risk also remains close to the gradient of population risk provided that $Y(\mathbf{x})$ has limited bandwidth and Fourier ℓ_1-norm. The two results together show that by assuming a labeling scheme with constrained bandwidth and Fourier ℓ_1-norm, the local minima found by a gradient descent optimization method will have good generalization behavior.

A. Population Risk With Sinusoidal Activation

Consider $f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x}) = \sum_{j=1}^d a_j\sin(2\pi\mathbf{w}_j^T\mathbf{x} + b_j)$ as a 2-layer neural network with d sinusoidal hidden units. Given the labeling scheme $Y(\mathbf{x})$ the population risk will be

$$\begin{aligned}&\mathbb{E}_{P_{\mathbf{X}}}\big[\ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x})\ Y(\mathbf{x})\big)\big]\\ &\quad= \mathbb{E}_{P_{\mathbf{X}}}\left[\left(Y(\mathbf{x}) - \sum_{j=1}^d a_j\sin(2\pi\mathbf{w}_j^T\mathbf{x} + b_j)\right)^2\right],\end{aligned} \tag{14}$$

where the expectation is according to the population density function $P_{\mathbf{X}}(\mathbf{x})$.

Lemma 1: Consider the population risk in (14). Assume $\mathbf{w}_j$ satisfies $\forall\, i \neq j :\ \min\{\|\mathbf{w}_i - \mathbf{w}_j\|_2, \|\mathbf{w}_i + \mathbf{w}_j\|_2\} > \mathcal{B}(P_{\mathbf{X}})$. Then, if $(\mathbf{a}, \mathbf{W}, \mathbf{b})$ is assumed to be a local minimum of the population risk,

$$|a_j| \leq 2\left|\widehat{Y} \star \widehat{P_{\mathbf{X}}}(\mathbf{w}_j)\right|. \tag{15}$$

Proof: We defer the proof to the Appendix. ∎

Lemma 1 shows that if the component $a_j\sin(2\pi\mathbf{w}_j^T\mathbf{x})$ is isolated from the other components at a local minimum, by which we mean there is no component $a_i\sin(2\pi\mathbf{w}_i^T\mathbf{x})$ with $\min\{\|\mathbf{w}_i - \mathbf{w}_j\|_2, \|\mathbf{w}_i + \mathbf{w}_j\|_2\}$ less than $P_{\mathbf{X}}$'s bandwidth, then a_j's value in that local minimum will be bounded based on the population distribution's bandwidth. This result leads to the following theorem.

Theorem 4: Consider the minimization problem of the population risk (14). If a local minimum $(\mathbf{a}^*, \mathbf{W}^*, \mathbf{b}^*)$ has isolated

components, i.e., for any $i \neq j$ we have $\min\{\|\mathbf{w}_i^* - \mathbf{w}_j^*\|_2, \|\mathbf{w}_i^* + \mathbf{w}_j^*\|_2\} > 2\,\mathcal{B}(P_{\mathbf{X}})$, then for the function $f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}$

- $\mathcal{B}(f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}) \leq \mathcal{B}(Y) + \mathcal{B}(P_{\mathbf{X}})$,
- $\|\widehat{f}_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}\|_1 \leq 2\,\|\widehat{Y}\|_1$.

Proof: We defer the proof to the Appendix. ∎

Theorem 4 implies that the bandwidth of the local minima of the population risk is less than the sum of bandwidths for Y and $P_{\mathbf{X}}$. Also, the Fourier ℓ_1-norm for the local minima of the population distribution is bounded by twice the Fourier ℓ_1-norm of Y.

Remark 1: In Theorem 4, the bandwidth of $P_{\mathbf{X}}$ is supposed to be smaller than half the smallest distance among $\mathbf{w}_i^*$'s. If we suppose that $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_{k\times k})$ has a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and scalar covariance matrix with standard deviation σ, then the result shows that if for any i, j we have $\min\{\|\mathbf{w}_i^* - \mathbf{w}_j^*\|_2, \|\mathbf{w}_i^* + \mathbf{w}_j^*\|_2\} > 2C/\sigma$ for some constant C, then

- $\mathcal{B}(f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}) \leq \mathcal{B}(Y) + O(\sqrt{k}/\sigma)$,
- $\|\widehat{f}_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}\|_1 \leq 2(1 + d\exp(-C^2/2))\,\|\widehat{Y}\|_1$.

Proof: We defer the proof to the Appendix. ∎

B. Fourier Analysis Applied to the Empirical Risk

Theorem 4 characterizes the Fourier properties of the local minima in the population risk. However, our main goal is to study the generalization properties of the local minima in the empirical risk for observed training samples $(\mathbf{x}_i, Y(\mathbf{x}_i))_{i=1}^n$, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x}_i)\ Y(\mathbf{x}_i)\big) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \sum_{j=1}^{d} a_j \sin(2\pi\mathbf{w}_j^T\mathbf{x}_i + b_j)\right)^2. \quad (16)$$

To address this question, it can be seen that the bandwidth and Fourier ℓ_1-norm of the loss's derivative with respect to each a_j are bounded in terms of the bandwidth and Fourier ℓ_1-norm of $Y(\mathbf{x})$ as

$$\left\| \nabla_{a_j} \ell\Big(f_{\widehat{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x})}, Y(\mathbf{x})\right\|_1 \leq \|\widehat{Y}\|_1 + \|\mathbf{a}\|_1, \quad (17)$$

$$\mathcal{B}\big(\nabla_{a_j}\ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x}), Y(\mathbf{x})\big)\big) \leq \mathcal{B}(Y) + 2\|\mathbf{W}\|_{2,\infty}. \quad (18)$$

We apply Corollary 1 to show that not only the empirical risk uniformly converges to the population risk, but also the gradient of the empirical risk remains close to the gradient of the population risk.

Corollary 3: Consider $f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x}) = \sum_{j=1}^{d} a_j \sin(\mathbf{w}_j^T\mathbf{x} + b_j)$ and squared error loss ℓ. Then, assuming that $\|\mathbf{X}\|_2 \leq C$ holds with probability 1, for any $\delta > 0$ with probability at least $1-\delta$ we have

$$\forall\, j, \mathbf{a}, \mathbf{W}, \mathbf{b} \text{ s.t. } \|\mathbf{a}\|_1 + \|\widehat{Y}\|_1 \leq V,\ 2\|\mathbf{W}\|_{2,\infty} + \mathcal{B}(Y) \leq B:$$
$$\left|\mathbb{E}\big[\big(\nabla_{a_j}\ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{X}), Y(\mathbf{X})\big)\big)\big] - \frac{1}{n}\sum_{i=1}^{n}\big[\big(\nabla_{a_j}\ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{x}_i), Y(\mathbf{x}_i)\big)\big)\big]\right| \leq O\left(V\sqrt{\frac{k\log(nBC/\delta)}{n}}\right).$$

Proof: The corollary is an immediate result of Corollary (1), (17), and (18). Note that the generalization bound holds with probability $1-\delta$ for the derivative with respect to all a_j's, since the bounds in (17) and (18) hold for every j. ∎

Note that to prove Theorem 4 we only need to analyze the risk function's derivative with respect to a_j's. Hence, generalization of the empirical risk's gradient with respect to a_j's, that is shown in the above corollary, is sufficient to apply an approximate version of Theorem 4 in Section VII-I to a local minimum $(\mathbf{a}^*, \mathbf{W}^*, \mathbf{b}^*)$ with isolated components and found by the gradient descent approach initialized at sufficiently small $\|\mathbf{a}\|_1$ and $\|\mathbf{W}\|_{2,\infty}$. Therefore, with probability at least $1-\delta$ the Fourier integral of $f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}$ outside the bandwidth ball with radius $\mathcal{B}(Y) + \mathcal{B}(P_{\mathbf{X}})$ is bounded by $O(dV\sqrt{\frac{k\log(nBC/\delta)}{n}})$, and also

$$\|\widehat{f}_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}\|_1 \leq 2\,\|\widehat{Y}\|_1 + O\left(dV\sqrt{\frac{k\log(nBC/\delta}{n}}\right).$$

Based on the above discussion, if a gradient descent method starts learning from $f_{\mathbf{a},\mathbf{W},\mathbf{b}}$ with small $\|\mathbf{a}\|_1$ and $\|\mathbf{W}\|_{2,\infty}$ and also we assume that the bandwidth and the Fourier ℓ_1-norm of $Y(\mathbf{x})$ are properly bounded, Theorem 4 combined with Corollary 1 will guarantee good generalization performance for the local minima found by the gradient descent method.

