ITAT 2016 Proceedings, CEUR Workshop Proceedings Vol. 1649, pp. 118–122
http://ceur-ws.org/Vol-1649, Series ISSN 1613-0073, © 2016 V. Kůrková

Multivariable Approximation by Convolutional Kernel Networks

Věra Kůrková

Institute of Computer Science, Academy of Sciences of the Czech,
vera@cs.cas.cz,
WWW home page: http://www.cs.cas.cz/∼vera

Abstract: Computational units induced by convolutional kernels together with biologically inspired perceptrons belong to the most widespread types of units used in neurocomputing. Radial convolutional kernels with varying widths form RBF (radial-basis-function) networks and these kernels with fixed widths are used in the SVM (support vector machine) algorithm. We investigate suitability of various convolutional kernel units for function approximation. We show that properties of Fourier transforms of convolutional kernels determine whether sets of input-output functions of networks with kernel units are large enough to be universal approximators. We compare these properties with conditions guaranteeing positive semidefinitness of convolutional kernels.

1 Introduction

Computational units induced by radial and convolutional kernels together with perceptrons belong to the most widespread types of units used in neurocomputing. In contrast to biologically inspired perceptrons [15], localized radial units [1] were introduced merely due to their good mathematical properties. Radial-basis-function units (RBF) computing spherical waves were followed by kernel units [7]. Kernel units in the most general form include all types of computational units, which are functions of two vector variables: an input vector and a parameter vector. However, often the term kernel unit is reserved merely for units computing symmetric positive semidefinite functions of two variables. Networks with these units have been widely used for classification with maximal margin by the support vector machine algorithm (SVM) [2] as well as for regression [21].

Other important kernel units are units induced by convolutional kernels in the form of translations of functions of one vector variable. Isotropic RBF units can be viewed as non symmetric kernel units obtained from convolutional radial kernels by adding a width parameter. Variability of widths is a strong property. It allows to apply arguments based on classical results on approximation of functions by sequences of their convolutions with scaled bump functions to prove universal approximation capabilities of many types of RBF networks [16, 17]. Moreover, some estimates of rates of approximation by RBF networks exploit variability of widths [9, 10, 13].

On the other hand, symmetric positive semidefinite kernels (which include some classes of RBFs with fixed widths parameters) benefit from geometrical properties of reproducing kernel Hilbert spaces (RKHS) generated by these kernels. These properties allow an extension of the maximal margin classification from finite dimensional spaces also to sets of data which are not linearly separable by embedding them into infinite dimensional spaces [2]. Moreover, symmetric positive semidefinite kernels generate stabilizers in the form of norms on RKHSs suitable for modeling generalization in terms of regularization [6] and enable characterizations of theoretically optimal solutions of learning tasks [3, 19, 11].

Arguments proving the universal approximation property of RBF networks using sequences of scaled kernels might suggest that variability of widths is necessary for the universal approximation. However, for the special case of the Gaussian kernel, the universal approximation property holds even when the width is fixed and merely centers are varying [14, 12].

On the other hand, it is easy to find some examples of positive semidefinite kernels such that sets of input-output functions of shallow networks with units generated by these kernels are too small to be universal approximators. For example, networks with product kernel units of the form $K(x,y) = k(x)k(y)$ generate as input-output functions only scalar multiples $ck(x)$ of the function k.

In this paper, we investigate capabilities of networks with one hidden layer of convolutional kernel units to approximate multivariable functions. We show that a crucial property influencing whether sets of input-output functions of convolutional kernel networks are large enough to be universal approximators is behavior of the Fourier transform of the one variable function generating the convolutional kernel. We give a necessary and sufficient condition for universal approximation of kernel networks in terms of the Fourier transforms of kernels. We compare this condition with properties of kernels guaranteeing their positive definitness. We illustrate our results by examples of some common kernels such as Gaussian, Laplace, parabolic, rectangle, and triangle.

The paper is organized as follows. In section 2, notations and basic concepts on one-hidden-layer networks and kernel units are introduced. In section 3, a necessary and sufficient condition on a convolutional kernel that guarantees that networks with units induced by the kernel have the universal approximation property. In section 4 this condition is compared with a condition guaranteeing that a kernel is positive semidefinite and some examples of kernels satisfying both or one of these conditions are given. Section 5 is a brief discussion.

