Neural Networks 57 (2014) 23–28

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Comparing fixed and variable-width Gaussian networks

Věra Kůrková [a,*], Paul C. Kainen [b]

[a] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic
[b] Department of Mathematics and Statistics, Georgetown University, 3700 Reservoir Rd., N.W., Washington, DC 20057, USA

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form 4 May 2014
Accepted 11 May 2014
Available online 20 May 2014

Keywords:
Gaussian radial and kernel networks
Functionally equivalent networks
Universal approximators
Stabilizers defined by Gaussian kernels
Argminima of error functionals

ABSTRACT

The role of width of Gaussians in two types of computational models is investigated: Gaussian radial-basis-functions (RBFs) where both widths and centers vary and Gaussian kernel networks which have fixed widths but varying centers. The effect of width on functional equivalence, universal approximation property, and form of norms in reproducing kernel Hilbert spaces (RKHS) is explored. It is proven that if two Gaussian RBF networks have the same input–output functions, then they must have the same numbers of units with the same centers and widths. Further, it is shown that while sets of input–output functions of Gaussian kernel networks with two different widths are disjoint, each such set is large enough to be a universal approximator. Embedding of RKHSs induced by "flatter" Gaussians into RKHSs induced by "sharper" Gaussians is described and growth of the ratios of norms on these spaces with increasing input dimension is estimated. Finally, large sets of argminima of error functionals in sets of input–output functions of Gaussian RBFs are described.

1. Introduction

Originally, artificial neural networks were built from biologically inspired computational units. These units, called perceptrons, compute functions in the form of plane waves. As an alternative, computational units in the form of spherical or elliptic waves were proposed mainly due to their good mathematical properties. Broomhead and Lowe (1988) introduced radial-basis-functions (RBFs) and Girosi and Poggio (1990) proposed more general kernel units. In particular, support vector machines (SVMs) built from units defined by symmetric positive semidefinite kernels became very popular (Cortes & Vapnik, 1995). Heaviside perceptrons cut input spaces into two halfspaces, with values of outputs equal to 0 on one half-space and 1 on the other, and so they are highly non-local. RBFs are geometrically opposite; they assign values close to 0 outside of spherical areas around their centers. Thus RBFs are localized.

Among localized computational units, a prominent position is occupied by units induced by the Gaussian function. Radial-basis-function units with the Gaussian radial function are the most common type of RBFs and Gaussian kernels with fixed widths are typical symmetric positive definite kernels. Both these computational models, the one with Gaussian RBF units having variable widths and the one with Gaussian units having fixed widths, have their advantages. Arbitrarily small widths of Gaussian RBFs were used in proofs of their universal approximation capability based on classical results on convolutions with sequences of scaled kernels (Park & Sandberg, 1991, 1993). Varying widths also play an important role in learning algorithms (see, e.g., Benoudjit, Archambeau, Lendasse, Lee, & Verleysen, 2002; Kecman, 2001; Verleysen & Hlaváčková, 1996; Wallace, Tsapatsoulis, & Kollias, 2005) and in some estimates of rates of approximation by Gaussian RBFs (see, e.g., Girosi, 1994; Girosi & Anzellotti, 1993; Kainen, Kůrková, & Sanguineti, 2009; Mhaskar, 2004). On the other hand, fixing the width allows one to fix the geometrical structure of a Hilbert space and apply the maximal margin classification algorithm (SVM) (Cortes & Vapnik, 1995). It also enables characterization of theoretically optimal solutions of learning tasks and modeling of generalization (see, e.g., Cucker & Smale, 2002; Girosi, 1998; Girosi, Jones, & Poggio, 1995; Kůrková, 2013; Poggio & Smale, 2003).

Some comparisons of capabilities of Gaussian networks with fixed and varying widths were obtained by Schmitt (2002) for the special case of input dimension equal to one. He proved that a Gaussian kernel network with a fixed width computing the same one-variable input–output function as a Gaussian RBF network with varying widths must be at least a factor of 1.5 larger.

In this paper, we investigate the role of widths of Gaussian functions in computational models which they generate. First, we

* Corresponding author. Tel.: +420 266053231.
E-mail addresses: vera@cs.cas.cz (V. Kůrková), kainen@georgetown.edu (P.C. Kainen).

http://dx.doi.org/10.1016/j.neunet.2014.05.005
0893-6080/

show that if input–output functions of two Gaussian RBF networks are equal, then the networks must have the same numbers of units and the same output weights, centers, and widths (up to a permutation of hidden units). This implies that possibilities of compressions of parameter spaces of Gaussian RBF networks are limited to equivalences induced by permutations. Our result holds for any input dimension d and any open domain in $\mathbb{R}^d$. Its proof takes advantage of the analyticity of the Gaussian function and properties of complex functions.

Further, we show that although sets of input–output functions of Gaussian kernel networks with two different widths are disjoint, each such set is large enough to be a universal approximator. In proving the density of Gaussian kernel networks, we use properties of Fourier transform of the Gaussian as an alternative to arguments of Mhaskar (1995), which are based on the form of derivatives of the Gaussian, and of Steinwart and Christmann (2008, p. 155), who use the Taylor series. Thus our results show that while no input–output function of a Gaussian RBF network whose units have at least two different widths can be exactly computed by a Gaussian kernel network with fixed width, each such function can be approximated with any required accuracy by Gaussian kernel networks having a given fixed width.

We also investigate how growth in the ratios of stabilizers induced by Gaussian kernels with two different widths depends on the input dimension. Finally, we describe multiple minima of empirical error functionals over sets of input–output functions computable by Gaussian RBFs. Some preliminary results appeared in the regional conference proceedings (Kůrková, 2013).

The paper is organized as follows. In Section 2, notations and basic concepts on one-hidden-layer RBF and kernel networks are introduced. In Section 3, it is shown that for two different widths, Gaussian kernel networks are not functionally equivalent. Section 4 shows that Gaussian kernel networks with fixed width are universal approximators. In Section 5, it is shown that the ratio of stabilizers with two different widths grows exponentially with increasing input dimension. Section 6 concludes the paper.

