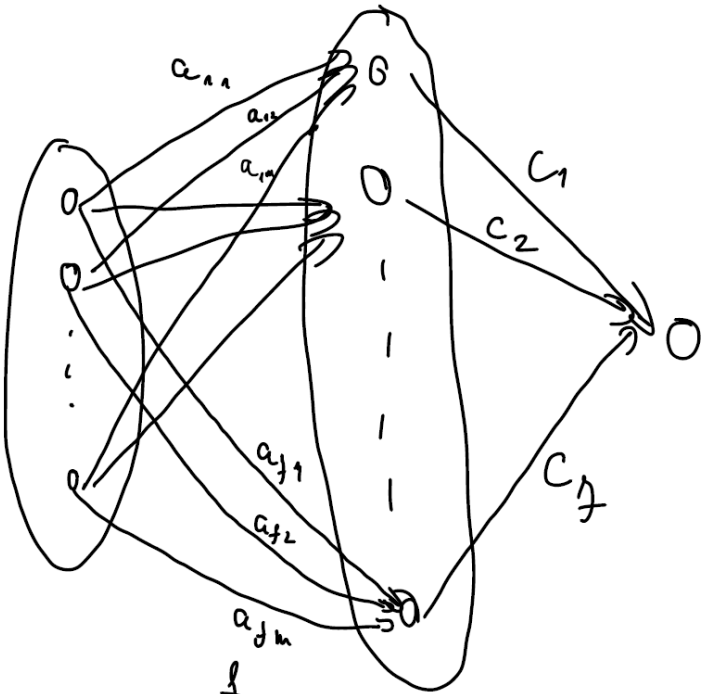
Ridgelet analysis of neural networks

Dániel Rácz
ELTE, Doctoral School of Mathematics
SZTAKI

January 20, 2021

Shallow networks



$$g(\underline{x}) = \sum_{j=1}^{f} c_j \cdot \sigma(\underline{a}_j^T \underline{x} - b_j)$$

$$ReLU(x) = \max(0, x)$$

$$sigmoid(x) = \frac{1}{1 + e^{-x}}$$

- weights can be learnt via backpropagation for a given dataset [3]
- any "sensible" function can be approximated this way
- approximation property depends on $\sigma(x)$
- deep vs. shallow

Universal approximation property

Cybenko, Hornik, 1990 [2] [4]

If the activation function is not a polynomial, then the family of all shallow networks (parametrized by (a_j, b_j, c_j)) is dense in $L^p(\mathbb{R})$ and $C^0(\mathbb{R})$.

Width can be arbitrarily large. Several variants have been proved since then, also for the deep case.

Zhou Lu et al, 2017 [5]

For any Lebesgue-integrable function $f : \mathbb{R}^n \to \mathbb{R}$ and any $\varepsilon > 0$ there exists a fully-connected ReLU network F having its width $d \leq n + 4$ such that

$$\int_{\mathbb{R}^n} \left| f(x) - F(x) \right| \, dx < \varepsilon$$

While width is bounded here, depth can be arbitrarily large.