VI. Numerical Experiments

For all experiments described in this section, we implemented and trained the two-layer neural network described in Figure 1a using TensorFlow 1.3.0. We used SGD to train the model for 2000 epochs with an initial learning rate of 0.01. The learning rate decayed slightly each epoch at a rate of 0.95 every 390 epochs. We used $h = 512$ hidden units and a batch size of 128. When working with CIFAR10 data, we preprocessed the data as described in [1], resulting in each training sample having dimension $d = 2352$. Initial weights from the first layer were sampled from $\mathcal{N}(0, 0.01/d)$ and initial weights from the second layer were sampled from $\mathcal{N}(0, 0.01/h)$.

A. SGD Gradually Learns Higher Fourier ℓ_1-Norm, Bandwidth Hypotheses

We first numerically demonstrate that how Fourier ℓ_1-norm and bandwidth both increases during training via SGD. Motivated by the analysis from Section V, we use the squared-error as our loss function and sine as our activation function. Our samples consist of cats and airplanes from the CIFAR10 dataset with the labels mapped to -1 and 1. We use 5000 and 2000 samples from each category for training and test, respectively. We arbitrarily chose two of the ten classes to accommodate our choice of loss function. We evaluate the network's performance for both random and true labels.

Figure 2a shows that without regularization, SGD learns to perfectly fit both the true and random labels, which is consistent with the results from [1]. Additionally, the random labels are harder to learn, requiring more epochs before achieving a perfect fit. Figures 2b and 2c confirm that both Fourier ℓ_1-norm and bandwidth consistently increase with training,

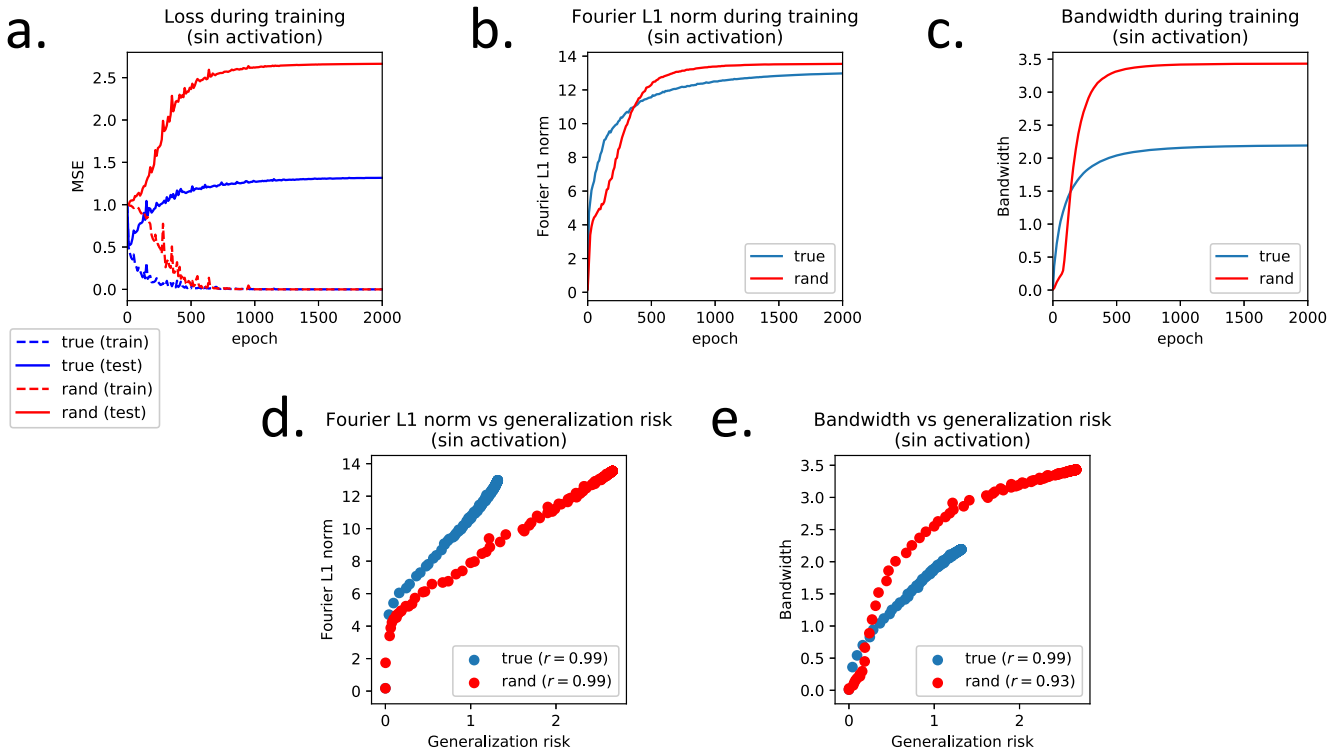


Fig. 2. Training an test performance on cat and airplane CIFAR10 images with true and random labels. Sine activation and mean-squared-error loss were used.

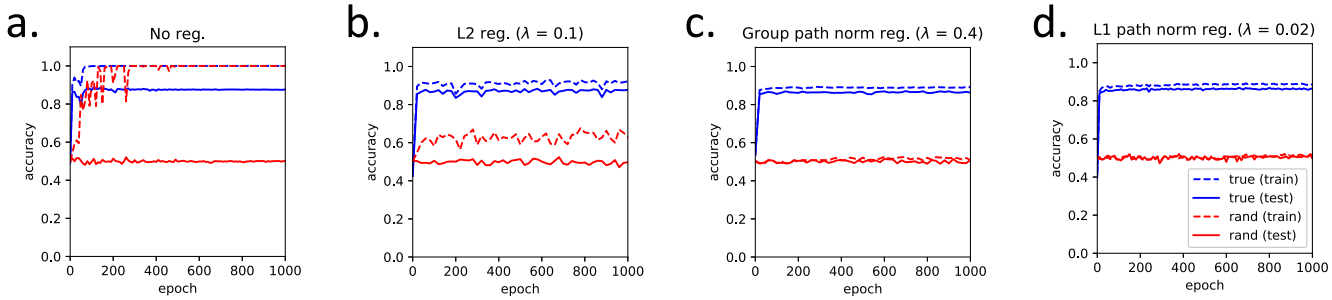


Fig. 3. Training and test performance on cat and airplane CIFAR10 images with true and random labels. ReLU activation and cross-entropy loss were used.

highlighting how SGD gradually finds more complex hypotheses in order to fit the data. Finally, we see in figures 2d and 2e how both Fourier ℓ_1-norm and bandwidth increase with generalization risk (the difference between test mean squared-error (MSE) and training MSE) with almost perfect correlation. This suggests that, as implied by the theory above, regularizing Fourier ℓ_1-norm and bandwidth could improve generalizability of the final learned model.

B. Group Path Norm Regularization for ReLU Activation

We regularize group path norm for ReLU activation as motivated by Theorem 3. Although χ_2-group path norm is not convex, it is differentiable and we can use it as an additive penalty and find a local minimum via SGD. Using the same experimental setup as from Section VI-A, we swap sine for ReLU and test the network's performance for both random and true labels.