2 Preliminaries

Radial-basis-function networks as well as kernel models belong to the class of one-hidden-layer networks with one linear output unit. Such networks compute input-output functions from sets of the form

$$\operatorname{span} G = \left\{ \sum_{i=1}^{n} w_i g_i \,|\, w_i \in \mathbb{R},\, g_i \in G, n \in \mathbb{N}_+ \right\},$$

where the set G is called a *dictionary* [8], and $\mathbb{R}$, $\mathbb{N}_+$ denote the sets of real numbers and positive integers, resp. Typically, dictionaries are parameterized families of functions modeling computational units, i.e., they are of the form

$$G_K(X,Y) = \{K(.,y) : X \to \mathbb{R} \,|\, y \in Y\}$$

where $K : X \times Y \to \mathbb{R}$ is a function of two variables, an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter $y \in Y \subseteq \mathbb{R}^s$. Such functions of two variables are called *kernels*. This term, derived from the German term "kern", has been used since 1904 in theory of integral operators [18, p.291].

An important class of kernels are *convolutional kernels* which are obtained by translations of one-variable functions $k : \mathbb{R}^d \to \mathbb{R}^d$ as

$$K(x,y) = k(x-y).$$

Radial convolutional kernels are convolutional kernels obtained as translations of radial functions, i.e., functions of the form

$$k(x) = k_1(\|x\|),$$

where $k_1 : \mathbb{R}_+ \to \mathbb{R}$.

The *convolution* is an operation defined as

$$f * g(x) = \int_{\mathbb{R}^d} f(y-x)g(y)dy = \int_{\mathbb{R}^d} f(y)g(x-y)dy$$

[20, p.170].

Recall, that a kernel $K : X \times X \to \mathbb{R}$ is called *positive semidefinite* if for any positive integer m, any $x_1, \dots, x_m \in X$ and any $a_1, \dots, a_m \in \mathbb{R}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(x_i, x_j) \geq 0.$$

Similarly, a function of one variable $k : \mathbb{R}^d \to \mathbb{R}$ is called positive semidefinite if for any positive integer m, any $x_1, \dots, x_m \in X$ and any $a_1, \dots, a_m \in \mathbb{R}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j k(x_i - x_j) \geq 0.$$

For symmetric positive semidefinite kernels K, the sets $\operatorname{span} G_K(X)$ of input-output functions of networks with units induced by the kernel K are contained in Hilbert spaces defined by these kernels. These spaces are called *reproducing kernel Hilbert spaces (RKHS)* and denoted $\mathscr{H}_K(X)$. They are formed by functions from

$$\operatorname{span} G_K(X) = \operatorname{span}\{K_x \,|\, x \in X\},$$

where

$$K_x(.) = K(x,.),$$

together with limits of their Cauchy sequences in the norm $\|.\|_K$. The norm $\|.\|_K$ is induced by the inner product $\langle .,. \rangle_K$, which is defined on

$$G_K(X) = \{K_x \,|\, x \in X\}$$

as

$$\langle K_x, K_y \rangle_K = K(x,y).$$

So $\operatorname{span} G_K(X) \subset \mathscr{H}_K(X)$.

3 Universal approximation capability of convolutional kernel networks

In this section, we investigate conditions guaranteeing that sets of input-output functions of convolutional kernel networks are large enough to be universal approximators.

The universal approximation property is formally defined as density in a normed linear space. A class of one-hidden-layer networks with units from a dictionary G is said to have the *universal approximation property in a normed linear space* $(\mathscr{X}, \|.\|_{\mathscr{X}})$ if it is *dense* in this space, i.e., $\operatorname{cl}_{\mathscr{X}} \operatorname{span} G = \mathscr{X}$, where $\operatorname{span} G$ denotes the *linear span* of G and $\operatorname{cl}_{\mathscr{X}}$ denotes the closure with respect to the topology induced by the norm $\|.\|_{\mathscr{X}}$. More precisely, for every $f \in X$ and every $\varepsilon > 0$ there exist a positive integer n, $g_1, \dots, g_n \in G$, and $w_1, \dots, w_n \in \mathbb{R}$ such that

