Automatically Gating Multi-Frequency Patterns through Rectified Continuous Bernoulli Units with Theoretical Principles

Zheng-Fan Wu[1,2*], **Yi-Nan Feng**[1,2*] and **Hui Xue**[1,2†]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, 210096, China
[2]MOE Key Laboratory of Computer Network and Information Integration (Southeast University), China
{zfwu, yinanf, hxue}@seu.edu.cn

Abstract

Different nonlinearities are only suitable for responding to different frequency signals. The locally-responding $ReLU$ is incapable of modeling high-frequency features due to the spectral bias, whereas the globally-responding sinusoidal function is intractable to represent low-frequency concepts cheaply owing to the optimization dilemma. Moreover, nearly all the practical tasks are composed of complex multi-frequency patterns, whereas there is little prospect of designing or searching a heterogeneous network containing various types of neurons matching the frequencies, because of their exponentially-increasing combinatorial states. In this paper, our contributions are three-fold: 1) we propose a general Rectified Continuous Bernoulli ($Re\mathcal{CB}$) unit paired with an efficient variational Bayesian learning paradigm, to automatically detect/gate/represent different frequency responses; 2) our numerically-tight theoretical framework proves that $Re\mathcal{CB}$-based networks can achieve the optimal representation ability, which is $\mathcal{O}(m^{\eta/d^2})$ times better than that of popular neural networks, for a hidden dimension of m, an input dimension of d, and a Lipschitz constant of η; 3) we provide comprehensive empirical evidence showing that $Re\mathcal{CB}$-based networks can keenly learn multi-frequency patterns and push the state-of-the-art performance.

1 Introduction

Deep neural networks have led to a series of remarkable breakthroughs. In addition to the deep compositional architectures, their representational properties depend heavily on the activation functions. *Different kinds of nonlinearities provide different response characteristics, and are only suitable for disposing of different frequency signals.*

Most activation functions typically used nowadays, e.g., $Sigmoid$ and $ReLU$, are locally-responding (i.e., monotonic), mimicking the binary activation/inhibition of the Heaviside function. Locally-responding neurons only altering their states in a local range make sense from intuitive points of view: 1) They are more likely attracted to noticeable/generalizable/low-frequency features; 2) Monotonic nonlinearity substantially smoothes the fluctuation of gradient in optimization. This plays an important role in the success of training deep neural networks containing hundreds of millions of parameters.

However, it is not nitpicking that their representation ability is defective. [Tancik *et al.*, 2020; Jacot *et al.*, 2018] comprehensively showed that standard neural networks are poorly suited for learning high-frequency content, a phenomenon referred to spectral bias caused by a rapid frequency falloff. [Sitzmann *et al.*, 2020b; Bond-Taylor and Willcocks, 2020; Sitzmann *et al.*, 2020a] further proved that quasiconvex activations are incapable of modeling potential information contained in higher-order derivatives of natural signals.

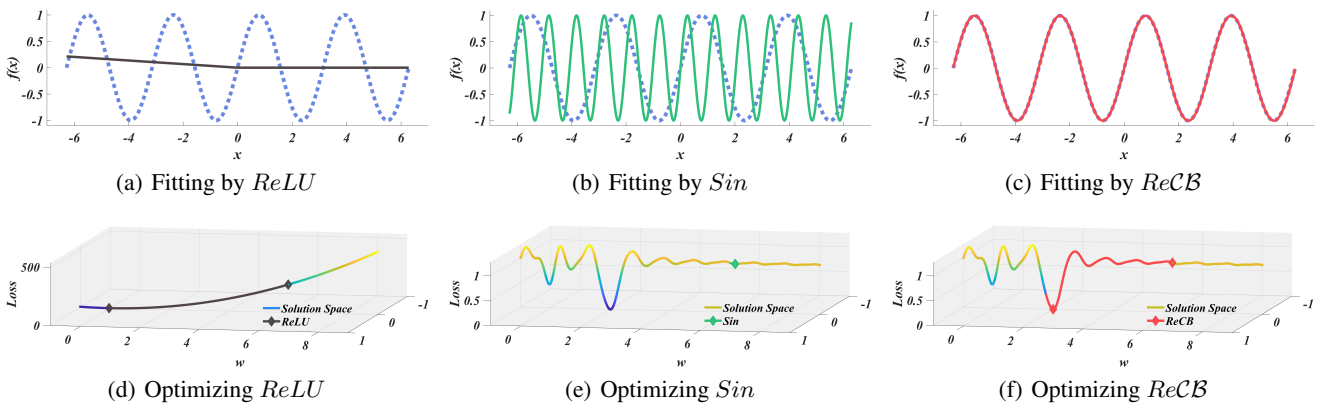
By comparison, globally-responding neurons activated by periodic nonlinearity, e.g., the sinusoidal function, are able to adjust their activation/inhibition states dynamically across the whole feature space. They are considered to be a competitive paradigm offering revolutionary benefits: 1) compactly characterizing complex high-frequency patterns; 2) precisely representing implicit high-order derivatives. They have the potential to reveal input-dependent and long-range characteristics [Xue *et al.*, 2019; Xue and Wu, 2020].

Nevertheless, they do have some drawbacks. As the correlation with the input increases, the state of nonlinearity will fluctuate between stronger activation and weaker inhibition, and thus is inappropriate to represent low-frequency concepts cheaply. Moreover, the periodic sinusoidal mapping has infinite *Vapnik-Chervonenkis (VC)* dimension leading to the optimization dilemma that the solution space has numerous poor and dense local minima [Parascandolo *et al.*, 2017]. Consequently, the improper use of globally-responding nonlinearities probably leads to extra complexity and severe risk.

