Applied Mathematics and Computation 218 (2012) 7481–7497

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

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

Accuracy of approximations of solutions to Fredholm equations by kernel methods ☆

Giorgio Gnecco [a], Věra Kůrková [b], Marcello Sanguineti [a,*]

[a] Department of Communications, Computer, and System Sciences (DIST), University of Genoa, Via Opera Pia 13, 16145 Genova, Italy
[b] Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic

ARTICLE INFO

Keywords:
Approximate solutions to integral equations
Radial and kernel-based networks
Gaussian kernels
Model complexity
Analysis of algorithms

ABSTRACT

Approximate solutions to inhomogeneous Fredholm integral equations of the second kind by radial and kernel networks are investigated. Upper bounds are derived on errors in approximation of solutions of these equations by networks with increasing model complexity. The bounds are obtained using results from nonlinear approximation theory. The results are applied to networks with Gaussian and kernel units and illustrated by numerical simulations.

1. Introduction

Fredholm integral equations play an important role in many problems in applied science and engineering. They arise in image restoration [1], differential problems with auxiliary boundary conditions, potential theory and elasticity [2, Chapter IV], and many other problems (see, e.g., [3]). Solving an *inhomogeneous Fredholm integral equation of the second kind* is an inverse problem of finding for a function f representing measured data a function ϕ which is mapped to f by a linear operator of the form $I - \lambda T_K$, where I is the identity operator, T_K is an integral operator with a kernel K defined as

$$T_K(\phi)(x) := \int_X \phi(y)K(x,y)dy$$

and λ is a parameter. So for every x in the domain of the functions f and ϕ, the solution ϕ of the Fredholm equation satisfies

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

The classical Fredholm Alternative Theorem from 1903 (see, e.g., [4, Section 1.3]) gives under suitable continuity and compactness assumptions a formula describing an exact solution of the Eq. (1) in terms of Liouville–Neumann series and resolvent kernel (see Appendix A.1 in the Appendix). But numerical calculations based on this theorem may be computationally demanding (see Appendix A.2). Thus various methods constructing so-called surrogate models [5] of solutions have been developed. A classical approach is based on polynomial interpolation of numerical solutions at certain collocation points in the domain of ϕ (see, e.g., [6, Chapter 11]).

☆ G. Gnecco and M. Sanguineti were partially supported by a PRIN Grant from the Italian Ministry for University and Research, project "Adaptive State Estimation and Optimal Control". V. Kůrková was partially supported by GA ČR Grant P202/11/1368, MŠMT Grant INTELLI OC10047, and the Institutional Research Plan AV0Z10300504. Collaboration of V. Kůrková with M. Sanguineti and G. Gnecco was partially supported by CNR - AV ČR project 2010–2012 "Complexity of Neural-Network and Kernel Computational Models".

* Corresponding author.
E-mail addresses: giorgio.gnecco@dist.unige.it (G. Gnecco), vera@cs.cas.cz (V. Kůrková), marcello@dist.unige.it (M. Sanguineti).

0096-3003/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2012.01.015

Recently, several authors proposed an alternative approach to classical surrogate modeling of solutions of Fredholm equations by replacing polynomials with various nonlinear computational models such as perceptron networks [7] and Gaussian and multiquadric radial-basis functions (RBF) [8,9]. In these methods, numerically calculated values of solutions at collocation points are used as training sets for various learning algorithms that optimize parameters of neural networks with various types of units (perceptrons or RBF). As neural networks have more free parameters than linear models, in many cases they achieve better accuracy of approximation with smaller model complexity than linear methods [10–13]. Thus one may expect that approximation of solutions of Fredholm integral equations by neural networks is more efficient than approximation by linear models. Recent experimental results [7–9] indicate usefulness of applications of neural networks to solutions of Fredholm equations.

In this paper, we theoretically investigate efficiency of neural-network surrogate modeling of these solutions. Using results from nonlinear approximation theory which have been motivated by problems from neurocomputing, we derive estimates of rates of approximation of solutions of Fredholm equations by networks with increasing number of units. We obtain estimates for networks with units of several types: Gaussian radial units, kernel units induced by degenerate kernels, and smooth kernels. Our results show that the number of computational units required for a desired accuracy of approximation depends on the VC-dimension of the set of functions computable by network units (which is in some literature called "dictionary" [11,12,14]), the Lebesgue measure of the domain where the solution is searched for, and the $\mathcal{L}^1$-norm of the function f describing the data in the Eq. (1). We apply our estimates to networks with Gaussian radial units and kernel units induced by degenerate kernels, for which we estimate VC-dimensions of induced dictionaries. These results are theoretical – they estimate approximation accuracies achievable by networks with increasing numbers of units. Then we address the issue of designing networks achieving our theoretical estimates. A preliminary version of some of the results appeared in conference proceedings [15].

The paper is organized as follows. In Section 2, Fredholm integral equations and methods of their approximate solutions are briefly reviewed. Section 3 recalls some results from a branch of nonlinear approximation theory, called variable-basis approximation. These results are used as tools in Sections 4 and 5 to derive estimates of accuracy of approximate solutions to Fredholm equations. Section 6 applies estimates to Gaussian RBF networks. Section 7 studies a constructive method to achieve the upper bounds and Section 8 analyzes learning methods for networks computing approximate solutions. Section 9 provides some numerical results and Section 10 is a short discussion. To make the paper self-contained, we have included an appendix (Appendix A) with some technical issues.

2. Approximate solutions to Fredholm integral equations

An *inhomogeneous Fredholm integral equation of the second kind* on $X \subset \mathbb{R}^d$ is defined by a kernel $K : X \times X \to \mathbb{R}$, a function $f : X \to \mathbb{R}$, and a parameter $\lambda \in \mathbb{R} \setminus \{0\}$. Its solution is a function $\phi : X \to \mathbb{R}$ satisfying for all $x \in X$,

$$\phi(x) - \lambda \int_X \phi(y) K(x,y) dy = f(x). \tag{1}$$

Solving the Eq. (1) is an inverse problem of finding a function ϕ which is mapped to f by the operator

$$I - \lambda T_K,$$

where I denotes the identity operator and T_K is the operator defined as

$$T_K(\phi)(x) := \int_X \phi(y) K(x,y) dy. \tag{2}$$

So a solution of the Eq. (1) is a function ϕ such that

$$(I - \lambda T_K)(\phi) = f.$$

Let $\mathcal{C}(X)$ denotes the space of continuous functions on X. When X is compact and K is continuous, then $T_K : \mathcal{C}(X) \to \mathcal{C}(X)$ is compact and thus by the Fredholm Alternative Theorem [4, Section 1.3] for every $\lambda \neq 0$ such that $1/\lambda$ is not an eigenvalue of the operator $T_K : \mathcal{C}(X) \to \mathcal{C}(X)$ and every $f \in \mathcal{C}(X)$, there exists a unique continuous solution $\phi : X \to \mathbb{R}$ of the Eq. (1). So under latter assumptions, the inverse problem defined by $I - \lambda T_K$ is well posed. Its solution has the form

$$\phi(x) = f(x) - \lambda \int_X R_K^\lambda(x,y) f(y) dy, \tag{3}$$

where $R_K^\lambda : X \times X$ is a continuous function called *resolvent kernel* (see Appendix A.1 in the Appendix).