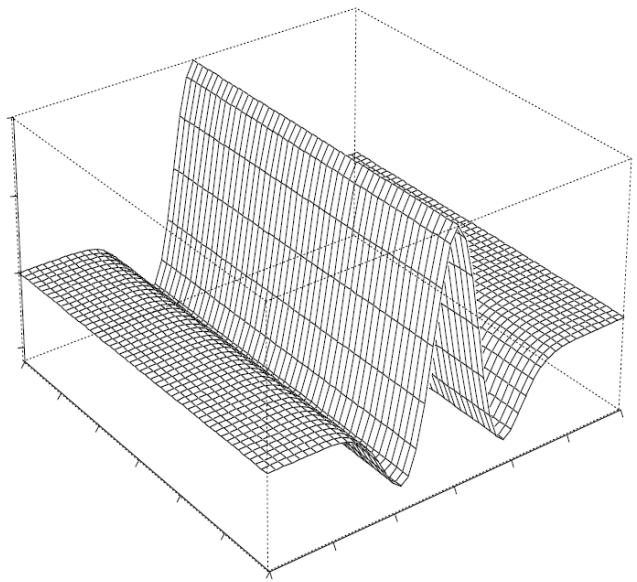
Universal approximation property

- Sonoda, Murata (2015, [8]) ridgelet analysis of shallow networks
- universal approximation property proved (again) for several families of activation functions (in particular ReLU)
- nontrivial relation between the weigths of trained network is discovered; suggested by theory, confirmed by numerical experiments
- they claim to find global optimum (no need for backpropagation)
- chance to extend theory to cover deep networks (and explain why they tend to work better in practice than the shallow counterparts)

Ridgelet analysis

Definition (Murata, [6])

An $F : \mathbb{R}^m \to \mathbb{R}$ is called a *ridge function* if there exists a $G : \mathbb{R} \to \mathbb{R}$ function and $a \in \mathbb{R}^m$ such that $F(x) = G(a^T x)$.



Source: [1]
A ridge is constant along hyperplanes whose normal is parallel to a, i.e. on the parallel hyperplanes $a^T x = c, c \in \mathbb{R}$

Ridgelet transform

Definition [8]

The classical ridgelet transform of an $f : \mathbb{R}^m \to \mathbb{C}$ function w.r.t $\tau : \mathbb{R} \to \mathbb{C}$ is given by

$$R_\tau f(\mathbf{a}, b) := \int_{\mathbb{R}^m} f(\mathbf{x})\overline{\tau(\mathbf{a}\mathbf{x} - b)}\|\mathbf{a}\|^s \, d\mathbf{x}$$

where $\mathbf{a} \in \mathbb{R}^m$ and $b \in \mathbb{R}$.

Definition [8]

The dual ridgelet transform of a $T : \mathbb{R}^{m+1} \to \mathbb{C}$ function w.r.t $\tau : \mathbb{R} \to \mathbb{C}$ is given by

$$R_\tau^+ T(\mathbf{x}) := \int_{\mathbb{R}^{m+1}} T(\mathbf{a}, b)\tau(\mathbf{a}\mathbf{x} - b)\|\mathbf{a}\|^{-s} \, d\mathbf{a}\, db$$

where $\mathbf{x} \in \mathbb{R}^m$.

Ridgelet transform

The terms $\|\mathbf{a}\|^{-s}$ and $\|\mathbf{a}\|^{s}$ are only for technical reasons.

Definition [8]

Two functions σ and τ said to be admissable if

$$K_{\sigma,\tau} := (2\pi)^{m-1} \int_{\mathbb{R}} \frac{\hat{\sigma}(\omega)\overline{\hat{\tau}(\omega)}}{|\omega|^{m}}\, d\omega$$

is finite and nonzero.

It can be shown that if f is "nice enough" and σ and τ are admissable, then the following reconstruction formula holds ([8])

$$R_{\tau}^{+} R_{\sigma} f = K_{\sigma,\tau} f$$

Ridgelet transform

By introducing variables $\mathbf{u} := \mathbf{a}/\|\mathbf{a}\|$, $\alpha := 1/\|\mathbf{a}\|$, $\beta := b/\|\mathbf{a}\|$ we have

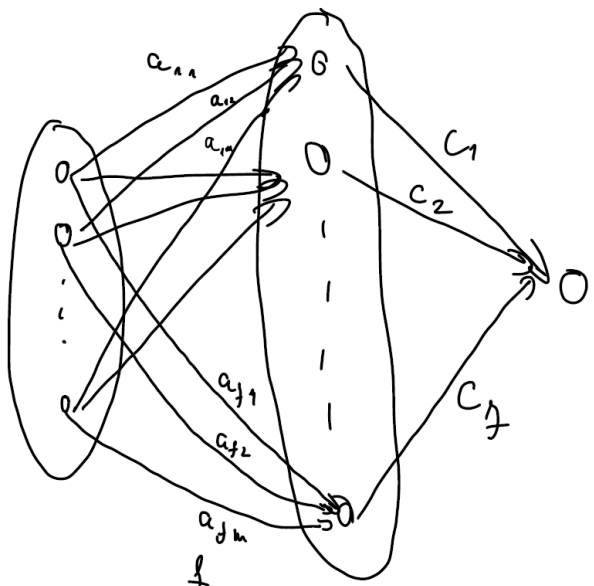
$$\tau_{\mathbf{u},\alpha,\beta}(\mathbf{x}) := \tau\left(\frac{\mathbf{u}\mathbf{x} - \beta}{\alpha}\right)\frac{1}{\alpha^s}$$