Therefore, despite the relatively vacuous uniform approximation theory, *there is actually a gap between the representational properties of neural networks and the frequency characteristics of practical tasks.* On the one hand, different nonlinearities with relative merits are only suitable for processing signals with different frequencies. On the other hand, nearly all the practical tasks are composed of complex multi-frequency patterns.

Ideally, the nonlinear mappings in hidden layers can be regarded as the approximate decomposition of some implicit

*Equal Contribution
†Contact Author



(a) Fitting by $ReLU$ (b) Fitting by Sin (c) Fitting by $ReCB$

(d) Optimizing $ReLU$ (e) Optimizing Sin (f) Optimizing $ReCB$

Figure 1: The results on fitting $\sin(2x), x \in [-2\pi, 2\pi]$, and the corresponding optimization traces in solution space.

functions. Each computational element in networks should be compatible with the frequency attributes of these functions. But realistically, there is little prospect of designing or searching such a heterogeneous network containing different kinds of neurons matching the frequency of patterns. In addition to the critical issue in identifying the frequency of implicit functions, another fundamental bottleneck is the exponentially-increasing combinatorial nature of $2^{|\sigma|}$ possible states, where $|\sigma|$ is the total number of neurons.

Hence, in this paper, we particularly focus on improving the representational properties of neural networks to compactly and efficiently learn complex multi-frequency patterns with a minimal compromise in computational overhead. **Our main contributions are three-fold:**

• We propose a general Rectified Continuous Bernoulli ($ReCB$) unit paired with a differentiable variational Bayesian learning paradigm, to automatically detect/gate/represent different frequency responses from locally/globally-responding nonlinearities.

• We present a theoretical framework to analyze the representational properties of locally/globally-responding nonlinearities. Our main result proves that $ReCB$-based networks can achieve the optimal representation ability, which is $\mathcal{O}(m^{\eta/d^2})$ times better than that of popular $ReLU$-based neural networks, for a hidden dimension of m, an input dimension of d, and a Lipschitz constant of η.

• We provide comprehensive empirical evidence showing that our theoretical findings are consistent with the practical observations, and the novel $ReCB$-based networks can keenly characterize multi-frequency patterns. They impressively outperform the related state-of-the-art networks (e.g., $ReCB$-based ResNet-20 outperforms plain ResNet-110).

2 Related Work

On the one side, [Tancik *et al.*, 2020; Sitzmann *et al.*, 2020b; Bond-Taylor and Willcocks, 2020; Sitzmann *et al.*, 2020a] pointed out that ordinary neural networks activated by locally-responding nonlinearities are inappropriate to represent high-frequency features and higher-order derivatives. By comparison, [Mildenhall *et al.*, 2020; Zhong *et al.*, 2019; Xue *et al.*, 2019; Xue and Wu, 2020; Ramachandran *et al.*, 2017; Vaswani *et al.*, 2017; Xu *et al.*, 2019; Kazemi *et al.*, 2019] further constructed a series of competitive models based on the globally-responding sinusoidal function for a broad range of applications.

On the other side, these conceptually-attractive globally-responding nonlinearities are also well-known for being hard to train [Xue and Wu, 2020; Parascandolo *et al.*, 2017]. It is probably responsible for this optimization dilemma that the periodicity gives rise to numerous poor and dense local minima. The direction and norm of gradient oscillate continually during error backpropagation. These globally-responding networks may be prematurely stuck in local minima, and leave out more effective error feedback.

To understand these issues intuitively, we conduct a synthetic experiment on fitting $\sin(2x), x \in [-2\pi, 2\pi]$ with three models defined in Section 5.1: 1) the locally-responding $ReLU$; 2) the globally-responding Sin; 3) the multi-frequency gated $ReCB$. They have only one neuron with a scalar weight w initialized at $w = 6$. The fitting results and optimization traces are shown in Figure 1. Firstly, the solution space of $ReLU$ is smooth. But owing to the structural limitation, $ReLU$ can only capture the mean values. Secondly, in contrast, the solution space of Sin has more poor and dense local minima. Sin is prematurely stuck in terrible solution and fits the wrong frequency. Thirdly, according to the projection of $ReCB$ in the solution space of Sin, $ReCB$ finds the optimal solution $w = 2$ and perfectly reconstructs the original function by rectifying these local minima. Consequently, by accurately controlling the proportion of locally/globally-responding nonlinearities, we can alleviate the optimization difficulty on the premise of preserving the structural superiority.

3 Gating Multi-Frequency Patterns

3.1 Multi-Frequency Gated Unit

To characterize the response from locally/globally-responding nonlinearities for each neuron, a multi-frequency gated unit is defined by

$$\sigma_z(\cdot) = z \odot \tilde{\sigma}(\cdot) + (1 - z) \odot \bar{\sigma}(\cdot), \quad z \in \{0, 1\}. \tag{1}$$

$\bar{\sigma}(\cdot)$ is a locally-responding activation function, and $\tilde{\sigma}(\cdot)$ is a globally-responding counterpart. Without loss of generality, we make an innocuous stipulation that $\bar{\sigma}(\cdot)$ and $\tilde{\sigma}(\cdot)$ are

the quasiconvex $ReLU$ and the periodic sinusoidal function, respectively. Whether the responding preference of $\sigma_z(\cdot)$ is low-frequency or high-frequency is determined by the discrete value of the binary gate z. But it should be emphasized that the approach is naturally compatible with other activation functions without modification.

By controlling the proportion of different kinds of nonlinearities, we can dynamically adjust the response characteristics of mappings for complex multi-frequency patterns. But the practical optimization under this naïve gating scheme is computationally intractable because of the non-differentiability and the exponentially-increasing combinatorial nature of $2^{|\sigma_z|}$ possible states, where $|\sigma_z|$ is the total number of neurons.

