Translation-Invariant Kernels for Multivariable Approximation

Věra Kůrková and David Coufal

***Abstract*—Suitability of shallow (one-hidden-layer) networks with translation-invariant kernel units for function approximation and classification tasks is investigated. It is shown that a critical property influencing capabilities of kernel networks is how the Fourier transforms of kernels converge to zero. The Fourier transforms of kernels suitable for multivariable approximation can have negative values, but must be almost everywhere nonzero. In contrast, the Fourier transforms of kernels suitable for maximal margin classification must be everywhere nonnegative, but can have large sets where they are equal to zero (e.g., they can be compactly supported). Behavior of Fourier transforms of multivariable kernels is analyzed using the Hankel transform. The general results are illustrated by examples of both uni- and multivariable kernels (such as Gaussian, Laplace, rectangle, sinc and cut power kernel).**

***Index Terms*—Translation-invariant kernels; radial kernels; function approximation; classification; Fourier and Hankel transforms.**

I. Introduction

Artificial neural networks were introduced as computational models composed from perceptrons representing simplified models of neurons [1]. Perceptrons compute highly nonlocal functions in the form of plane waves with shapes defined by activation functions (originally mostly sigmoidal ones, recently also piecewise linear rectifiers). A disadvantage of perceptron networks is the lack of transparency of representations in the form of plane waves as clearly expressed in the classical monograph [2, p. 676]: "But always the use of plane waves fails to exhibit clearly the domains of dependence and the role of characteristics."

As an alternative to biologically inspired perceptrons, various types of computational units were proposed because of their good mathematical properties. Radial-basis-functions (RBF) in the form of spherical waves [3] were followed by more general hyper-basis-functions [4] induced by Green functions of differential operators playing roles of smoothing in regularization techniques. Later, units formed by symmetric positive definite kernels became popular. Geometrical properties of Hilbert spaces induced by these kernels play a crucial role in classification by the support vector machine (SVM) algorithm [5], in regularization [6], [7], [8], and a variety of learning techniques [9], [10], [11], [12], [13].

V. Kůrková and D. Coufal are with the Department of Machine Learning, Institute of Computer Science of the Czech Academy of Sciences, 182 07 Prague, Pod Vodárenskou věží, 2, Czech Republic (e-mail: vera@cs.cas.cz, david.coufal@cs.cas.cz).

This work was partially supported by the Czech Science Foundation grant GA18-23827S and institutional support of the Institute of Computer Science RVO 6798580.

Manuscript received Month XX, 2019; revised Month XX, 2020.

The term "kernel" was introduced by Hilbert in 1904 [14, p. 291] for functions of two variables $K(x, y)$ forming "kernels of integral operators" $\int f(y)K(x, y)\,dy$, which model many phenomena investigated in physics. Some of these kernels are named for the mathematicians who studied them - e.g., Gauss, Weierstrass, Abel, Laplace, Poisson. Also computational units being functions of two vector variables (input and parameter) can be seen as kernels. Many kernels used in various branches of applied mathematics turned out to be suitable for generating computational units. Some of them became popular in applications. In particular, translation-invariant kernels defined as translations of one-variable functions generate computational units possessing many useful properties. Translation-invariant kernels formed by radial kernels in the shape of "bump" functions have localized character. The most widely used kernel is the Gaussian. With varying widths, it is the most common kernel in RBF networks and with fixed widths in the SVM algorithm. Also, inverse multiquadric and thin-plate spline have been popular in kernel models. Other translation-invariant kernels with interesting mathematical properties are Laplace and cut power kernels, rectangular pulse and sinc function [15], [16].

In this paper, we explore general translation-invariant kernels with the aim of selecting those that are suitable for use as computational units. The first important factor influencing this selection is capability of kernel networks to express sufficiently large sets as input-output functions, ideally dense enough to find an input-output function arbitrarily close to any reasonable function. The second factor is the applicability of methods for regularization that made RBF networks and SVM so popular. The essential properties of kernels needed for theoretically justified application of these methods is symmetry and positive definitness. Kernels with these properties induce reproducing kernel Hilbert spaces whose geometrical structure is essential in mathematical theory of regularization and SVM.

We analyze suitability of kernel computational units for function approximation and/or classification in terms of properties of their Fourier transforms. For typical kernels, the transforms are real, uniformly continuous and converging with increasing frequencies to zero. They can be classified according to the form of this convergence: some of them are band-limited (equal zero above certain frequency), some are positive monotonically decreasing, while others are oscillating between positive and negative values.

We show that for function approximation, kernels with almost everywhere nonzero Fourier transforms are suitable, while for classification with maximal margin (SVM), kernels with nonnegative Fourier transforms are needed as such kernels are positive definite.

In neurocomputing, functions typically depend on many variables. Thus we investigate Fourier transforms of multivariable kernels. Fourier transforms of radial functions are radial, but often they cannot be expressed analytically. Their expressions use the Hankel transform, which is an integral transform formulated in terms of special functions, called Bessel functions of the first kind. Using the Hankel transform, we analyze multivariable Fourier transforms of kernels of interest such as Laplace and cut power kernels or circ function.

Our analysis shows that one cannot take for granted that computational units which are suitable for classification tasks performed by SVM [5] with good generalization capabilities are also suitable for function approximation. Necessary and sufficient conditions for these two tasks are different and we present examples of kernels which satisfy merely one of these conditions. We also characterize translation-invariant kernels, which generate networks with both benefits: sufficient expressibility needed for function approximation and symmetry and positive definiteness needed for maximal margin classification. A preliminary version of some results from this paper appeared in a Czech-Slovak conference proceedings [17] and in a conference proceedings [18]. In [17], a sketch of an alternative proof of Theorem 3.1 was given and in both [17] and [18], some examples of kernels with various properties of Fourier transforms were demonstrated. In this paper, results from these conference papers are presented with more mathematical rigor. They are combined to create a unifying framework characterizing all four classes of kernels with respect to universality and positive definiteness. Analysis of the case of multivariate kernels, which requires expressions of their Fourier transforms in terms of the Hankel transform, is included.

The paper is organized as follows. Section II contains notation and a background material on kernel units, one-hidden-layer networks, and Fourier and Hankel transforms. Section III is devoted to approximation of functions by kernel networks. In Section IV, conditions for function approximation are compared with conditions on maximal margin classification and regularization. In Section V, properties of multivariable kernels are analyzed. Section VI contains some concluding remarks. Appendix contains our alternative argument of the Wiener Closure Theorem.

II. Preliminaries

The set of input-output functions of a one-hidden-layer network with one linear output unit has the form

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

where the set G is called a *dictionary* [19], and $\mathbb{R}$, $\mathbb{N}_+$ denote the sets of real numbers and positive integers, respectively. Recently, one-hidden-layer networks became called *shallow* to distinguish them from deep ones with more than one hidden layers.

Dictionaries can be formally described as

$$G_\psi = G_\psi(X, Y) := \{\psi(x, y) : X \to \mathbb{R} \,|\, y \in Y\}, \quad (2)$$

where $\psi : X \times Y \to \mathbb{R}$ is a function of two variables, $x \in X \subseteq \mathbb{R}^d$ an input vector and $y \in Y \subseteq \mathbb{R}^s$ a parameter.

In this paper, we focus on dictionaries of *translation-invariant (shift-invariant) kernel units*, which are generated by translations of suitable one-variable functions. For a function $k : \mathbb{R}^d \to \mathbb{R}$, we denote by $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the function of two variables defined as

$$K(x, y) = k(x - y). \quad (3)$$

We denote by

$$G_K = G_K(X) := \{K(x, y) : X \to \mathbb{R} \,|\, y \in X\}, \quad (4)$$

the dictionary of kernel units induced by the kernel K.