Figure 3a confirms that, like before, the network can fit both true and random labels. The generalization gap, however, remains large for random labels. By regularizing the ℓ_2-norm of all the weights, we see that the generalization gap closes for both the true labels and the random labels without compromising test accuracy significantly (Figure 3b). This result is further improved when we use the χ_2-group path norm and ℓ_1-path norm (Figure 3c and 3d), demonstrating that direct regularization of Fourier ℓ_1-norm leads to better generalization.

We cross-validated the value of λ for each regularization technique, and we chose the λ that resulted in the smallest generalization gap with comparable validation performance. To fairly compare different regularization strategies, we tested five lambda values for each strategy and then reported the performance on the test set for the lambda value that resulted in the best performance on the validation set.

We repeated the experiment using all 50000 CIFAR10 training samples (and 10000 test samples). We included all 10 classes and switched to cross-entropy loss. The results are shown in Figure 1b. Again, we see that while all regularization techniques give similar test performance, the generalization gap is closed significantly for the χ_2-group path norm and ℓ_1-path norm.

VII. Related Work

Generalization has been a topic of central interest in statistical learning theory [13]. Generalization error bounds can be derived using different tools such as the stability of learning algorithms [14] and complexity measures of a function space including VC-dimension [15] and Rademacher complexity [10]. Reference [16] develops a stability-based generalization result for SGD as the learning algorithm, which holds for non-convex loss functions. For kernel methods, generalization bounds can be shown by bounding the Rademacher complexity of the kernel space [17], [18]. Moreover, the role of overparameterization in the generalization properties of a learned over-parameterized function has been recently studied in the context of linear regression [19], [20], [21], [22], [23]. The goal pursued by these works is to demonstrate that overparameterization can result in harmless interpolation in the linear regression context.

We note that Fourier analysis has provided a powerful framework for analyzing neural networks. Reference [24] uses a Fourier-based approach to prove the universal approximation theorem for 2-layer neural networks. Reference [25] applies Fourier analysis to extend Barron's result to a general feed-forward neural network. Also, our Fourier-based approach to analyze SGD's performance for 2-layer neural networks follows the same principles as the analysis performed in [26] to prove the hardness of fitting periodic labeling schemes via gradient-based methods. We should note that in this work we use only periodic activation functions and not periodic labeling schemes. Therefore, the hardness result shown in [26] does not affect our numerical experiments.

In general, theoretical studies of deep neural networks can be categorized into three general categories: 1) Approximation: Neural networks have been proven to be powerful in expressing rich spaces of functions [27] and in general deeper networks need fewer neurons to express the same class of functions [28], [29]. Recently, [30], [31] build an approximation analysis based on spline theory for deep neural networks. 2) Generalization: Tight bounds have been shown on the VC-dimesnion of feed-forward neural networks [32], [33]. Also, norm-based Rademacher complexity bounds have been proved at [10], [34]. Sharpness of local minima and its connection to their generalizibility have been the focus of several recent works [35], [36]. Reference [37] introduces a compression approach to further improve the margin-based bounds presented by [9], [38]. Also, [39] develops stronger generalization bounds for over-parameterized 2-layer neural networks. 3) Optimization: theoretical studies have shown both positive [40] and negative [41] results about the performance of gradient-based methods in training neural networks. Reference [42] further connects the learning problem over convolutional neural networks (CNNs) to a graphical model-based maximum likelihood problem, providing a probabilistic framework for deep learning.

Appendix

A. Fourier Series and Periodic Functions

Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous periodic function with period T, i.e., $f(x) = f(x+T)$. The Fourier series provides a sinusoidal basis to express f as

$$f(x) = \sum_{n=-\infty}^{+\infty} a_n \exp(2\pi i n x/T),$$

where

$$\forall n\text{: } a_n = \frac{1}{T}\int_{-T/2}^{T/2} f(x)\exp(-2\pi i n x/T)\,\mathrm{d}x. \tag{19}$$

The above result also characterizes f's Fourier transform as $\widehat{f}(\xi) = \sum_{n=-\infty}^{+\infty} a_n \delta(\xi - 2\pi i n/T)$. Therefore, $\|\widehat{f}\|_1 = \sum_{n=-\infty}^{+\infty} |a_n|$. If we further assume that f is continuous and piecewise smooth, not only f's Fourier series converges to f, but also the Fourier series for f's derivative can be derived by element-wise differentiation of f's Fourier series as

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \sum_{n=-\infty}^{+\infty} \frac{2\pi i n a_n}{T}\exp(2\pi i n x/T). \tag{20}$$

B. Application of Theorem 2 to Shift-Invariant Kernels

Kernel methods provide a popular approach to learn over non-linear function spaces. Here, we learn a prediction rule linear in a feature mapping $\psi(\mathbf{x})$ with ψ mapping $\mathbf{x}$ to a high-dimensional space specified by the kernel function. To efficiently learn the optimal prediction rule, the kernel trick can be applied for kernel κ defined as $\kappa(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^T\psi(\mathbf{x}')$. Given training samples $(\mathbf{x}_i, y_i)_{i=1}^n$, the solution f^* has the form $f(\mathbf{x}) = \sum_{i=1}^n a_i^* \kappa(\mathbf{x}_i, \mathbf{x})$ for a vector $\mathbf{a}^* \in \mathbb{R}^n$. Therefore, the kernel trick is to solve the risk minimization problem over $\mathcal{F}_\kappa = \{\sum_{i=1}^n a_i \kappa(\mathbf{x}_i, \mathbf{x}) : \mathbf{a} \in \mathbb{R}^n\}$.

Here, we apply Theorem 2 to bound the generalization risk of learning over the space of B-band-limited shift-invariant kernel functions. Note that a kernel function κ is called shift-invariant if $\kappa(\mathbf{x}, \mathbf{x}')$ is only a function of the difference $\mathbf{x} - \mathbf{x}'$. For example, the Gaussian kernel $\kappa_\sigma(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2/2\sigma^2)$ is shift-invariant. Bochner's theorem characterizes the Fourier transform of shift-invariant kernels.

Theorem 5 (Bochner's theorem, [43]): The shift-invariant $\kappa(\mathbf{x} - \mathbf{x}')$ is a kernel function if and only if the Fourier transform $\widehat{\kappa}$ is real and non-negative everywhere.

Applying Bochner's Theorem to Theorem 2, we obtain the following corollary.

Corollary 4: Consider function space $\mathcal{F}_B = \{f(\mathbf{x}) = \sum_{i=1}^n a_i \kappa(\mathbf{x}_i, \mathbf{x}) : \|\mathbf{a}\|_1 \le A,\ \kappa \in \Theta_B\}$ where Θ_B is the space of all B-band-limited shift-invariant kernels. We also assume every κ in Θ_B is normalized, i.e., $\forall \mathbf{x} : \kappa(\mathbf{x}, \mathbf{x}) = 1$. Then,

supposing that $\|\mathbf{x}_i\|_2 \leq C$ holds for every $\mathbf{x}_i \in \mathbb{R}^k$,

$$\mathcal{R}_n^{\text{emp}}(\mathcal{F}_B) \leq O\left(A\sqrt{\frac{k\log(nBC)}{n}}\right). \tag{21}$$

The above complexity bound is further applicable to the random Fourier features scheme proposed in [44] for approximating shift-invariant kernels. According to this scheme, i.i.d. sample $\boldsymbol{\omega}_i$'s are drawn according to the Fourier transform of the normalized shift-invariant kernel κ. Then, the scheme fits a linear model using the random Fourier features $z_i(\mathbf{x}) = \cos(\boldsymbol{\omega}_i^T\mathbf{x})$. Corollary 4 bounds the generalization risk of using random features for B-band-limited shift-invariant kernels.

C. Proof of Theorem 2

We use a high-dimensional grid in the Fourier domain to approximate the Fourier transform of a B-band-limited function. Consider the ball $\{\boldsymbol{\xi} : \|\boldsymbol{\xi}\|_2 \leq B\}$. Using the bounds on the covering number for ℓ_2-norm, for any $0 < \epsilon < B$ we can find a set of points $\{\boldsymbol{\xi}_j : 1 \leq j \leq (3B/\epsilon)^k\}$ such that for any $\boldsymbol{\xi}$ with $\|\boldsymbol{\xi}\|_2 \leq B$, there exists some $\boldsymbol{\xi}_j$ with $\|\boldsymbol{\xi} - \boldsymbol{\xi}_j\|_2 \leq \epsilon$.