and

$$R_\tau f(\mathbf{u},\alpha,\beta) := \int_{\mathbb{R}^m} f(\mathbf{x})\overline{\tau_{\mathbf{u},\alpha,\beta}(\mathbf{x})}\,d\mathbf{x}$$

The name *ridgelet* comes from $\tau_{\mathbf{u},\alpha,\beta}(\mathbf{x})$ as it is constant on $(\mathbb{R}\mathbf{u})^\perp$ and behaves as a wavelet function on $\mathbb{R}\mathbf{u}$.
It can also be shown that ridgelet transform is actually the application of a wavelet transform to the slices of the Radon transform.

Integral representation of NN



$$g(\underline{x}) = \sum_{j=1}^{f} c_j \cdot \sigma(\underline{a}_j^T \underline{x} - b_j)$$

$$ReLU(x) = \max(0, x)$$

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

The *integral representation* [7] of this shallow NN is given by

$$S[\gamma](x) := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\mathbf{a}, b)\sigma(\mathbf{a}\mathbf{x} - b)\, d\lambda(\mathbf{a}, b),$$

where $\gamma \in L^2(\mathbb{R}^{m+1})$.
Informally, this is an infinite weighted sum of hidden units.
The trick is that we want to examine the weights of the output layer as a function of the weights of the previous layer.

Integral representation of NN

$$S[\gamma](x) := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\mathbf{a}, b)\sigma(\mathbf{a}\mathbf{x} - b)\, d\lambda(\mathbf{a}, b),$$

This is actually the dual ridgelet transform of γ w.r.t σ.
However, results related to classical ridgelet analysis do not cover popular activation functions, such as the ReLU.
Solution ([8]): defining the ridgelet transform w.r.t distributions:

$$R_\psi f(\mathbf{u}, \alpha, \beta) := \int_{\mathbb{R}} Rad(f)(\mathbf{u}, \alpha z + \beta)\overline{\psi(z)}\, dz$$

Ridgelet transform w.r.t distributions

The dual transform and the admissibility condition can be extended in a similar manner. Without going into much details (see [8]), the following holds:

Reconstruction formula

Let $f \in L^1(\mathbb{R}^m)$ satisfying $\hat{f} \in L^1(\mathbb{R}^m)$ and let $(\psi, \tau) \in S(\mathbb{R}) \times S_0'(\mathbb{R})$, where $S(\mathbb{R})$ denotes the space of rapidly decreasing functions and $S_0'(\mathbb{R})$ denotes the space of so-called Lizorkin distributions (for definition see [8]). Then the following holds almost everywhere

$$R_\tau^+ R_\psi f = K_{\psi,\tau}' f$$

where $K_{\psi,\tau}'$ is defined in a similar manner to $K_{\psi,\tau}$.

This covers ReLU as it belongs to $S_0'(\mathbb{R})$.
They also show that there exists an $l \in \mathbb{N}$ such that a suitable back-projection (dual Radon transform) of the $G^{(l)}$ derivative of the function $G(z) = exp(-z^2/2)$ is admissible with ReLU.