An important subclass of translation-invariant kernels are *radial translation-invariant kernels*. A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is called *radial* if it can be expressed as

$$\Phi(x) = \varphi(\|x\|), \quad (5)$$

where $\varphi : [0, \infty) \to \mathbb{R}$ is a one-variable function and $\|x\|$ is a norm on $\mathbb{R}^d$, usually the Euclidean one. Thus a radial translation-invariant kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ has the form

$$K(x, y) = \varphi(\|x - y\|), \quad (6)$$

where $\varphi : [0, \infty) \to \mathbb{R}$.

Radial-basis-function units (RBF) compute translations of radial functions. When they have a *fixed width*, they have the form

$$\varphi(\|x - a\|), \quad (7)$$

where $a \in \mathbb{R}^d$ is called a *center* and $\varphi : [0, \infty) \to \mathbb{R}$ is a one-variable function, typically such that $\lim_{r \to \infty} \varphi(r) = 0$. *RBF with variable widths* $b > 0$ compute functions

$$\varphi_b(\|x - a\|) = \varphi(\|x - a\|/b). \quad (8)$$

Recall [20, p. 30], that a kernel $K : X \times X \to \mathbb{R}$, where $X \subseteq \mathbb{R}^d$, is called *positive definite* if for any positive integer $n \in \mathbb{N}_+$, any $x_1, \ldots, x_n \in X$ and any $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) \geq 0. \quad (9)$$

Similarly, a real function $k : X \to \mathbb{R}$, where $X \subseteq \mathbb{R}^d$, is called *positive definite* if for any positive integer $n \in \mathbb{N}_+$, any $x_1, \ldots, x_n \in X$ and any $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j k(x_i - x_j) \geq 0. \quad (10)$$

A kernel is called *strictly positive definite* if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) > 0. \quad (11)$$

Note that in some literature (e.g. in [15], see p. 65), the terms positive semi-definite and positive definite are used instead of positive definite and strictly positive definite, resp. (see Sec. 3.1 in [16, p. 28] or Remark 2.6 in [20, p. 30]). Moreover, indefinite kernels are also studied [21].

When K is symmetric positive definite, then the set $\operatorname{span} G_K(X)$ of input-output functions of networks with units

from the dictionary $G_K(X)$ induced by the kernel K generate a *reproducing kernel Hilbert space (RKHS)* denoted $\mathcal{H}_K(X)$. It is the completion of the preHilbert space

$$\operatorname{span}\{K_x \,|\, x \in X\}, \tag{12}$$

where $K_x(y) = K(x, y)$, by adding the limits of the Cauchy sequences in the norm $\|\cdot\|_K$. This norm is induced by the inner product

$$\langle K_x, K_y \rangle_K = K(x, y). \tag{13}$$

The *convolution* is an operation defined as

$$f * g(x) = \int_{\mathbb{R}^d} f(x-y)g(y)\,dy = \int_{\mathbb{R}^d} f(y)g(x-y)\,dy \tag{14}$$

[22, p. 170].

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

$$\hat{f}(s) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\mathrm{i}x\cdot s} f(x)\,dx \tag{15}$$

and extended to $\mathcal{L}^2(\mathbb{R}^d)$ [22, p. 183].

Recall that the Fourier transform of a radial function is also a radial function which can be expressed in terms of the Hankel transform. The *Hankel transform of order* ν of a function $f : [0, \infty) \to \mathbb{R}$ is defined as

$$\mathscr{H}_\nu\{f(r)\}(s) = \int_0^\infty f(r) J_\nu(sr) r\,dr \tag{16}$$

where $J_\nu : [0, \infty) \to \mathbb{R}$ is the *Bessel function of the first kind of order* $\nu \geq -\frac{1}{2}$ [23]. Bessel functions are obtained as solutions of certain differential equations and play an important role in problems related to wave propagation.

III. Translation-invariant kernel networks for function approximation

When networks are used for function approximation, it is desirable that sets of their input-output functions are large enough so that to any function from a set of interest there exists a sufficiently close input-output function of a network from that class. Formally, a class of networks is said to have the *universal approximation property in a normed linear space* $(\mathcal{X}, \|\cdot\|_\mathcal{X})$ if it is *dense* in this space. In particular, for shallow networks with single linear output and hidden units from a dictionary G this means that $\operatorname{cl}_\mathcal{X} \operatorname{span} G = \mathcal{X}$, where $\operatorname{span} G$ denotes the *linear span* of G and $\operatorname{cl}_\mathcal{X}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_\mathcal{X}$. So for every $f \in \mathcal{X}$ and every $\varepsilon > 0$ there exist a positive integer $n \in \mathbb{N}_+$, $g_1, \ldots, g_n \in G$ and $w_1, \ldots, w_n \in \mathbb{R}$ such that

$$\|f - \sum_{i=1}^n w_i g_i\|_\mathcal{X} < \varepsilon. \tag{17}$$

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

$$\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)| \tag{18}$$

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

$$\|f\|_{\mathcal{L}^p} = \left(\int_{\mathbb{R}^d} |f(y)|^p\,dy \right)^{1/p}. \tag{19}$$

A simple example of a positive definite kernel which is not suitable for function approximation is any product kernel of the form $K(x, y) = k(x)k(y)$, where $k : X \to \mathbb{R}$ is a function of one variable. The set of input-output functions of networks with units induced by a product kernel K contains only scalar multiples of the function k

$$\operatorname{span} G_K(X) = \{c\,k(x) : X \to \mathbb{R} \,|\, c \in \mathbb{R}\} \tag{20}$$

and thus it cannot be dense in $\mathcal{L}^2(X)$.

Expressibility power of RBF networks with varying width has long been known. The universal approximation property in $\mathcal{L}^2(\mathbb{R}^d)$ of one-hidden-layer networks with units of the form

$$\varphi(\|x - a\|/b) \tag{21}$$

was proven in [24], [25] under a rather mild condition $0 \neq \int_\mathbb{R} \varphi(t)\,dt < \infty$ on the "shape" function φ. The proof in [24], [25] is based on a classical result on approximation of functions by sequences of their convolutions with scaled kernels and thus it suggests that variability of widths might be needed for density of RBF networks. RBF networks with varying width parameters can compute much larger sets of functions than RBF networks with fixed widths. So, it is not surprising that variability of width plays an important role in estimates of rates of approximation [26], [27]. However, RBF units with varying widths $\varphi(\|x - a\|/b)$ cannot be expressed as symmetric kernels as the input vector has dimension d and the parameter vector has dimension $d + 1$. In RBF units with fixed width $\varphi(\|x - a\|)$ both input and parameter vector have dimension d and thus they can be expressed as symmetric kernels. Symmetry is an important feature because together with positive definiteness induces the Hilbert space structure of RKHSs. Properties of RKHS enable to prove that the SVM algorithm minimizes the margin between two classes and to characterize theoretically optimal solutions of minimization of regularized empirical error functionals [5], [6], [7], [10].

Nevertheless, in the special case of the Gaussian kernel with a fixed width, the universal approximation capability holds in spaces of continuous functions on compact subsets of $\mathbb{R}^d$. Its proof given in [28] exploits special properties of the Gaussian function (its derivatives are products of Hermite polynomials with the Gaussian) and thus it cannot be extended to other kernels.

