# Accuracy of Surrogate Solutions of Integral Equations by Feedforward Networks

**Věra Kůrková**

**Abstract** Surrogate solutions of Fredholm integral equations by feedforward neural networks are investigated theoretically. Convergence of surrogate solutions computable by networks with increasing numbers of computational units to theoretically optimal solutions is proven and upper bounds on rates of convergence are derived. The results hold for a variety of computational units, they are illustrated by examples of perceptrons and Gaussian radial units.

**Keywords** Surrogate modeling by neural networks · Approximate solutions of integral equations · Feedforward neural networks · Model complexity · Rates of approximation

## 1 Introduction

One of successful applications of feedforward neural networks is surrogate modelling of functional relationships. It has been successfully used for modelling of empirical functions, i.e., functions for which no mathematical formulas are known and thus their values can only be obtained experimentally. Often such experimental evaluations are too expensive or time consuming and so they are performed merely for samples of points in the domains of the empirical functions and the obtained values are used for training feedforward networks. The networks trained on such training sets play roles of surrogate models of these empirical functions. For example, input–output functions of feedforward networks have been used in chemistry as surrogate models of empirical functions assigning to compositions of chemicals measures of quality of catalyzers produced by reactions of these chemicals, in biology as models of

V. Kůrková (✉)
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic
e-mail: vera@cs.cas.cz

L. T. Kóczy et al. (eds.), *Issues and Challenges of Intelligent Systems and Computational Intelligence*, Studies in Computational Intelligence 530,
DOI: 10.1007/978-3-319-03206-1_7, 

empirical functions classifying structures of RNA, and in economy as models of functions assigning credit ratings to companies [1, 2]. It should be emphasized that results obtained by surrogate modelling of empirical functions can only be used as suggestions to be confirmed by additional experiments as no other than empirical knowledge of the functions is available. Also suitable types of network architectures and computational units have to be found experimentally.

In contrast to the case of empirical functions, for functions with known, although complicated, analytical descriptions, there is a potential for theoretical analysis of quality of surrogate models. When numerical computations of complicated analytical formulas are too time-consuming, relatively small samples of data obtained by such numerical computations can be sufficient for training feedforward networks. Investigation of mathematical properties of analytical formulas and their comparison with input-output functions of feedforward networks of various types can lead to estimates of accuracies of approximations and their dependence on types of computational units, their numbers, and input dimensions.

Many types of feedforward networks (including all standard types that are popular in applications as well as many others that may not have been considered by experimentalists) are known to be universal approximators. It means that it is possible to adjust their parameters so that they approximate to any desired accuracy a wide variety of mappings between subsets of multidimensional spaces. In particular, the universal approximation property has been proven for approximation of continuous functions on compact subsets of $d$-dimensional Euclidean spaces by one-hidden layer networks with almost all types of reasonable computational units (see, e.g., [3, 4]). It should be emphasizes that the universal approximation property requires potentially unlimited number of network units. Thus one can conclude that when a function with a complicated analytical description is continuous, surrogate models formed by input–output functions of networks of various types converge with increasing numbers of units to this function. However, a critical factor influencing whether a given type of network units is suitable for the task is the speed of the convergence. Such speed can differ considerably for various types of computational units. For some choices of network units, a sufficient accuracy can be achieved within a feasible bound on the number of network units, while for others, it might require numbers of units that are too large for a practical implementation. In particular for some high-dimensional tasks, the numbers of units of some types can grow with the input dimension exponentially, while choice of other types can lead to quadratic or even linear growth [5].

A large class of functions expressed by formulas, whose numerical calculations are difficult, is formed by solutions of Fredholm integral equations. These equations play an important role in many problems in applied science and engineering. They arise in image restoration, heat conduction, population modelling, potential theory and elasticity, etc. (see, e.g., [6–8]). Mathematical descriptions of solutions of Fredholm equations following from classical Fredholm theorem [9, p. 499] involve complicated expressions in terms of infinite Liouville–Neumann series with coefficients in the forms of integrals. Thus numerical calculations of these expressions are time consuming. Recently, several authors [10, 11] explored experimentally possi-

bilities of surrogate modelling of solutions of Fredholm equations by perceptron and kernel networks. Motivated by these experimental studies, Gnecco et al. [12] initiated a theoretical analysis of surrogate solutions of Fredholm equations computable by neural networks. In Refs. [12, 13], estimates of rates of approximation with increasing numbers of network units were derived for networks with kernel units induced by the same kernels as the kernels defining the equations and extended to certain smooth kernels.