3.2 Variational Bayesian Learning Paradigm

We propose a more efficient differentiable learning paradigm to jointly optimize the gates with original network parameters, utilizing variational Bayesian inference as theoretical basis.

Given some observed data $\mathcal{D}$, a group of random variables $\boldsymbol{z}$ gating the low/high-frequency responses, and a collection of activations $\boldsymbol{\sigma_z}$ regarded as random variables reparameterized by $\boldsymbol{z}$. According to Bayesian inference, a general learning problem can be defined as minimizing the negative Evidence Lower BOund (ELBO) $-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z})$.

$$-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z}) = -\int \log \frac{\mathbb{P}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z})}{q(\boldsymbol{\sigma_z}, \boldsymbol{z})} q(\boldsymbol{\sigma_z}, \boldsymbol{z}) d\boldsymbol{\sigma_z} d\boldsymbol{z}, \tag{2}$$

where q is the approximate posterior over $\boldsymbol{\sigma_z}$ and $\boldsymbol{z}$.

Furthermore, suppose p is a spike and slab prior over $\boldsymbol{\sigma_z}$ and $\boldsymbol{z}$. It is defined as a mixture of a delta spike at zero and a continuous distribution over the real line.

$$\begin{aligned} &p(\boldsymbol{z}) = \text{Bernoulli}(\boldsymbol{\rho}), \\ &p(\boldsymbol{\sigma_z}|\boldsymbol{z} = \mathbf{0}) = \delta(\boldsymbol{\sigma_z}), \\ &p(\boldsymbol{\sigma_z}|\boldsymbol{z} \neq \mathbf{0}) = \mathcal{N}(\boldsymbol{\sigma_z}|\mathbf{0}, \mathbf{1}). \end{aligned} \tag{3}$$

Since the true posterior distribution under this prior is intractable, we let $q(\boldsymbol{\sigma_z}, \boldsymbol{z})$ be a spike and slab approximate posterior over $\boldsymbol{\sigma_z}$ and $\boldsymbol{z}$. $-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z})$ under the spike and slab prior and approximate posterior can be rewritten as

$$\begin{aligned} &-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z}) \\ &= -\mathbb{E}_{q(\boldsymbol{\sigma_z}, \boldsymbol{z})}\big[\log \mathbb{P}(\mathcal{D}|\boldsymbol{\sigma_z}, \boldsymbol{z})\big] + KL(q(\boldsymbol{\sigma_z}, \boldsymbol{z})\|p(\boldsymbol{\sigma_z}, \boldsymbol{z})). \end{aligned} \tag{4}$$

We assume that the multi-frequency gated units are independent of each other. p and q factorize over the dimensionality of $\boldsymbol{\sigma_z}$ and $\boldsymbol{z}$ in an element-wise way. Furthermore, according to the chain rule of KL-divergence, we have

$$\begin{aligned} &-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z}) \\ &= -\mathbb{E}_{q(\boldsymbol{z})q(\boldsymbol{\sigma_z}|\boldsymbol{z})}\big[\log \mathbb{P}(\mathcal{D}|\boldsymbol{\sigma_z})\big] + \sum_{i=1}^{|\boldsymbol{\sigma_z}|} KL(q(z_i)\|p(z_i)) \\ &\quad + \sum_{i=1}^{|\boldsymbol{\sigma_z}|} q(z_i = 0) KL(q(\sigma_{z,i}|z_i = 0)\|p(\sigma_{z,i}|z_i = 0)) \\ &\quad + \sum_{i=1}^{|\boldsymbol{\sigma_z}|} q(z_i \neq 0) KL(q(\sigma_{z,i}|z_i \neq 0)\|p(\sigma_{z,i}|z_i \neq 0)). \end{aligned} \tag{5}$$

Since

$$\begin{aligned} &KL(q(z_i)\|p(z_i)) \geq 0, \\ &KL(q(\sigma_{z,i}|z_i = 0)\|p(\sigma_{z,i}|z_i = 0)) = 0, \\ &KL(q(\sigma_{z,i}|z_i \neq 0)\|p(\sigma_{z,i}|z_i \neq 0)) = \gamma, \end{aligned} \tag{6}$$

where γ is a weighting factor for explicitly penalizing the globally-responding nonlinearities for introducing extra model complexity and structural risk. We have

$$\begin{aligned} &-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z}) \\ &\geq -\mathbb{E}_{q(\boldsymbol{z})q(\boldsymbol{\sigma_z}|\boldsymbol{z})}\big[\log \mathbb{P}(\mathcal{D}|\boldsymbol{\sigma_z})\big] + \gamma \sum_{i=1}^{|\boldsymbol{\sigma_z}|} q(z_i \neq 0). \end{aligned} \tag{7}$$

As long as we apply a differentiable approximate posterior $q(\boldsymbol{z}|\boldsymbol{\rho})$ allowing for the reparameterization trick $\boldsymbol{z} = f(\boldsymbol{\rho}, \boldsymbol{\epsilon})$ over the parameters $\boldsymbol{\rho}$, a deterministic differentiable function f, and a parameter-free noise distribution $\tau(\boldsymbol{\epsilon})$, we can reformulate the optimization objective $-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z})$ and solve it by Monte Carlo approximation.

$$\begin{aligned} &-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z}) \\ &\geq -\mathbb{E}_{\tau(\boldsymbol{\epsilon})}\big[\log \mathbb{P}(\mathcal{D}|\boldsymbol{\sigma}_{\boldsymbol{f}(\boldsymbol{\rho}, \boldsymbol{\epsilon})})\big] + \gamma \sum_{i=1}^{|\boldsymbol{\sigma_z}|} q(z_i \neq 0|\rho_i), \\ &\approx -\sum_{k=1}^{K} \log \mathbb{P}(\mathcal{D}|\boldsymbol{\sigma}_{\boldsymbol{f}(\boldsymbol{\rho}, \boldsymbol{\epsilon}^{(k)})}) + \gamma \sum_{i=1}^{|\boldsymbol{\sigma_z}|} q(z_i \neq 0|\rho_i). \end{aligned} \tag{8}$$

Crucially, the learning objective is now differentiable with respect to the parameters $\boldsymbol{\rho}$, thus enabling for efficient stochastic gradient based optimization. The parameters of the distribution over the gates can then be jointly optimized with the original network parameters. Moreover, the requisite frequency information can also be perceived implicitly during error backpropagation.