Although the Fredholm Alternative Theorem gives a formula expressing the solution ϕ of the Eq. (1), it may be of limited practical use as numerical calculations based on the formula (3) are sometimes too computationally demanding (see Section A.2 in the Appendix). Thus various approximate methods have been developed which require these calculations only in certain *collocation points* [6, Chapter 11]. In all other points of the domain, merely approximations of values of the solution are calculated using suitable surrogate models [5]. Such models are chosen in such a way that they provide functions interpolating or closely approximating the numerically calculated values of the solution in collocation points. Traditional

surrogate models used for solutions of Fredholm equations have been formed by functions from *linear* subspaces of $C(X)$, such as polynomials up to a certain degree.

Recently, alternative surrogate models in the form of perceptron [7] and Gaussian radial-basis function networks [9] were explored experimentally [7,9]. It is important to guarantee that surrogate solutions approximate the solution given by the formula (3) sufficiently well also in non collocation points. It is known that perceptron and Gaussian RBF networks are universal approximators (see, e.g., [16,17]). So any continuous function on a compact subset of $\mathbb{R}^d$ (in particular the solution ϕ) can be approximated arbitrarily well by input–output functions of these networks with sufficiently large numbers of units. However, to evaluate efficiency of neural-network based surrogate modelling of solutions of Fredholm equations one has to investigate the trade-off between model complexity (measured by the number of network units) and accuracy of approximation which such networks can provide.

3. Approximation from a dictionary

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

$$\mathrm{span}_n G := \left\{ \sum_{i=1}^{n} w_i g_i | w_i \in \mathbb{R}, g_i \in G \right\}, \tag{4}$$

where the set G is sometimes called a *dictionary* [14] and n is the *number of computational units* in so called *hidden layer*. This number can be interpreted as a measure of *model complexity* of the network. Approximation by sets of the form $\mathrm{span}_n G$ is called *variable-basis approximation* in contrast to traditional *linear approximation*, where a fixed linear ordering $\{g_n | n \in \mathbb{N}_+\}$ of the set G is given and approximating families are n-dimensional subspaces of the form

$$\mathrm{span}\{g_1, \ldots, g_n\} := \left\{ \sum_{i=1}^{n} w_i g_i | w_i \in \mathbb{R} \right\}.$$

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

$$G_K(X, Y) := \{K(\cdot, y) : X \to \mathbb{R} | y \in Y\}, \tag{5}$$

where $K : X \times Y \to \mathbb{R}$ is a function of two variables, an input vector $x \in X \subseteq \mathbb{R}^d$ and a parameter $y \in Y \subseteq \mathbb{R}^s$. When $X = Y$, we write briefly $G_K(X)$ and when $X = Y = \mathbb{R}^d$, we write merely G_K. In some contexts, K 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

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

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product on $\mathbb{R}^d$, and $\sigma : \mathbb{R} \to \mathbb{R}$ a sigmoidal function, the computational scheme (4) describes one-hidden-layer *perceptron networks*. *RBF units* with an activation function $\beta : \mathbb{R} \to \mathbb{R}$ are modeled by the kernel

$$K(x, y) = K(x, (v, b)) = \beta(v \| x - b \|),$$

where $\beta : \mathbb{R} \to \mathbb{R}$ is an even function. A typical choice of β is the Gaussian function.

Estimates of model complexity of one-hidden layer networks can be obtained by inspection of upper bounds on rates of decrease of errors in approximation of families of functions of interest by sets $\mathrm{span}_n G$ with n increasing. Such rates have been studied in mathematical theory of neurocomputing for various types of computational units and norms measuring approximation errors such as Hilbert-space norms [18,10,19], $\mathcal{L}^p$-norms, $p \in (1, \infty)$ [20], and the supremum norm [21,22]. Typically, these estimates were derived for approximating sets of form

$$\mathrm{conv}_n G := \left\{ \sum_{i=1}^{n} w_i g_i | w_i \in [0, 1], \sum_{i=1}^{n} w_i = 1, g_i \in G \right\} \tag{6}$$

and then extended to $\mathrm{span}_n G$.

To estimate accuracy of surrogate solutions of the Eq. (1) over the whole input domain X uniformly, we shall exploit an upper bound on approximation error in the *supremum norm*, defined for a bounded function f on a set X as

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

The next sup-norm estimate of rates of approximation is a slight reformulation of a theorem by Girosi [21]. It holds for functions which can be represented as images of functions from the space

$$\mathcal{L}^1(X) := \{f : X \to \mathbb{R} | \int_X |f(x)| dx < \infty\}$$

under the operator T_K. Girosi's estimate is formulated in terms of the VC-dimension of the dictionary $G_K(X)$. Recall that the *Vapnik–Chervonenkis dimension* (*VC-dimension*) of a set F of real-valued functions on a set X is the maximal cardinality h of a

set of points $\{y_i \in X | i = 1,\ldots,h\}$ that can be separated in all 2^h possible ways into two classes H_1 and H_2 by means of functions $f(\cdot) - \alpha$, with $f \in F$ and $\alpha \in \mathbb{R}$, where a pair (f,α) classifies y_i as belonging to H_1 if $f(y_i) - \alpha \geqslant 0$ and to H_2 if $f(y_i) - \alpha < 0$ [23].

For a function g in a normed linear space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ and a subset A of $\mathcal{X}$ we denote by $\|g - A\|_{\mathcal{X}} := \inf_{f\in A}\|g - f\|_{\mathcal{X}}$ the distance of g from A. For $c \in \mathbb{R}$, we write $cA := \{cf | f \in A\}$.