$$\|f - \sum_{i=1}^{n} w_i g_i\|_{\mathscr{X}} < \varepsilon.$$

Function spaces where the universal approximation property has been of interest are spaces $(\mathbb{C}(X), \|.\|_{\sup})$ of continuous functions on subsets X of $\mathbb{R}^d$ (typically compact) with the supremum norm

$$\|f\|_{\sup} = \sup_{x \in X} |f(x)|$$

and spaces $(\mathscr{L}^p(\mathbb{R}^d), \|.\|_{\mathscr{L}^p})$ of functions on $\mathbb{R}^d$ with finite $\int_{\mathbb{R}^d} |f(y)|^p dy$ and the norm

$$\|f\|_{\mathscr{L}^p} = \left(\int_{\mathbb{R}^d} |f(y)|^p dy \right)^{1/p}.$$

Recall that the *d-dimensional Fourier transform* is an isometry on $\mathscr{L}^2(\mathbb{R}^d)$ defined on $\mathscr{L}^2(\mathbb{R}^d) \cap \mathscr{L}^1(\mathbb{R}^d)$ as

$$\hat{f}(s) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-ix \cdot s} f(x)\, dx$$

and extended to $\mathscr{L}^2(\mathbb{R}^d)$ [20, p.183].

Note that the Fourier transform of an even function is real and the Fourier transform of a radial function is radial. If $k \in cL^1(\mathbb{R}^d)$, then $\hat{k}$ is uniformly continuous and with increasing frequencies converges to zero, i.e.,

$$\lim_{\|s\|\to\infty} \hat{k}(s) = 0.$$

The following theorem gives a necessary and sufficient condition on a convolutional kernel that guarantees that the class of input-output functions computable by networks with units induced by the kernel can approximate arbitrarily well all functions in $\mathscr{L}^2(\mathbb{R}^d)$. The condition is formulated in terms of the size of the set of frequencies for which the Fourier transform is equal to zero. By λ is denoted the Lebesgue measure.

Theorem 1. *Let d be a positive integer, $k \in \mathscr{L}^1(\mathbb{R}^d) \cap \mathscr{L}^2(\mathbb{R}^d)$ be even, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be defined as $K(x,y) = k(x-y)$, and $X \subseteq \mathbb{R}^d$ be Lebesgue measurable. Then $\operatorname{span} G_K(X)$ is dense in $(\mathscr{L}^2(X), \|.\|_{\mathscr{L}^2})$ if and only if $\lambda(\{s \in \mathbb{R}^d \,|\, \hat{k}(s) = 0\}) = 0$.*

Proof. First, we prove the necessity. To prove it by contradiction, assume that $\lambda(S) \neq 0$. Take any function $f \in \mathscr{L}^2(\mathbb{R}^d) \cap \mathscr{L}^1(\mathbb{R}^d)$ with a positive Fourier transform (for example, f can be the Gaussian). Let $\varepsilon > 0$ be such that

$$\varepsilon < \int_{\mathbb{R}^d} \hat{f}(s)^2 ds.$$

Assume that there exists n, $w_i \in \mathbb{R}$, and $y_i \in \mathbb{R}^d$ such that

$$\|f - \sum_{j=1}^n w_i k(.-y_i)\|_{\mathscr{L}_2} < \varepsilon.$$

Then by the Plancherel Theorem [20, p.188],

$$\|\hat{f} - \sum_{j=1}^n w_i \widehat{k(.-y_i)}\|^2_{\mathscr{L}_2} = \|\hat{f} - \sum_{j=1}^n \bar{w}_i \hat{k}\|^2_{\mathscr{L}_2},$$

where $\bar{w}_i = w_i e^{iy_i}$. Hence

$$\|\hat{f} - \sum_{j=1}^n \bar{w}_i \hat{k}\|^2_{\mathscr{L}_2} =$$

$$\int_{\mathbb{R}^d \setminus S} \left(\hat{f}(s) - \sum_{j=1}^n \bar{w}_i \hat{k}(s)\right)^2 ds + \int_S \hat{f}(s)^2 ds > \varepsilon,$$

which is a contradiction.