In this chapter, we investigate surrogate solutions of Fredholm integral equations by networks with general computational units. Taking advantage of results from nonlinear approximation theory and suitable integral representations of functions in the form of "infinite" networks, we estimate how well surrogate solutions computable by feedforward networks can approximate exact solutions of Fredholm equations. We derive estimates of approximation errors measured in $\mathcal{L}^2$-norm. The estimates depend on relationships of kernels of the equations to types of computational units. We apply general results to networks with the most common computational units—sigmoidal perceptrons and Gaussian radial units. A preliminary version of this chapter appeared in a conference proceedings [14].

The chapter is organized as follows. In Sect. 2, we describe approximation of functions by feedforward neural networks. In Sect. 3, we introduce Fredholm integral equations and recall theoretical approach to their solutions. In Sect. 4, we apply some results from nonlinear approximation theory to approximation of solutions of Fredholm equations by feedforward networks. We illustrate our results by examples of surrogate solutions of Fredholm equations with the Gaussian kernel by networks with perceptrons and with Gaussian radial units. Section 5 is a discussion.

## 2 Approximation of Functions by Feedforward Neural Networks

A traditional approach to approximation of functions known only by samples of data points was based on linear methods such as polynomial interpolation. For suitable points $x_1, \ldots, x_m$ from the domain $X \subset \mathbb{R}^d$ of a function $\phi$ to be approximated, samples of empirically or numerically obtained approximations $\bar{\phi}(x_1), \ldots, \bar{\phi}(x_m)$ of its values $\phi(x_1), \ldots, \phi(x_m)$ are interpolated by functions from suitable $n$-dimensional function spaces. Such spaces are often generated as *linear spans*

$$\operatorname{span}\{g_1, \ldots, g_n\} := \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R} \right\}, \tag{1}$$

where the functions $g_1, \ldots, g_n$ are the first $n$ elements from a set $G = \{g_n \mid n \in \mathbb{N}_+\}$ with a *fixed linear ordering*. Typical examples of linear approximators are algebraic or trigonometric polynomials. They are obtained by linear combinations of powers of increasing degrees or trigonometric functions with increasing frequencies, respectively.

Feedforward neural networks have more adjustable parameters than linear models as in addition to coefficients of linear combinations of basis functions, also inner coefficients of computational units are optimized during learning. Thus they are sometimes called *variable-basis-approximation schemas* in contrast to traditional linear approximators which are called *fixed-basis-approximation schemas*. In some cases, especially in approximation of functions of large numbers of variables, it was proven that neural networks achieve better approximation rates than linear models with much smaller model complexities [15, 16].

*One-hidden-layer networks with one linear output unit* compute input–output functions from sets of the form

$$\operatorname{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i \mid w_i \in \mathbb{R}, g_i \in G \right\}, \tag{2}$$

where the set $G$ is sometimes called a *dictionary* [17] and $n$ is the *number of hidden computational units*. This number can be interpreted as a measure of *model complexity* of the network. In contrast to the case of linear approximation, the dictionary $G$ has no fixed ordering.

Often, dictionaries are parameterized families of functions modeling computational units, i.e., they are of the form

$$G_F(X, Y) := \{F(\cdot, y) : X \to \mathbb{R} \mid y \in Y\}, \tag{3}$$

where $F : X \times Y \to \mathbb{R}$ is a function of two variables, an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter $y \in Y \subseteq \mathbb{R}^s$. When $X = Y$, we write briefly $G_F(X)$. So one-hidden-layer networks with $n$ units from a dictionary $G_F(X, Y)$ compute functions from the set

$$\operatorname{span}_n G_F(X, Y) := \left\{ \sum_{i=1}^{n} w_i F(x, y_i) \mid w_i \in \mathbb{R}, y_i \in Y \right\}.$$

In some contexts, $F$ is called a *kernel*. However, the above-described computational scheme includes fairly general computational models, such as functions computable by perceptrons, radial or kernel units, Hermite functions, trigonometric polynomials, and splines. For example, with

$$F(x, y) = F(x, (v, b)) := \sigma(\langle v, x \rangle + b)$$

and $\sigma : \mathbb{R} \to \mathbb{R}$ a sigmoidal function, the dictionary $G_F(X, Y)$ describes a set of functions computable by *perceptrons*. *Radial (RBF) units* with an activation function $\beta : \mathbb{R} \to \mathbb{R}$ are modelled by the kernel

$$F(x, y) = F(x, (v, b)) := \beta(v\|x - b\|).$$

Typical choice of $\beta$ is the Gaussian function. *Kernel units* used in support vector machine (SVM) have the form $F(x, y)$ where $F : X \times X \to \mathbb{R}$ is a symmetric positive semidefinite function [9].