A different approach based on properties of the Fourier transform of kernels allows us to characterize general translation invariant-kernels which generate networks possessing the universal approximation capability. The next theorem shows that when kernels with fixed width satisfy certain conditions on sets of frequencies for which their Fourier transforms are equal to zero, then they form networks which are suitable for function approximation.

Recall that the Fourier transform of an even function is real and the Fourier transform of a radial function with respect

to the Euclidean norm is also radial [29]. If $k \in \mathcal{L}^1(\mathbb{R}^d)$, then $\hat{k}$ is uniformly continuous and with increasing frequencies converges to zero, i.e., $\lim_{\|s\|\to\infty} \hat{k}(s) = 0$. By λ^d we denote the Lebesgue measure on $\mathbb{R}^d$.

Theorem 3.1: Let d be a positive integer, $k \in \mathcal{L}^1(\mathbb{R}^d) \cap \mathcal{L}^2(\mathbb{R}^d)$ be even, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the kernel induced by k, i.e., $K(x, y) = k(x - y)$, and $X \subseteq \mathbb{R}^d$ be Lebesgue measurable. Then $\operatorname{span} G_K(X)$ is dense in $(\mathcal{L}^2(X), \|\cdot\|_{\mathcal{L}^2})$ if and only if $\lambda^d(\{s \in \mathbb{R}^d \,|\, \hat{k}(s) = 0\}) = 0$.

Theorem 3.1 is a version of the Wiener Closure Theorem proven in [30, Chapter I]. As Wiener's proof is difficult to follow, in the Appendix we give a more transparent alternative argument. It is based on fundamental theorems from functional analysis (Hahn-Banach and Riesz Representation Theorems) and basic properties of convolution and Fourier transform (Young Inequality, Plancherel Theorem).

Theorem 3.1 characterizes translation-invariant kernels which can be used as units in shallow networks having the universal approximation property. It proves that kernel networks can approximate with any accuracy all $\mathcal{L}^2$-functions if and only if the Fourier transform of the kernel function k is almost everywhere nonzero. In particular, it implies that networks with kernels having Fourier transforms with discrete sets of zeros are suitable for function approximation. On the other hand, shallow networks with kernel units having band-limited Fourier transforms ($\hat{k}(s) = 0$ for all s such that $\|s\| \geq r$ for some $r > 0$) are too small to express sufficiently large sets of functions for the universal approximation capability. Inspection of our proof of Theorem 3.1 given in the Appendix shows that shallow networks with such kernel units cannot approximate arbitrarily well functions with positive Fourier transforms (e.g., the Gaussian function).

IV. Positive definiteness and/or universal approximation property

In this section, we compare properties of Fourier transforms of positive definite functions with those characterizing the universal approximation property from Theorem 3.1. We illustrate general results by examples of one-dimensional kernels. The more complicated case of multivariable kernels requiring the Hankel transform is deferred to the next section.

Using the expression

$$K(x, y) = k(x - y) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{k}(s) e^{\mathrm{i}(x-y)\cdot s}\, ds \quad (22)$$

of the inverse Fourier transform of a translated function, it is easy to prove the following well-known proposition (see, e.g., [31]).

Proposition 4.1: Let $k \in \mathcal{L}^1(\mathbb{R}^d) \cap \mathcal{L}^2(\mathbb{R}^d)$ be even and $\hat{k}(s) \geq 0$ for all $s \in \mathbb{R}^d$. Then k is positive definite and therefore so is the kernel $K(x, y) = k(x - y)$.

A complete characterization of complex-valued positive definite continuous kernels in terms of Fourier transforms of finite Borel measure was proven by Bochner (see, e.g., [32, p. 220] or [16, p. 31]).

Theorem 4.2 (Bochner): Let $k : \mathbb{R}^d \to \mathbb{C}$ be continuous. Then k is positive definite if and only if there exists a nonnegative finite Borel measure μ such that k is its Fourier transform, i.e.,

$$k(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-\mathrm{i}x\cdot s} \mu(ds). \quad (23)$$

Theorem 4.2 implies that when a Borel measure μ has a density w.r.t. the Lebesgue measure λ^d then the condition on the Fourier transform of an even function assumed in Proposition 4.1 is both sufficient and necessary. The following theorem from [16, Theorem 3.5, p. 33] characterizes strictly positive definite continuous integrable functions.

Theorem 4.3: Let $k \in \mathcal{L}^1(\mathbb{R}^d)$ be continuous. Then k is strictly positive definite if and only if k is bounded and its Fourier transform is nonnegative and not identically equal to zero.

For our purposes, we would need a result addressing positive definite functions, not only the strict ones. Although in [15, Theorem 6.11, p. 74] the result is proven for the strict variant, an inspection of the converse part of its proof shows that nonnegativity of $\hat{k}$ is also necessary condition for (nonstrict) positive definiteness of k. Thus an even, continuous, and bounded function $k \in \mathcal{L}^1(\mathbb{R}^d) \cap \mathcal{L}^2(\mathbb{R}^d)$ whose Fourier transform has some negative values, cannot be positive definite. On the other hand, networks with units induced by such k can posses the universal approximation property if the set of zeros of its Fourier transform is negligible (has Lebesgue measure zero).

A comparison of the two conditions on a kernel, the one from Theorem 3.1 for the universal approximation and the one from Theorem 4.3 on positive definiteness implies characterization of classes of kernels suitable for function approximation, but not for SVM and on those suitable for SVM, but not for approximation. In the sequel, we illustrate these general results by some examples.

We start with the paradigmatic example of the *Gaussian kernel*. For a width $b > 0$, we denote

$$g_b(x) = e^{-\frac{x^2}{b^2}}. \quad (24)$$

Its Fourier transform is real and it is well-known that

$$\widehat{g_b}(s) = \frac{b}{\sqrt{2}} e^{-\frac{b^2 s^2}{4}}. \quad (25)$$

Proposition 4.1 and Theorem 3.1 show the good properties of the Gaussian kernel. It is suitable for both SVM and function approximation as it is positive definite and networks with Gaussian units with any fixed width are universal approximators in $\mathcal{L}^2(\mathbb{R}^d)$.

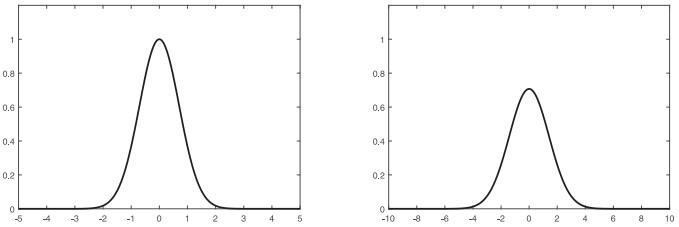


Fig. 1. The Gaussian kernel for $b = 1$ and its Fourier transform.

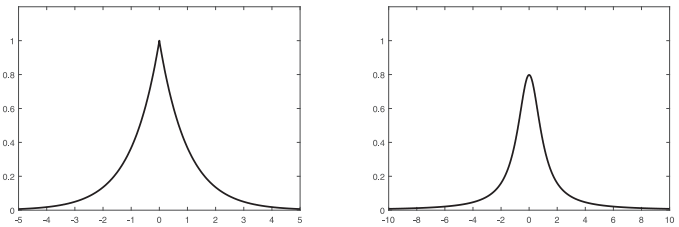


Fig. 2. The Laplace kernel for $b = 1$ and its Fourier transform.