To prove the sufficiency, we first assume that $X = \mathbb{R}^d$. We prove it by contradiction, so we suppose that

$$\mathrm{cl}_{\mathscr{L}^2} \operatorname{span} G_K(\mathbb{R}^d) = \mathrm{cl}_{\mathscr{L}^2} \operatorname{span}\{K(.,y) \,|\, y \in \mathbb{R}^d\} \neq \mathscr{L}^2(\mathbb{R}^d).$$

Then by the Hahn-Banach Theorem [20, p. 60] there exists a bounded linear functional l on $\mathscr{L}^2(\mathbb{R}^d)$ such that for all $f \in \mathrm{cl}_{\mathscr{L}^2} \operatorname{span} G_K(\mathbb{R}^d)$, $l(f) = 0$ and for some $f_0 \in \mathscr{L}^2(\mathbb{R}^d) \setminus \mathrm{cl}_{\mathscr{L}^2} \operatorname{span} G_K(\mathbb{R}^d)$, $l(f_0) = 1$. By the Riesz Representation Theorem [5, p.206], l can be expressed as an inner product with some $h \in \mathscr{L}^2(\mathbb{R}^d)$.

As k is even, for all $y \in \mathbb{R}^d$,

$$\langle h, K(.,y)\rangle = \int_{\mathbb{R}^d} h(x)k(x-y)dx =$$

$$\int_{\mathbb{R}^d} h(x)k_1(y-x)dx = h * k_1(x) = 0.$$

By the Young Inequality for convolutions $h * k \in \mathscr{L}^2(\mathbb{R}^d)$ and so by the Plancherel Theorem [20, p.188],

$$\|\widehat{h * k_1}\|_{\mathscr{L}^2} = 0.$$

As

$$\widehat{h * k_1} = \frac{1}{(2\pi)^{d/2}} \hat{h}\hat{k}$$

[20, p.183], we have $\|\hat{h}\hat{k}\|_{\mathscr{L}^2} = 0$ and so

$$\int_{\mathbb{R}^d} (\hat{h}(s)\hat{k}(s))^2 ds = 0.$$

As the set

$$S = \{s \in \mathbb{R}^d \,|\, \hat{k}(s) = 0\}$$

has Lebesgue measure zero we have

$$\int_{\mathbb{R}^d} \hat{h}(s)^2 \hat{k}(s)^2 ds = \int_{\mathbb{R}^d \setminus S} \hat{h}(s)^2 \hat{k}(s)^2 ds = 0.$$

As for all $s \in \mathbb{R}^d \setminus S$, $\hat{k}(s)^2 > 0$, we have $\|\hat{h}\|^2_{\mathscr{L}^2} ds = 0$. So $\|h\|_{\mathscr{L}^2} = 0$ and hence by the Cauchy-Schwartz Inequality we get

$$1 = l(f_0) = \int_{\mathbb{R}^d} f_0(y)\, h(y) dy \leq \|f_0\|_{\mathscr{L}^2} \, \|h\|_{\mathscr{L}^2} = 0,$$

which is a contradiction.

Extending a function f from $\mathscr{L}^2(X)$ to $\bar{f}$ from $\mathscr{L}^2(\mathbb{R}^d)$ by setting its values equal to zero outside of X and restricting approximations of $\bar{f}$ by functions from $\operatorname{span} G_K(\mathbb{R}^d)$ to X, we get the statement for any Lebesgue measurable subset X of $\mathbb{R}^d$.

□

Theorem 1 shows that sets of input-output functions of convolutional kernel networks are large enough to approximate arbitrarily well all $\mathscr{L}^2$-functions if and only if the Fourier transform of the function k is almost everywhere non-zero.

Theorem 1 implies that when $\hat{k}(s)$ is equal to zero for all s such that $\|s\| \geq r$ for some $r > 0$ (the Fourier transform is band-limited), then the set $\operatorname{span} G_K(\mathbb{R}^d)$ is too small to have the universal approximation capability. In the next section we show, that some of such kernels are positive semidefinite. So they can be used for classification by the SVM algorithm but they are not suitable for function approximation.

4 Positive semidefinitness and universal approximation property

In this section, we compare a condition on positive semidefinitness of a convolutional kernel with the condition on the universal approximation property derived in the previous section.

As the inverse Fourier transform of a convolutional kernel can be expressed as

$$K(x,y) = k(x-y) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{k}(s) e^{i(x-y)\cdot s} ds$$

it is easy to verify that when $\hat{k}$ is positive or non negative than K defined as $K(x, y = k(x-y)$ is positive definite, semidefinite, resp.