Let $S_j = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\xi}_j\|_2 \leq \epsilon\}$ for each $1 \leq j \leq (3B/\epsilon)^k$. Note that $\{\boldsymbol{\xi} : \|\boldsymbol{\xi}\|_2 \leq B\} \subset \cup_j S_j$. We then define $S_j' = S_j \setminus \cup_{t=1}^{j-1} S_t$ to have a group of disjoint sets S_j' covering $\{\boldsymbol{\xi} : \|\boldsymbol{\xi}\|_2 \leq B\}$. Since any $f \in \mathcal{F}$ is assumed to be B-bandlimited , for $f \in \mathcal{F}$

$$\begin{aligned} f(\mathbf{x}) &= + \int \widehat{f}(\boldsymbol{\xi}) \exp\big(2\pi i \boldsymbol{\xi}^T\mathbf{x}\big)\, \mathrm{d}\boldsymbol{\xi} \\ &= \sum_{j=1}^{(3B/\epsilon)^k} \int_{\boldsymbol{\xi}\in S_j'} \widehat{f}(\boldsymbol{\xi}) \exp\big(2\pi i \boldsymbol{\xi}^T\mathbf{x}\big)\, \mathrm{d}\boldsymbol{\xi}. \end{aligned} \tag{22}$$

Then, for any $f \in \mathcal{F} = \{f : \mathcal{B}(f) \leq B, \|\widehat{f}\|_1 \leq V\}$ we have

$$\begin{aligned} &\left| f(\mathbf{x}) - \sum_{j=1}^{(3B/\epsilon)^k} \left[\exp\Big(2\pi i \boldsymbol{\xi}_j^T\mathbf{x}\Big) \int_{\boldsymbol{\xi}\in S_j'} \widehat{f}(\boldsymbol{\xi})\, \mathrm{d}\boldsymbol{\xi} \right] \right| \\ &\overset{(a)}{=} \left| \sum_{j=1}^{(3B/\epsilon)^k} \int_{\boldsymbol{\xi}\in S_j'} \widehat{f}(\boldsymbol{\xi}) \Big[\exp\big(2\pi i \boldsymbol{\xi}^T\mathbf{x}\big) - \exp\Big(2\pi i \boldsymbol{\xi}_j^T\mathbf{x}\Big) \Big]\, \mathrm{d}\boldsymbol{\xi} \right| \\ &\leq \sum_{j=1}^{(3B/\epsilon)^k} \int_{\boldsymbol{\xi}\in S_j'} \left| \widehat{f}(\boldsymbol{\xi}) \Big[\exp\big(2\pi i \boldsymbol{\xi}^T\mathbf{x}\big) - \exp\Big(2\pi i \boldsymbol{\xi}_j^T\mathbf{x}\Big) \Big] \right| \mathrm{d}\boldsymbol{\xi} \\ &\overset{(b)}{\leq} \sum_{j=1}^{(3B/\epsilon)^k} \int_{\boldsymbol{\xi}\in S_j'} \big|\widehat{f}(\boldsymbol{\xi})\big| 2\pi \|\mathbf{x}\|_2 \, \|\boldsymbol{\xi} - \boldsymbol{\xi}_j\|_2 \, \mathrm{d}\boldsymbol{\xi} \\ &\leq 2\pi \|\mathbf{x}\|_2 \sum_{j=1}^{(3B/\epsilon)^k} \int_{\boldsymbol{\xi}\in S_j'} \big|\widehat{f}(\boldsymbol{\xi})\big| \, \|\boldsymbol{\xi} - \boldsymbol{\xi}_j\|_2 \, \mathrm{d}\boldsymbol{\xi} \\ &\overset{(c)}{\leq} 2\pi\, \epsilon\, \|\mathbf{x}\|_2 \sum_{j=1}^{(3B/\epsilon)^k} \int_{\boldsymbol{\xi}\in S_j'} \big|\widehat{f}(\boldsymbol{\xi})\big|\, \mathrm{d}\boldsymbol{\xi} \\ &= 2\pi\, \epsilon\, \|\mathbf{x}\|_2 \int \big|\widehat{f}(\boldsymbol{\xi})\big|\, \mathrm{d}\boldsymbol{\xi} \\ &\leq 2\pi\, \epsilon\, \|\mathbf{x}\|_2 V. \end{aligned}$$

Here, (a) is a direct application of (22). (b) holds as $\exp(ibz) = \cos(bz) + i\sin(bz)$ is b-Lipschitz as a function of $z \in \mathbb{R}$ for any real number $b > 0$. (c) holds because according to our definitions $S_j' \subseteq S_j$ and $S_j = \{\boldsymbol{\xi} : \|\boldsymbol{\xi} - \boldsymbol{\xi}_j\|_2 \leq \epsilon\}$.

Therefore, the following function space $\mathcal{F}_\epsilon$ can approximate any $f \in \mathcal{F} = \{f : \mathcal{B}(f) \leq B, \|\widehat{f}\|_1 \leq V\}$ within $2\pi\epsilon CV$ accuracy for any $\|\mathbf{x}\|_2 \leq C$. Here $\mathbf{a}$ is, in general, a vector of complex numbers, and $\|\mathbf{a}\|_1 \coloneqq \sum_j |a_j|$ where $|z|$ denotes the absolute value of complex number z,

$$\mathcal{F}_\epsilon = \left\{ f(\mathbf{x}) = \sum_{j=1}^{(3B/\epsilon)^k} a_j \exp(2\pi i \boldsymbol{\xi}_j^T\mathbf{x}) : \|\mathbf{a}\|_1 \leq V \right\}. \tag{23}$$

Then, $\mathcal{F}_\epsilon$ is the space of ℓ_1-norm bounded linear functions in terms of the input vector $[\exp(2\pi i \boldsymbol{\xi}_j^T\mathbf{x})]_j$. Now, we can apply a well-known bound [13] on the Rademacher complexity of ℓ_1-norm bounded linear space $\mathcal{F}_{\text{lin},1} = \{f : \mathbb{R}^k \to \mathbb{R} \text{ s.t. } f(\mathbf{x}) = \mathbf{a}^T\mathbf{x},\ \|\mathbf{a}\|_1 \leq A\}$ as

$$\mathcal{R}_n^{\text{emp}}(\mathcal{F}_{\text{lin},1}) \leq A \max_i \|\mathbf{x}_i\|_\infty \sqrt{\frac{2\log(2k)}{n}}. \tag{24}$$

Applying the above bound, we can bound the Rademacher complexity of $\mathcal{F}_\epsilon$ as

$$\mathcal{R}_n^{\text{emp}}(\mathcal{F}_\epsilon) \leq V\sqrt{\frac{2k\log(6B/\epsilon)}{n}}. \tag{25}$$

Since for each $f \in \mathcal{F}$ there exists $\tilde{f} \in \mathcal{F}_\epsilon$ such that $\forall \|\mathbf{x}\|_2 \leq C : |f(\mathbf{x}) - \tilde{f}(\mathbf{x})| \leq 2\pi\epsilon CV$,

$$\begin{aligned} \mathcal{R}_n^{\text{emp}}(\mathcal{F}) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \sigma_i f(\mathbf{x}_i) \right] \\ &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\tilde{f}\in\mathcal{F}_\epsilon} \frac{1}{n}\sum_{i=1}^n \sigma_i \tilde{f}(\mathbf{x}_i) \right] + 2\pi\epsilon V \max_i \|\mathbf{x}_i\|_2 \\ &= \mathcal{R}_n^{\text{emp}}(\mathcal{F}_\epsilon) + 2\pi\epsilon V \max_i \|\mathbf{x}_i\|_2. \end{aligned} \tag{26}$$