Various learning algorithms optimize parameters $y_1, \ldots, y_n$ of computational units as well as coefficients $w_1, \ldots, w_n$ of their linear combinations so that network input–output functions

$$\sum_{i=1}^{n} w_i \, F(., y_i)$$

from the set $\text{span}_n \, G_F(X, Y)$ fit well to training samples $\{(x_i, \bar{\phi}(x_i) \,|i = 1, \ldots, m\}$.

## 3 Fredholm Integral Equations

Solving an *inhomogeneous Fredholm integral equation of the second kind* on a domain $X \subseteq \mathbb{R}^d$ for a given $\lambda \in \mathbb{R} \setminus \{0\}$, $K : X \times X \to \mathbb{R}$, and $f : X \to \mathbb{R}$ is a task of finding a function $\phi : X \to \mathbb{R}$ such that for all $x \in X$

$$\phi(x) - \lambda \int_X \phi(y) K(x, y) \, dy = f(x). \tag{4}$$

The function $\phi$ is called *solution*, $f$ *data*, $K$ *kernel*, and $\lambda$ *parameter* of the equation (4).

Fredholm equations can be described in terms of theory of inverse problems. Formally, an *inverse problem* is defined by a linear operator $A : \mathcal{X} \to \mathcal{Y}$ between two function spaces. It is a task of finding for $f \in \mathcal{Y}$ (called *data*) some $\phi \in \mathcal{X}$ (called *solution*) such that

$$A(\phi) = f.$$

Let $T_K$ denotes the integral operator with a kernel $K : X \times X \to \mathbb{R}$ defined for every $\phi$ in a suitable function space $\mathcal{X}$ as

$$T_K(\phi)(x) := \int_X \phi(y) \, K(x, y) \, dy \tag{5}$$

and $I_{\mathcal{X}}$ denotes the identity operator. Then the Fredholm equation (4) can be represented as an inverse problem defined by the linear operator $I_{\mathcal{X}} - \lambda T_K$. So it is a problem of finding for a given data $f$ a solution $\phi$ such that

$$(I_{\mathcal{X}} - \lambda T_K)(\phi) = f. \tag{6}$$

The classical Fredholm alternative theorem from 1903 proved existence and uniqueness of solutions of Fredholm equations for continuous one-variable functions on intervals. A modern version holding for general Banach spaces is stated in the next theorem from [9, p. 499]. Recall that an operator $T : (\mathcal{X}, \|.\|_{\mathcal{X}}) \to (\mathcal{Y}, \|.\|_{\mathcal{Y}})$ between two Banach spaces is called *compact* if it maps bounded sets to precompact sets (i.e., sets whose closures are compact).

**Theorem 1** *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Banach space, $T : (\mathcal{X}, \|.\|_{\mathcal{X}}) \to (\mathcal{X}, \|.\|_{\mathcal{X}})$ be a compact operator, and $I_{\mathcal{X}}$ be the identity operator. Then the operator $I_{\mathcal{X}} + T : (\mathcal{X}, \|.\|_{\mathcal{X}}) \to (\mathcal{X}, \|.\|_{\mathcal{X}})$ is one-to-one if and only if it is onto.*

A straightforward corollary of Theorem 1 guarantees existence and uniqueness of solutions of the inverse problem (6) when $T$ is a compact operator and $1/\lambda$ is not its eigenvalue (i.e., there is no $\phi \in \mathcal{X}$ for which $T(\phi) = \frac{\phi}{\lambda}$).

**Corollary 1** *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Banach space, $T : (\mathcal{X}, \|.\|_{\mathcal{X}}) \to (\mathcal{X}, \|.\|_{\mathcal{X}})$ be a compact operator, $I_{\mathcal{X}}$ be the identity operator, and $\lambda \neq 0$ be such that $1/\lambda$ is not an eigenvalue of $T$. Then the operator $I_{\mathcal{X}} - \lambda T$ is invertible (one-to-one and onto).*