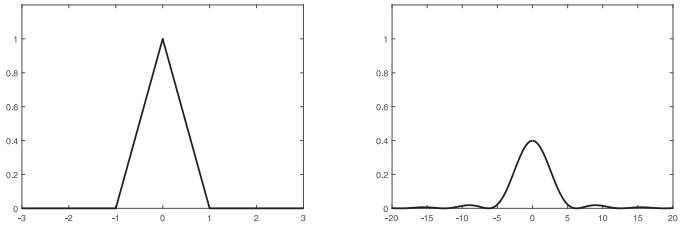


Fig. 3. The triangle kernel for $b = 1$ and its Fourier transform.

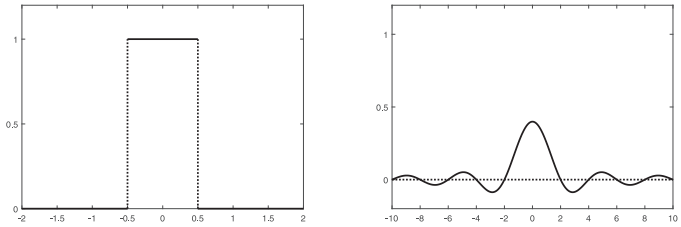


Fig. 4. The rectangle kernel for $b = 1$ and its Fourier transform.

Another example of a translation-invariant kernel possessing both properties assumed in Proposition 4.1 and those assumed in Theorem 3.1 is the *Laplace kernel*. It is defined for $b > 0$ as

$$l_b(x) = e^{-\frac{|x|}{b}}. \tag{26}$$

Its Fourier transform is

$$\hat{l}_b(s) = \sqrt{\frac{2}{\pi}} \frac{b}{1 + b^2 s^2} \tag{27}$$

and thus it is everywhere positive.

For $b > 0$, the *triangle* and the *rectangle kernel* are defined as

$$\begin{aligned} \mathrm{tri}_b(x) &= \max\{0, 1 - |x|/b\} = (1 - |x|/b)_+ \\ \mathrm{rect}_b(x) &= \begin{cases} 1 & \text{for } x \in [-b/2, b/2], \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The Fourier transforms of the triangle and the rectangle kernels are expressed in terms of the *sinc function* defined for all $x \in \mathbb{R}$, $x \neq 0$ as

$$\mathrm{sinc}(x) = \frac{\sin(x)}{x} \quad \text{and} \quad \mathrm{sinc}(0) = 1. \tag{28}$$

The Fourier transforms of tri_b and rect_b have forms

$$\begin{aligned} \widehat{\mathrm{tri}_b}(s) &= \frac{b}{\sqrt{2\pi}} \mathrm{sinc}^2\left(\frac{bs}{2}\right), \\ \widehat{\mathrm{rect}_b}(s) &= \frac{b}{\sqrt{2\pi}} \mathrm{sinc}\left(\frac{bs}{2}\right). \end{aligned}$$

The Fourier transform of the triangle kernel is nonnegative with a discrete set of zeros which is equal to the set of zeros of the sinc function. Therefore the triangle kernel is positive definite and induces networks with the universal approximation property.

An example of a kernel which induces networks with the universal approximation property, but is not positive definite is the rectangle kernel. It is not continuous, but the set of frequencies for which its Fourier transform is zero is discrete and thus its Lebesgue measure is zero. To show directly that it is not positive definite we use the following simple example. Let $n = 3$, $x_1 = 0$, $x_2 = 1/2$, $x_3 = -1/4$, then the 3×3 matrix A with entries $A_{ij} = \mathrm{rect}_1(x_i - x_j)$ is not positive definite as it has a negative eigenvalue, namely $\lambda^* = -0.4142$. Indeed, for the eigenvector $v = (-\sqrt{2}/2, 1/2, 1/2)^T$, we have $\lambda^* = v^T A\, v$ and therefore for $c_i = v_i$, $i, j = 1, 2, 3$, one has $\sum_{i=1}^3 \sum_{j=1}^3 c_i\, c_j \mathrm{rect}_1(x_i - x_j) < 0$.

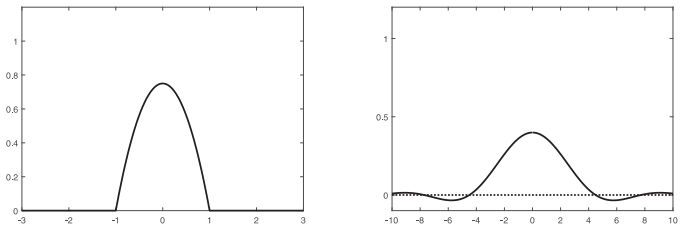


Fig. 5. The Epanechnikov kernel for $b = 1$ and its Fourier transform.

Another example of a kernel which is not positive definite, but induces networks with the universal approximation property is the *cut parabolic (Epanechnikov) kernel*. It is a special case of the cut power kernel. Cut parabolic kernel is defined as

$$\mathrm{epi}_b(x) = \frac{3}{4}(1 - (x/b)^2)_+ \tag{29}$$

For $b = 1$, its Fourier transform can be expressed as

$$\widehat{\mathrm{epi}_1}(s) = \begin{cases} 1 & \text{for } s = 0, \\ \frac{3}{\sqrt{2\pi}} \frac{\sin(s) - s\cos(s)}{s^3} & \text{otherwise,} \end{cases} \tag{30}$$

and for $b > 0$,

$$\widehat{\mathrm{epi}_b}(s) = b\, \widehat{\mathrm{epi}_1}(bs). \tag{31}$$

So the Fourier transform has negative values, the kernel is continuous and therefore by Theorem 4.3 it cannot be positive definite. As the set of zeros of the Fourier transform is discrete, by Theorem 3.1 the cut parabolic kernel induces networks possessing the universal approximation property.

Another interesting kernel is the *squared sinc* kernel - sinc^2. Its Fourier transform is a properly scaled and stretched triangular pulse which is nonnegative and band limited (see Fig. 3 for the Fourier transform pair). Thus sinc^2 is positive definite, but kernel networks induced by sinc^2 are not universal approximators. Note that sinc has similar properties as sinc^2, but as $\mathrm{sinc} \notin \mathcal{L}^1(\mathbb{R})$ Theorem 3.1 cannot be applied.

To construct a kernel which is neither positive definite nor induces networks with the universal approximation property we employ spherical Bessel functions. The *spherical Bessel kernel* spB with the width $b = 1$ is defined as

$$\mathrm{spB}(x) = j_0(|x|) + \frac{5}{2} j_2(|x|) = \frac{-3[(x^2-5)\sin(x) + 5x\cos(x)]}{2x^3} \tag{32}$$

where j_0 and j_2 are the *spherical Bessel functions of the first kind* [33, p. 2779]. The Fourier transform of this kernel is

$$\widehat{\mathrm{spB}}(s) = \sqrt{\frac{\pi}{2}} \left(P_0(s) - \frac{5}{2} P_2(s) \right) = \sqrt{\frac{\pi}{2}} \frac{3}{4} (3 - 5s^2) \cdot 1_{[-1,1]} \tag{33}$$

where P_n are the Legendre orthogonal polynomials and $1_{[-1,1]}$ is the indicator function of the interval $[-1, 1]$. This formula follows from [34, Vol. I, p. 123 (8)] and the identity $j_n(x) = \sqrt{\pi/(2x)} J_{n+1/2}(x)$.

The Fourier transform $\widehat{\mathrm{spB}}$ of the spherical Bessel kernel has negative values and is compactly supported (hence it is band limited). Moreover, $\mathrm{spB} \in \mathcal{L}^2(\mathbb{R})$ as

$$\int_{\mathbb{R}} (\mathrm{spB}(x))^2 \, dx = \int_{\mathbb{R}} (\widehat{\mathrm{spB}}(s))^2 \, ds = \frac{9\pi}{4}. \tag{34}$$

However, $\mathrm{spB} \notin \mathcal{L}^1(\mathbb{R})$ and thus Theorem 3.1 cannot be applied to find out whether the spherical Bessel kernel induces networks with the universal approximation property.