Finally, combining (25) and (26) we obtain:

$$\forall \epsilon > 0 : \mathcal{R}_n^{\text{emp}}(\mathcal{F}) \leq V\sqrt{\frac{2k\log(6B/\epsilon)}{n}} + 2\pi\epsilon V \max_i \|\mathbf{x}_i\|_2. \tag{27}$$

If we choose the value $\epsilon = \frac{1}{2\pi n \max_i \|\mathbf{x}_i\|_2}$, then we get

$$\begin{aligned} \mathcal{R}_n^{\text{emp}}(\mathcal{F}) &\leq V\left(\sqrt{\frac{2k\log(12\pi nB\max_i \|\mathbf{x}_i\|_2)}{n}} + \frac{1}{n}\right) \\ &\leq V\sqrt{\frac{4k\log(12\pi nB\max_i \|\mathbf{x}_i\|_2) + 2/n}{n}} \\ &\leq V\sqrt{\frac{4k\log(64\, nB\max_i \|\mathbf{x}_i\|_2)}{n}}, \end{aligned} \tag{28}$$

where the last inequality follows from the fact that $1 \leq k, n$. Therefore, the proof is complete.

D. Proof of Corollary 2

First, we prove the following lemma.

Lemma 2: Given function $f : \mathbb{R}^k \to \mathbb{R}$ and matrix $\mathbf{W} \in \mathbb{R}^{k\times k}$, we define $g(\mathbf{x}) = f(\mathbf{W}\mathbf{x})$. Then,

- $\mathcal{B}(g) \le \|\mathbf{W}\|_2\, \mathcal{B}(f)$ with $\|\mathbf{W}\|_2$ denoting the spectral norm of $\mathbf{W}$,
- $\|\widehat{g}\|_1 = \|\widehat{f}\|_1$.

Proof: From the properties of the Fourier transform we know

$$\widehat{g}(\boldsymbol{\xi}) = \frac{1}{|\det(\mathbf{W})|}\widehat{f}\big(\mathbf{W}^{-T}\boldsymbol{\xi}\big). \tag{29}$$

Therefore, $\widehat{g}(\mathbf{W}^T\boldsymbol{\xi}') = \frac{1}{|\det(\mathbf{W})|}\widehat{f}(\boldsymbol{\xi}')$ and if $\|\boldsymbol{\xi}'\|_2 \le \mathcal{B}(f)$, then $\|\mathbf{W}^T\boldsymbol{\xi}'\|_2 \le \|\mathbf{W}\|_2\mathcal{B}(f)$ gives an upperbound on $\mathcal{B}(g)$. Also,

$$\begin{aligned}
\|\widehat{g}\|_1 &= \int |\widehat{g}(\boldsymbol{\xi})|\,\mathrm{d}\boldsymbol{\xi} \\
&= \int \frac{1}{|\det(\mathbf{W})|}\big|\widehat{f}\big(\mathbf{W}^{-T}\boldsymbol{\xi}\big)\big|\,\mathrm{d}\boldsymbol{\xi} \\
&= \frac{1}{|\det(\mathbf{W})|}\int \big|\widehat{f}\big(\mathbf{W}^{-T}\boldsymbol{\xi}\big)\big|\,\mathrm{d}\boldsymbol{\xi} \\
&= \frac{1}{|\det(\mathbf{W})|}\int \big|\widehat{f}\big(\boldsymbol{\xi}'\big)\big|\frac{1}{|\det(\mathbf{W}^{-T})|}\,\mathrm{d}\boldsymbol{\xi}' \\
&= \int \big|\widehat{f}(\boldsymbol{\xi}')\big|\,\mathrm{d}\boldsymbol{\xi}' \\
&= \|\widehat{f}\|_1.
\end{aligned} \tag{30}$$

It can be seen that this result remains valid even if $\mathbf{W}$ is not an invertible matrix, which will complete the proof for Corollary 2. However, we continue proving Corollary 2 without using this fact.

As shown in the above lemma, Fourier ℓ_1-norm and bandwidth are invariant to an orthonormal transformation $\mathbf{W}$. Given $f_i(\mathbf{x}) = a_i\phi(\mathbf{w}_i^T\mathbf{x})$, we define $g_i(\mathbf{x}) = f_i(\mathbf{A}_i\mathbf{x})$ where A_i is an orthonormal matrix with $\mathbf{w}_i$ as an eigenvector. Note that $\|\widehat{f_i}\|_1 = \|\widehat{g_i}\|_1$ and $\mathcal{B}(f_i) = \mathcal{B}(g_i)$. However, $g_i(\mathbf{x})$ is a function of only one of the coordinates, which we can assume, without loss of generality, to be the first coordinate. Hence, $g_i(\mathbf{x}) = a_i\phi(\|\mathbf{w}_i\|_2 x_1)$ for the first coordinate x_1, implying $\widehat{g_i}(\boldsymbol{\xi}) = \frac{a_i}{\|\mathbf{w}_i\|_2}\widehat{\phi}(\frac{\xi_1}{\|\mathbf{w}_i\|_2}).\boldsymbol{\delta}_2(\xi_2)\ldots\boldsymbol{\delta}_k(\xi_k)$ where $\boldsymbol{\delta}_j$ is the Dirac delta function across the jth dimension. Hence, we can use the above lemma in the 1-dimensional case to show $\|\widehat{g_i}\|_1 = |a_i|\|\widehat{\phi}\|_1$ and $\mathcal{B}(g_i) = \|\mathbf{w}_i\|_2\mathcal{B}(\phi)$. As a result,

$$\|\widehat{f_i}\|_1 = |a_i|\|\widehat{\phi}\|_1, \quad \mathcal{B}(f_i) \le \|\mathbf{w}_i\|_2\mathcal{B}(\phi). \tag{31}$$

Hence, for $f(\mathbf{x}) = \mathbf{a}^T\phi(\mathbf{W}\mathbf{x}+\mathbf{b}) = \sum_{i=1}^d a_i\phi(\mathbf{w}_i^T\mathbf{x}+b_i)$ we have

$$\|\widehat{f}\|_1 \le \|\mathbf{a}\|_1\|\widehat{\phi}\|_1, \quad \mathcal{B}(f) \le \|\mathbf{W}\|_{2,\infty}\mathcal{B}(\phi). \tag{32}$$

The corollary is then a direct application of Theorem 2.

E. Proof of Theorem 3

Given a ReLU-type activation function $\phi_\alpha(z) = \max\{z, \alpha z\}$,

$$\phi_\alpha\big(\mathbf{w}^T\mathbf{x}\big) = \|\mathbf{w}\|_q\phi_\alpha\bigg((\frac{\mathbf{w}}{\|\mathbf{w}\|_q})^T\mathbf{x}\bigg). \tag{33}$$

Since $\|\frac{\mathbf{w}}{\|\mathbf{w}\|_q}\|_q = 1$, if $\|\mathbf{x}\|_p \le C$, then $|(\frac{\mathbf{w}}{\|\mathbf{w}\|_q})^T\mathbf{x}| \le C$ and hence the input to ϕ_α in the R.H.S. of (33) is always between $-C$ and C.

Suppose that function ψ_α satisfies $\psi_\alpha(z) = \phi_\alpha(z)$ for $z \in [-C, C]$. Then, based on the above discussion, we can bound the Rademacher complexity of $\mathcal{F}_{\phi_\alpha}$ by finding a bound on the Rademacher complexity of $\mathcal{F}_{\psi_\alpha} = \{f_{\mathbf{v},\mathbf{U}}(\mathbf{x}) = \mathbf{v}^T\psi_\alpha(\mathbf{U}\mathbf{x}) : \|\mathbf{v}\|_1 \le V,\ \forall i : \|\mathbf{u}_i\|_q = 1\}$.