3.3 Rectified Continuous Bernoulli Unit

Based on the differentiable learning paradigm, we further refine the multi-frequency gated units by utilizing a continuously differentiable distribution allowing for the reparameterization trick. Assume that we have a continuous Bernoulli random variable v distributed in the $(0, 1)$ interval with probability density function $q_v(v|\rho)$ and cumulative distribution function $Q_v(v|\rho)$. The parameter $0 < \rho < 1$ implies the degree that v is more likely closer to 1 than 0. We can calculate $q_v(v|\rho)$ and $Q_v(v|\rho)$ in closed forms.

$$q_v(v|\rho) = \begin{cases} 2\rho^v(1-\rho)^{1-v}, & \rho = \frac{1}{2} \\ \frac{2\tanh^{-1}(1-2\rho)\rho^v(1-\rho)^{1-v}}{1-2\rho}, & \rho \neq \frac{1}{2} \end{cases}, \tag{9}$$

and

$$Q_v(v|\rho) = \begin{cases} v, & \rho = \frac{1}{2} \\ \frac{\rho^v(1-\rho)^{1-v}+\rho-1}{2\rho-1}, & \rho \neq \frac{1}{2} \end{cases}. \tag{10}$$

Here, we stretch the continuous Bernoulli distribution to the (ξ, ζ) interval, with $\xi \leq 0$ and $\zeta \geq 1$, and further bound it in $[0, 1]$ by applying a min-max rectifier.

$$\begin{aligned} &\breve{v} = v(\zeta - \xi) + \xi, \\ &z = \min(1, \max(0, \breve{v})). \end{aligned} \tag{11}$$

This would then induce a rectified continuous Bernoulli distribution serving as a better approximation of the discrete Bernoulli distribution: 1) the probability mass of $q_{\breve{v}}(\breve{v}|\rho)$ on the negative values, $Q_{\breve{v}}(0|\rho)$ is folded to a delta peak at zero; 2) the probability mass on values larger than one, $1 - Q_{\breve{v}}(1|\rho)$ is folded to a delta peak at one; 3) the original distribution $q_{\breve{v}}(\breve{v}|\rho)$ is truncated to the $(0, 1)$ interval. The rectified continuous Bernoulli distribution includes $\{0, 1\}$ in its support, while still allowing for gradient based optimization of its parameters due to the continuous probability mass that connects these two values.

Considering $q(z \neq 0|\rho) = 1 - Q_{\breve{v}}(0|\rho)$, we define the optimization objective by minimizing the total risk $\mathcal{R}(\mathcal{D})$.

$$\mathcal{R}(\mathcal{D}) := -\log \mathbb{P}(\mathcal{D}|\boldsymbol{\sigma_z}) + \gamma \sum_{i=1}^{|\boldsymbol{\sigma_z}|} \left[1 - Q_{\breve{v}_i}(0|\rho_i)\right], \quad (12)$$

where

$$Q_{\breve{v}_i}(0|\rho_i) = Q_{v_i}(\frac{-\xi}{\zeta - \xi}|\rho_i). \quad (13)$$

In training, paired with a parameter-free noise random variable $\epsilon \sim \mathcal{U}(0, 1)$, z can be sampled efficiently.

$$\begin{aligned} v &= \begin{cases} \epsilon, & \rho = \frac{1}{2} \\ \frac{\log(\epsilon(2\rho-1)+(1-\rho))-\log(1-\rho)}{\log\rho-\log(1-\rho)}, & \rho \neq \frac{1}{2} \end{cases}, \\ z &= \min(1, \max(0, v(\zeta - \xi) + \xi)). \end{aligned} \quad (14)$$

In prediction, we apply the following unbiased estimator.

$$\begin{aligned} \bar{v} &= \begin{cases} \frac{1}{2}, & \rho = \frac{1}{2} \\ \frac{\rho}{2\rho-1} + \frac{1}{2\tanh^{-1}(1-2\rho)}, & \rho \neq \frac{1}{2} \end{cases}, \\ \bar{z} &= \min(1, \max(0, \bar{v}(\zeta - \xi) + \xi)). \end{aligned} \quad (15)$$

The total risk $\mathcal{R}(\mathcal{D})$ is a special case of the negative ELBO $-\mathcal{L}(\mathcal{D}, \boldsymbol{\sigma_z}, \boldsymbol{z})$ by setting the sampling number of $K = 1$. The reason for optimizing $\mathcal{R}(\mathcal{D})$ is that we focus on efficiently learning complex multi-frequency patterns under large-scale network architectures, instead of revealing the uncertainty of gates. As the training continues, $ReCB$-based networks can converge very well even if sampling only once.