2. Dictionaries and kernels

The most widespread computational model used in neurocomputing is a *one-hidden-layer network with one linear output unit*. Such networks compute linear combinations of functions computable by a given type of computational units. The coefficients of linear combinations are called *output weights* and sets of functions computable by various types of units are called *dictionaries*. Networks with n units from a dictionary G compute functions from the set

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R}, g_i \in G \right\}.$$

The set of input–output functions of networks with any number of hidden units is denoted

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

where $\mathbb{N}_+$ denotes the set of positive integers.

Typically, dictionaries are given as parameterized families of functions. Let $K : X \times Y \to \mathbb{R}$ be a function of two variables representing an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter vector $y \in Y \subseteq \mathbb{R}^s$. We denote by

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

the dictionary of computational units computing K. When Y is clear from the context, we write shortly $G_K(X)$ (for symmetric kernels, $X = Y$).

In mathematics, various functions of two variables are called *kernels* (from the German term "kern", introduced by Hilbert in the context of theory of integral operators (Pietsch, 1987, p. 291)). In neurocomputing and learning theory, the term kernel is often reserved for a *symmetric positive semidefinite* function. This is a kernel $K : X \times Y \to \mathbb{R}$ such that $X = Y$, $K(x, y) = K(y, x)$ for all x, $y \in X$ and for any positive integer m, any $x_1, \ldots, x_m \in X$, and any $a_1, \ldots, a_m \in \mathbb{R}$,

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

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. Such spaces are called *reproducing kernel Hilbert spaces (RKHSs)* and denoted $\mathcal{H}_K(X)$. These spaces are formed by functions from $\operatorname{span} G_K(X)$ together with limits of their Cauchy sequences with respect to the norm $\|\cdot\|_K$, so $\operatorname{span} G_K(X) \subset \mathcal{H}_K(X)$. Usually, elements of $G_K(X)$ are denoted

$$K_x(\cdot) := K(x, .).$$

The norm $\|\cdot\|_K$ is induced by the inner product $\langle \cdot, \cdot \rangle_K$, which is defined on $G_K(X) = \{K_x \mid x \in X\}$ as

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

In this paper, we focus on dictionaries of three types defined in terms of the Gaussian function. The first one, $G_{F_d}(X)$ is induced by the nonsymmetric function $F_d : X \times Y \to \mathbb{R}$ (where $X \subseteq \mathbb{R}^d$, $Y = \mathbb{R}_+ \times \mathbb{R}^d$, and $\mathbb{R}_+$ denotes the set of positive real numbers) defined for every $x \in X$ and $(a, c) = (a, c_1, \ldots, c_d) \in \mathbb{R}_+ \times \mathbb{R}^d$ as

$$F_d(x, (a, c)) := e^{-\|a(x-c)\|^2}.$$

So

$$G_{F_d}(X) := \left\{ F_d(., (a, c)) : X \to \mathbb{R} \mid a > 0,\ c \in \mathbb{R}^d \right\}.$$

We call networks from the set $\operatorname{span} G_{F_d}(X)$ *Gaussian RBF networks* to distinguish them from *Gaussian kernel networks* which are induced by dictionaries $G_{K_d^a}(X)$ defined for each fixed $a > 0$ corresponding to *width* $\frac{1}{a}$ as

$$G_{K_d^a}(X) := \left\{ K_d^a(., c) : X \to \mathbb{R} \mid c \in \mathbb{R}^d \right\},$$

where $K_d^a : X \times \mathbb{R}^d \to \mathbb{R}$ satisfies for every $x \in X$ and $c \in \mathbb{R}^d$

$$K_d^a(x, c) := e^{-\|a(x-c)\|^2}.$$

So $G_{K_d^a}(X)$ consists of functions on X computable by units induced by the d-variable Gaussian with a fixed width $\frac{1}{a}$. Thus we can express the dictionary $G_{F_d}(X)$ as the union of the dictionaries $G_{K_d^a}(X)$, i.e.,

$$G_{F_d}(X) := \bigcup_{a \in \mathbb{R}_+} G_{K_d^a}(X).$$

We also consider the dictionary $G_{L_d}(X)$ induced by anisotropic *elliptic Gaussian units* with widths varying in each coordinate, where the kernel $L_d : X \times \mathbb{R}_+^d \times \mathbb{R}^d \to \mathbb{R}$ is defined for each $x = (x_1, \ldots, x_d) \in X$, $a = (a_1, \ldots, a_d) \in \mathbb{R}_+^d$, and $c = (c_1, \ldots, c_d) \in \mathbb{R}^d$ as

$$L_d(x, (a, c)) := e^{-\sum_{i=1}^{d} (a_i(x_i - c_i))^2}.$$

So

$$G_{L_d}(X) := \left\{ L_d(., (a, c)) : X \to \mathbb{R} \mid a \in \mathbb{R}_+^d, c \in \mathbb{R}^d \right\}.$$

3. Functionally equivalent Gaussian RBFs

In this section, we show that for any two different widths, $a \neq b$, on every open subset X of $\mathbb{R}^d$, the sets of input–output functions computable by Gaussian kernel networks with fixed widths a and b are disjoint.

The relationship of sets of input–output functions of Gaussian kernel networks with different fixed widths can be formulated in terms of a functional equivalence. Two neural networks are called *functionally equivalent* if they compute the same input–output function. This concept has been investigated for perceptron networks (Albertini & Sontag, 1993; Kůrková & Kainen, 1994, 1996; Sussman, 1992), RBF and fuzzy inference systems (Jang & Sun, 1993; Kůrková & Neruda, 1994).

A simple example of a dictionary for which all one-hidden-layer networks with units from this dictionary are functionally equivalent is any dictionary induced by a *product kernel*, i.e., a kernel $P : X \times X \to \mathbb{R}$ of the form $P(x, y) = p(x)p(y)$, where $p : X \to \mathbb{R}$ is any function on $X \subseteq \mathbb{R}^d$. Obviously, any network from span $G_P(X)$ computes a multiple of the function p and thus each two of them are functionally equivalent.