To find a good candidate for ψ_α, we use a symmetrization trick to define

$$\psi_\alpha(z) = \begin{cases} -\alpha C & \text{if } z < -C, \\ \phi_\alpha(z) & \text{if } -C \le z < C, \\ \phi_\alpha(2C-z) & \text{if } C \le z < 3C, \\ -\alpha C & \text{if } 3C \le z. \end{cases} \tag{34}$$

Note that $\psi_\alpha(z) = (1-\alpha)C\,h(\frac{z-C}{C}) + 2\alpha C\,h(\frac{z-C}{2C}) - \alpha C$ where $h(z) = \max\{0, 1-|z|\}$. It can be seen that $\widehat{h}(\xi) = (\frac{\sin(\pi\xi)}{\pi\xi})^2$ which is real and positive everywhere. Therefore, $\|\widehat{h}\|_1 = h(0) = 1$ which means that $\|\widehat{\psi_\alpha}\|_1 \le C(1+2\alpha) \le 3C$.

Since $|\widehat{h}(\xi)| \le \frac{1}{\xi^2}$, we have $|\widehat{\psi_\alpha}(\xi)| \le \frac{1}{\xi^2}$. For $B > 0$, we let the B-filtered $\psi_{\alpha,B}$ be a function with the following Fourier transform:

$$\widehat{\psi_{\alpha,B}}(\xi) = \begin{cases} \widehat{\psi_\alpha}(\xi) & \text{if } |\xi| \le B \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

Then, since $|\widehat{\psi_\alpha}(\xi)| \le \frac{1}{\xi^2}$ we have

$$\forall\, z \in \mathbb{R}: \quad \big|\psi_\alpha(z) - \psi_{\alpha,B}(z)\big| \le \int_{|\xi|\ge B}\big|\widehat{\psi_\alpha}(\xi)\big|\,\mathrm{d}\xi \le \frac{2}{B}. \tag{36}$$

Thus, for any $B > 0$ the defined $\psi_{\alpha,B}$ approximates ϕ_α with a maximum error of $\frac{2}{B}$ uniformly over $[-C, C]$. $\psi_{\alpha,B}$ also satisfies $\|\widehat{\psi_{\alpha,B}}\|_1 \le 3C$ and $\mathcal{B}(\psi_{\alpha,B}) = B$. Applying Corollary 2, we get

$$\forall\, B: \mathcal{R}_n^{\text{emp}}\big(\mathcal{F}_{\phi_\alpha}\big) \le O\bigg(VC\sqrt{\frac{k\log(nB\max\|\mathbf{x}_i\|_2)}{n}}\bigg) + \frac{2}{B}.$$

Here we can bound $\max_i\|\mathbf{x}_i\|_2 \le \sqrt{k}\max_i\|\mathbf{x}_i\|_\infty \le \sqrt{k}C$, and choose $B = n$ to get

$$\mathcal{R}_n^{\text{emp}}\big(\mathcal{F}_{\phi_\alpha}\big) \le O\bigg(VC\sqrt{\frac{k\log(nkC)}{n}} + \frac{1}{n}\bigg), \tag{37}$$

which completes the proof.

F. Proof of Lemma 1

Note that

$$\begin{aligned}
&\nabla_{a_j}\mathbb{E}_{P_\mathbf{X}}\big[\ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{X}), Y(\mathbf{X})\big)\big] \\
&= \mathbb{E}_{P_\mathbf{X}}\big[\nabla_{a_j}\ell\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{X}), Y(\mathbf{X})\big)\big] \\
&= \mathbb{E}_{P_\mathbf{X}}\Big[\nabla_{a_j}\big(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{X}) - Y(\mathbf{X})\big)^2\Big] \\
&= \mathbb{E}_{P_\mathbf{X}}\Bigg[2\sin\Big(2\pi\mathbf{w}_j^T\mathbf{X} + b_j\Big) \\
&\qquad\qquad \times\bigg(\sum_{t=1}^d a_t\sin\big(2\pi\mathbf{w}_t^T\mathbf{X} + b_t\big) - Y(\mathbf{X})\bigg)\Bigg] \\
&= \mathbb{E}_{P_\mathbf{X}}\Big[2a_j\sin^2\Big(2\pi\mathbf{w}_j^T\mathbf{X} + b_j\Big)\Big] \\
&\quad - \mathbb{E}_{P_\mathbf{X}}\Big[2\sin\Big(2\pi\mathbf{w}_j^T\mathbf{X} + b_j\Big)Y(\mathbf{X})\Big]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}_{P_{\mathbf{X}}}\left[\sum_{t\neq j} 2a_t \sin\big(2\pi \mathbf{w}_t^T\mathbf{X}+b_t\big)\sin\Big(2\pi \mathbf{w}_j^T\mathbf{X}+b_j\Big)\right]\\
&= a_j - 2\big[\cos\big(b_j\big)\,\mathrm{Im}\big\{\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big\}\\
&\qquad + \sin\big(b_j\big)\,\mathrm{Re}\big\{\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big\}\big].
\end{aligned}
$$

To show the last equality, we use the isolatedness assumption for $\mathbf{w}_j$, i.e., $\forall t\neq j$: $\min\{\|\mathbf{w}_t-\mathbf{w}_j\|_2, \|\mathbf{w}_t+\mathbf{w}_j\|_2\} > \mathcal{B}(P_{\mathbf{X}})$, and also $\|\mathbf{w}_j\|_2 \geq \mathcal{B}(P_{\mathbf{X}})/2$. Then, for each t

$$
\begin{aligned}
&\mathbb{E}_{P_{\mathbf{X}}}\Big[2\sin\big(2\pi \mathbf{w}_t^T\mathbf{X}+b_t\big)\sin\Big(2\pi \mathbf{w}_j^T\mathbf{X}+b_j\Big)\Big]\\
&= 2\int P_{\mathbf{X}}(\mathbf{x})\sin\big(2\pi \mathbf{w}_t^T\mathbf{x}+b_t\big)\sin\Big(2\pi \mathbf{w}_j^T\mathbf{x}+b_j\Big)\,\mathrm{d}\mathbf{x}\\
&= \int P_{\mathbf{X}}(\mathbf{x})\Big[\cos\Big(2\pi\big(\mathbf{w}_t-\mathbf{w}_j\big)^T\mathbf{x}+b_t-b_j\Big)\\
&\qquad\qquad - \cos\Big(2\pi\big(\mathbf{w}_t+\mathbf{w}_j\big)^T\mathbf{x}+b_t+b_j\Big)\Big]\,\mathrm{d}\mathbf{x}\\
&= 0.5\exp\big(j\big(b_t-b_j\big)\big)\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t-\mathbf{w}_j\big)\\
&\quad + 0.5\exp\big(j\big(b_j-b_t\big)\big)\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j-\mathbf{w}_t\big)\\
&\quad - 0.5\exp\big(j\big(b_t+b_j\big)\big)\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t+\mathbf{w}_j\big)\\
&\quad - 0.5\exp\big(-j\big(b_t+b_j\big)\big)\widehat{P_{\mathbf{X}}}\big(-\mathbf{w}_t-\mathbf{w}_j\big)\\
&= \begin{cases} 0 & \text{if } t\neq j,\\ 1 & \text{if } t=j.\end{cases}
\end{aligned}
\tag{38}
$$

Also, by applying the convolution property of Fourier transform we can show

$$
\begin{aligned}
&\mathbb{E}_{P_{\mathbf{X}}}\Big[\sin\Big(2\pi \mathbf{w}_j^T\mathbf{X}+b_j\Big)Y(\mathbf{X})\Big]\\
&= \int P_{\mathbf{X}}(\mathbf{x})Y(\mathbf{x})\sin\Big(2\pi \mathbf{w}_j^T\mathbf{x}+b_j\Big)\,\mathrm{d}\mathbf{x}\\
&= \int (P_{\mathbf{X}}\times Y)(\mathbf{x})\\
&\quad\times\Big[\cos\big(b_j\big)\sin\Big(2\pi \mathbf{w}_j^T\mathbf{X}\Big)+\sin\big(b_j\big)\cos\Big(2\pi \mathbf{w}_j^T\mathbf{X}\Big)\Big]\,\mathrm{d}\mathbf{x}\\
&= \cos\big(b_j\big)\,\mathrm{Im}\big\{\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big\}+\sin\big(b_j\big)\,\mathrm{Re}\big\{\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big\}.
\end{aligned}
$$