4 Theoretical Framework

The main insights in our theoretical results are characterized chiefly: 1) The globally-responding networks have the conceptually-attractive ability in approximating 2π-periodic p-order Lebesgue-integrable functions, which is $\mathcal{O}(m^{\eta/d^2})$ times better than that of popular locally-responding networks, for a hidden dimension of m, an input dimension of d, and a Lipschitz constant of η; 2) $ReCB$-based networks consisting finely of different kinds of nonlinearities can also achieve the theoretically-optimal representation ability; 3) The deep compositional architectures can significantly improve the representational properties by reducing the exponential approximation errors to polynomial ones.

4.1 Approximation under Shallow Architectures

Definition 1 (Notation)**.** *For $p \geq 1$, let Ψ be the space of 2π-periodic p-order Lebesgue-integrable functions $L^p_{2\pi}(\mathbb{R}^d)$ (p is bounded) or 2π-periodic continuous functions $C_{2\pi}(\mathbb{R}^d)$ ($p = \infty$). Let $f \in \Psi$. Define the Ψ-norm by*

$$\|f\|_\Psi = \left[(2\pi)^{-d} \int_{-\pi}^{\pi} |f(\boldsymbol{x})|^p d\boldsymbol{x}\right]^{\frac{1}{p}}. \quad (16)$$

Definition 2 (Modulus of Continuity)**.** *Let $f \in \Psi$ and $\delta \geq 0$. The 1-order modulus of continuity $\omega(f, \delta)_\Psi$ of f under Ψ-norm is defined by*

$$\omega(f, \delta)_\Psi = \sup_{\|\Delta\| \leq \delta} \{\|f(\boldsymbol{x} + \Delta) - f(\boldsymbol{x})\|_\Psi, \quad \forall \Delta \in \mathbb{R}^d\}. \quad (17)$$

In particular, if $f \in Lip^\eta_C$ satisfies a Lipschitz condition with a constant of $C > 0$ and an exponent of $\eta > 0$ under Ψ-norm, then $\omega(f, \delta)_\Psi$ is bounded by $M\delta^\eta$.

Theorem 1 (Approximation Bound for $\mathcal{NN}[\tilde{\sigma}]^{(1)}$)**.** *Let the shallow globally-responding mapping $\mathcal{NN}[\tilde{\sigma}]^{(1)} = \sum_{i=1}^m \tilde{\sigma}_i$ where the hidden dimension is $m = (\lambda + 1)^d - 1$. Let $f \in \Psi$ be the target function. The approximation error under Ψ-norm is estimated by*

$$\inf_{\mathcal{NN}[\tilde{\sigma}]^{(1)}} \sup_f \|\mathcal{NN}[\tilde{\sigma}]^{(1)} - f\|_\Psi \leq D\omega_\lambda, \quad (18)$$

where

$$D = 1 + \frac{\pi^2}{2}\sqrt{d}, \quad \omega_\lambda = \omega(f, \frac{1}{\lambda + 2})_\Psi. \quad (19)$$

Theorem 2 (Approximation Bound for $\mathcal{NN}[\bar{\sigma}]^{(1)}$)**.** *Let the shallow locally-responding mapping $\mathcal{NN}[\bar{\sigma}]^{(1)} = \sum_{i=1}^m \bar{\sigma}_i$ where the hidden dimension is $m = 4\sigma\lambda(\lambda + 1)^d$. Let $f \in \Psi$ be the target function. The approximation error under Ψ-norm is estimated by*

$$\inf_{\mathcal{NN}[\bar{\sigma}]^{(1)}} \sup_f \|\mathcal{NN}[\bar{\sigma}]^{(1)} - f\|_\Psi \leq D\omega_\lambda + \Phi, \quad (20)$$

and the remainder Φ is

$$\Phi = 4\|f\|_\Psi \left[\frac{\sqrt{d}\pi}{(2\pi)^{d-1}\sigma}\left(\frac{4\sigma}{\pi} - \cot\frac{\pi}{4\sigma}\right)\right]^{\frac{1}{p}}. \quad (21)$$

Theorem 3 (Approximation Bound for $\mathcal{NN}[\sigma_z]^{(1)}$)**.** *Let the shallow multi-frequency gated mapping $\mathcal{NN}[\sigma_z]^{(1)} = \sum_{i=1}^m \sigma_{zi} = \sum_{i=1}^{m_h} \tilde{\sigma}_i + \sum_{i=1}^{m_l} \bar{\sigma}_i$ where the hidden dimension is $m = m_h + m_l$. m_h and m_l are the dimension of $\tilde{\sigma}$ and $\bar{\sigma}$, respectively. Let $f \in \Psi$ be the target function. The approximation error under Ψ-norm is estimated by*

$$\inf_{\mathcal{NN}[\sigma_z]^{(1)}} \sup_f \|\mathcal{NN}[\sigma_z]^{(1)} - f\|_\Psi \leq D\omega_m + \Theta, \quad (22)$$

where

$$\omega_m = \omega(f, \frac{1}{(m+1)^{\frac{1}{d}} + 1})_\Psi, \quad (23)$$

and the remainder Θ is

$$\begin{aligned} \Theta = 4\|f\|_\Psi \Big[&\frac{4\sqrt{d}}{(2\pi)^{d-1}} - \frac{4\sqrt{d}\pi[(m_l+1)^{\frac{1}{d}} - 1](m_l + 1)}{(2\pi)^{d-1} m_l} \\ &\times \cot\frac{\pi[(m_l+1)^{\frac{1}{d}} - 1](m_l+1)}{m_l}\Big]^{\frac{1}{p}}. \end{aligned} \quad (24)$$