Another example of a dictionary inducing functionally equivalent networks with different numbers of hidden units is the dictionary $F_\beta(\mathbb{R})$ obtained by scalings and translations of the convolution kernel $K_\beta(x, y) = \beta(x-y)$, where $\beta : \mathbb{R} \to \mathbb{R}$ is the "triangle wave" defined by $\beta(x) = x+1$ on $[-1, 0]$, $\beta(x) = 1-x$ on $[0, 1]$, $\beta(x) = 0$ elsewhere. So $F_\beta(\mathbb{R}) = \{\beta(a(. - c)) : \mathbb{R} \to \mathbb{R} \mid a \in \mathbb{R}_+, c \in \mathbb{R}\}$. It is easy to check that $\beta(x) = \beta(2x) + \frac{1}{2}\beta(2x + \frac{1}{2}) + \frac{1}{2}\beta(2x - \frac{1}{2})$. Thus there exist two different networks computing the same input–output function, namely the network with one unit computing $\beta(x)$ and the network with three units computing $\beta(2x) + \frac{1}{2}\beta(2x + \frac{1}{2}) + \frac{1}{2}\beta(2x - \frac{1}{2})$.

Functional equivalences of neural networks can be studied in terms of linear dependences of dictionaries. A set of functions F is *linearly independent* if for any finite subset of its elements $\{f_1, \ldots, f_m\}$ and real numbers $w_1, \ldots, w_m$, $\sum_{i=1}^m w_i f_i = 0$ implies $w_i = 0$ for all $i = 1, \ldots, m$ (e.g., Friedman, 1982, p. 124). If a dictionary is linearly independent, then two networks are functionally equivalent only when they have the same number of units with the same parameters which can only differ by a permutation.

The following theorem shows that on any open subset X of $\mathbb{R}^d$ the dictionary $G_{L_d}(X)$ of elliptic Gaussian RBFs is linearly independent. Thus also its subset $G_{F_d}(X)$ formed by spherical Gaussian RBFs is linearly independent.

Theorem 3.1. *For every positive integer d and every open subset X of $\mathbb{R}^d$, the dictionary $G_{L_d}(X)$ is linearly independent.*

We prove Theorem 3.1 in several steps. Using fast convergence to zero of the values of the Gaussian, we first prove linear independence of the dictionary $G_{F_1}(\mathbb{R})$ formed by one-dimensional Gaussian RBFs on the whole real line $\mathbb{R}$. To show that linear independence also holds for the dictionary $G_{F_1}(X)$ when the domain X is an arbitrary open subset X of $\mathbb{R}$, we use the Identity Theorem from theory of complex functions. Then we verify a simple lemma on linear independence of products of functions. Finally applying this lemma together with an expression of the d-dimensional Gaussian as a product of d one-dimensional Gaussians, we prove the theorem for an arbitrary dimension d.

Theorem 3.2. *The dictionary*

$$G_{F_1}(\mathbb{R}) = \left\{e^{-a^2(.-c)^2} : \mathbb{R} \to \mathbb{R} \mid a \in \mathbb{R}_+, c \in \mathbb{R}\right\}$$

is linearly independent.

Proof. We show that no nontrivial linear combination of elements of $G_{F_1}(\mathbb{R})$ is the zero-function. Let m be a positive integer, $w_1, \ldots, w_m$ be nonzero real numbers, and $\{(a_j, c_j) \in \mathbb{R}_+ \times \mathbb{R} \mid j = 1, \ldots, m\}$ be a set of distinct pairs of widths and centers. Arguing by contradiction, assume that for all $x \in \mathbb{R}$

$$\sum_{j=1}^m w_j\, e^{-a_j^2(x-c_j)^2} = 0. \tag{1}$$

Without loss of generality one can further suppose that

(a) $1 = a_1 = \cdots = a_k < a_{k+1} \leq \cdots \leq a_m$, and
(b) $c_1 > c_j$ for all $j = 2, \ldots, k$.

Indeed, first reorder all the terms in Eq. (1) so that the initial k Gaussian functions have the same minimal width $a_j = a_1$ for $j = 2, \ldots, k$. Second, change the scale so that $a_1 = 1$. As the pairs $(1, c_1), \ldots, (1, c_k)$ are distinct, so are $c_1, \ldots, c_k$. Third, reorder the first k terms so that $c_1 > c_j$ for $j = 2, \ldots, k$.

Multiplying Eq. (1) by $e^{(x-c_1)^2}$ one finds that for all $x \in \mathbb{R}$,

$$w_1 + \sum_{j=2}^k \bar{w}_j e^{-2x(c_1-c_j)} + \sum_{j=k+1}^m \bar{w}_j e^{x^2(1-a_j^2)+2x(a_j^2c_j-c_1)} = 0,$$

where $\bar{w}_j = w_j e^{(c_1^2-c_j^2)}$ for $j = 2, \ldots, k$, and $\bar{w}_j = w_j e^{(c_1^2-a_j^2c_j^2)}$ for $j = k+1, \ldots, m$. By (a) and (b), both the exponential sums go to zero asymptotically as $x \to \infty$. Hence, $w_1 = 0$ violating the assumption that $w_1 \neq 0$. □

The proof of the next extension of the statement of Theorem 3.2 to any open subset X of $\mathbb{R}$ uses some properties of complex functions. A function f on $\mathbb{R}$ or $\mathbb{C}$ is *real analytic*, *complex analytic*, resp., if it is locally representable by a power series. Recall that a subset U of $\mathbb{C}$ is connected if and only if each pair of its points can be joined by a piecewise-linear path (i.e., a finite number of straight-line segments joined end to end) which lies entirely in U. An open set that is connected is called a *domain*. An *arc* A in the complex plane $\mathbb{C}$ is a set $A = \{(x(t), y(t)) \mid x : [0, 1] \to \mathbb{R}$ continuous, $y : [0, 1] \to \mathbb{R}$ continuous$\}$. We use a basic result from complex analysis (Churchill, Brown, & Verhey, 1974, p. 284) stating that a complex analytic function is uniquely determined by its values in a domain or along an arc.