Finally if $(\mathbf{a},\mathbf{W},\mathbf{b})$ is a local minimum for the population risk, for all t's we have $\nabla_{a_t}\mathbb{E}_{P_{\mathbf{X}}}[\ell(f_{\mathbf{a},\mathbf{W},\mathbf{b}}(\mathbf{X}),Y(\mathbf{X}))]=0$. Therefore, due to the isolatedness assumption of $\mathbf{w}_j$ we have

$$
\begin{aligned}
\frac{|a_j|}{2} &= \big|\cos\big(b_j\big)\,\mathrm{Im}\big\{\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big\}+\sin\big(b_j\big)\,\mathrm{Re}\big\{\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big\}\big|\\
&\leq \big|\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_j\big)\big|.
\end{aligned}
$$

G. Proof of Theorem 4

Since the isolatedness assumption holds for all j's, by Lemma 1,

$$
\forall j:\quad \big|a_j^*\big| \leq 2\Big|\widehat{Y}\star\widehat{P_{\mathbf{X}}}\Big(\mathbf{w}_j^*\Big)\Big|. \tag{39}
$$

If $\|\mathbf{w}_t^*\|_2 > \mathcal{B}(Y)+\mathcal{B}(P_{\mathbf{X}})$ holds for some t , (39) implies

$$
\big|a_t^*\big| \leq 2\big|\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*\big)\big| = 0. \tag{40}
$$

Hence, a_t^* will be 0, implying there will be no component in $f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}$ with $\|\mathbf{w}_t^*\|_2 > \mathcal{B}(Y)+\mathcal{B}(P_{\mathbf{X}})$. This discussion proves the first part of Theorem, i.e., $\mathcal{B}(f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}) \leq \mathcal{B}(Y)+\mathcal{B}(P_{\mathbf{X}})$.

To show the second part, note that

$$
\begin{aligned}
\|\widehat{f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}}\|_1 &= \|\mathbf{a}^*\|_1\\
&\overset{(a)}{\leq} 2\sum_{t=1}^{d}\big|\widehat{Y}\star\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*\big)\big|\\
&= 2\sum_{t=1}^{d}\left|\int \widehat{Y}(\boldsymbol{\xi})\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*-\boldsymbol{\xi}\big)\,\mathrm{d}\boldsymbol{\xi}\right|\\
&\leq 2\sum_{t=1}^{d}\int\big|\widehat{Y}(\boldsymbol{\xi})\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*-\boldsymbol{\xi}\big)\big|\,\mathrm{d}\boldsymbol{\xi}\\
&= 2\sum_{t=1}^{d}\int\big|\widehat{Y}(\boldsymbol{\xi})\big|\,\big|\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*-\boldsymbol{\xi}\big)\big|\,\mathrm{d}\boldsymbol{\xi}\\
&= 2\int\big|\widehat{Y}(\boldsymbol{\xi})\big|\left[\sum_{t=1}^{d}\big|\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*-\boldsymbol{\xi}\big)\big|\right]\mathrm{d}\boldsymbol{\xi}\\
&\overset{(b)}{\leq} 2\sum_{t=1}^{d}\int\big|\widehat{Y}(\boldsymbol{\xi})\big|\,\mathrm{d}\boldsymbol{\xi}\\
&= 2\|\widehat{Y}\|_1.
\end{aligned}
\tag{41}
$$

Here, (a) comes from Lemma 1. Also, since $\min\{\|\mathbf{w}_t^*-\mathbf{w}_r^*\|_2, \|\mathbf{w}_t^*+\mathbf{w}_r^*\|_2\} > 2\,\mathcal{B}(P_{\mathbf{X}})$ is assumed for any $t\neq r$, for any $\boldsymbol{\xi}$ at most one element in $[\widehat{P_{\mathbf{X}}}(\mathbf{w}_t^*-\boldsymbol{\xi})]_{t=1}^d$ can be nonzero. Because if both $\widehat{P_{\mathbf{X}}}(\mathbf{w}_t^*-\boldsymbol{\xi})$ and $\widehat{P_{\mathbf{X}}}(\mathbf{w}_r^*-\boldsymbol{\xi})$ are nonzero for $r\neq t$, then $\|\mathbf{w}_t^*-\boldsymbol{\xi}\| \leq \mathcal{B}(P_{\mathbf{X}})$ and also $\|\mathbf{w}_r^*-\boldsymbol{\xi}\| \leq \mathcal{B}(P_{\mathbf{X}})$ which results in $\|\mathbf{w}_t^*-\mathbf{w}_r^*\| \leq 2\mathcal{B}(P_{\mathbf{X}})$ which is a contradiction. Hence,

$$
\sum_{t=1}^{d}\big|\widehat{P_{\mathbf{X}}}\big(\mathbf{w}_t^*-\boldsymbol{\xi}\big)\big| \leq \max_{\boldsymbol{\xi}'}\big|\widehat{P_{\mathbf{X}}}\big(\boldsymbol{\xi}'\big)\big| \leq \int|P_{\mathbf{X}}(\mathbf{x})|\,\mathrm{d}\mathbf{x} = 1, \tag{42}
$$

which proves (b) and completes the proof.

H. Applying Theorem 4 to Multivariate Gaussian $\mathbf{X}$

Assume $\mathbf{X}\sim\mathcal{N}(\boldsymbol{\mu},\sigma^2\mathbf{I}_{k\times k})$ has a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and standard deviation σ. Then, the Fourier transform $\hat{P}_{\mathbf{X}}$ has a Gaussian shape with mean $\mathbf{0}$ and standard deviation $1/\sigma$. Hence, if for any i,j we have $\min\{\|\mathbf{w}_i^*-\mathbf{w}_j^*\|_2, \|\mathbf{w}_i^*+\mathbf{w}_j^*\|_2\} > 2C/\sigma$ for some constant C, the approximation error term which should be added to the upperbound in Equation (41) is $2\|\widehat{Y}\|_1 d\exp(-C^2/2)$. Also, given any $\epsilon>0$ the Fourier ℓ_1 norm outside the bandwidth $O(\sqrt{k}\log(1/\epsilon)/\sigma)$ is at most ϵ. Therefore, Theorem 4 implies

- $\mathcal{B}(f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}) \leq \mathcal{B}(Y)+O(\sqrt{k}/\sigma)$,
- $\|\widehat{f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}}\|_1 \leq 2(1+d\exp(-C^2/2))\,\|\widehat{Y}\|_1$.

I. Approximate Version of Theorem 4

Here we show an approximate version of Theorem 4 which applies to approximate population local minima.

Theorem 6: Consider minimizing the population risk (14). Consider an approximate local minimum $(\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*)$ where $|\nabla_{a_j}\mathbb{E}[\ell(f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}(\mathbf{X}),Y(\mathbf{X})))]| \leq \epsilon$ for all j's. If for any two different i,j we have $\min\{\|\mathbf{w}_i^*-\mathbf{w}_j^*\|_2, \|\mathbf{w}_i^*+\mathbf{w}_j^*\|_2\} > 2\,\mathcal{B}(P_{\mathbf{X}})$,

then the Fourier ℓ_1-norm of $f_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}$ outside the bandwidth $\mathcal{B}(Y)+\mathcal{B}(P_{\mathbf{X}})$ is bounded by $d\,\epsilon$ and

$$\|\widehat{f}_{\mathbf{a}^*,\mathbf{W}^*,\mathbf{b}^*}\|_1 \leq 2\,\|\widehat{Y}\|_1 + d\,\epsilon.$$

Proof: Since the isolated components condition holds, we can apply Lemma 1's proof to show under the above assumptions

$$\forall\, j: \quad |a_j^*| \leq 2\left|\hat{Y} \star \hat{P}_{\mathbf{X}}\left(\mathbf{w}_j^*\right)\right| + \epsilon.$$

Then, a simple modification of Theorem 4's proof according to the above inequality proves the above theorem. ■

REFERENCES

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," 2016. [Online]. Available: arXiv:1611.03530.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[5] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[6] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "Exploring generalization in deep learning," 2017. [Online]. Available: arXiv:1706.08947.