If $1/\lambda$ is not an eigenvalue of $T$, then $I_{\mathcal{X}} - \lambda T_K$ is one-to-one and so by Theorem 1 it is also onto. Thus for any data $f$, there is a unique solution $\phi$ of the equation $I_{\mathcal{X}} - \lambda T_K)(\phi) = f$. Corollary 1 can be applied to a Fredholm integral equation with a kernel $K$ inducing a compact operator $T_K$. The following proposition gives conditions guaranteeing compactness of operators $T_K$ in spaces $(\mathcal{C}(X), \|.\|_{\sup})$ of bounded continuous functions on $X \subseteq \mathbb{R}^d$ with the supremum norm $\|f\|_{\sup} = \sup_{x \in X} |f(x)|$ and in spaces $(\mathcal{L}^2(X), \|.\|_{\mathcal{L}^2})$ of square integrable functions with the norm $\|f\|_{\mathcal{L}^2} = \left(\int_X f(x)^2\, dx\right)^{1/2}$. The proof is well-known and easy to check (see, e.g., [18, p. 112]).

**Proposition 1** *(i) If $X \subset \mathbb{R}^d$ is compact and $K : X \times X \to \mathbb{R}$ is continuous, then $T_K : (\mathcal{C}(X), \|.\|_{\sup}) \to (\mathcal{C}(X), \|.\|_{\sup})$ is a compact operator.*
*(ii) If $X \subset \mathbb{R}^d$ and $K \in \mathcal{L}^2(X \times X)$, then $T_K : (\mathcal{L}^2(X), \|.\|_{\mathcal{L}^2}) \to (\mathcal{L}^2(X), \|.\|_{\mathcal{L}^2})$ is a compact operator.*

So by Corollary 1, when the assumptions of the Proposition 1(i) or (ii) are satisfied and $1/\lambda$ is not an eigenvalue of $T_K$, then for every $f$ in $\mathcal{C}(X)$ or $\mathcal{L}^2(X)$, resp., there exists a unique solution $\phi$ of the Eq. (4). It is known (see, e.g, [19]) that the solution $\phi$ can be expressed as

$$\phi(x) = f(x) - \lambda \int_X f(y)\, R_K^{\lambda}(x, y)\, dy\,, \tag{7}$$

where $R_K^{\lambda} : X \times X \to \mathbb{R}$ is called a *resolvent kernel*. However, the formula expressing the resolvent kernel is not suitable for efficient computation as it is expressed as an infinite Neumann series in powers of $\lambda$ with coefficients in the form of integrals with iterated kernels [20, p. 140]. So numerical calculations of values of solutions of Fredholm equations based on (7) are quite computationally demanding. Thus various

methods of finding surrogate solutions of (4) have been used [10, 11]. Traditional methods employed polynomial interpolation. Recently, approximations of solutions by feedforward networks were explored experimentally. Such networks were trained on samples of input–output pairs $\{(x_1, \bar{\phi}(x_1)), \ldots, (x_m, \bar{\phi}(x_m)\}$, where $\{x_1, \ldots, x_m\}$ are selected points from the domain $X$ and $\{\bar{\phi}(x_1), \ldots, \bar{\phi}(x_m)\}$ are numerically computed approximations of values $\{\phi(x_1), \ldots, \phi(x_m)\}$ of the solution $\phi$. In these experiments, one-hidden-layer networks with perceptrons and Gaussian radial units were used. However, without a theoretical analysis, it is not clear how to choose a proper type and number $n$ of network units to guarantee that input–output functions approximate well the solution and the networks are not too large to make their implementation unfeasible.

## 4 Rates of Convergence of Surrogate Solutions

Estimates of numbers of network units needed to guarantee a required accuracies of surrogate solutions of Fredholm equations by neural networks of various types can be obtained from inspection of upper bounds on rates of variable-basis approximation. Some such bounds have the form $\frac{\xi(h,G)}{\sqrt{n}}$, where $n$ is the number of network units and $\xi(h, G)$ depends on a certain norm of the function $h$ to be approximated and the dictionary $G$.

For our purposes we need a reformulation of this theorem in terms of a norm tailored to the dictionary $G$. The norm is defined quite generally for any bounded nonempty subset $G$ of a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$. It is called *$G$-variation*, denoted $\|.\|_G$, and defined for all $f \in \mathcal{X}$ as

$$\|f\|_{G,\mathcal{X}} := \inf \{c > 0 \mid f/c \in \mathrm{cl}_{\mathcal{X}} \,\mathrm{conv}\,(G \cup -G)\},$$

where the closure $\mathrm{cl}_{\mathcal{X}}$ is taken with respect to the topology generated by the norm $\|.\|_{\mathcal{X}}$ and conv denotes the *convex hull*. So $G$-variation depends on the ambient space norm, but when it is clear from the context, we write merely $\|f\|_G$ instead of $\|f\|_{G,\mathcal{X}}$.