Theorem 1. *Let $X \subseteq \mathbb{R}^d$, $K : X \times Y \to \mathbb{R}$ be a bounded kernel, $\tau_K := \sup_{x\in X,\ y\in Y}|K(x,y)|$, h the VC-dimension of $G_K(X,Y)$, and g a function on X such that $g = T_K(w)$ for some $w \in \mathcal{L}^1(Y)$. Then for every positive integer $n \geqslant h/2$*

$$\|g - \mathrm{span}_n G_K(X,Y)\|_{\sup} \leqslant \left\|g - \|w\|_{\mathcal{L}^1(Y)}\mathrm{conv}_n(G_K(X,Y) \cup -G_K(X,Y))\right\|_{\sup} \leqslant 4\tau_K\|w\|_{\mathcal{L}^1(Y)}\sqrt{\frac{h\ln\frac{2en}{h} + \ln 4}{n}}.$$

In the next sections, we apply Theorem 1 to some dictionaries for which we estimate their VC-dimensions. We consider dictionaries induced by Gaussian and degenerate kernels. We denote by $S_d : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the *d-dimensional Gaussian kernel*, defined as

$$S_d(x,y) := e^{-\|x-y\|^2}$$

and for every $b > 0$ we denote by $S_d^b : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the d-dimensional Gaussian kernel with the width b, defined as

$$S_d^b(x,y) := e^{-b\|x-y\|^2}.$$

Note that Fredholm integral equations of the second kind with Gaussian kernels arise, e.g., in image restoration problems [1].

Recall that a kernel K is called *degenerate* when it can be represented as

$$K(x,y) = \sum_{j=1}^{m} \xi_j(x)\eta_j(y), \tag{7}$$

where m is finite and $\{\xi_j\}$ and $\{\eta_j\}$ are two sets of linearly-independent functions on X [24, Section 2.3]. Degenerate kernels are of interest because every $\mathcal{L}^2$-kernel can be arbitrarily well approximated by a sequence of degenerate kernels (with m increasing) [24, Section 2.6]. Some estimates of sup-norm differences between solutions of two Fredholm equations in dependence on sup norm distance between kernels of these equations are given in [24, Section 2.6].

The next proposition gives bounds on the VC-dimensions of the dictionaries associated with Gaussian and degenerate kernels.

Proposition 1. *Let d be a positive integer and $X \subseteq \mathbb{R}^d$. Then*

(i) *for every $b > 0$, the VC-dimensions of the dictionaries $G_{S_d}(X,\mathbb{R}^d)$ and $G_{S_d^b}(X,\mathbb{R}^d)$ are bounded from above by $d + 1$;*
(ii) *for every kernel $K : X \times Y \to \mathbb{R}$ such that $K(x,y) = \sum_{j=1}^m \xi_j(x)\eta_j(y)$, where $\{\xi_j\}$ is a linearly independent set of functions on X, the VC-dimension of $G_K(X)$ is bounded from above by $m + 1$.*

Proof

(i) Dudley [25] proved that the VC-dimension of the set of balls in $\mathbb{R}^d$ is equal to $d + 1$. Thus the VC-dimensions of the dictionaries $G_{S_d(X,\mathbb{R}^d)}$ and $G_{S_d^b(X,\mathbb{R}^d)}$ for all $b > 0$ are bounded from above by $d + 1$.
(ii) Let $M := \{\xi_1,\ldots,\xi_m\}$. As $\{\xi_j\}$ is linearly independent, the dimension of span$M$ is equal to m. So by [26, Theorem 1],[1] the VC-dimension of span$M$ is bounded from above by $m + 1$. As $G_K(X) := \{K(\cdot,y) | y \in Y\} \subseteq$ span M, also the VC-dimension of $G_K(X)$ is bounded from above by $m + 1$. □

4. Accuracy estimates for dictionaries generated by kernels of integral equations

In this section, we estimate rates of approximation of solutions of Fredholm integral equation by networks with units from dictionaries induced by kernels of the equations.

Theorem 2. *Let $X \subset \mathbb{R}^d$ be compact, $K : X \times X \to \mathbb{R}$ a continuous kernel, $\tau_K := \sup_{x,y\in X}|K(x,y)|$, $\rho_K := \int_X \sup_{y\in X}|K(x,y)|dx$, $\lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of $T_K : \mathcal{C}(X) \to \mathcal{C}(X)$, $|\lambda| < \frac{1}{\rho_K}$, and $c_1(K,f,\lambda) = \frac{4\tau_K|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}$. Then the solution ϕ of the* Eq. (1) *with f continuous satisfies*

[1] It has to be remarked that in [26] the VC-dimension of a set F of functions is defined in a slightly different way. More specifically, referring to the definition that we have given before Theorem 1, in [26] one has $\alpha = 0$. Both definitions appear in the literature (e.g., they are both used in [23]) and are easily related when F is of the form span$M$. Indeed, in such a case α can interpreted as a weight associated with a constant function, which can be included in the set M. This implies that the upper bound on VC-dimension given in [26, Theorem 1] has to be increased by 1.

(i) *for the Gaussian kernel* S_d *and* $n \geqslant (d+1)/2$

$$\|\phi - f - \mathrm{span}_n G_{S_d}(X, \mathbb{R}^d)\|_{\sup} \leqslant c_1(K,f,\lambda)\sqrt{\frac{(d+1)\ln(2en) + \ln 4}{n}}; \tag{8}$$

(ii) *for a degenerate kernel* K *such that* $K(x,y) = \sum_{j=1}^{m} \xi_j(x)\eta_j(y)$ *for all* $x,y \in X$ *and* $n \geqslant (m+1)/2$

$$\|\phi - f - \mathrm{span}_n G_K(X)\|_{\sup} \leqslant c_1(K,f,\lambda)\sqrt{\frac{(m+1)\ln(2en) + \ln 4}{n}}. \tag{9}$$

Proof. By Proposition 1, in both cases (i) and (ii), the dictionaries have finite VC-dimensions bounded from above by $d+1$ and $m+1$, resp. Thus we can apply Theorem 1 to $g = \phi - f$ and $w = \lambda\phi$. As ϕ satisfies the Eq. (1), for every $x \in X$ we have