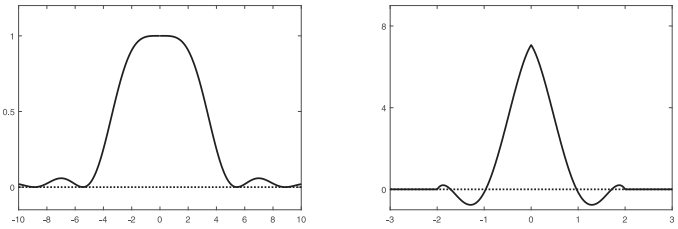


Fig. 6. The squared spherical Bessel kernel and its Fourier transform.

Instead of the spherical Bessel kernel, consider its second power spB^2. Clearly, $\mathrm{spB}^2 \in \mathcal{L}^1(\mathbb{R})$ by the above formula (34). Further, the Fourier transform of the second power is the second convolution power of the Fourier transform, i.e., $\widehat{\mathrm{spB}^2} = \widehat{\mathrm{spB}} * \widehat{\mathrm{spB}}$, so $\widehat{\mathrm{spB}^2}$ is compactly supported as well as $\widehat{\mathrm{spB}}$ (see Fig. 6 for the graphs of both spB^2 and $\widehat{\mathrm{spB}^2}$). As $\widehat{\mathrm{spB}^2}$ is continuous, one has $||\widehat{\mathrm{spB}^2}||_{\mathcal{L}^2} < \infty$ and therefore $\mathrm{spB}^2 \in \mathcal{L}^2(\mathbb{R})$.

By Theorem 4.3, spB^2 is not positive definite. As $\widehat{\mathrm{spB}^2}$ is compactly supported on the interval $[-2, 2]$, by Theorem 3.1, shallow networks with the squared spherical Bessel kernel units do not have the universal approximation property.

Table I presents several one-dimensional kernels including examples of kernels from all four classes defined by the two conditions on the Fourier transforms: the condition needed for capability to approximate functions and the one needed for application of SVM.

	SVM yes	SVM no
approximation yes	Gaussian Laplace triangle	rectangle cut parabolic
approximation no	sinc^2	spherical Bessel^2

TABLE I
ONE-DIMENSIONAL KERNELS AND THEIR PROPERTIES.

V. PROPERTIES OF MULTIVARIABLE RADIAL TRANSLATION-INVARIANT KERNELS

In this section, we analyze properties of Fourier transforms of multivariable translation-invariant kernels. First, let us consider a simple case of product kernels and then a more complicated case of general radial kernels, analysis of which requires the Hankel transform.

The multiplicative form of functions of several variables which can be expressed as products of one-variable functions

$$K(x, y) = \prod_{i=1}^{d} K_i(x_i, y_i) = \prod_{i=1}^{d} k_i(x_i - y_i) \tag{35}$$

induces separability of variables. Fourier transforms of such multivariable functions can be expressed as products of the Fourier transforms of one-variable functions. It follows from the construction of the product Lebesgue measure that if for a measurable $S \subset \mathbb{R}$, $\lambda(S) = 0$ then also $\lambda^d(S^d) = 0$ and if $\lambda(S) > 0$, then also $\lambda^d(S^d) > 0$. Hence, characterization of multivariable product kernels inducing networks having the universal approximation property can be reduced to analysis of properties of one-dimensional kernels.

Similarly, suitability of kernels of the form (35) for SVM can be reduced to the one-dimensional case. By the Schur Product Theorem [35], products of positive definite matrices are positive definite and hence products of positive definite kernels are positive definite (see also [36, Proposition 3.22, p. 75]).

Radial kernels are obtained by applying one-variable radial functions to norms of multivariable arguments. For the Euclidean norm $||\cdot||_2$, they are rotationally invariant. The Gaussian kernel, being both product and radial kernel, establishes a link between these two classes.

The following theorem provides a representation of the Fourier transform of a radial function with the Euclidean norm in terms of the Hankel transform [29, Theorem 3.3].

Theorem 5.1: Let $\Phi \in \mathcal{L}^1(\mathbb{R}^d)$ be continuous and $\Phi(x) = \varphi(||x||_2)$. Then its Fourier transform $\widehat{\Phi} : \mathbb{R}^d \to \mathbb{R}$ is also radial and $\widehat{\Phi}(\cdot) = \varphi^{\mathscr{H}}(||\cdot||_2)$ where $\varphi^{\mathscr{H}}(s) =$

$$= s^{-\nu} \mathscr{H}_\nu \left(\varphi(r) r^\nu \right)(s) = s^{-\frac{d-2}{2}} \int_0^\infty \varphi(r) r^{\frac{d}{2}} J_{(d-2)/2}(sr) \, dr, \tag{36}$$

$\nu = (d-2)/2$, and J_ν is the Bessel function of the first kind of order ν.

By Theorem 5.1, the Fourier transform of a radial function $\varphi(\|x\|_2)$ can be represented as a multiple of the Hankel transform of the one-variable function $\varphi(r)$ scaled by factor r^ν, where ν depends on the number of variables d ($\nu = -\frac{1}{2}, 0, \frac{1}{2}, 1, \dots$ for $d = 1, 2, 3, 4 \dots$). See the definition (16) of the Hankel transform $\mathscr{H}_\nu$. It is defined in terms of the *Bessel functions of the first kind* J_ν [33, p. 198].

For more details on Hankel transform see, e.g., [23], [34]. Note that in some literature (e.g., [34]), an alternative definition of the Hankel transform

$$h_\nu\{f(x)\}(y) = \int_0^\infty f(x) J_\nu(xy)(xy)^{1/2}\, dx \tag{37}$$

is used. For $\mathscr{H}_\nu(\varphi(r)r^\nu)$ defined in (16) the following relation holds

$$\mathscr{H}_\nu\{\varphi(r)r^\nu\}(s) = s^{-1/2} h_\nu\Big\{x^{\nu+1/2}\varphi(x)\Big\}(s). \tag{38}$$

It follows directly from the definition (16) that

$$\mathscr{H}_\nu\{\varphi(r/b)r^\nu\}(s) = b^{\nu+2}\mathscr{H}_\nu\{\varphi(r)r^\nu\}(bs) \tag{39}$$

for a scaling factor $b > 0$.

For some one-variable kernels from Section IV, Table II presents Hankel transforms of scaled versions of shape functions φ obtained using formulas from [34, Vol. II] together with the above equations.

In Table III, the Fourier transforms of scaled multivariable Gaussian, Laplace, cut power and circ kernels are compiled. The expressions follow from Theorem 5.1 with $\nu = (d-2)/2 = d/2 - 1$ for d being the number of variables. Note that $b^{\nu+1} = b^{d/2}$ and $b^{2(\nu+1)} = b^d$.

Combining the expressions for multivariable Fourier transforms specified in Table III with Theorems 3.1 and 4.3, we obtain the following results on suitability for classification and/or function approximation by networks with multivariate Gaussian, Laplace, cut power and circ kernel units.

Gaussian kernel. Since the multivariable Gaussian can be expressed as the product of one-variable Gaussians, its Fourier transform is the product of the respective one-dimensional Fourier transforms. Hence the analysis of the one-variable case from Section IV can be directly extended to the multivariable one.