Indeed, to verify that $\sum_{j,l=1}^{n} a_j a_l K(x_j, x_l) \geq 0$ we express K in terms of the inverse Fourier transform. Thus we get

$$\sum_{j,l=1}^{n} a_j a_l K(x_j, x_l) = \sum_{j,l=1}^{n} a_j a_l \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{k}(s) e^{i(x_j - x_l)\cdot s} ds =$$

$$\frac{1}{(2\pi)d/2} \int_{\mathbb{R}^d} \left(\sum_{j}^{n} a_j e^{i(x_j)\cdot s} \right) \left(\sum_{l}^{n} a_k e^{-i(x_l)\cdot s} \right) \hat{k}(s) ds =$$

$$\frac{1}{(2\pi)d/2} \int_{\mathbb{R}^d} \left| \sum_{j}^{n} a_j e^{i(x_j)\cdot s} \right|^2 \hat{k}(s) ds \geq 0.$$

The following proposition is well-known (see, e.g., [4]).

Proposition 2. *Let $k \in \mathscr{L}^1(\mathbb{R}^d) \cap \mathscr{L}^2(\mathbb{R}^d)$ be an even function such that $\hat{k}(s) \geq 0$ for all $s \in \mathbb{R}^d$. Then $K(x,y) = k(x-y)$ is positive semidefinite.*

A complete characterization of positive semidefinite bounded continuous kernels follows from the Bochner Theorem.

Theorem 3 (Bochner)**.** *A bounded continuous function $k : \mathbb{R}^d \to \mathbb{C}$ is positive semidefinite iff k is the Fourier transform of a nonnegative finite Borel measure μ, i.e.,*

$$k(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-x\cdot s} \mu(ds).$$

The Bochner Theorem implies that when the Borel measure μ has a distribution function then the condition in Proposition 2 is both sufficient and necessary.

Comparison of the characterization of kernels for which by Theorem 1 one-hidden-layer kernel networks are universal approximators with the condition on positive semidefinitness from Proposition 2 shows that there are positive semidefinite kernels which do not generate networks possessing the universal approximation capability and there also are kernels which are not positive definite but induce networks with the universal approximation property. The first ones are suitable for SVM but not for regression, while the second ones can be used for regression but are not suitable for SVM. In the sequel, we give some examples of such kernels.

A paradigmatic example of a convolutional kernel is the *Gaussian kernel* $g_a : \mathbb{R}^d \to \mathbb{R}$ defined for a width $a > 0$ as

$$\mathrm{g_a} = e^{-a^2 \|.\|^2}.$$

For any fixed width a and any dimension d,

$$\widehat{g_a} = (\sqrt{2}a)^{-d} e^{-1/a^2 \|.\|^2}.$$

So the Gaussian kernel is positive definite and the class of Gaussian kernel networks have the universal approximation property.

The *rectangle kernel* is defined as

$$\mathrm{rect}(x) = 1 \text{ for } x \in (-1/2, 1/2),$$
$$\text{otherwise } \mathrm{rect}(x) = 0.$$

Its Fourier transform is the sinc function

$$\widehat{\mathrm{rect}}(s) = \mathrm{sinc}(s) = \frac{\sin(\pi s)}{\pi s}.$$

So the Fourier transform of rect is not non negative but its zeros form a discrete set of the Lebesgue measure zero. Thus the rectangle kernel is not positive semidefinite but induces class of networks with the universal approximation property. On the other hand, the Fourier transform of sinc is the rectangle kernel and thus it is positive semidefinite, but does not induce networks with the universal approximation property.

The *Laplace kernel* is defined for any $a > 0$ as

$$l(x) = e^{-a|x|}.$$

Its Fourier transforms is positive as

$$\hat{l}(s) = \frac{2a}{a^2 + (2\pi s)^2}.$$

The *triangle kernel* is defined as

$$\mathrm{tri}(x) = 2x - 1/2 \text{ for } x \in (-1/2, 0),$$
$$\mathrm{tri}(x) = -2(x + 1/2) \text{ for } x \in (0, 1/2),$$
$$\text{otherwise } \mathrm{tri}(x) = 0.$$

Its Fourier transforms is positive as

$$\widehat{\mathrm{tri}}(s) = \mathrm{sinc}(s)^2 = \left(\frac{\sin(\pi s)}{\pi s} \right)^2.$$

Thus both the Laplace and the triangle kernel are positive definite and induce networks having the universal approximation property.