[7] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, "Path-SGD: Path-normalized optimization in deep neural networks," in *Proc. 28th Int. Conf. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2422–2430.

[8] A. Krizhevsky and G. Hinton, *Learning Multiple Layers of Features from Tiny Images*, Univ. Toronto, Toronto, ON, Canada, 2009.

[9] P. Bartlett, D. J. Foster, and M. Telgarsky, "Spectrally-normalized margin bounds for neural networks," 2017. [Online]. Available: arXiv:1706.08498.

[10] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, Nov. 2002.

[11] S. M. Kakade, K. Sridharan, and A. Tewari, "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 793–800.

[12] S. Mei, Y. Bai, and A. Montanari, "The landscape of empirical risk for non-convex losses," 2016. [Online]. Available: arXiv:1607.06534.

[13] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge Univ. Press, 2014.

[14] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, Mar. 2002.

[15] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer, 2013.

[16] M. Hardt, B. Recht, and Y. Singer, "Train faster, generalize better: Stability of stochastic gradient descent," 2015. [Online]. Available: arXiv:1509.01240.

[17] O. Bousquet and D. J. Herrmann, "On the complexity of learning the kernel matrix," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 415–422.

[18] C. Cortes, M. Mohri, and A. Rostamizadeh, "Generalization bounds for learning kernels," in *Proc. 27th Int. Conf. Mach. Learn. (ICML-10)*, 2010, pp. 247–254.

[19] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine learning practice and the bias-variance trade-off," 2018. [Online]. Available: arXiv:1812.11118.

[20] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," 2019. [Online]. Available: arXiv:1903.08560.

[21] V. Muthukumar, K. Vodrahalli, and A. Sahai, "Harmless interpolation of noisy data in regression," 2019. [Online]. Available: arXiv:1903.09139.

[22] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and double descent curve," 2019. [Online]. Available: arXiv:1908.05355.

[23] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," 2019. [Online]. Available: arXiv:1906.11300.

[24] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[25] H. Lee, R. Ge, A. Risteski, T. Ma, and S. Arora, "On the ability of neural nets to express distributions," 2017. [Online]. Available: arXiv:1702.07028.

[26] O. Shamir, "Distribution-specific hardness of learning neural networks," 2016. [Online]. Available: arXiv:1609.01037.

[27] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signals Syst. (MCSS)*, vol. 2, no. 4, pp. 303–314, 1989.

[28] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 907–940.

[29] S. Liang and R. Srikant, "Why deep neural networks for function approximation?" 2016. [Online]. Available: arXiv:1610.04161.

[30] R. Balestriero and R. Baraniuk, "A spline theory of deep learning," in *Proc. 35th Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 374–383.

[31] R. Balestriero and R. Baraniuk, "Mad max: Affine spline insights into deep learning," 2018. [Online]. Available: arXiv:1805.06576.

[32] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[33] N. Harvey, C. Liaw, and A. Mehrabian, "Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks," 2017. [Online]. Available: arXiv:1703.02930.

[34] B. Neyshabur, R. Tomioka, and N. Srebro, "Norm-based capacity control in neural networks," in *Proc. 28th Conf. Learn. Theory (COLT)*, 2015, pp. 1376–1401.

[35] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," 2016. [Online]. Available: arXiv:1609.04836.

[36] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio, "Sharp minima can generalize for deep nets," 2017. [Online]. Available: arXiv:1703.04933.

[37] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," 2018. [Online]. Available: arXiv:1802.05296.

[38] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, "A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks," 2017. [Online]. Available: arXiv:1707.09564.

[39] B. Neyshabur, Z. Li, S. Bhojanapalli, Y. LeCun, and N. Srebro, "Towards understanding the role of over-parametrization in generalization of neural networks," 2018. [Online]. Available: arXiv:1805.12076.

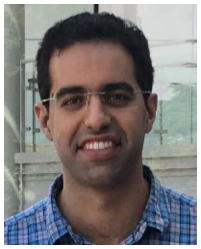
[40] A. Andoni, R. Panigrahy, G. Valiant, and L. Zhang, "Learning polynomials with neural networks," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, 2014, pp. 1908–1916.

[41] S. Shalev-Shwartz, O. Shamir, and S. Shammah, "Failures of gradient-based deep learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3067–3075.

[42] A. B. Patel, M. T. Nguyen, and R. Baraniuk, "A probabilistic framework for deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2558–2566.

[43] W. Rudin, *Fourier Analysis on Groups*. Hoboken, NJ, USA: Wiley, 2011.

[44] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 3, 2007, pp. 1177–1184.



Farzan Farnia received the first and second bachelor's degrees in electrical engineering and mathematics from the Sharif University of Technology in 2013, the master's degree in electrical engineering from Stanford University in 2015, and the Ph.D. degree from the Electrical Engineering Department, Stanford University, where he was a member of David Tse's Group. He is a Postdoctoral Scholar with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology. His research interests include information theory, statistical learning theory, and convex optimization. He was a recipient of the Stanford Graduate Fellowship (Sequoia Capital Fellowship) from 2013 to 2016 and the Numerical Technology Founders Prize as the Second Top Performer of the Stanford Electrical Engineering Ph.D. qualifying exams in 2014.

Jesse M. Zhang received the Ph.D. degree in electrical engineering from Stanford University, where he was a member of David Tse's Group, and his research revolved around analysis of single-cell RNA-Seq datasets and understanding deep learning models. He is the Co-Founder of the Beacons AI. From Winter 2016 to Spring 2018 quarters, he helped design and teach a new Stanford course on data science for high-throughput sequencing. He has also had the privilege of working for several organizations during his graduate and undergraduate careers, including Grail, Cellular Research, MC10, MIT Lincoln Laboratory, and Dana-Farber.

David N. Tse (Fellow, IEEE) received the B.A.Sc. degree in systems design engineering from the University of Waterloo in 1989, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology in 1991 and 1994, respectively.

From 1994 to 1995, he was a Postdoctoral Member of Technical Staff with AT&T Bell Laboratories. From 1995 to 2014, he was a faculty with the University of California at Berkeley (UC Berkeley). He is currently the Thomas Kailath and Guanghan Xu Professor with Stanford University. He has coauthored the text *Fundamentals of Wireless Communication* (with Pramod Viswanath), which has been used in over 60 institutions around the world. He is the Inventor of the proportional-fair scheduling algorithm used in all third and fourth-generation cellular systems, serving 2.7 billion subscribers around the world. He received the NSF CAREER Award in 1998, the Erlang Prize from the INFORMS Applied Probability Society in 2000, the Gilbreth Lectureship from the National Academy of Engineering in 2012, the IEEE Claude E. Shannon Award in 2017, and the IEEE Richard W. Hamming Medal in 2019. He received multiple best paper awards, including the Information Theory Society Paper Award in 2003, the IEEE Communications Society and Information Theory Society Joint Paper Awards in 2000, 2013, and 2015, the Signal Processing Society Best Paper Award in 2012, and the IEEE Communications Society Stephen O. Rice Prize in 2013. For his contributions to education, he received the Outstanding Teaching Award from the Department of Electrical Engineering and Computer Sciences, UC Berkeley in 2008 and the Frederick Emmons Terman Award from the American Society for Engineering Education in 2009. He was an Elected Member of the U.S. National Academy of Engineering in 2018.