Corollary 1 (For Lipschitz). *Suppose $f \in Lip_C^{\eta}$. We have*

$$\begin{aligned}\inf_{\mathcal{NN}[\tilde{\sigma}]^{(1)}} \sup_f \|\mathcal{NN}[\tilde{\sigma}]^{(1)} - f\|_\Psi \leq & \mathcal{O}\Big(m^{-\frac{\eta}{d}}\Big),\\ \inf_{\mathcal{NN}[\bar{\sigma}]^{(1)}} \sup_f \|\mathcal{NN}[\bar{\sigma}]^{(1)} - f\|_\Psi \leq & \mathcal{O}\Big(m^{-\frac{\eta}{d+1}}\Big),\\ \inf_{\mathcal{NN}[\sigma_z]^{(1)}} \sup_f \|\mathcal{NN}[\sigma_z]^{(1)} - f\|_\Psi \leq & \mathcal{O}\Big(m^{-\frac{\eta}{d}}\Big).\end{aligned} \tag{25}$$

These numerically-tight results clarify the power of $\mathcal{NN}[\sigma_z]^{(1)}$ consisting finely of different kinds of nonlinearities. It can achieve the conceptually-optimal approximation bound paired with a small remainder [Feinerman and Newman, 1975], if the number of globally-responding elements $\tilde{\sigma}$ is not significantly less than others.

4.2 Approximation under Deep Architectures

Definition 3 (Compositional Functions [Poggio *et al.*, 2017]). *Let $\mathcal{G}$ be a directed acyclic graph (DAG), with the set of source nodes S and the set of vertexes V. For each vertex $v \in V$, d_v is the number of in-edges of v. Let f be a compositional $\mathcal{G}$-function defined by the compositional structure corresponding to $\mathcal{G}$. f is recursively constructed by a class of constituent functions $\{f_v\}_{v\in V}$ layer-by-layer. f_v with inputs $\{f_{v_i}\}_{i=1}^{d_v}$ corresponds to v with precursor vertexes $\{v_i\}_{i=1}^{d_v}$. Hence, f corresponds to the whole $\mathcal{G}$ with $|S|$-dimensional inputs.*

Theorem 4 (Approximation Bound for $\mathcal{NN}[\tilde{\sigma}]$). *Follow the notations in Theorem 1. Let the target function $f \in \Psi$ be a compositional $\mathcal{G}$-function including the constituent functions $\{f_v \in \Psi\}_{v\in V}$. Let the deep globally-responding mapping $\mathcal{NN}[\tilde{\sigma}]$ have the same architecture as $\mathcal{G}$. For each $v \in V$, let $\mathcal{NN}[\tilde{\sigma}]_v^{(1)}$ correspond to v. The hidden dimension of $\mathcal{NN}[\tilde{\sigma}]_v^{(1)}$ is $m_v = (\lambda_v + 1)^{d_v} - 1$. The approximation error under Ψ-norm is estimated by*

$$\begin{aligned}\inf_{\mathcal{NN}[\tilde{\sigma}]} \sup_f \|\mathcal{NN}[\tilde{\sigma}] - f\|_\Psi \leq & \sum_{v\in V} \Big\{ D_v \omega_{\lambda_v} \\ & + \Big(1 + (\lambda_v + 2)\sum_{i=1}^{d_v} D_{v_i}\omega_{\lambda_{v_i}}\Big)\omega_{\lambda_v}\Big\}.\end{aligned} \tag{26}$$

Theorem 5 (Approximation Bound for $\mathcal{NN}[\bar{\sigma}]$). *Follow the notations in Theorem 2. Let the target function $f \in \Psi$ be a compositional $\mathcal{G}$-function including the constituent functions $\{f_v \in \Psi\}_{v\in V}$. Let the deep locally-responding mapping $\mathcal{NN}[\bar{\sigma}]$ have the same architecture as $\mathcal{G}$. For each $v \in V$, let $\mathcal{NN}[\bar{\sigma}]_v^{(1)}$ correspond to v. The hidden dimension of $\mathcal{NN}[\bar{\sigma}]_v^{(1)}$ is $m_v = 4\sigma_v\lambda_v(\lambda_v + 1)^{d_v}$. The approximation error under Ψ-norm is estimated by*

$$\begin{aligned}\inf_{\mathcal{NN}[\bar{\sigma}]} \sup_f \|\mathcal{NN}[\bar{\sigma}] - f\|_\Psi \leq & \sum_{v\in V} \Big\{ D_v \omega_{\lambda_v} + \Phi_v \\ & + \Big(1 + (\lambda_v + 2)\sum_{i=1}^{d_v} (D_{v_i}\omega_{\lambda_{v_i}} + \Phi_{v_i})\Big)\omega_{\lambda_v}\Big\}.\end{aligned} \tag{27}$$

Theorem 6 (Approximation Bound for $\mathcal{NN}[\sigma_z]$). *Follow the notations in Theorem 3. Let the target function $f \in \Psi$ be a compositional $\mathcal{G}$-function including the constituent functions $\{f_v \in \Psi\}_{v\in V}$. Let the deep multi-frequency gated mapping $\mathcal{NN}[\sigma_z]$ have the same architecture as $\mathcal{G}$. For each $v \in V$, let $\mathcal{NN}[\sigma_z]_v^{(1)}$ correspond to v. The hidden dimension of $\mathcal{NN}[\sigma_z]_v^{(1)}$ is $m_v = m_{h_v} + m_{l_v}$. The approximation error under Ψ-norm is estimated by*

$$\begin{aligned}\inf_{\mathcal{NN}[\sigma_z]} \sup_f \|\mathcal{NN}[\sigma_z] - f\|_\Psi \leq & \sum_{v\in V} \Big\{ D_v \omega_{m_v} + \Theta_v \\ & + \Big(1 + ((m_v + 1)^{\frac{1}{d_v}} + 1)\sum_{i=1}^{d_v} (D_{v_i}\omega_{m_{v_i}} + \Theta_{v_i})\Big)\omega_{m_v}\Big\}.\end{aligned} \tag{28}$$