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

Hence $\|\phi\|_{\mathcal{L}^1} \leqslant |\lambda|\rho_K\|\phi\|_{\mathcal{L}^1} + \|f\|_{\mathcal{L}^1}$ and so $\|\phi\|_{\mathcal{L}^1}(1 - |\lambda|\rho_K) \leqslant \|f\|_{\mathcal{L}^1}$. This inequality is non trivial only when $|\lambda| < \frac{1}{\rho_K}$ as we assume. Thus we get

$$\|w\|_{\mathcal{L}^1} = |\lambda|\|\phi\|_{\mathcal{L}^1} \leqslant \frac{|\lambda|\|f\|_{\mathcal{L}^1}}{1 - |\lambda|\rho_K}.$$

The statements (i) and (ii) follow then from Theorem 1, Proposition 1, and the upper bound $h\ln\left(\frac{2en}{h}\right) \leqslant h\ln(2en)$. □

Theorem 2 shows that surrogate solutions computable by networks with n kernel units can theoretically achieve accuracy of approximation in all points of the domain X within $c_1(K,f,\lambda)\sqrt{\frac{(d+1)\ln(2en)+\ln 4}{n}}$ in the case of the Gaussian kernel and $c_1(K,f,\lambda)\sqrt{\frac{(m+1)\ln(2en)+\ln 4}{n}}$ in the case of a degenerate kernel representable in terms of m linearly independent functions. For a given Fredholm Eq. (1), the term $c_1(K,f,\lambda)$ is a constant, while the terms $\sqrt{\frac{(d+1)\ln(2en)+\ln 4}{n}}$ and $\sqrt{\frac{(m+1)\ln(2en)+\ln 4}{n}}$ resp., converge to zero with increasing number n of network units.

The estimates given in Theorem 2 hold when $|\lambda|$ is sufficiently small. Denoting by μ_d the Lebesgue measure on $\mathbb{R}^d$, we get $\rho_K \leqslant \tau_K\mu_d(X)$. So Theorem 2 applies to every λ satisfying $|\lambda| \leqslant \frac{1}{\tau_K\mu_d(X)} \leqslant \frac{1}{\rho_K}$. For instance, in the case of the Gaussian kernel S_d and X the unit ball in $\mathbb{R}^d$, our estimates hold even for $|\lambda|$ growing with d exponentially fast. The constraint $|\lambda| < \frac{1}{\rho_K}$ can be removed at the cost of getting an upper bound on $\|\phi\|_{\mathcal{L}^1}$ looser than $\frac{\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}$. Such an upper bound can be obtained starting from the representation (3) of the solution ϕ and proceeding as in Section 5 to obtain an upper bound on the quantity $\sup_{x,y\in X}|R_K^\lambda(x,y)|$.

Note that theoretical results on linear approximation of solutions of Fredholm equations by the method of successive approximations (see, e.g., [24, Section 2.1]) assume an upper bound on $|\lambda|$. More specifically, convergence of Neumann's series is proven provided that $|\lambda| < \frac{1}{\|K\|_{\mathcal{L}^2(X\times X)}}$ (see [24, Section 2.1, formulas (6) and (9)]).

Theorem 2 extends to other kernels whose associated dictionaries have finite VC-dimensions. Among such kernels, we cite the ones considered in [9, Example 2] and in [27, Table 2, Examples I-III]. In particular, in Section 9 we shall give numerical results for the kernels $K(x,y) = e^{-2xy}$ and

$$K(x,y) = \begin{cases} e^{2x}, 0 \leqslant y < 0.5, \\ e^{-2x}, 0.5 \leqslant y \leqslant 1 \end{cases}$$

with $X = [0,1]$, for which the VC-dimensions of the associated dictionaries $G_K(X)$ are bounded from above by 1 and 2, resp. For instance for the kernel $K(x,y) = e^{-2xy}$, by the monotonicity properties of the functions $K(\cdot,y)$, it is easy to show that any set of two points can be separated in $3 < 2^2$ possible ways into two classes H_1 and H_2 by means of functions $K(\cdot,y) - \alpha$, with $\alpha \in \mathbb{R}$. Note that the second kernel considered above is bounded but not continuous. However, when the solution ϕ of the integral Eq. (1) is continuous, the technique used to prove Theorem 2 works also for a bounded kernel which is not continuous.

5. Accuracy estimates for dictionaries generated by resolvent kernels

In this section, we derive an upper bound on decrease of approximation errors with increasing number of computational units from dictionaries generated by resolvent kernels. For a continuous kernel $K : X \times X \to \mathbb{R}$ on compact $X \subseteq \mathbb{R}^d$ and $\lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of T_K, we denote by

$$G_{R_K^\lambda}(X) := \{R_K^\lambda(\cdot,y) | y \in X\} \tag{10}$$

the *dictionary induced by the resolvent kernel* R_K^λ (defined in the Eq. (36) in the Appendix).

Theorem 3. *Let $X \subset \mathbb{R}^d$ be compact, $K : X \times X \to \mathbb{R}$ be a continuous degenerate kernel representable for every $x, y \in X$ as $K(x,y) = \sum_{j=1}^{m} \xi_j(x)\eta_j(y)$ with $\{\xi_j\}$ and $\{\eta_j\}$ linearly independent sets of functions, $\lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of $T_K : \mathcal{C}(X) \to \mathcal{C}(X)$, and R_K^λ the resolvent kernel associated with K. Then the solution ϕ of the* Eq. (1) *with f continuous satisfies for every positive integer $n \geqslant (m+1)/2$*

$$\|\phi - f - \mathrm{span}_n G_{R_K^\lambda}(X)\|_{\sup} \leqslant 4 \sup_{x,y \in X} |R_K^\lambda(x,y)||\lambda| \|f\|_{\mathcal{L}^1(X)} \sqrt{\frac{(m+1)\ln(2en) + \ln 4}{n}}. \tag{11}$$

Proof. To apply Theorem 1 to the representation

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

we have to verify that the VC-dimension of $G_{R_K^\lambda}(X)$ is finite and to estimate it from above. As the kernel K of the Eq. (1) is degenerate, by [24, Section 2.3]) we get

$$R_K^\lambda(x,y) = -\frac{1}{D(\lambda)} \sum_{k=1}^{m} [D_{1,k}(\lambda)\eta_1(y) + D_{2,k}(\lambda)\eta_2(y) + \cdots + D_{m,k}(\lambda)\eta_m(y)]\xi_k(x),$$

where R_K^λ is defined in Eq. (36) in the Appendix and the $D_{i,k}(\lambda)$ are coefficients depending on λ (see, e.g., [24, pp. 56-57] for their expressions for $d = 1$). By Proposition 1, the VC-dimension $G_{R_K^\lambda}(X)$ is bounded from above by $m+1$. So the statement follows by Theorem 1 and the upper bound $h \ln\left(\frac{2en}{h}\right) \leqslant h \ln(2en)$. □

Theorem 3 shows that accuracy of approximation of the solutions of the Fredholm Eq. (1) by networks with units generated by resolvent kernels depends on $|\lambda|$, the $\mathcal{L}^1$-norm of f, the number m of functions in the spectral representation of the degenerate kernel K, and on the supremum norm of the resolvent kernel R_K^λ on $X \times X$. Note that in contrast to Theorem 2, in Theorem 3 there is no restriction on size of $|\lambda|$.