Example [8]

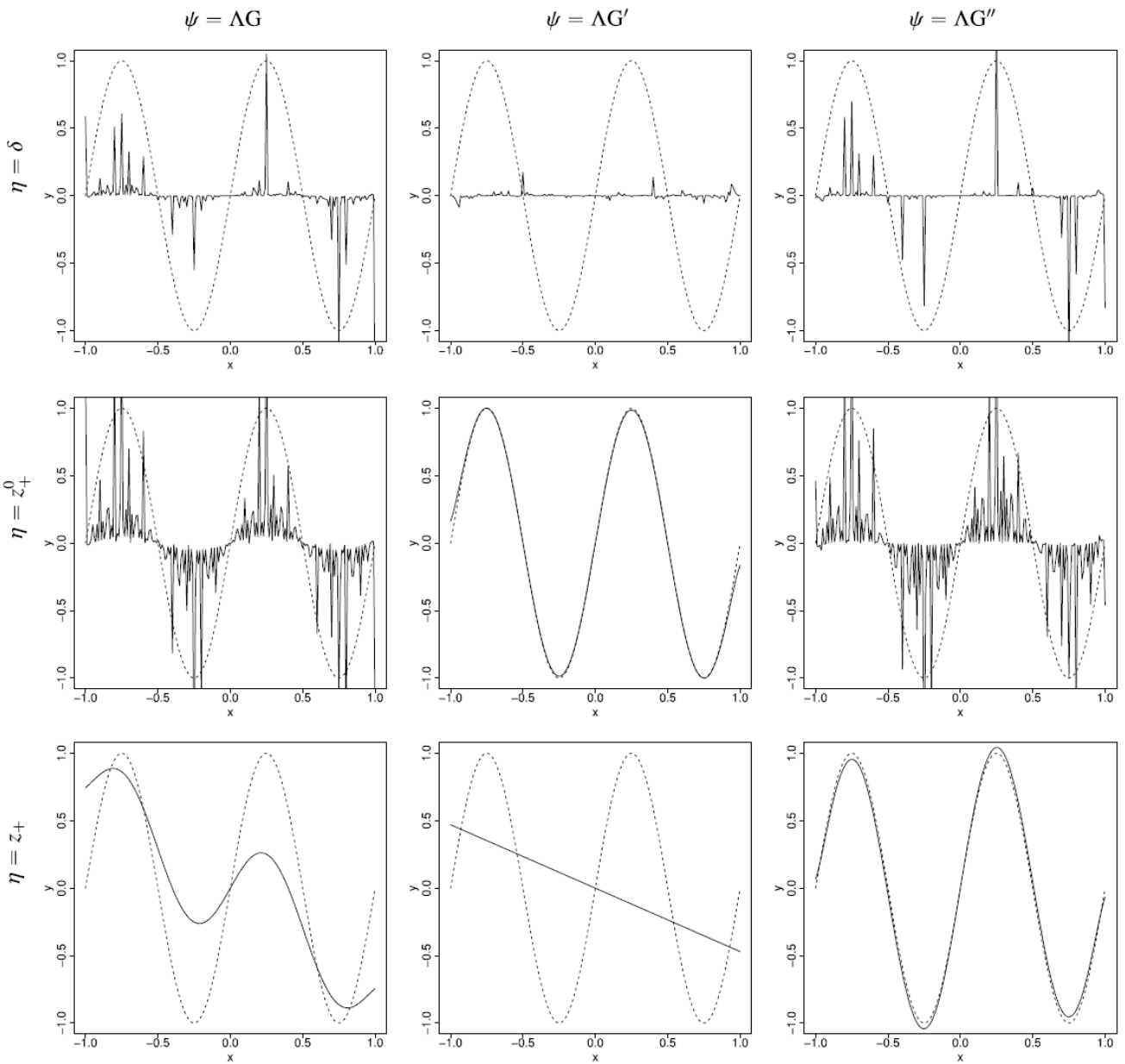


Fig. 4. Reconstruction with truncated power functions — Dirac's δ, unit step z_+^0, and ReLU z_+. The solid line plots the reconstruction result; the dotted line plots the original signal.