Corollary 2 (For Lipschitz). *Suppose $f \in Lip_C^{\eta}$ and $\{f_v \in Lip_C^{\eta}\}_{v\in V}$. Let $\{\mathcal{NN}[\tilde{\sigma}]_v^{(1)}, \mathcal{NN}[\bar{\sigma}]_v^{(1)}, \mathcal{NN}[\sigma_z]_v^{(1)}\}_{v\in V}$ have the same hidden dimension of m_v. We have*

$$\begin{aligned}\inf_{\mathcal{NN}[\tilde{\sigma}]} \sup_f \|\mathcal{NN}[\tilde{\sigma}] - f\|_\Psi \leq & \sum_{v\in V} \mathcal{O}\Big(m_v^{-\frac{2\eta}{d_v}}\Big),\\ \inf_{\mathcal{NN}[\bar{\sigma}]} \sup_f \|\mathcal{NN}[\bar{\sigma}] - f\|_\Psi \leq & \sum_{v\in V} \mathcal{O}\Big(m_v^{-\frac{2\eta}{d_v+1}}\Big),\\ \inf_{\mathcal{NN}[\sigma_z]} \sup_f \|\mathcal{NN}[\sigma_z] - f\|_\Psi \leq & \sum_{v\in V} \mathcal{O}\Big(m_v^{-\frac{2\eta}{d_v}}\Big).\end{aligned} \tag{29}$$

These results further emphasize that the superiority of the deeper $\mathcal{NN}[\sigma_z]$ is inherited from that of the shallower $\mathcal{NN}[\sigma_z]^{(1)}$. $ReCB$ units can improve the representational properties of neural networks, whether in shallow or deep architectures.

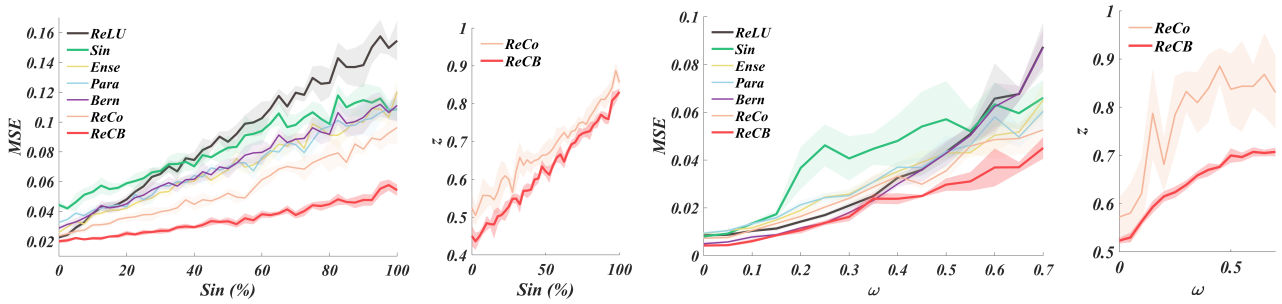
5 Experiments

5.1 Experimental Models

They are denoted by the nonlinearities and gating schemes.
- $ReLU$: as the baseline of locally-responding nonlinearity.
- Sin: as the baseline of globally-responding nonlinearity.
- $Ense$: as a weighted Ensemble of $ReLU$ and Sin.
- $Para$: setting the gates as a group of learnable Parameters.
- $Bern$: sampling the gates from the Bernoulli distribution.
- $ReCo$: gating via Rectified Concrete (ReCo) distribution.
- $ReCB$: gating multi-frequency patterns through the proposed Rectified Continuous Bernoulli ($ReCB$) units.

5.2 Learning Heterogeneous Patterns

The experiment is designed to evaluate the sensitivity of $ReCB$ units in gating heterogeneous patterns. The implicit patterns to be learned are represented by a single-layer neural network $\mathcal{G}$ with the hidden dimension of 200. $\mathcal{G}$ consists of the locally-responding $ReLU$ and the globally-responding sinusoid. The proportion of the sinusoidal elements linearly increase from 0% to 100%. We uniformly get 20000 samples $\{(\boldsymbol{x}_i, \mathcal{G}(\boldsymbol{x}_i))\}_{i=1}^{20000}, \boldsymbol{x}_i \in [-2\pi, 2\pi]^{20}$, and randomly divide them into two non-overlapping training and test sets that are equal in size. The division, training and test processes are repeated 5 times, and then we assess the average performance on Mean Squared Error (MSE). All compared models have



(a) MSE (b) $\mathbb{E}[z]$ (c) MSE (d) $\mathbb{E}[z]$

Figure 2: Learning heterogeneous patterns (left) and frequency-shifting patterns (right).

the same architecture as $\mathcal{G}$, and are optimized by Adam. The results are shown in Figure 2.

Besides the apparent observation, there are two in-depth insights. Firstly, the expectation value of gates is linearly correlated to the proportion of pattern components. Because the variational Bayesian learning paradigm sensitively captures the changes of components implied in complex patterns, and then feeds the error information back to the gates. Secondly, $ReCB$ performs much better than $ReCo$ even though their gates have the very close expectation value. It is a major reason that $ReCB$ units concentrate probability mass better around the extrema, and do not cause too much unpleasant sampling uncertainty. These properties play a major role in modeling discrete gating behaviours under stochastic gradient based optimization.

5.3 Learning Frequency-Shifting Patterns

Furthermore, another experiment is designed to evaluate the effectiveness of $ReCB$ units in capturing the changes of frequencies. The implicit patterns are also represented by $\mathcal{G}$ consisting of 10 $ReLU$ neurons and 190 sinusoidal neurons. The frequencies ω of these sinusoidal neurons linearly increase from 0.0 to 0.7 paired with extra noise $\mathcal{N}(0, 0.1^2)$. All compared models have the same single-layer network architecture containing 500 neurons. The results are shown in Figure 2.