The term most difficult to estimate is the supremum of R_K^λ on $X \times X$. The following proposition gives an upper bound in terms of the eigenvalues of the operator T_K.

Proposition 2. *Let $X \subset \mathbb{R}^d$ be compact, $K \in \mathcal{L}^2(X \times X)$ symmetric and bounded with $\tau_K = \sup_{x,y \in X} |K(x,y)|$, and $1/\lambda_j$ be the eigenvalues of the operator $T_K : \mathcal{L}^2(X) \to \mathcal{L}^2(X)$, and suppose that at most a finite number of eigenvalues are negative. Then for every $\lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of T_K*

$$\sup_{x,y \in X} |R_K^\lambda(x,y)| \leqslant \frac{\sum_{l=0}^{\infty} \frac{|\lambda|^l}{l!} \mu(X)^l \tau_K^{l+1} (l+1)^{\frac{1}{2}(l+1)}}{e^{-\lambda\left(2\int_X K(x,x)dx - \sum_{j=1}^{N_u(\lambda)} \frac{1}{\lambda_j}\right)} \prod_{j=1}^{N_u(\lambda)} \left|1 - \frac{\lambda}{\lambda_j}\right| e^{\frac{\lambda}{\lambda_j}}},$$

where $N_u(\lambda)$ is a positive integer such that for every $j > N_u(\lambda), \frac{\lambda}{\lambda_j} \leqslant \frac{1}{2}$.

Proof. By the formula (36) in the Appendix,

$$R_K^\lambda(x,y) = \frac{\mathcal{N}(x,y,\lambda)}{\mathcal{D}(\lambda)}.$$

First we estimate $|\mathcal{N}(x,y,\lambda)|$ from above. By Hadamard's theorem (see, e.g., [24, Appendix II]) we get

$$\int_X \int_X \cdots \int_X \mathcal{K}\begin{pmatrix} x, \xi_1, \xi_2, \ldots, \xi_l \\ y, \xi_1, \xi_2, \ldots, \xi_l \end{pmatrix} d\xi_1 d\xi_2 \ldots d\xi_l \leqslant \mu(X)^l \tau_K^{l+1} (l+1)^{\frac{1}{2}(l+1)}. \tag{12}$$

So

$$\sup_{x,y \in X} |\mathcal{N}(x,y,\lambda)| = \sup_{x,y \in X} \left| \sum_{l=0}^{\infty} \frac{(-\lambda)^l}{l!} \int_X \int_X \cdots \int_X \mathcal{K}\begin{pmatrix} x, \xi_1, \xi_2, \ldots, \xi_l \\ y, \xi_1, \xi_2, \ldots, \xi_l \end{pmatrix} d\xi_1 d\xi_2 \ldots d\xi_l \right| \leqslant \sum_{l=0}^{\infty} \frac{|\lambda|^l}{l!} \mu(X)^l \tau_K^{l+1} (l+1)^{\frac{1}{2}(l+1)},$$

where the series converges by the ratio test [28, Section 6.2]. As $K \in \mathcal{L}^2(X \times X), T_K : \mathcal{L}^2(X) \to \mathcal{L}^2(X)$ is compact (see, e.g., [4, Section 1.2]) and as K is symmetric, T_K is self-adjoint. Thus by the spectral theorem, there exist a finite or infinite countable orthonormal family $\{\psi_j\}$ of eigenfunctions of T_K and corresponding eigenvalues $\frac{1}{\lambda_j}$ such that for all $x, y \in X$,

$$K(x,y) = \sum_{j=1}^{N} \frac{1}{\lambda_j} \psi_j(x)\psi_j(y) \tag{13}$$

and for N infinite, $\lim_{j \to \infty} \left|\frac{1}{\lambda_j}\right| = 0$.

To derive a lower bound on $|\mathcal{D}(\lambda)|$, we take advantage of a representation of $\mathcal{D}(\lambda)$ in terms of the eigenvalues of T_K

$$\mathcal{D}(\lambda) = e^{a\lambda} \prod_{j=1}^{N} \left(1 - \frac{\lambda}{\lambda_j}\right) e^{\frac{\lambda}{\lambda_j}}, \tag{14}$$

where $a := -\int_X K(x,x)dx$ (see [28, Theorem 7, Chapter 6]). By (13) we have (see, e.g., [28, Corollary 2, p. 92] and [29, Theorem 2.10])

$$\sum_{i=1}^{N} \frac{1}{\lambda_i} = \int_X K(x,x)dx. \tag{15}$$

When N is infinite, $\lim_{j\to\infty} \frac{1}{\lambda_j} = 0$ and so there exists a positive integer $N_u(\lambda)$ such that for every $j > N_u(\lambda)$ we have $\frac{\lambda}{\lambda_j} \leqslant \frac{1}{2}$. So, a simple calculation shows that for every $j > N_u(\lambda)$

$$1 - \frac{\lambda}{\lambda_j} \geqslant e^{-2\frac{\lambda}{\lambda_j}}. \tag{16}$$

If $\frac{\lambda}{\lambda_j} > 2$ for every j, let $N_l(\lambda) := 1$, otherwise define $N_l(\lambda)$ as the first index j such that $\frac{\lambda}{\lambda_j} \leqslant 2$. By combining (14)–(16) we get

$$|\mathcal{D}(\lambda)| \geqslant e^{a\lambda} \prod_{j=1}^{N_u(\lambda)} \left|1 - \frac{\lambda}{\lambda_j}\right| e^{\frac{\lambda}{\lambda_j}} \prod_{j=N_u(\lambda)+1}^{\infty} e^{-\frac{\lambda}{\lambda_j}} = e^{-\lambda\left(2\int_X K(x,x)dx - \sum_{j=1}^{N_u(\lambda)} \frac{1}{\lambda_j}\right)} \prod_{j=1}^{N_u(\lambda)} \left|1 - \frac{\lambda}{\lambda_j}\right| e^{\frac{\lambda}{\lambda_j}}. \tag{17}$$

The lower bound (17) requires to compute the first $N_u(\lambda) - 1$ eigenvalues of the operator T_K.