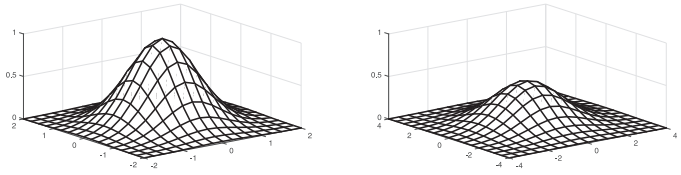


Fig. 7. The 2D Gaussian kernel and its Fourier transform.

Laplace kernel. The Hankel transform of the exponential function is an inverse multiquadric, which is a rational function. As this function is everywhere positive, also its composition with any norm is positive. Hence networks with multivariable Laplace kernel units have the universal approximation property. Moreover, Laplace kernel is an even, integrable function with positive Fourier transform and so it is positive definite. Thus similarly as the multivarible Gaussian kernel, also the Laplace kernel is suitable for both function approximation and SVM classification tasks in any dimension.

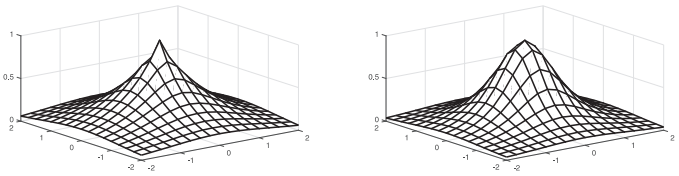


Fig. 8. The 2D Laplace kernel and its Fourier transform.

Cut power kernel. In contrast to Gaussian and Laplace kernels, the cut power is compactly supported. Thus its Fourier transform is analytic, which cannot be compactly supported unless it is a constant equal to zero. The Hankel transform of the cut power involves the Bessel function of the first kind J_ν, which is not positive, but the set of its zeros has Lebesgue measure zero as it is countable. Thus networks with the cut power kernel units are suitable for function approximation. On the other hand, by Theorem 4.3 the cut power cannot be positive definite, because on some subsets of $\mathbb{R}^d$ its Fourier transform is negative. So networks with cut power units have the universal approximation property, but they are not suitable for the SVM algorithm.

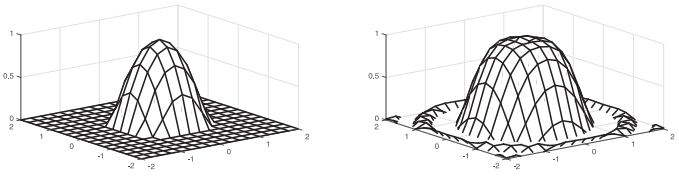


Fig. 9. The 2D cut power and its Fourier transform for $\mu = 2$.

Circ kernel. The circ kernel is based on the circ function which is the characteristic function of the unit ball in $\mathbb{R}^d$. In the one-dimensional case, it corresponds to the rectangular pulse $1_{[-1,1]}$. Its Hankel transform equals to the Bessel function of the first kind multiplied by the factor s^{-1}. So its Fourier transform has a nonempty set of zeros whose Lebesgue measure is zero. As the circ function is not continuous, Theorem 4.3 cannot be applied. However, an analogous argument as the one used in Section IV for the one-variable case, shows that also the multivariable circ function is not positive definite.

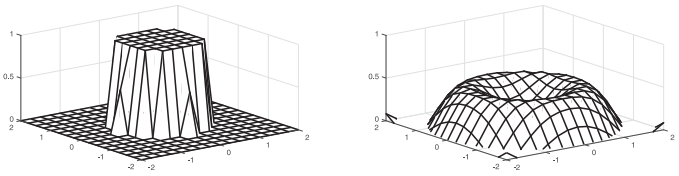


Fig. 10. The 2D circ function and its Fourier transform.

Inspecting formulas in Table III for $d = 1$, we get the same results as derived earlier for one-dimensional Fourier transforms. For the Gaussian and Laplace kernels, a transition

	$\varphi_b(r) = \varphi(r/b)$	$\mathscr{H}_\nu\left(\varphi_b(r) r^\nu\right)(s)$	pages in [34, Vol. II]
Gaussian	$\exp(-(r/b)^2)$	$\frac{b^{2(\nu+1)} s^\nu}{2^{\nu+1}} \exp(-\frac{1}{4}(bs)^2)$	p. 29 (10)
exponential	$\exp(-r/b)$	$\frac{2^{\nu+1}\Gamma(\nu+3/2) b^{2(\nu+1)}}{\pi^{1/2}(1+b^2 s^2)^{\nu+3/2}}\, s^\nu$	p. 29 (4)
cut power	$(a^2 - (r/b)^2)_+^\mu$	$\frac{b^{\nu-\mu+1} 2^\mu \Gamma(\mu+1) a^{\nu+\mu+1}}{s^{\mu+1}}\, J_{\nu+\mu+1}(abs)$	p. 26 (33)
rect. pulse $1_{[0,1]}$	$\begin{cases} 1 & 0 \le r/b \le 1 \\ 0 & \text{otherwise} \end{cases}$	$b^{\nu+1} s^{-1} J_{\nu+1}(bs)$	p. 22 (6)

TABLE II
HANKEL TRANSFORMS OF SCALED ONE-VARIABLE FUNCTIONS $\varphi_b(r)\, r^\nu$.

	$\varphi_b(\|\|x\|\|_2)$	$\hat{\varphi}_b(s)$
Gaussian	$\exp(-\|\|x\|\|_2^2/b^2)$	$(b^2/2)^{d/2} \exp(-\frac{b^2}{4}\|\|s\|\|_2^2)$
Laplace	$\exp(-\|\|x\|\|_2/b)$	$\frac{2^{d/2}\Gamma(d/2+1/2) b^d}{\pi^{1/2}(1+b^2\|\|x\|\|_2^2)^{(d+1)/2}}$
cut power	$(a^2 - \|\|x\|\|_2^2/b^2)_+^\mu$	$\frac{2^\mu \Gamma(\mu+1)\, a^{d/2+\mu}\, b^{d/2-\mu}}{\|\|s\|\|_2^{d/2+\mu}}\, J_{d/2+\mu}(ab\|\|s\|\|_2)$
circ	$\begin{cases} 1 & 0 \le \|\|x\|\|_2/b \le 1 \\ 0 & \text{otherwise} \end{cases}$	$(b/\|\|s\|\|_2)^{d/2} J_{d/2}(b\|\|s\|\|_2)$

TABLE III
SCALED MULTIVARIABLE RADIAL FUNCTIONS AND THEIR FOURIER TRANSFORMS.

from the multidimensional case to the one-dimensional one is straightforward. For the cut power, however, it might look a bit unclear how to get the version of the transform for the Epanechnikov kernel ($\mu = 1$, $a = 1$, $b = 1$ and multiplication by 3/4) presented in Section IV. The key lies in the fact that the Bessel function of the first kind $J_{3/2}$ admits the closed expression $J_{3/2}(|s|) = (2/\pi)^{1/2}|s|^{-3/2}(\sin(|s|) - |s|\cos(|s|))$.