The concept of variational norm was introduced by Barron [21] for sets of characteristic functions, in particular for the set of characteristic functions of half-spaces corresponding to the dictionary of functions computable by Heaviside perceptrons. Barron's concept was generalized in [22, 23] to variation with respect to an arbitrary bounded set of functions and applied to various dictionaries of computational units such as Gaussian RBF units or kernel units [24].

The following theorem on rates of approximation by sets of the form $\mathrm{span}_n G$ is a reformulation from [23] of results by Maurey [25], Jones [26], Barron [27] in terms of $G$-variation. For a normed linear space $(\mathcal{X}, \|.\|_{\mathcal{X}})$, $g \in \mathcal{X}$ and $A \subset \mathcal{X}$, we denote by

$$\|g - A\|_{\mathcal{X}} := \inf_{f \in A} \|g - f\|_{\mathcal{X}}$$

the *distance* of $g$ from $A$.

**Theorem 2** *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a Hilbert space, $G$ its bounded nonempty subset, $s_G = \sup_{g\in G} \|g\|_{\mathcal{X}}$, $f \in \mathcal{X}$, and $n$ be a positive integer. Then*

$$\|h - \mathrm{span}_n G\|_{\mathcal{X}}^2 \leq \frac{s_G^2 \|h\|_G^2 - \|h\|_{\mathcal{X}}^2}{n}.$$

Theorem 2 guarantees that for every $\varepsilon > 0$ and $n$ satisfying

$$n \geq \left(\frac{s_G \, \|h\|_G}{\varepsilon}\right)^2,$$

a network with $n$ units computing functions from the dictionary $G$ approximates the function $h$ within $\varepsilon$. So the size of $G$-variation of the function $h$ to be approximated is a critical factor influencing model complexities of networks with units from the dictionary $G$ approximating $h$. Generally, it is not easy to estimate $G$-variation. However, the following theorem from [28] shows that for the special case of functions with integral representations in the form of "infinite networks", variational norms are bounded from above by the $\mathcal{L}^1$-norms of "output-weight" functions of these networks.

**Theorem 3** *Let $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}^s$, $w \in \mathcal{L}^1(Y)$, $K : X \times Y \to \mathbb{R}$ be such that $G_K(X, Y) = \{K(., y) \mid y \in Y\}$ is a bounded subset of $(\mathcal{L}^2(X), \|.\|_{\mathcal{L}^2})$, and $h \in \mathcal{L}^2(X)$ be such that for all $x \in X$, $h(x) = \int_Y w(y)\, K(x, y)\, dy$. Then*

$$\|h\|_{G_K(X,Y)} \leq \|w\|_{\mathcal{L}^1}.$$

In experiments with surrogate solutions of Fredholm equations [10, 11], common computational units such as perceptrons and Gaussian RBFs were used to approximate solutions of Fredholm equations with a variety of kernels $K$. Thus to apply Theorem 3 to these cases, we need estimates of $G$-variations for dictionaries $G$ of general computational units in terms of $G_K$-variations induced by various kernels $K$. The next proposition from [29] describes a relationship between variations with respect to two sets, $G$ and $F$.

**Proposition 2** *Let $(\mathcal{X}, \|.\|_{\mathcal{X}})$ be a normed linear space, $F$ and $G$ its bounded subsets such that $c_{G,F} := \sup_{g\in F} \|g\|_G \infty$. Then for all $h \in \mathcal{X}$, $\|h\|_G \leq c_{G,F}\, \|h\|_F$.*

Combining Theorems 2, 3, and Proposition 2, we obtain the next corollary on rates of approximation of functions which can be expressed as $h = T_K(w)$ by one-hidden-layer networks with units from a dictionary of computational units $G$.