Optimum of BP training

We consider an m-in-1-out shallow network with p hidden unit and activation function σ:

$$g(x) = \sum_{j=1}^{p} c_j \sigma(a_j x - b_j),$$

where $(a_j, b_j) \in \mathbb{R}^m \times \mathbb{R}$ are called the hidden parameters and $c_j \in \mathbb{R}$ are called the output parameters.
Multidimensional output can be handled similarly.
Integral representation again:

$$S[\gamma](x) := \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\mathbf{a}, b)\sigma(\mathbf{a}\mathbf{x} - b)\, d\lambda(\mathbf{a}, b)$$

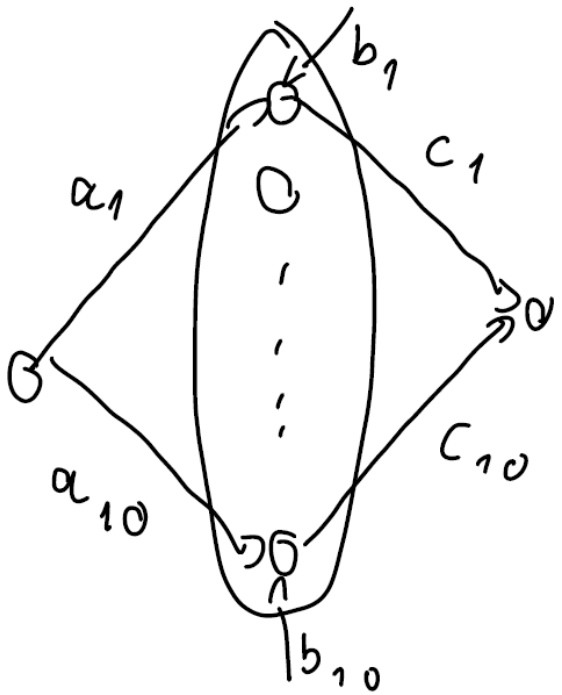
Optimum of BP training

Above described results suggest that γ should be $R_\tau f$ for a suitable τ, i.e. a particular solution to

$$S[\gamma] = f$$

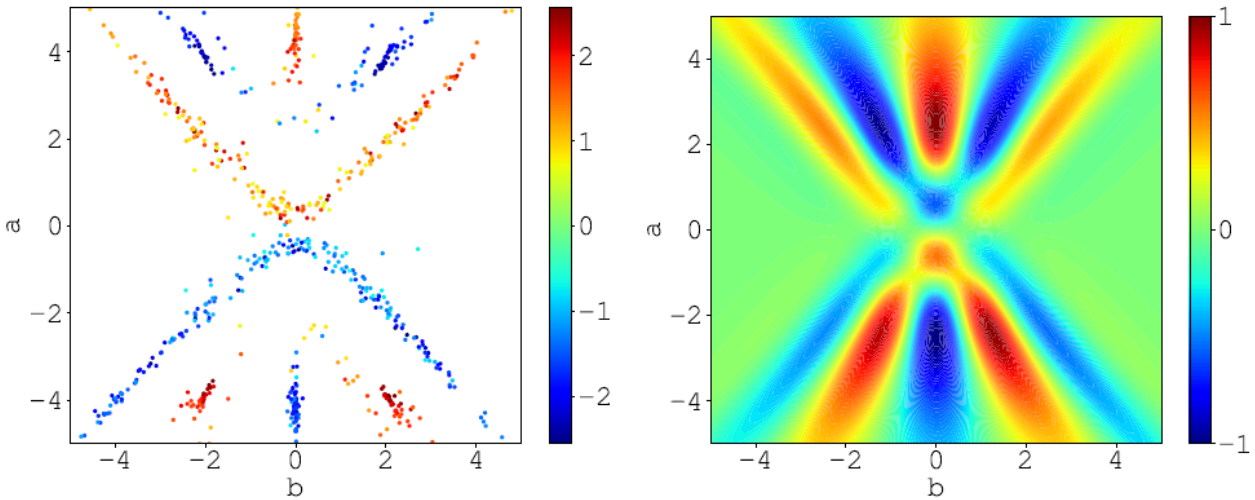
is given by $R_\tau f$.