Similarly, from the last row of Table III we get

$$\widehat{\mathrm{rect}_b}(s) = \widehat{\mathrm{circ}}_{b/2}(s) = \sqrt{\frac{b}{2|s|}} J_{1/2}\left(\frac{b|s|}{2}\right). \tag{40}$$

Using the well-known identities for Bessel functions (see [33, p. 2779])

$$j_0(x) = \sqrt{\pi/(2x)} J_{1/2}(x), \quad j_0(x) = \mathrm{sinc}(x) \tag{41}$$

and the fact that sinc is even, we obtain

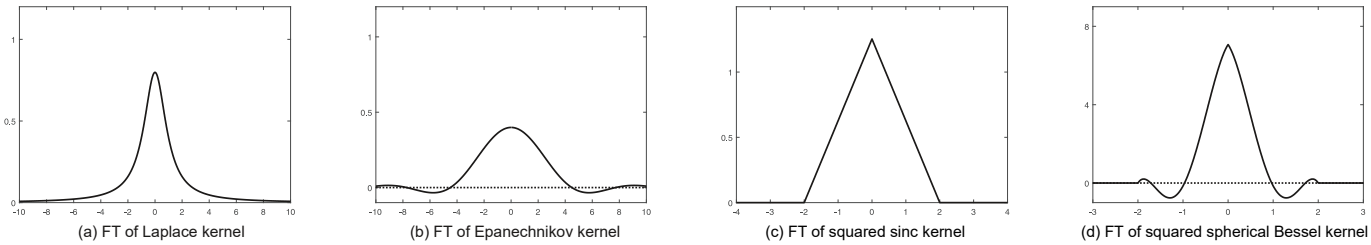
$$\sqrt{\frac{b}{2|s|}} J_{1/2}\left(\frac{b|s|}{s}\right) = \frac{b}{\sqrt{2\pi}}\, j_0\left(\frac{b|s|}{2}\right) = \frac{b}{\sqrt{2\pi}} \mathrm{sinc}\left(\frac{bs}{2}\right). \tag{42}$$

Other kernels. Table I in Section IV presents seven one-dimensional kernels and properties of their Fourier transforms. Multidimensional versions of four of these kernels are investigated in this section using the Hankel transforms of the associated shape functions. For the multivariate counterparts of the remaining three kernels (triangular, sinc^2 and spB^2) simple closed forms of the corresponding Hankel transforms of these kernels for general ν are not known to the authors of this article. However, some insights can be gained using numerical computations.

VI. CONCLUSION

We presented a unifying framework for exploration of capabilities of shallow networks with translation-invariant kernel units. Fourier transforms of even integrable functions (in particular of integrable translation-invariant kernels) are real, uniformly continuous, and converging with increasing frequencies to zero. Their sets of zeros have various forms: they can be equal to zero above certain frequencies (band limited), have discrete sets of zeros, or be everywhere positive, or be negative on some subset (see Fig. 11 for some paradigmatic examples).

Key properties of the Fourier transforms influencing suitability of kernels for function approximation and/or classification are their nonnegativity and size of their sets of zeros. For SVM and RKHSs, Fourier transforms must be nonnegative, but they can have large sets of zeros, they can even be compactly supported. On contrary, an existence of nonnegligible sets of frequencies, for which the values of the Fourier transforms are zero, limits approximation capabilities of kernel networks. For the universal approximation property, the Fourier transform of a kernel can be negative, but it cannot be zero on any set of frequencies of positive Lebesgue measure.



(a) FT of Laplace kernel (b) FT of Epanechnikov kernel (c) FT of squared sinc kernel (d) FT of squared spherical Bessel kernel

Fig. 11. Four types of Fourier transforms of even continuous integrable functions.

We derived an alternative proof of the classical Wiener Closure Theorem with basic tools from functional analysis. For analysis of Fourier transforms of multivariate radial kernels, we employed the Hankel transform, which does not have an analytic expression but can be represented in terms of Bessel functions. Their properties have been studied in applied science in connection with wave propagation and thus their sets of zeros are known.

We illustrated our results by concrete examples of several well-known kernels. We presented examples of kernels suitable for both function approximation and SVM (Gaussian, Laplace, and triangle), kernels merely suitable for function approximation (rectangle and cut power), a kernel merely suitable for SVM and regularization (squared sinc), and a kernel which is neither suitable for function approximation nor for classification by SVM (squared spherical Bessel kernel).

It might seem that it is always advantageous to use kernels that possess both properties (positive definitness and universal approximation property of induced networks) in all dimensions (such as Gaussian and Laplace). However, the price paid for universality is computational complexity connected with unbounded supports of such kernels. In applications of neural computing, we have recently witnessed a shift towards using compactly supported functions as their derivatives equals zero outside their supports, which greatly simplifies learning by gradient-based methods. Another advantage of compactly supported kernels is that their supports can be localized. The message of our theoretical results for practical applications is that while use of compactly supported kernels brings lower computational cost, their approximation/classification capabilities might be limited. For example, the compactly supported Epanechnikov kernel is not suitable for SVM tasks. Our paper provides theoretical guidelines for choice of kernels in terms of their Fourier transforms.

APPENDIX A
PROOF OF THEOREM 3.1

First, suppose that $\lambda^d(S) = \lambda^d(\{s \in \mathbb{R}^d \,|\, \hat{k}(s) = 0\}) \neq 0$. As Lebesgue measure is inner regular, there exists a compact set $S_\mathcal{K} \subseteq S$ such that $\lambda^d(S_\mathcal{K}) > 0$. Let $f \in \mathcal{L}^1(\mathbb{R}^d) \cap \mathcal{L}^2(\mathbb{R}^d)$ be such that $\hat{f}(s) > 0$ for all $s \in S$ (e.g., f can be the Gaussian function) and let $\varepsilon > 0$ be such that $\varepsilon < \int_{S_\mathcal{K}} \hat{f}(s)^2\, ds$. Assume that $f \in \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_K(X)$. Denoting $K_y(x) = k(x-y)$, we have

$$\|f - \sum_{j=1}^{n} w_j K_{y_j}\|^2_{\mathcal{L}_2} < \varepsilon, \tag{43}$$

for some $n \in \mathbb{N}_+$, $w_j \in \mathbb{R}$, $y_j \in \mathbb{R}^d$. Setting $u_j(s) = w_j e^{\mathrm{i} y_j \cdot s}$, we get by the isometry of the Fourier transform (Plancherel Theorem [22, p. 188])

$$\begin{aligned} \|f - \sum_{j=1}^{n} w_j K_{y_j}\|^2_{\mathcal{L}_2} &= \|\hat{f} - \sum_{j=1}^{n} w_j \widehat{K_{y_j}}\,\|^2_{\mathcal{L}_2} \\ &= \|\hat{f} - \sum_{j=1}^{n} u_j \hat{k}\|^2_{\mathcal{L}_2}. \end{aligned} \tag{44}$$

Hence $\|\hat{f} - \sum_{j=1}^{n} u_j \hat{k}\|^2_{\mathcal{L}_2} =$

$$= \int_{\mathbb{R}^d \setminus S_\mathcal{K}} \left(\hat{f}(s) - \sum_{j=1}^{n} u_j(s) \hat{k}(s) \right)^2 ds + \int_{S_\mathcal{K}} \hat{f}(s)^2 ds > \varepsilon, \tag{45}$$

which contradicts (43).