Theorem 3.3. *Let $D \subseteq \mathbb{C}$ be a domain and $f : D \to \mathbb{C}$ be analytic such that $f(z) = 0$ for all z in some domain $B \subseteq D$ or some arc $A \subseteq D$, then $f(z) = 0$ for all $z \in D$.*

Theorem 3.4. *For every nonempty open subset $X \subset \mathbb{R}$, the dictionary*

$$G_{F_1}(X) = \left\{e^{-a^2\|.-c\|^2} : X \to \mathbb{R} \mid a \in \mathbb{R}_+, c \in \mathbb{R}\right\}$$

is linearly independent.

Proof. Assume that for all $x \in X$, $\sum_{j=1}^m w_j e^{-a_j^2\|x-c_j\|^2} = 0$, where m is a positive integer, $w_1, \ldots, w_m$ are real numbers, and $\{(a_j, c_j) \in \mathbb{R}_+ \times \mathbb{R}^d \mid j = 1, \ldots, m\}$ a set of distinct pairs. Consider the function $f : \mathbb{C} \to \mathbb{R}$ defined as $f(z) = \sum_{j=1}^m w_j e^{-a_j^2\|z-c_j\|^2}$. As f is a linear combination of Gaussians, it is complex analytic. Assume that $f(z) = 0$ for all $z \in X$. As X is a nonempty open subset of $\mathbb{R}$, it must contain an open interval, which contains an arc A in $\mathbb{C}$. By Theorem 3.3, f must be equal to zero on the whole of $\mathbb{C}$ and hence also on its subset $\mathbb{R}$. Thus by Theorem 3.2, $w_i = 0$ for all $i = 1, \ldots, m$ and so the dictionary $G_{F_1}(X)$ is linearly independent. □

To extend the statement of Theorem 3.4 to any dimension d and elliptic Gaussian RBF, we use the following lemma.

Lemma 3.1. *Let d be a positive integer, $X, Y \subseteq \mathbb{R}^d$, $\{f_i : X \to \mathbb{R} \mid i = 1, \ldots, m\}$ and $\{g_j : Y \to \mathbb{R} \mid j = 1, \ldots, n\}$ be two families of linearly independent functions. Then $\{h_{i,j} : X \times Y \to \mathbb{R} \mid i = 1, \ldots, m;\ j = 1, \ldots, n\}$ defined for all $(x, y) \in X \times Y$ as $h_{i,j}(x, y) = f_i(x)g_j(y)$ is linearly independent on $X \times Y$.*

Proof. Assume that $\sum_{i=1}^m \sum_{j=1}^n w_{ij}h_{ij}(x, y) = \sum_{i=1}^m \sum_{j=1}^n w_{ij}f_i(x)g_j(y) = 0$ for all $(x, y) \in X \times Y$. Thus $\sum_{i=1}^m \left(\sum_{j=1}^n w_{ij}g_j(y)\right) f_i(x) = 0$ for all $x \in X$ and all $y \in Y$. As $\{f_i \mid i = 1, \ldots, m\}$ is linearly independent on X, for all $y \in Y$, and all $i = 1, \ldots, m$, $\sum_{j=1}^n w_{ij}g_j(y) = 0$. So by linear independence of the set $\{g_j \mid j = 1, \ldots, n\}$ on Y, we get $w_{ij} = 0$ for all i and all j. □

Proof of Theorem 3.1. As X is an open subset of $\mathbb{R}^d$, it contains an open cube $\prod_{i=1}^d X_i$. Linear independence on $\prod_{i=1}^d X_i$ implies linear independence on X, so it is sufficient to prove it on $\prod_{i=1}^d X_i$.

We proceed by induction. For $d = 1$, the dictionary $G_{L_1}(X_1)$ is equal to the dictionary $G_{F_1}(X_1)$ and thus the statement follows from Theorem 3.4. Assume that the statement holds for $d - 1$. An elliptic d-dimensional Gaussian unit computes a tensor product of one-dimensional scaled Gaussians. For all $x = (x_1, \ldots, x_d) \in \prod_{i=1}^d X_i$, $a = (a_1, \ldots, a_d) \in \mathbb{R}_+^d$, and $c = (c_1, \ldots, c_d) \in \mathbb{R}^d$, we have $L_d(x, (a, c)) = e^{-\sum_{i=1}^d a_i^2(x_i - c_i)^2} = \prod_{i=1}^d e^{-a_i^2(x_i-c_i)^2} = \prod_{i=1}^d K_1^{a_i}(x_i, c_i)$. Thus $L_d(x, (a, c)) = L_{d-1}(\bar{x}, (\bar{a}, \bar{c}))\, L_1(x_d, (a_d, c_d))$, where $\bar{x} = (x_1, \ldots, x_{d-1})$, $\bar{a} = (a_1, \ldots, a_{d-1})$, and $\bar{c} = (c_1, \ldots, c_{d-1})$. By hypothesis, $G_{L_{d-1}}(\prod_{i=1}^{d-1} X_i)$ is linearly independent and by Theorem 3.4, $G_{F_1}(X_d)$ is also linearly independent. So the statement follows from Lemma 3.1. □

Theorem 3.1 shows that two Gaussian RBF networks on any open subset of $\mathbb{R}^d$ can compute the same input–output function only when they have the same numbers of hidden units, which have up to a permutation the same centers, widths, and output weights. It also shows that two Gaussian kernels with different widths determine disjoint sets of input–output functions.

Theorem 3.1 implies that the only reduction of parameter spaces of Gaussian RBF networks based on their functional equivalences is that induced by permutations of hidden units. Search in such reduced parameter spaces might be implementable for genetic algorithms which operate with strings of vectors of parameters.