If $\lambda < 0$, then a looser but much simpler lower bound can be obtained, as

$$\mathcal{D}(\lambda) \geqslant e^{-\lambda \int_X K(x,x)dx} e^{\lambda \sum_{i=j}^{N} \frac{1}{\lambda_j}} = e^{-\lambda \int_X K(x,x)dx} e^{\lambda \int_X K(x,x)dx} = 1.$$

The statement then follows by combining (17) and (5) together with (13). □

6. Accuracy estimates for approximation by Gaussian RBF networks

In this section, we investigate approximation of solutions of Fredholm equations with sufficiently smooth kernels by networks with Gaussian computational units with varying widths and centers. Such networks were used in [8,9] as surrogate models to derive approximate solutions to the Fredholm Eq. (1).

We take advantage of integral representations from [30] of smooth functions in terms of such Gaussians. These representations were obtained by combining two integrals: the first one expresses a smooth function as a convolution of a Bessel potential of a suitable degree and the second one expresses the Bessel potential as an integral of Gaussians with varying widths.

We denote by

$$F_d(x,(v,b)) := e^{-b\|x-v\|^2} : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R}_+) \to \mathbb{R}$$

the Gaussian kernel corresponding to the Gaussian computational unit (RBF) with parameters $b \in \mathbb{R}$ and $v \in \mathbb{R}^d$ representing widths and centers, resp. So,

$$G_{F_d}(X, \mathbb{R}^d) := \bigcup_{b\in\mathbb{R}_+} G_{S_d^b}(X, \mathbb{R}^d)$$

is the dictionary induced by F_d.

We denote the *d-dimensional Fourier transform* on $\mathcal{L}^2(\mathbb{R}^d) \cap \mathcal{L}^1(\mathbb{R}^d)$ by

$$\mathcal{F}(f)(s) := \hat{f}(s) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle x,s\rangle} f(x)dx. \tag{18}$$

To prove the next theorem, we introduce some definitions and notations. For a positive integer d and $r > 0$, the *Bessel potential* of the order r, denoted by β_r, is the function on $\mathbb{R}^d$ with the Fourier transform

$$\hat{\beta}_r(s) = (1 + \|s\|^2)^{-r/2}.$$

For $r > 0$, β_r is non-negative, radial, exponentially decreasing at infinity, analytic except at the origin, and belongs to $\mathcal{L}^1(\mathbb{R}^d)$ [31, p. 132]. The Bessel potential can be expressed by the integral formula

$$\beta_r(u) = c_1(r,d) \int_0^{\infty} e^{-t/(4\pi)} t^{-d/2+r/2-1} e^{-(\pi/t)\|u\|^2} dt, \tag{19}$$

where

$$c_1(r,d) := (2\pi)^{d/2}(4\pi)^{-r/2}/\Gamma(r/2)$$

and for $z>0$ we let $\Gamma(z) := \int_0^\infty t^{z-1}e^{-t}dt$ the Gamma function (see [32, p. 296] or [31]). The factor $(2\pi)^{d/2}$ occurs since our choice of Fourier transform (18) includes the factor $(2\pi)^{-d/2}$.

We denote by

$$\mathcal{L}^{r,1}(\mathbb{R}^d) := \{f : \mathbb{R}^d \to \mathbb{R} | f = w * \beta_r, w \in \mathcal{L}^1(\mathbb{R}^d)\},$$

the *Bessel potential space* of the order r, where $*$ denotes convolution. The norm on $\mathcal{L}^{r,1}(\mathbb{R}^d)$ is defined as

$$\|f\|_{\mathcal{L}^{r,1}} := \|w\|_{\mathcal{L}^1(\mathbb{R}^d)}$$

for $f = w * \beta_r$.

For an open set $\Omega \subseteq \mathbb{R}^d$ and a positive integer r, let $\mathcal{W}^{r,1}(\Omega)$ denote the *Sobolev space* of the order r, which is formed by functions on Ω with iterated partial derivatives up to the order r in $\mathcal{L}^1(\Omega)$.

For $X \subseteq \mathbb{R}^d$, we denote by int (X) the interior of the set X and by $\mathcal{C}^r(X)$ the space of continuous functions on X with continuous iterated partial derivatives up to the order r.

The following theorem estimates rates of approximation of solutions of Fredholm integral equations by Gaussian RBF networks. It gives an upper bound on speed of decrease of approximation errors with increasing number of network units formulated in terms of smoothness of the solution ϕ expressed in terms of a function $w \in \mathcal{L}^1(X)$ such that $f - \phi = w * \beta_r|_X$. So the result is merely existential.

Theorem 4. *Let $X \subset \mathbb{R}^d$ be compact and such that int (X) is bounded and convex, $r > d$ be a positive even integer, $K \in \mathcal{C}^r(X \times X)$ a kernel, and $\lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of $T_K : \mathcal{C}(X) \to \mathcal{C}(X)$. Then there exists $w \in \mathcal{L}^1(\mathbb{R}^d)$ such that the solution ϕ of the Fredholm integral* Eq. (1) *with $f \in \mathcal{C}(X)$ satisfies $f - \phi = w * \beta_r|_X$ and for all $n \geqslant (d+1)/2$*

$$\|\phi - f - \mathrm{span}_n G_{F_d}(X, \mathbb{R}^d)\|_{\sup} \leqslant 2^{-\frac{d}{2}+2}\frac{\Gamma\left(\frac{r-d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\|w\|_{\mathcal{L}^1(\mathbb{R}^d)}\sqrt{\frac{(d+1)\ln\frac{2en}{d+1} + \ln 4}{n}}. \tag{20}$$

Proof. As $f \in \mathcal{C}(X)$, $\phi \in \mathcal{C}(X)$, and $K \in \mathcal{C}^r(X \times X)$, it follows by (1) that $\phi - f \in \mathcal{C}^r(X)$. As X is compact, $(\phi - f)|_{\mathrm{int}X} \in \mathcal{W}^{r,1}(\mathrm{int}X)$. By Sobolev's extension theorem [31, Theorem 5, p. 181 and Example 2, p. 189], the space $\mathcal{W}^{r,1}(\mathrm{int}X)$ can be extended to $\mathcal{W}^{r,1}(\mathbb{R}^d)$ by a bounded extension operator. As r is even, by [31, Remark 6.6 (b), p. 160] we have $\mathcal{W}^{r,1}(\mathbb{R}^d) \subset \mathcal{L}^{r,1}(\mathbb{R}^d)$. Thus there exists $w \in \mathcal{L}^1(\mathbb{R}^d)$ such that $f - \phi = w * \beta_r|_X$.