We prove sufficiency of the condition $\lambda^d(S) = 0$ using a standard method of proving density of sets of functions based on the Hahn-Banach Theorem (see, e.g., [22, p. 60]). We can assume that $X = \mathbb{R}^d$ (otherwise we embed $\mathcal{L}^2(X)$ into $\mathcal{L}^2(\mathbb{R}^d)$ by setting all functions equal to zero outside of X). Assuming that $\mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_K(\mathbb{R}^d)$ is a proper subset of $\mathcal{L}^2(\mathbb{R}^d)$, we would get a bounded linear functional L on $\mathcal{L}^2(\mathbb{R}^d)$ that vanishes for $f \in \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_K(\mathbb{R}^d)$, but for some $f_0 \in \mathcal{L}^2(\mathbb{R}^d) \setminus \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_K(\mathbb{R}^d)$, $L(f_0) = 1$. As by the Riesz Representation Theorem [37, p. 206], all linear functionals on Hilbert spaces are expressible as inner products, $L(f) = \langle f, h \rangle$ for some $h \in \mathcal{L}^2(\mathbb{R}^d)$. Then for even k, we have

$$\begin{aligned} \langle h, K_y \rangle &= \int_{\mathbb{R}^d} h(x) k(x-y)\, dx \\ &= \int_{\mathbb{R}^d} h(x) k(y-x)\, dx = h * k(y) = 0. \end{aligned} \tag{46}$$

Using properties of the Fourier transform and the Young Inequality [22, p. 183, p. 188], we get $h * k \in \mathcal{L}^2(\mathbb{R}^d)$ and $\|\widehat{h * k}\|_{\mathcal{L}^2} = 0$. As

$$\widehat{h * k} = \frac{1}{(2\pi)^{d/2}} \hat{h}\, \hat{k}, \tag{47}$$

we have $\|\hat{h}\, \hat{k}\|_{\mathcal{L}^2} = 0$ and so

$$\int_{\mathbb{R}^d} (\hat{h}(s)\, \hat{k}(s))^2 ds = 0. \tag{48}$$

By the assumption that $\lambda^d(S) = 0$, we get

$$\int_{\mathbb{R}^d} \hat{h}(s)^2 \hat{k}(s)^2 ds = \int_{\mathbb{R}^d \setminus S} \hat{h}(s)^2 \hat{k}(s)^2 ds = 0. \tag{49}$$

As for all $s \in \mathbb{R}^d \setminus S$, $\hat{k}(s)^2 > 0$, we have $\|\hat{h}\|^2_{\mathcal{L}^2} = 0$ and so $\|h\|_{\mathcal{L}^2} = 0$. Finally, using the Cauchy-Schwartz inequality we get a contradiction

$$1 = L(f_0) = \int_{\mathbb{R}^d} f_0(x)\, h(x)\, dx \leq \|f_0\|_{\mathcal{L}^2}\, \|h\|_{\mathcal{L}^2} = 0. \quad (50)$$

□

REFERENCES

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

[2] R. Courant and D. Hilbert, *Methods of Mathematical Physics*. New York: Wiley, 1962, vol. 2.

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

[4] F. Girosi and T. Poggio, "Regularization algorithms for learning that are equivalent to multilayer networks," *Science*, vol. 247, no. 4945, pp. 978–982, 1990.

[5] C. Cortes and V. N. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

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

[7] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of AMS*, vol. 39, pp. 1–49, 2002.

[8] V. Kůrková, "Neural network learning as an inverse problem," *Logic Journal of IGPL*, vol. 13, pp. 551–559, 2005.

[9] V. Kecman, *Learning and Soft Computing*. Cambridge: MIT Press, 2001.

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

[11] J. A. Suykens, M. Signoretto, and A. Argyriou, *Regularization, Optimization, Kernels, and Support Vector Machines*. Chapman and Hall/CRC, 2014.

[12] H. Chen, P. Tino, and X. Yao, "Efficient probabilistic classification vector machine with incremental basis function selection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25(2), pp. 356–369, 2014.

[13] F. Schleif, A. Gisbrecht, and P. Tino, "Supervised low rank indefinite kernel approximation using minimum enclosing balls," *Neurocomputing*, vol. 318, pp. 213–226, 2018.

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

[15] H. Wendland, *Scattered Data Approximation*. Cambridge University Press, 2005.

[16] G. E. Fasshauer, *Meshfree Approximation Methods with MATLAB*. World Scientific Publishing Co., Inc., 2007.

[17] V. Kůrková, "Multivariable approximation by convolutional kernel networks," in *Information Technologies - Applications and Theory, CEUR Workshop Proceedings*, B. Brejová, Ed., vol. V-1649. CreateSpace Independent Publishing Platform, 2016, pp. 118–122.

[18] D. Coufal, "Kernel Networks for Function Approximation," in *Engineering Applications of Neural Networks EANN2016, Communications in Computer and Information Science, 629*, C. Jayne and L. Iliadis, Eds. Springer, 2016, pp. 295–306.

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

[20] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, ser. Adaptive computation and machine learning. MIT Press, 2002.

[21] G. Gnecco, "Approximation and estimation bounds for subsets of reproducing kernel Krein spaces," *Neural Processing Letters*, vol. 39, pp. 137–153, 2014.

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

[23] L. Debnath and D. Bhatta, *Integral Transforms and Their Applications*. Chapman & Hall/CRC, 2007.

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

[25] ——, "Approximation and radial basis function networks," *Neural Computation*, vol. 5, pp. 305–316, 1993.

[26] V. Kůrková and M. Sanguineti, "Approximate minimization of the regularized expected error over kernel models," *Mathematics of Operations Research*, vol. 33, pp. 747–756, 2008.

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

[28] H. N. Mhaskar, "Versatile Gaussian networks," in *Proceedings of IEEE Workshop of Nonlinear Image Processing*, 1995, pp. 70–73.

[29] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton University Press, 1971.

[30] N. Wiener, "Tauberian Theorems," *Annals of Math.*, vol. 33, no. 1, pp. 1–100, 1932.

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

[32] M. A. Pinsky, *Introduction to Fourier Analysis and Wavelets*. American Mathematical Society, 2008.

[33] E. W. Weisstein, *CRC Concise Encyclopedia of Mathematics, Second Edition*. CRC Press, 2002.

[34] H. Bateman, *Tables of Integral Transforms [Volumes I & II]*. McGraw-Hill Book Company, New York, 1954, http://authors.library.caltech.edu/43489.

[35] J. Schur, "Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen," *Journal für die reine und angewandte Mathematik*, vol. 140, pp. 1–28, 1911. [Online]. Available: http://eudml.org/doc/149352

[36] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

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

PLACE PHOTO HERE

Věra Kůrková received PhD. in mathematics from the Charles University, Prague, and DrSc. (Prof.) in theoretical computer science from the Czech Academy of Sciences. She is a senior scientist at the Department of Machine Learning, Institute of Computer Science of the Czech Academy of Sciences. Her research interests are in mathematical theory of neurocomputing and machine learning. Her work includes analysis of capabilities and limitations of shalow and deep networks, dependence of network complexity on increasing dimensionality of computational tasks, connections between theory of inverse problems and generalization in machine learning, and nonlinear approximation theory. She was awarded the Bolzano Medal for her contribution to mathematical sciences by the Czech Academy of Sciences (2010). Since 2008, she has been a member of the Board of the European Neural Network Society (ENNS), in 2017-2019 its president. She is a member of the editorial boards of the journals Neural Networks and Neural Processing Letters and was a guest editor of special issues of the journals Neural Networks and Neurocomputing. She was the general chair of conferences ICANNGA 2001 and ICANN 2008, and co-chair or honorary chair of ICANN 2017, ICANN 2018, ICANN 2019, and ICANN 2020.

PLACE PHOTO HERE

David Coufal received PhD in Probability and Statistics from the Charles University, Prague and in Technical Cybernetics from the University of Pardubice. Since 1996 he has been working with the Institute of Computer Science of the Czech Academy of Science, Prague, where he is a researcher at the Department of Machine Learning. His research interests are machine learning, neuro and fuzzy computing and stochastic filtering. Namely, he is interested in research on using radial functions and kernel methods for function representation and approximation.