**Corollary 2** *Let $X \subseteq \mathbb{R}^d$, $K : X \times Y \to \mathbb{R}$ be a bounded kernel, and $h \in \mathcal{L}^2(X)$ such that $h = T_K(w) = \int_Y w(y) K(., y)\, dy$ for some $w \in \mathcal{L}^1(Y)$, where $G_K(X, Y)$*

*is a bounded subset of* $\mathcal{L}^2(X)$. *Let* $G$ *be a bounded subset of* $\mathcal{L}^2(X)$ *with* $s_G = \sup_{g\in G} \|g\|_{\mathcal{L}^2}$ *such that* $c_{G,K} = \sup_{y\in Y} \|K(.,y)\|_G$ *is finite. Then for all* $n > 0$,

$$\left\|h - \text{span}_n\, G\right\|_{\mathcal{L}^2} \leq \frac{s_G\, c_{G,K}\, \|w\|_{\mathcal{L}^1}}{\sqrt{n}}.$$

A critical factor in the estimate given in Corollary 2 is the $\mathcal{L}^1$-norm of the "output-weight function" $w$ in the representation of the function $h$ to be approximated as an "infinite network" with units from the dictionary $G_K$ in the form $h(x) = T_K(w) = \int_Y w(y)\, K(x,y)\, dy$. We apply Corollary 2 to the representation

$$\phi - f = T_K(\lambda\,\phi) = \lambda \int_X \phi(y) K(x,y)\, dy\,,$$

where $\lambda\,\phi$ plays the role of the "output-weight" function in the infinite network $\int_X \lambda\,\phi(y)\, K(x,y)\, dy$.

**Theorem 4** *Let* $X \subset \mathbb{R}^d$ *be compact,* $K : X \times X \to \mathbb{R}$ *be a bounded kernel such that* $K \in \mathcal{L}^2(X \times X)$, $\rho_K = \int_X \sup_{y\in X} |K(x,y)|dx$ *be finite,* $G$ *be a bounded subset of* $\mathcal{L}^2(X)$ *with* $s_G = \sup_{g\in G} \|g\|_{\mathcal{L}^2}$ *such that* $c_{G,K} = \sup_{y\in Y} \|K(.,y)\|_G$ *is finite, and* $\lambda \neq 0$ *be such that* $\frac{1}{\lambda}$ *is not an eigenvalue of* $T_K$ *and* $|\lambda|\,\rho_K < 1$. *Then the solution* $\phi$ *of the Eq. (4) satisfies for all* $n > 0$,

$$\|\phi - f - \text{span}_n\, G\|_{\mathcal{L}^2} \leq \frac{s_G\, c_{G,K}\, |\lambda|\, \|f\|_{\mathcal{L}^1}}{(1 - |\lambda|\,\rho_K)\,\sqrt{n}}\,.$$

**Proof** As $\phi - f$ satisfies the Fredholm equation (4), we have for every $x \in X$,

$$|\phi(x)| \leq |\lambda|\, \|\phi\|_{\mathcal{L}^1} \sup_{y\in X} |K(x,y)| + |f(x)|.$$

Integrating over $X$ we get

$$\|\phi\|_{\mathcal{L}^1} \leq |\lambda|\,\rho_K\, \|\phi\|_{\mathcal{L}^1} + \|f\|_{\mathcal{L}^1}$$

and so $\|\phi\|_{\mathcal{L}^1}\,(1 - |\lambda|\,\rho_K) \leq \|f\|_{\mathcal{L}^1}$. This inequality is non trivial only when $|\lambda| < \frac{1}{\rho_K}$. Thus we get $\|w\|_{\mathcal{L}^1} = |\lambda|\|\phi\|_{\mathcal{L}^1} \leq \frac{|\lambda|\,\|f\|_{\mathcal{L}^1}}{1-|\lambda|\,\rho_K}$. The statement then follows from Corollary 2. □

Theorem 4 estimates rates of approximation of the function

$$\phi - f = \lambda \int_X f(y)\, R_K^\lambda(x,y)\, dy$$

by functions computable by networks with units from a dictionary $G$. As $f$ plays a role of a constant function, we can consider a surrogate solution formed by an input–output function of the network with $n$ units from the dictionary $G$ and one unit assigning to an input $x \in X$ the value $f(x)$.

With an increasing number of network units, the upper bound on rate of approximation decreases with $1/\sqrt{n}$. The speed of decrease depends on the $\mathcal{L}^1$-norm of the function $f$ representing data in the Fredholm equation, bound $c_{G,K}$ on $G$-variations of functions from the dictionary $G_K$ and $\rho_X$ depending on the size of the domain $X$ where the solution is approximated. For $|\lambda| < \frac{1}{\rho_K}$ and any bounded dictionary $G$ with finite bound $c_{G,K}$ on $G_{K(X)}$-variations on its elements, input–output functions of networks with increasing numbers of units from $G$ converge to the function $\phi - f$. When for a reasonable number $n$ of network units, the upper bound from Theorem 4 is sufficiently small, the network can serve as a good surrogate model of the solution of the Fredholm equation.