The *parabolic (Epinechnikov) kernel* is defined

$$\mathrm{epi}(x) = \tfrac{3}{4}(1 - x^2) \text{ for } x \in (-1, 1),$$
$$\text{otherwise } \mathrm{epi}(x) = 0.$$

Its Fourier transforms is

$$\widehat{\mathrm{epi}}(s) = \frac{3}{s^3}(\sin(s) - \frac{1}{2}s\cos(s)) \text{ for } s \neq 0,$$
$$\widehat{\mathrm{epi}}(s) = 1 \text{ for } s = 0.$$

So the parabolic kernel is not positive semidefinite but induces networks with the universal approximation property.

5 Discussion

We investigated effect of properties of the Fourier transform of a kernel function on suitability of the convolutional kernel for function approximation (universal approximation property) and for maximal margin classification algorithm (positive semidefinitness). We showed that these properties depend on the way how the Fourier transform converges with increasing frequencies to infinity. For the universal approximation property, the Fourier transform can be negative but cannot be zero on any set of frequencies of non-zero Lebesgue measure. On the other hand, functions with non-negative Fourier transforms are positive semidefinite even if they are compactly supported. We illustrated our results by the paradigmatic example of the multivariable Gaussian kernel and by some one-dimensional examples. Multivariable Gaussian is a product of one variable functions and thus its multivariable Fourier transform can be computed using transforms of one-variable Gaussians. Fourier transforms of other radial multivariable kernels are more complicated, their expressions include Bessel functions and the Hankel transform. Investigation of properties of Fourier transforms of multivariable radial convolutional kernels is subject of our future work.

Acknowledgments. This work was partially supported by the grant GAČR 15-181085 and institutional support of the Institute of Computer Science RVO 67985807.

References

[1] D. S. Broomhead and D. Lowe. Error bounds for approximation with neural networks. *Complex Systems*, 2:321–355, 1988.

[2] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[3] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1–49, 2002.

[4] F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, Cambridge, 2007.

[5] A. Friedman. *Modern Analysis*. Dover, New York, 1982.

[6] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10:1455–1480 (AI memo 1606), 1998.

[7] F. Girosi and T. Poggio. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247(4945):978–982, 1990.

[8] R. Gribonval and P. Vandergheynst. On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. *IEEE Trans. on Information Theory*, 52:255–261, 2006.

[9] P. C. Kainen, V. Kůrková, and M. Sanguineti. Complexity of Gaussian radial basis networks approximating smooth functions. *J. of Complexity*, 25:63–74, 2009.

[10] P. C. Kainen, V. Kůrková, and M. Sanguineti. Dependence of computational models on input dimension: Tractability of approximation and optimization tasks. *IEEE Transactions on Information Theory*, 58:1203–1214, 2012.

[11] V. Kůrková. Neural network learning as an inverse problem. *Logic Journal of IGPL*, 13:551–559, 2005.

[12] V. Kůrková and P. C. Kainen. Comparing fixed and variable-width gaussian networks. *Neural Networks*, 57:23–28, 2014.

[13] V. Kůrková and M. Sanguineti. Model complexities of shallow networks representing highly varying functions. *Neurocomputing*, 171:598–604, 2016.

[14] H. N. Mhaskar. Versatile Gaussian networks. In *Proceedings of IEEE Workshop of Nonlinear Image Processing*, pages 70–73, 1995.

[15] M. Minsky and S. Papert. *Perceptrons*. MIT Press, 1969.

[16] J. Park and I. Sandberg. Universal approximation using radial–basis–function networks. *Neural Computation*, 3:246–257, 1991.

[17] J. Park and I. Sandberg. Approximation and radial basis function networks. *Neural Computation*, 5:305–316, 1993.

[18] A. Pietsch. *Eigenvalues and s-Numbers*. Cambridge University Press, Cambridge, 1987.

[19] T. Poggio and S. Smale. The mathematics of learning: dealing with data. *Notices of AMS*, 50:537–544, 2003.

[20] W. Rudin. *Functional Analysis*. Mc Graw-Hill, 1991.

[21] B. Schölkopf and A. J. Smola. *Learning with Kernels – Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, 2002.