Example [7]



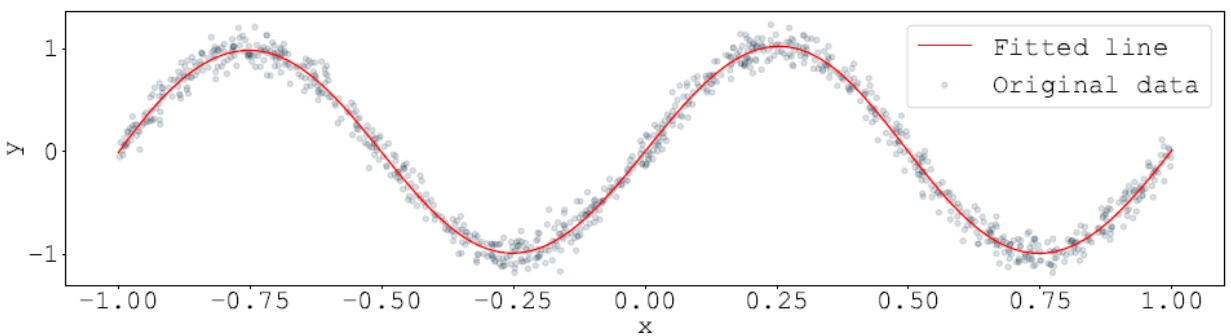
$$\sigma(z) = \tanh(z)$$

Example [7]



(a) BP trained parameters

(b) ridgelet spectrum



(c) dataset and an example of training results

Figure 1: Motivating example: Scatter plot (a) and ridgelet spectrum (b) were obtained from the same dataset (c) and bear an intriguing resemblance to each other, despite the fact that they were obtained from different procedures—numerical optimization and numerical integration.

Optimum of BP training

BP training is the minimization of

$$L(\theta) = \mathbb{E}_X |f(X) - g(X, \theta)|^2 + \Omega(\theta),$$

where f is the ground truth function and Ω is a regularizing term.
We reformulate the BP problem as

$$L(\gamma) = \mathbb{E}_X |f(X) - S[\gamma](X)|^2 + \Omega(\gamma)$$

Optimum of BP training

Let $f \in L^2(\mu)$ and $\beta > 0$, also let $L_{f,\beta}(\gamma) = \left\| |f - S[\gamma]| \right\|^2_{L^2(\mu)} + \beta \|\gamma\|^2_{L^2(\lambda)}$.

Theorem [7]

For all such f there exists a ρ admissable function such that

$$\arg\min_{\gamma \in L^2(\lambda)} L_{f,\beta}(\gamma) = R_\rho[f]$$

Possible improvements

- given an activation function, how to find the "best" (admissible) ridgelet function?
- besides toy examples, there is no empirical proof that this works in practice (in progress)
- can we extend this to deep networks?
- can we say more if we have some restrictions on the input data

Bibliography I

Emmanuel Jean Candes.

Ridgelets: theory and applications.

PhD thesis, Stanford University Stanford, 1998.

George Cybenko.

Approximation by superpositions of a sigmoidal function.

Mathematics of control, signals and systems, 2(4):303–314, 1989.

Geoffrey E Hinton.

Connectionist learning procedures.

In *Machine learning*, pages 555–610. Elsevier, 1990.

Kurt Hornik.

Approximation capabilities of multilayer feedforward networks.

Neural networks, 4(2):251–257, 1991.

Bibliography II

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang.

The expressive power of neural networks: A view from the width.

In *Advances in neural information processing systems*, pages 6231–6239, 2017.

Noboru Murata.

An integral representation of functions using three-layered networks and their approximation bounds.

Neural Networks, 9(6):947–956, 1996.

Sho Sonoda, Isao Ishikawa, Masahiro Ikeda, Kei Hagihara, Yoshihiro Sawano, Takuo Matsubara, and Noboru Murata.

The global optimum of shallow neural network is attained by ridgelet transform.

arXiv preprint arXiv:1805.07517, 2018.

Bibliography III

Sho Sonoda and Noboru Murata.

Neural network with unbounded activation functions is universal approximator.

Applied and Computational Harmonic Analysis, 43(2):233–268, 2017.