4. Universal approximation of Gaussian kernel networks

In this section, we show that although Gaussian kernel units with fixed widths have fewer free parameters than Gaussian radial units with varying widths, fixed-width Gaussian kernel networks still generate classes of input–output functions large enough to be universal approximators. Recall that 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* $(\mathcal{X}, \|.\|_{\mathcal{X}})$ if it is *dense* in this space, i.e., $\mathrm{cl}_{\mathcal{X}} \operatorname{span} G = \mathcal{X}$, where $\mathrm{cl}_{\mathcal{X}}$ denotes the closure with respect to the topology induced by the norm $\|\cdot\|_{\mathcal{X}}$ (see, e.g., Kůrková, 2002; Pinkus, 1999). A subset A of a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is *dense* if for all $f \in \mathcal{X}$ and all $\varepsilon > 0$, there exists $g \in A$ such that $\|f - g\|_{\mathcal{X}} < \varepsilon$.

Function spaces where the universal approximation has been of interest are spaces $(\mathcal{C}(X), \|.\|_{\sup})$ of continuous functions on subsets X of $\mathbb{R}^d$ (typically compact) with the supremum norm and the space $(\mathcal{L}^2(\mathbb{R}^d), \|.\|_{\mathcal{L}^2})$ of square integrable functions on $\mathbb{R}^d$ with the norm $\|f\|_{\mathcal{L}^2} = \left(\int_{\mathbb{R}^d} f(y)^2 dy\right)^{1/2}$.

Note that the capability to approximate arbitrarily well all real-valued functions is much stronger than the capability of classification, which merely needs approximation up to a certain accuracy of functions with finite domains.

For RBF networks with radial functions satisfying rather mild conditions (which hold for the Gaussian), the universal approximation property was proven by Park and Sandberg (1991). Their proof exploits varying widths—it is based on a classical result on approximation of functions by sequences of their convolutions with scaled kernels. This proof might suggest that variability of widths is essential for the universal approximation. However using special properties of Hermite functions (which are derivatives of the Gaussian function), Mhaskar (1995) proved the universal approximation capability of Gaussian kernel networks in spaces of continuous functions on compact subsets of $\mathbb{R}^d$.

A related notion to the concept of universal approximation property of a class of networks is the "universal kernel" defined in Steinwart and Christmann (2008, p. 152) for the case of continuous kernels on compact metric spaces. Such a kernel K is called *universal* if the RKHS $\mathcal{H}_K(X)$ induced by K is dense in the space $(\mathcal{C}(X), \|\cdot\|_{\sup})$ of continuous functions with the supremum norm. As $\operatorname{span} G_K(X)$ is dense in $(\mathcal{H}_K(X), \|\cdot\|_K)$ and $\|\cdot\|_{\sup} \leq \|\cdot\|_K \sup_{x \in X} K(x, x)$, it follows that, for bounded kernels (in particular for continuous kernels on compact sets), the density of $\mathcal{H}_K(X)$ in $(\mathcal{C}(X), \|\cdot\|_{\sup})$ is equivalent to the density of $\operatorname{span} G_K(X)$.

Obviously, not all positive semidefinite kernels are universal. For example, the product kernel $K(x, y) = xy$ is not universal. The above mentioned result of Mhaskar (1995) shows that a Gaussian kernel with any fixed width is universal. The universality of an arbitrary fixed-width Gaussian kernel was also established in Steinwart and Christmann (2008, p. 155) using the Stone–Weierstrass theorem and a Taylor series.

Here, we prove the universal approximation property of Gaussian kernel networks with any fixed width in $\mathcal{L}^2(\mathbb{R}^d)$. Our argument is based on properties of the Fourier transform of the Gaussian and on the Hahn–Banach theorem.

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

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

and extended to $\mathcal{L}^2(\mathbb{R}^d)$ (Rudin, 1991, p. 183).

Theorem 4.1. *Let $d \in \mathbb{N}_+$, $a > 0$. Then both of the following hold:*

(i) *for $X \subseteq \mathbb{R}^d$ Lebesgue measurable, $\operatorname{span} G_{K_d^a}(X)$ is dense in $(\mathcal{L}^2(X), \|\cdot\|_{\mathcal{L}^2})$;*

(ii) *for $X \subset \mathbb{R}^d$ compact, $\operatorname{span} G_{K_d^a}(X)$ is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$.*

Proof. First assume that $X = \mathbb{R}^d$. Suppose $\mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d) \neq \mathcal{L}^2(\mathbb{R}^d)$. Then by the Hahn–Banach Theorem (Rudin, 1991, p. 60) there is a linear functional l on $\mathcal{L}^2(\mathbb{R}^d)$ such that for all $f \in \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d)$, $l(f) = 0$ and for some $f_0 \in \mathcal{L}^2(\mathbb{R}^d) \setminus \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d)$, $l(f_0) = 1$. By the Riesz Representation Theorem (Friedman, 1982), there exists $h \in \mathcal{L}^2(\mathbb{R}^d)$, such that for all $g \in \mathcal{L}^2(\mathbb{R}^d)$,

$$l(g) = \int_{\mathbb{R}^d} g(y)h(y)dy.$$

Thus for all $f \in \mathrm{cl}_{\mathcal{L}^2} \operatorname{span} G_{K_d^a}(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} f(y)h(y)dy = 0$. Defining $k^a(x) := e^{-a^2\|x\|^2}$, we get for all $x \in \mathbb{R}^d$, $\int_{\mathbb{R}^d} h(y)k^a(x - y)dy = (h * k^a)(x) = 0$. Thus by Plancherel's Theorem (Rudin, 1991, p. 188), $\|\widehat{h * k^a}\|_{\mathcal{L}^2} = 0$. As $\widehat{h * k^a} = \frac{1}{(2\pi)^{d/2}} \hat{h}\, \hat{k}^a$ (Rudin, 1991, p. 183), we have $\|\hat{h}\, \hat{k}^a\|_{\mathcal{L}^2} = 0$. As $\widehat{e^{-a^2\|\cdot\|^2}} = (\sqrt{2}a)^{-d} e^{-(1/a^2)\|\cdot\|^2}$ (Rudin, 1991, p. 186), we obtain $\|\hat{h}\|_{\mathcal{L}^2} = 0$. So again by Plancherel's Theorem, $\|h\|_{\mathcal{L}^2} = 0$. Hence we get

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

which is a contradiction.