Firstly, the curve slope of $ReCB$ is the flattest, which means that $ReCB$ has the potential to achieve greater performance advantages as the frequencies increase. Secondly, the performance depends heavily on the distribution characteristics of gates. Compared to the optimal $ReCB$, $ReCo$ performs worse due to its unpleasant uncertainty in sampling process, and thus is not suitable for gating and optimizing the crucial nonlinearities. Lastly, automatically detecting/gating/representing different frequency responses is not a trivial behaviour mimicking random selection. The gate is even as important as the nonlinearity. $ReLU$ and Sin achieve the worst performance in learning the neural network $\mathcal{G}$ composed entirely of themselves.

5.4 Learning Image Classification

Moreover, to demonstrate that various networks can benefit from the proposed $ReCB$ units, we conduct an experiment to learn **CIFAR10** [Krizhevsky *et al.*, 2009] classification. All models are optimized by Stochastic Gradient Descent (SGD) with a mini-batch size of 128, a weight decay of 10^{-4}, and a Nesterov momentum of 0.9 [Paszke *et al.*, 2017; Sutskever *et al.*, 2013; Goodfellow *et al.*, 2016]. The learning rate is adjusted by a cosine annealing schedule with warm restarts [Loshchilov and Hutter, 2016]. The results are collected in Table 1.

Top-1 Error(%)	LeNet-5		ResNet-20		ResNet-56		ResNet-110	
	conv	**best**	**conv**	**best**	**conv**	**best**	**conv**	**best**
$ReLU$	30.63	30.49	8.87	8.76	7.27	7.06	7.41	7.30
Sin	29.48	29.36	11.52	11.35	-	-	-	-
$Ense$	24.81	24.78	9.30	9.02	-	-	-	-
$Para$	29.25	29.16	9.34	9.20	10.61	9.79	-	-
$Bern$	34.15	33.79	12.20	11.53	9.62	8.92	-	-
$ReCo$	25.43	25.43	9.86	7.89	7.57	6.65	-	-
$ReCB$	**23.47**	**23.19**	**6.66**	**6.52**	**5.80**	**5.64**	**5.55**	**5.54**

Table 1: Top-1 error(%) on the CIFAR10 dataset. Conv means the convergent error in the last epoch and best means the best error in all epochs. The best results are highlighted in bold.

In all the experiments $ReCB$ remarkably achieves the best performance. The top-1 error 6.52% of $ReCB$-based ResNet-20 ($ReCB$) is even more competitive compared with the official record 6.61% of ResNet-110 in the publication [He *et al.*, 2016]. Other models failed to train under the giant ResNet-110 architecture due to the unsolved optimization dilemma of the globally-responding sinusoidal nonlinearity, and behaved as poorly as random guessing. It suggests that $ReCB$ units are capable of improving the computationally intractable optimization of globally-responding elements.

6 Conclusion

In this paper, we propose a novel $ReCB$ unit paired with variational Bayesian learning paradigm, to automatically detect/gate/represent different frequency responses. A theoretical framework is also presented to analyze the representational properties of locally/globally-responding nonlinearities. Our main result characterizes that $ReCB$-based networks can achieve the conceptually-attractive approximation bound. Furthermore, we provide comprehensive empirical evidence showing that $ReCB$-based networks can keenly learn multi-frequency patterns.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No.62076062). Furthermore, the work was also supported by Collaborative Innovation Center of Wireless Communications Technology.

References

[Bond-Taylor and Willcocks, 2020] Sam Bond-Taylor and Chris G Willcocks. Gradient origin networks. *arXiv preprint arXiv:2007.02798*, 2020.

[Feinerman and Newman, 1975] Robert P. Feinerman and Donald J. Newman. Polynomial approximation. *Bull. Amer. Math. Soc.*, 81:28–30, 1975.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Jacot *et al.*, 2018] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[Kazemi *et al.*, 2019] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus Brubaker. Time2vec: Learning a vector representation of time. *arXiv preprint arXiv:1907.05321*, 2019.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[Loshchilov and Hutter, 2016] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[Mildenhall *et al.*, 2020] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *arXiv preprint arXiv:2003.08934*, 2020.

[Parascandolo *et al.*, 2017] Giambattista Parascandolo, Heikki Huttunen, and Tuomas Virtanen. Taming the waves: Sine as activation function in deep neural networks. *ICLR 2017 conference submission*, 2017.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. *NIPS 2017 Autodiff Workshop*, 2017.

[Poggio *et al.*, 2017] Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

[Ramachandran *et al.*, 2017] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[Sitzmann *et al.*, 2020a] Vincent Sitzmann, Eric R Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *arXiv preprint arXiv:2006.09662*, 2020.

[Sitzmann *et al.*, 2020b] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33, 2020.

[Sutskever *et al.*, 2013] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

[Tancik *et al.*, 2020] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[Xu *et al.*, 2019] Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Self-attention with functional time representation learning. In *Advances in Neural Information Processing Systems*, pages 15915–15925, 2019.

[Xue and Wu, 2020] Hui Xue and Zheng-Fan Wu. Bakernets: Bayesian random kernel mapping networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3073–3079. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

[Xue *et al.*, 2019] Hui Xue, Zheng-Fan Wu, and Wei-Xiang Sun. Deep spectral kernel learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4019–4025. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[Zhong *et al.*, 2019] Ellen D Zhong, Tristan Bepler, Joseph H Davis, and Bonnie Berger. Reconstructing continuous distributions of 3d protein structure from cryo-em images. In *International Conference on Learning Representations*, 2019.