Note that the $\mathcal{L}^1$-norm of the data $f$ does not depend on the choice of a dictionary of computational units. Also $\rho_K = \int_X \sup_{y \in X} |K(x, y)| dx$ is determined by the Fredholm equation to be solved. It depends on the Lebesgue measure of the domain $X$ and properties of the kernel $K$ of the equation. For large dimensions $d$, choice of the domain has a strong effect on the upper bound from Theorem 4. For example, the Lebesgue measure of the unit cube $[0, 1]^d$ is equal to 1 for all dimensions $d$, while Lebesgue measures of cubes od sizes larger than 1 grow exponentially with $d$ increasing, and Lebesgue measures of the $d$-dimensional unit balls decrease exponentially quickly to zero. The only factor that can be influenced by a choice of a type of computational units is $c_{G,K}$ expressing a bound on $G$-variations of functions induced by the kernel $K$ of the equation.

To illustrate our results, consider approximation of Fredholm equations with the Gaussian kernel

$$K_b(x, y) = e^{-b\|x-y\|}$$

with the width $b$ by surrogate solutions in the form of input–output functions of networks with two types of popular units: sigmoidal perceptrons and Gaussian radial units. Note that Fredholm equations with Gaussian kernels arise, e.g., in image restoration problems [8]. By $\mu$ is denoted the *Lebesgue measure* on $\mathbb{R}^d$ and by $P_d^\sigma(X)$ the *dictionary of functions on X computable by sigmoidal perceptrons*.

**Corollary 3** *Let $X \subset \mathbb{R}^d$ be compact, $b > 0$, $K_b(x, y) = e^{-b\|x-y\|^2}$, $\lambda \neq 0$ be such that $\frac{1}{\lambda}$ is not an eigenvalue of $T_{K_b}$ and $|\lambda| < 1$. Then the solution $\phi$ of the Eq. (4) with $f$ continuous satisfies for all $n > 0$*

$$\|\phi - f - \text{span}_n G_{K_b}(X)\|_{\mathcal{L}^2} \leq \frac{\mu(X)\,|\lambda|\,\|f\|_{\mathcal{L}^1}}{(1 - |\lambda|\,\mu(X))\,\sqrt{n}}$$

*and*

$$\|\phi - f - \text{span}_n P_d^\sigma(X)\|_{\mathcal{L}^2} \leq \frac{\mu(X)\,2d\,|\lambda|\,\|f\|_{\mathcal{L}^1}}{(1 - |\lambda|\,\mu(X))\,\sqrt{n}}\,.$$

**Proof** It was shown in Ref. [30] that variation of the $d$-dimensional Gaussian with respect to the dictionary formed by sigmoidal perceptrons is bounded from above by $2d$ and thus by Proposition 2, $c_{P_d^\sigma, K_b} \leq 2d$. The statement then follows by Theorem 4, an estimate $s_{G_{K_b}} \leq \mu(X)$ and equalities $s_{P_d^\sigma} = \mu(X)$ and $\rho_{K_b} = \mu(X)$. □

## 5 Discussion

Taking advantage of results from mathematical theory of neurocomputing holding for functions representable as "infinite neural networks" we derived estimates of rates of convergence of surrogate solutions of Fredholm equations computable by feedforward neural networks. Our estimates decrease with increasing number of network units $n$, they are smaller than $\frac{1}{\sqrt{n}}$ multiplied by a product of two factors, the first one depending on the parameters of the equation $f$, $K$, $\lambda$ and the domain $X$, and the second one depending on combination of the kernel $K$ and the dictionary of computational units $G$. Thus our results show that a proper choice of a type of computational units can influence speed of convergence of surrogate solutions, however for high dimensions, a choice of the domain can have a stronger impact.

**Acknowledgments** This work was partially supported by GA ČR grant P202/11/1368, MŠMT grant COST LD13002, and institutional support of the Institute of Computer Science 67985807.