So, by (19) for all $x \in X$ we get

$$f(x) - \phi(x) = c_1(r,d)\int_{\mathbb{R}^d}\int_0^\infty w(y)e^{-t/(4\pi)}t^{-d/2+r/2-1}e^{-(\pi/t)\|y-x\|^2}dydt.$$

Hence

$$f(x) - \phi(x) = \frac{(2\pi)^{d/2}(4\pi)^{-r/2}}{\Gamma(r/2)}\int_{\mathbb{R}^d}\int_0^\infty F_d(x,(y,-\pi/t))\Lambda(t,y)dtdy, \tag{21}$$

where

$$\Lambda(t,y) := e^{-t/(4\pi)}t^{\frac{r-d}{2}-1}w(y).$$

It is easy to check that $\Lambda \in \mathcal{L}^1(\mathbb{R}_+ \times X)$. Indeed as $r > d$ we get

$$\int_0^\infty\int_{\mathbb{R}^d}|\Lambda(t,y)|dtdy = (4\pi)^{(r-d)/2}\Gamma((r-d)/2)\|w\|_{\mathcal{L}^1(\mathbb{R}^d)}.$$

Thus

$$f(x) - \phi(x) = 2^{-\frac{d}{2}}\frac{\Gamma\left(\frac{r-d}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}\int_{\mathbb{R}^d}F_d(x,(y,-\pi/t))dy. \tag{22}$$

By Proposition 1, the VC-dimension of $G_{F_d}(X,\mathbb{R}^d)$ is bounded from above by $d+1$. So, the statement follows by Theorem 1 applied to the integral representation (21). □

For $r = d+1$ when d is odd or $r = d+2$ when d is even, the coefficient $\Gamma\left(\frac{r-d}{2}\right)/\Gamma\left(\frac{d}{2}\right)$ becomes $\Gamma(1/2)/\Gamma(d/2)$ or $1/\Gamma(d/2)$, resp. So, it decreases exponentially fast with d. Although the term $\|w\|_{\mathcal{L}^1(\mathbb{R}^d)}$ is difficult to estimate, Theorem 4 suggests that even quite large values of $\|w\|_{\mathcal{L}^1(\mathbb{R}^d)}$ can be compensated by $2^{-d/2}\Gamma\left(\frac{1}{2}\right)/\Gamma\left(\frac{d}{2}\right)$ or $2^{-d/2}/\Gamma\left(\frac{d}{2}\right)$, resp. Note that the input dimension d is often an important factor in estimates of accuracy of approximation by neural networks [33,34].

Inspection of the proof of Theorem 4 shows that, with the same assumptions on the set X, the estimate (20) can be extended to nonlinear integral equations of the form

$$\phi(x) - \lambda \int_X K(x,y,\phi(y))dy = f(x), \tag{23}$$

where $f \in \mathcal{C}^r(X)$ and $K : X \times X \times \mathbb{R} \to \mathbb{R}$ is bounded and of class $\mathcal{C}^r(X \times X \times \mathbb{R})$, provided that there exists a solution $\phi \in \mathcal{C}(X)$. This holds, e.g., if there exists $L > 0$ such that K satisfies a Lipschitz condition of the form

$$|K(x,y,z_1) - K(x,y,z_2)| \leqslant L\|z_1 - z_2\|$$

and $|\lambda| < \frac{1}{L\mathrm{vol}(X)}$ (where vol (X) denotes the volume of X), since under these conditions the existence of a (unique) continuous solution follows by the Banach Contraction Mapping Theorem (see also [35] for this and other sufficient conditions for the existence of a continuous solution to Eq. (23)).

7. Achieving the bounds

Theorems 2–4 guarantee existence of elements of $\mathrm{span}_n G_K(X)$, $\mathrm{span}_n G_{K_K^z}(X)$, and $\mathrm{span}_n G_{F_d}(X, \mathbb{R}^d)$, respectively, which approximate with certain accuracies solutions to Fredholm Eq. (1). In this section, we consider a way to find networks that achieve the bounds up to multiplicative constants independent of the number n of computational units. In particular, we focus on the estimates provided by Theorem 2.

Let $\tau_K = \sup_{x,y\in X}|K(x,y)|, \rho_K = \int_X \sup_{y\in X}|K(x,y)|dx$, and

$$c_1(K,f,\lambda) := \frac{4\tau_K|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}.$$

The following theorem shows that a network achieving an accuracy of the order specified by Theorem 2 can be obtained by minimizing the functional $\|(I - \lambda T_K)(\cdot) - f\|_{\sup}$ over the set

$$\Upsilon_n := \{f + g | g \in c_1(K,f,\lambda)\mathrm{conv}_n(G_K(X) \cup -G_K(X))\}$$

of functions. For every $\varepsilon > 0$, let

$$\begin{aligned}&\mathrm{argmin}_\varepsilon(\|(I - \lambda T_K)(\cdot) - f\|_{\sup}, \Upsilon_n)\\ &:= \{\phi_{n,\varepsilon} \in \Upsilon_n | \Upsilon_n(\phi_{n,\varepsilon}) < \inf_{\psi_n\in\Upsilon_n} \|(I - \lambda T_K)(\psi_n) - f\|_{\sup} + \varepsilon\}\end{aligned}$$

denote the set of *ε-near minimum points* of the functional $\|(I - \lambda T_K)(\cdot) - f\|_{\sup}$ over Υ_n.