Now let $X \subset \mathbb{R}^d$ be an arbitrary Lebesgue measurable set. We obtain (i) by extending functions from $\mathcal{L}^2(X)$ to $\mathcal{L}^2(\mathbb{R}^d)$, setting their values equal to zero outside of X, and restricting their approximations from span $G_K(\mathbb{R}^d)$ to X. For X compact, $\mathcal{C}(X) \subset \mathcal{L}^2(X)$ and the statement (ii) follows directly from (i). □

5. Minimization of error functionals over Gaussian networks

In this section, we investigate minimization of error functionals over Gaussian radial and kernel networks and dependence of stabilizers defined by norms on RKHSs on widths of Gaussian kernels. The spaces $\mathcal{H}_{K_d^a}(X)$, induced by Gaussian kernels with fixed positive width a, are formed by functions from span $G_{K_d^a}(X)$ together with limits of their Cauchy sequences in the norm $\|\cdot\|_{K_d^a}$. By Theorems 4.1 and 3.1, for all $a > 0$, the sets of input–output functions span $G_{K_d^a}(\mathbb{R}^d)$ are dense subspaces of $\mathcal{L}^2(\mathbb{R}^d)$, but for different widths $a \neq b$, they are disjoint.

The next theorem shows that RKHSs induced by Gaussian kernels K_d^a are nested. It gives the same upper bound $\frac{\|f\|_{K_d^a}}{\|f\|_{K_d^b}} \leq \left(\frac{a}{b}\right)^{d/2}$ on the ratio between the norms on RKHSs induced by the Gaussians with widths $\frac{1}{a}$ and $\frac{1}{b}$, which was derived in Steinwart and Christmann (2008, p. 143) by an argument using a semigroup of convolution integral operators. We give an alternative proof based on characterization of norms on RKHSs induced by convolution kernels in terms of Fourier transforms first observed in Girosi (1998) and rigorously proven in Loustau (2008).

Theorem 5.1. *Let d be a positive integer and $a, b > 0$ such that $b \leq a$. Then*

(i) $\mathcal{H}_{K_d^b}(\mathbb{R}^d) \subseteq \mathcal{H}_{K_d^a}(\mathbb{R}^d)$;

(ii) *the inclusion* $J_{b,a} : (\mathcal{H}_{K_d^b}(\mathbb{R}^d), \|\cdot\|_{K_d^b}) \to (\mathcal{H}_{K_d^a}(\mathbb{R}^d), \|\cdot\|_{K_d^a})$ *is continuous;*

(iii) *for all* $f \in \mathcal{H}_{K_d^b}(\mathbb{R}^d)$, $\|f\|_{K_d^a} \leq \left(\frac{a}{b}\right)^{d/2} \|f\|_{K_d^b}$.

Proof. By Loustau (2008), for a convolution kernel $K(x, y) = k(x-y)$ such that $\hat{k} > 0$,

$$\|f\|_K^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{\hat{f}(s)^2}{\hat{k}(s)}\, ds.$$

For all $a > 0$, $\widehat{e^{-a^2\|\cdot\|^2}} = (\sqrt{2}a)^{-d} e^{-(1/a^2)\|\cdot\|^2}$ and thus

$$\|f\|_{K_d^a}^2 = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{f}(s)^2 \left((\sqrt{2}a)^{-d} e^{-(1/a^2)\|\cdot\|^2}\right)^{-1} ds$$
$$= \frac{(\sqrt{2}a)^d}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{f}(s)^2 \left(e^{-(1/a^2)\|\cdot\|^2}\right)^{-1} ds.$$

As $b \leq a$ implies $\left(e^{-(1/a^2)\|\cdot\|^2}\right)^{-1} \leq \left(e^{-(1/b^2)\|\cdot\|^2}\right)^{-1}$, we have $\frac{\|f\|_{K_d^a}^2}{\|f\|_{K_d^b}^2} \leq \left(\frac{a}{b}\right)^d$ and so $\frac{\|f\|_{K_d^a}}{\|f\|_{K_d^b}} \leq \left(\frac{a}{b}\right)^{d/2}$. □

Theorem 5.1 shows that Hilbert spaces induced by "flatter" Gaussians are embedded in spaces induced by "sharper" Gaussians. For $0 < b < a$, the whole space $\mathcal{H}_{K_d^b}(\mathbb{R}^d)$ and hence also its subset span $G_{K_d^b}(\mathbb{R}^d)$ is contained in the space $\mathcal{H}_{K_d^a}(\mathbb{R}^d)$. However by Theorem 3.2, when $a \neq b$, the sets of input–output functions of Gaussian kernel networks with widths a and b are disjoint, i.e.,

$$\text{span } G_{K_d^a}(\mathbb{R}^d) \bigcap \text{span } G_{K_d^a}(\mathbb{R}^d) = \emptyset.$$

So the set span $G_{K_d^b}(\mathbb{R}^d)$ is contained in the subset of the space $\mathcal{H}_{K_d^a}(\mathbb{R}^d)$ formed by limits of Cauchy sequences from span $G_{K_d^a}(\mathbb{R}^d)$.

An *empirical error functional* $\mathcal{E}_z$ is determined by some training sample, $z = \{(u_i, v_i) \in \mathbb{R}^d \times \mathbb{R} \mid i = 1, \ldots, m\}$ of input–output pairs of data, by setting

$$\mathcal{E}_z(f) := \frac{1}{m} \sum_{i=1}^m (f(u_i) - v_i)^2.$$

Girosi and Poggio (1990); Poggio and Girosi (1990) initiated mathematical modeling of generalization in terms of Tikhonov regularization which adds to the empirical error a functional called the "stabilizer" penalizing undesired properties of solutions. Girosi et al. (1995) considered as stabilizers suitably weighted Fourier transforms; later Girosi (1998) realized that such stabilizers are squares of norms on spaces induced by kernels. We denote

$$\mathcal{E}_{z,\alpha,K} := \mathcal{E}_z + \alpha \|\cdot\|_K^2,$$

the *regularized empirical error* with the *stabilizer* $\|\cdot\|_K^2$ induced by a symmetric positive semidefinite kernel K and the parameter α controlling the trade-off.