## References

1. Forrester, A., Sobester, A., Keane, A.: Engineering Design via Surrogate Modelling: A Practical Guide. Wiley, Chichester (2008)
2. Baerns, M., Holeňa, M.: Combinatorial Development of Solid Catalytic Materials. Imperial College Press, London (2009)
3. Park, J., Sandberg, I.W.: Approximation and radial-basis-function networks. Neural Comput. **5**, 305–316 (1993)
4. Pinkus, A.: Approximation theory of the MLP model in neural networks. Acta Numerica **8**, 143–195 (1999)
5. Kainen, P.C., Kůrková, V., Sanguineti, M.: Dependence of computational models on input dimension: Tractability of approximation and optimization tasks. IEEE Trans. Inf. Theory **58**(2), 1203–1214 (2012)
6. Lovitt, W.V.: Linear Integral Equations. Dover, New York (1950)
7. Lonseth, A.T.: Sources and applications of integral equations. SIAM Rev. **19**, 241–278 (1977)
8. Lu, Y., Shen, L., Xu, Y.: Integral equation models for image restoration: high accuracy methods and fast algorithms. Inverse Prob. **26**, 045006 (2010)
9. Steinwart, I., Christmann, A.: Support Vector Machines. Springer, New York (2008)
10. Golbabai, A., Seifollahi, S.: Numerical solution of the second kind integral equations using radial basis function networks. Appl. Math. Comput. **174**, 877–883 (2006)
11. Effati, S., Buzhabadi, R.: A neural network approach for solving Fredholm integral equations of the second kind. Neural Comput. Appl. **21**(5), 843–852 (2012)

12. Gnecco, G., Kůrková, V., Sanguineti, M.: Bounds for approximate solutions of Fredholm integral equations using kernel networks. In: Honkela, T., et al. (eds.) Lecture Notes in Computer Science (Proceedings of ICANN 2011), vol. 6791, pp. 126–133. Springer, Heidelberg (2011)
13. Gnecco, G., Kůrková, V., Sanguineti, M.: Accuracy of approximations of solutions to Fredholm equations by kernel methods. Appl. Math. Comput. **218**, 7481–7497 (2012)
14. Kůrková, V.: Surrogate modeling of solutions of integral equations by neural networks. In: Iliadis, L., et al., (eds.), AIAI 2012—Artificial Intelligence Applications and Innovations, IFIP AICT 381, IFIP International Federation for Information Processing (2012), pp. 88-96
15. Gnecco, G., Kůrková, V., Sanguineti, M.: Some comparisons of complexity in dictionary-based and linear computational models. Neural Netw. **24**, 171–182 (2011)
16. Gnecco, G., Kůrková, V., Sanguineti, M.: Can dictionary-based computational models outperform the best linear ones? Neural Netw. **24**, 881–887 (2011)
17. Gribonval, R., Vandergheynst, P.: On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. IEEE Trans. Inf. Theory **52**, 255–261 (2006)
18. Rudin, W.: Functional Analysis. McGraw-Hill, Boston (1991)
19. Atkinson, K.: The Numerical Solution of Integral Equations of the Second Kind. Cambridge University Press, Cambridge (1997)
20. Courant, R., Hilbert, D.: Methods of Mathematical Physics, vol. I. Wiley, New York (1989)
21. Barron, A.R.: Neural net approximation. In: Narendra, K. (ed.) Proceedings of 7th Yale Workshop on Adaptive and Learning Systems. Yale University Press, New Haven (1992)
22. Kůrková, V.: Dimension-independent rates of approximation by neural networks. In: Warwick, K., Kárný, M. (eds.) Computer-Intensive Methods in Control and Signal Processing. The Curse of Dimensionality, pp. 261–270. Birkhäuser, Boston (1997)
23. Kůrková, V.: High-dimensional approximation and optimization by neural networks. In: Suykens, J., Horváth, G., Basu, S., Micchelli, C., Vandewalle, J. (eds.), Advances in Learning Theory: Methods, Models and Applications (Chapter 4), pp. 69–88. IOS Press, Amsterdam (2003)
24. Kainen, P.C., Kůrková, V., Sanguineti, M.: Complexity of Gaussian radial-basis networks approximating smooth functions. J. Complexity **25**, 63–74 (2009)
25. Pisier, G.: Remarques sur un résultat non publié de B. Maurey. In: Séminaire d'Analyse Fonctionnelle 1980–1981, vol. I, no. 12, École Polytechnique, Centre de Mathématiques, Palaiseau, France (1981)
26. Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. Ann. Stat. **20**, 608–613 (1992)
27. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory **39**, 930–945 (1993)
28. Kůrková, V.: Complexity estimates based on integral transforms induced by computational units. Neural Netw. **30**, 160–167 (2012)
29. Kůrková, V., Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation. IEEE Trans. Inf. Theory **47**, 2659–2665 (2001)
30. Kainen, P.C., Kůrková, V., Vogt, A.: A Sobolev-type upper bound for rates of approximation by linear combinations of Heaviside plane waves. J. Approx. Theory **147**, 1–10 (2007)