XXII Congresso U.M.I.
Pisa, 4–9 settembre 2023

On the functional optimality of neural networks

Michael Unser
Biomedical Imaging Group, EPFL, Lausanne, Switzerland

Let L be a linear shift-invariant and isotropic operator that is characterized by its radial Fourier profile $\widehat{L}_{\mathrm{rad}} : \mathbb{R} \to \mathbb{R}$. We further assume that $\widehat{L}_{\mathrm{rad}}$ is non-vanishing, except for a zero of order $(n_0 - 1)$ at the origin. This operator is in one-to-one correspondence with the activation function $\sigma_{\mathrm{L}} = \mathcal{F}^{-1}\{1/\widehat{L}_{\mathrm{rad}}\} : \mathbb{R} \to \mathbb{R}$ where $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. We define the corresponding Radon domain regularization operator $\mathrm{L}_{\mathrm{R}} = \mathrm{K}_{\mathrm{rad}}\mathrm{R}\mathrm{L} : \mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{even}}(\mathbb{R} \times \mathbb{S}^{d-1})$ where R is the Radon transform, $\mathrm{K}_{\mathrm{rad}}$ is the filtering operator of computed tomography (such that $\mathrm{R}^*\mathrm{K}_{\mathrm{rad}}\mathrm{R} = \mathrm{Id}$), and $\mathcal{M}_{\mathrm{even}}$ is the space of even hyper-spherical bounded measures (see [1] for the precise definition of these elements).

Given the data points $(\boldsymbol{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$ with $m = 1, \dots, M$, we then consider the functional optimization problem

$$\arg \min_{f \in \mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)} \sum_{m=1}^{M} \left| y_m - f(\boldsymbol{x}_m) \right|^2 + \lambda \|\mathrm{L}_{\mathrm{R}} f\|_{\mathcal{M}(\mathbb{R} \times \mathbb{S}^{d-1})} \tag{1}$$

where $\mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)$ is a suitable function space such that $\|\mathrm{L}_{\mathrm{R}} f\|_{\mathcal{M}} < +\infty$. The purpose of this presentation is to specify the native space $\mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)$ such that the global optimum of (1) is achieved by a "shallow" neural network of the form

$$f(\boldsymbol{x}) = \sum_{k=1}^{K_0} a_k \sigma_{\mathrm{L}}(\boldsymbol{\xi}_k^{\mathsf{T}} \boldsymbol{x} - t_k) + \sum_{|\boldsymbol{n}| \le n_0} b_{\boldsymbol{n}} \boldsymbol{x}^{\boldsymbol{n}} \tag{2}$$

for some $K_0 \le M - \dim(\mathcal{P}_{n_0})$, and an appropriate set of weights $(\boldsymbol{\xi}_k, t_k, a_k) \in \mathbb{S}^{d-1} \times \mathbb{R} \times \mathbb{R}$, $k \in \{1, \dots, K_0\}$, and $b_{\boldsymbol{n}} \in \mathbb{R}$, $|\boldsymbol{n}| \le n_0$. Specifically, by identifying the underlying native space $\mathcal{M}_{\mathrm{L}_{\mathrm{R}}}(\mathbb{R}^d)$ as a Banach space that is isometrically isomorphic to $\mathcal{M}_{\mathrm{even}}(\mathbb{R} \times \mathbb{S}^{d-1}) \times \mathcal{P}_{n_0}$ where $\mathcal{P}_{n_0}$ is the space of polynomials of degree n_0, we shall prove that the claimed optimality result holds under very mild conditions on $\widehat{L}_{\mathrm{rad}}$ or, equivalently, on σ_{L} (for the variational interpretation of given class of neural networks). In addition, we shall demonstrate that the expansion (2) is universal in the sense that any continuous function can be represented to an arbitrary degree of precision by a superposition of neurons (ridges) plus a polynomial term.

Bibliografia

[1] M. Unser, "From Kernel Methods to Neural Networks: A Unifying Variational Formulation," arXiv:2206.14625, 2022.

E–mail: `michael.unser@epfl.ch`.
Sezione 11: Teoria dell'approssimazione ed applicazioni.