Theorem 5.1 shows that "sharpening" of the Gaussian kernel increases the penalty represented by the stabilizer $\|.\|_{K_d^a}^2$ at most by a^d.

An *argminimum* of a functional is a function for which the functional attains its minimum. The next theorem characterizes argminima of $\mathcal{E}_z$ and $\mathcal{E}_{z,\alpha,K}$ (see, e.g., Cucker & Smale, 2002; Kůrková, 2013; Poggio & Smale, 2003). Let $\mathcal{K}[u]$ denote the matrix $\mathcal{K}[u]_{i,j} := K(u_i, u_j)$, $\mathcal{K}_m[u] = \frac{1}{m}\mathcal{K}[u]$, and $\mathcal{K}[u]^+$ denote the Moore–Penrose pseudoinverse of the matrix $\mathcal{K}[u]$.

Theorem 5.2. *Let $X \subseteq \mathbb{R}^d$, $K : X \times X \to \mathbb{R}$ a symmetric positive semidefinite kernel, m a positive integer, and $z = \{(u_i, v_i) \in \mathbb{R}^d \times \mathbb{R} \mid i = 1, \ldots, m\}$. Then*

(i) *there exists an argminimum f^+ of $\mathcal{E}_z$ over $\mathcal{H}_K(X)$ which satisfies*

$$f^+ = \sum_{i=1}^m c_i K_{u_i}, \quad \text{where } c = (c_1, \ldots, c_m) = \mathcal{K}[u]^+ v,$$

and for all $f^o \in \text{argmin}(\mathcal{H}_K(X), \mathcal{E}_z)$, $\|f^+\|_K \leq \|f^o\|_K$;

(ii) *for all $\alpha > 0$, there exists a unique argminimum f^α of $\mathcal{E}_{z,\alpha,K}$ over $\mathcal{H}_K(X)$ which satisfies for $v = (v_1, \ldots, v_m)$*

$$f^\alpha = \sum_{i=1}^m c_i^\alpha K_{u_i}, \text{ where } c^\alpha = (c_1^\alpha, \ldots, c_m^\alpha)$$
$$= (\mathcal{K}_m[u] + \alpha\, \mathit{I}_m)^{-1} v;$$

(iii) $\lim_{\alpha \to 0} \|f^\alpha - f^+\|_K = 0$.

Note that both argminima, f^+ and f^α, are computable by networks with m kernel units from $G_K(X)$. The argminima differ merely in coefficients of linear combinations (output weights) of kernel units with parameters corresponding to the data $u_1, \ldots, u_m$. Thus in the case of theoretically optimal solutions, generalization is achieved merely by modification of output weights which is influenced by the choice of a stabilizer.

The following theorem shows that in the space of continuous functions $\mathcal{C}(X)$ on a compact $X \subset \mathbb{R}^d$, for any training sample z and any width of the Gaussian, the empirical error functional $\mathcal{E}_z$ has an argminimum over $\mathcal{C}(X)$ formed by a linear combination of Gaussians of this width.

Theorem 5.3. *Let X be a compact subset of $\mathbb{R}^d$, m be a positive integer, and $z = \{(u_i, v_i) \in \mathbb{R}^d \times \mathbb{R} \mid i = 1, \ldots, m\}$. Then the set of argminima of $\mathcal{E}_z$ in $\mathcal{C}(X)$ contains the convex hull* conv $\{f_a^+ \mid a > 0\}$, *where $f_a^+ = \sum_{i=1}^m c_i^a K_d^a(., u_i)$ with $c^a = (c_1^a, \ldots, c_m^a) = \mathcal{K}_d^a[u]^+ v$ and $v = (v_1, \ldots, v_m)$.*

Proof. By Theorem 4.1(ii) for any $a > 0$, $\operatorname{span} G_{K_d^a}(X)$ is dense in $(\mathcal{C}(X), \|\cdot\|_{\sup})$. It is easy to show that $\mathcal{E}_z$ is continuous on $\mathcal{C}(X)$, an argminimum of a continuous functional over a dense subset is an argminimum over the whole space, and a convex combination of argminima is an argminimum. The statement then follows from Theorem 5.2. □

Thus for any training sample z, the empirical error $\mathcal{E}_z$ over the set of $\operatorname{span} F_d(X)$ of input–output functions of Gaussian RBF networks has a large convex set of argminima containing linear combinations of Gaussians of all widths. This result suggests why problems with multiple minima are encountered during learning. Some approaches to this problem were suggested in Bianchini, Frasconi, and Gori (1995); Gori and Tesi (1992).

6. Conclusion

We have compared capabilities of two popular computational models: Gaussian radial-basis function networks with varying widths and Gaussian kernel networks with fixed widths. Using methods from functional analysis we investigated the effect of width on functional equivalence, universal approximation property and norms in Hilbert spaces induced by Gaussian kernels. We proved that if two Gaussian RBF networks compute the same input–output functions, then they must have the same numbers of units with the same parameters (output weights, widths and centers); hence, the possibility of compressing parameter spaces is limited to the equivalences induced by permutations. We also gave a proof of the universal approximation property of Gaussian kernel networks based on properties of the Fourier transform. Our results show that input–output functions of Gaussian RBF networks with units having at least two different widths cannot be *exactly* represented as input–output functions of Gaussian kernel networks with one fixed width. We proved that networks with any fixed width can approximate arbitrarily well all $\mathcal{L}^2$-functions on $\mathbb{R}^d$ and so in particular any linear combination of Gaussians with varying widths. Rates of such approximation might be studied using integral representations in terms of convolutions with Gaussians studied in Girosi and Anzellotti (1993), Kainen et al. (2009) and Kainen, Kůrková, and Vogt (2007). We also investigated the role of width in Hilbert spaces induced by Gaussian kernels and proved that spaces induced by flatter Gaussians are embedded in spaces induced by narrower Gaussians.

Acknowledgments

V.K. was partially supported by MŠMT grant LD13002 and the institutional support of the Institute of Computer Science RVO 67985807.

References

Albertini, F., & Sontag, E. D. (1993). For neural networks, function determines form. *Neural Networks*, *6*(7), 975–990.
Benoudjit, N., Archambeau, C., Lendasse, A., Lee, J., & Verleysen, M. (2002). Width optimization of the Gaussian kernels in radial basis function networks, In *European symposium on artificial neural neworks ESANN*, Vol. 2 (pp. 425–432).
Bianchini, M., Frasconi, P., & Gori, M. (1995). Learning without local minima in radial basis function networks. *IEEE Transactions on Neural Networks*, *6*, 749–756.
Broomhead, D. S., & Lowe, D. (1988). Error bounds for approximation with neural networks. *Complex Systems*, *2*, 321–355.
Churchill, R. V., Brown, J. W., & Verhey, R. F. (1974). *Complex variables and applications*. New York: McGraw Hill.
Cortes, C., & Vapnik, V. N. (1995). Support vector networks. *Machine Learning*, *20*, 273–297.
Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of American Mathematical Society*, *39*, 1–49.
Friedman, A. (1982). *Modern analysis*. New York: Dover.
Girosi, F. (1994). Regularization theory, radial basis functions and networks. In V. Cherkassky, J. H. Friedman, & H. Wechsler (Eds.), *From statistics to neural networks* (pp. 166–187). Berlin, Heidelberg: Springer-Verlag.
Girosi, F. (1998). (AI memo 1606). An equivalence between sparse approximation and support vector machines. *Neural Computation*, *10*, 1455–1480.
Girosi, F., & Anzellotti, G. (1993). Rates of convergence for radial basis functions and neural networks. In R. J. Mammone (Ed.), *Artificial neural networks for speech and vision* (pp. 97–113). Chapman & Hall.
Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, *7*, 219–269.
Girosi, F., & Poggio, T. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*(4945), 978–982.
Gori, M., & Tesi, M. (1992). On the problem of local minima in backpropagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *14*, 76–86.
Jang, J.-S. R., & Sun, C.-T. (1993). Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks*, *4*(1), 156–159.
Kainen, P. C., Kůrková, V., & Sanguineti, M. (2009). Complexity of Gaussian radial basis networks approximating smooth functions. *Journal of Complexity*, *25*, 63–74.
Kainen, P. C., Kůrková, V., & Vogt, A. (2007). A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. *Journal of Approximation Theory*, *147*, 1–10.
Kecman, V. (2001). *Learning and soft computing*. Cambridge: MIT Press.
Kůrková, V. (2002). Universality and complexity of approximation of multivariable functions by feedforward networks. In R. Roy, M. Koeppen, S. Ovaska, T. Furuhashi, & F. Hoffmann (Eds.), *Softcomputing and industry: recent applications* (pp. 13–24). London: Springer-Verlag.
Kůrková, V. (2013). Gaussian radial and kernel networks with varying and fixed widths. In P. van Emde Boas, F. C. A. Groen, G. F. Italiano, J. Nawrocki, & H. Sack (Eds.), *SOFSEM 2013: theory and practice of computer science* (pp. 95–102). Prague: Institute of Computer Science, pp. II.
Kůrková, V., & Kainen, P. C. (1994). Functionally equivalent feedforward neural networks. *Neural Computation*, *6*(3), 543–558.
Kůrková, V., & Kainen, P. C. (1996). Singularities of finite scaling functions. *Applied Math Letters*, *9*(2), 33–37.
Kůrková, V., & Neruda, R. (1994). Uniqueness of functional representations by Gaussian basis function networks. In *Proceedings of ICANN'94* (pp. 471–474). London: Springer.
Loustau, S. (2008). Aggregation of SVM classifiers using Sobolev spaces. *Journal of Machine Learning Research*, *9*, 1559–1582.
Mhaskar, H. N. (1995). Versatile Gaussian networks. In *Proceedings of IEEE workshop of nonlinear image processing* (pp. 70–73).
Mhaskar, H. N. (2004). When is approximation by Gaussian networks necessarily a linear process? *Neural Networks*, *17*, 989–1001.
Park, J., & Sandberg, I. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, *3*, 246–257.
Park, J., & Sandberg, I. (1993). Approximation and radial basis function networks. *Neural Computation*, *5*, 305–316.
Pietsch, A. (1987). *Eigenvalues and s-numbers*. Cambridge: Cambridge University Press.
Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*, *8*, 143–195.
Poggio, T., & Girosi, F. (1990). Extensions of a theory of networks for approximation and learning: dimensionality reduction and clustering. *AI Memo*, 1167.
Poggio, T., & Smale, S. (2003). The mathematics of learning: dealing with data. *Notices of the American Mathematical Society*, *50*, 537–544.
Rudin, W. (1991). *Functional analysis*. Mc Graw-Hill.
Schmitt, M. (2002). Descartes rule of signs for radial basis function neural networks. *Neural Computation*, *14*, 2997–3011.
Steinwart, I., & Christmann, A. (2008). *Support vector machines*. New York: Springer-Verlag.
Sussman, H. J. (1992). Uniqueness of the weights for minimal feedforward nets with a given input–output map. *Neural Networks*, *5*(4), 589–593.
Verleysen, M., & Hlaváčková, K. (1996). Learning in RBF networks. In *International conference on nerual networks - ICNN* (pp. 199–204).
Wallace, M., Tsapatsoulis, N., & Kollias, S. (2005). Intelligent initialization of resource allocating RBF networks. *Neural Networks*, *18*(2), 117–122.