Theorem 5. *Let $X \subset \mathbb{R}^d$ be compact, $K : X \times X \to \mathbb{R}$ a continuous kernel, $\tau_K := \sup_{x,y\in X}|K(x,y)|, ,\rho_K := \int_X \sup_{y\in X}|K(x,y)|dx, \lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of $T_K, |\lambda| < \frac{1}{\rho_K}$, and f continuous. Let $c_1(K,f,\lambda) := \frac{4\tau_K|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}, c_1(K,\lambda) := \|I - \lambda T_K\|_{\sup}$, and $c_2(K,\lambda) := \|(I - \lambda T_K)^{-1}\|_{sup}$. Then for every $\varepsilon > 0$, every positive integer n, and every $\phi_{n,\varepsilon} \in argmin_\varepsilon(\|(I - \lambda T_K)(\cdot) - f\|_{sup}, \Upsilon_n)$ the following hold.*

(i) *For the Gaussian kernel $S_d : X \times X \to \mathbb{R}$ and $n \geqslant (d+1)/2$*

$$\|(I - \lambda T_{S_d})(\phi_{n,\varepsilon}) - f\|_{\sup} \leqslant c_1(S_d,\lambda)c_1(S_d,f,\lambda)\sqrt{\frac{(d+1)\ln(2en) + \ln 4}{n}} + \varepsilon,$$

$$\|\phi - \phi_{n,\varepsilon}\|_{\sup} \leqslant c_2(S_d,\lambda)\left[c_1(S_d,\lambda)c_1(S_d,f,\lambda)\sqrt{\frac{(d+1)\ln(2en) + \ln 4}{n}} + \varepsilon\right].$$

(ii) *For a degenerate kernel $K : X \times X \to \mathbb{R}$ such that $K(x,y) = \sum_{j=1}^m \xi_j(x)\eta_j(y)$ for all $x,y \in X$ and $n \geqslant (m+1)/2$*

$$\|(I - \lambda T_K)(\phi_{n,\varepsilon}) - f\|_{\sup} \leqslant c_1(K,\lambda)c_1(K,f,\lambda)\sqrt{\frac{(m+1)\ln(2en) + \ln 4}{n}} + \varepsilon,$$

$$\|\phi - \phi_{n,\varepsilon}\|_{\sup} \leqslant c_2(K,\lambda)\left[c_1(K,\lambda)c_1(K,f,\lambda)\sqrt{\frac{(m+1)\ln(2en) + \ln 4}{n}} + \varepsilon\right].$$

Proof. By compactness of $T_K : \mathcal{C}(X) \to \mathcal{C}(X)$ and the Fredholm Alternative Theorem, both $(I - \lambda T_K)$ and $(I - \lambda\ T_K)^{-1}$ are bounded linear operators on $\mathcal{C}(X)$. So, $c_1(K,\lambda)$ and $c_2(K,\lambda)$ are well-defined. We have

$$\begin{aligned}\|(I - \lambda T_K)(\phi_{n,\varepsilon}) - f\|_{\sup} &< \inf_{\psi\in\Upsilon_n} \|(I - \lambda T_K)(\psi) - f\|_{\sup} + \varepsilon = \inf_{\psi\in\Upsilon_n} \|(I - \lambda T_K)(\phi - \psi)\|_{\sup} + \varepsilon \leqslant \|I - \lambda T_K\|_{\sup} \inf_{\psi\in\Upsilon_n} \|\phi - \psi\|_{\sup}\\ &= c_1(K,\lambda) \inf_{\psi\in\Upsilon_n} \|\phi - \psi\|_{\sup} + \varepsilon \leqslant c_1(K,\lambda)\|\phi - f - \mathrm{span}_n G_K(X)\|_{\sup} + \varepsilon\end{aligned}$$

and

$$\|\phi-\phi_{n,\varepsilon}\|_{\sup}=\|(I-\lambda T_K)^{-1}(I-\lambda T_K)(\phi-\phi_{n,\varepsilon})\|_{\sup}=\|(I-\lambda T_K)^{-1}((I-\lambda T_K)(\phi_{n,\varepsilon})-f)\|_{\sup}$$
$$\leqslant\|(I-\lambda T_K)^{-1}\|_{\sup}\|((I-\lambda T_K)(\phi_{n,\varepsilon})-f)\|_{\sup}=c_2(K,\lambda)\|(I-\lambda T_K)(\phi_{n,\varepsilon})-f\|_{\sup}$$

So, (i) and (ii) follow by Theorem 2. □

For every $\varepsilon>0$, Theorem 5 estimates speed of convergence of ε-near minimum points of the functional $\|(I-\lambda T_K)(\cdot)-f\|_{\sup}$ over Υ_n to the solution of the Fredholm integral Eq. (1).

Theorem 5 requires to minimize a supremum norm. In practice, differentiability requirements make algorithms based on the minimization of $\mathcal{L}^2$-norm much more appealing. The following theorem shows that networks with the same order of accuracy can be obtained by minimizing over Υ_n the $\mathcal{L}^2$- norm $\|(I-\lambda T_K)(\cdot)-f\|_{\mathcal{L}^2(X)}$.

Theorem 6. *Let $X\subset\mathbb{R}^d$ be compact, $K:X\times X\to\mathbb{R}$ a continuous kernel, $\tau_K:=\sup_{x,y\in X}|K(x,y)|$, $\rho_K:=\int_X\sup_{y\in X}|K(x,y)|dx$, $\lambda\neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of T_K, $|\lambda|<\frac{1}{\rho_K}$, and f continuous. Let $c_1(K,f,\lambda):=\frac{4\tau_K|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}$, $c_1'(K,\lambda):=\sqrt{\mu(X)}\|I-\lambda T_K\|_{\mathcal{L}^2(X)}$, and $c_2'(K,\lambda):=\|(I-\lambda T_K)^{-1}\|_{\mathcal{L}^2(X)}$. Then for every $\varepsilon>0$, every positive integer n, and every $\phi_{n,\varepsilon}\in\operatorname{argmin}_\varepsilon(\|(I-\lambda T_K)(\cdot)-f\|_{\mathcal{L}^2(X)},\Upsilon_n)$ the following hold.*

(i) *For the Gaussian kernel $S_d:X\times X\to\mathbb{R}$ and $n\geqslant(d+1)/2$*

$$\|(I-\lambda T_{S_d})(\phi_{n,\varepsilon})-f\|_{\mathcal{L}^2(X)}\leqslant c_1'(S_d,\lambda)c_1(S_d,f,\lambda)\sqrt{\frac{(d+1)\ln(2en)+\ln 4}{n}}+\varepsilon,$$
$$\|\phi-\phi_{n,\varepsilon}\|_{\mathcal{L}^2(X)}\leqslant c_2'(S_d,\lambda)\left[c_1'(S_d,\lambda)c_1(S_d,f,\lambda)\sqrt{\frac{(d+1)\ln(2en)+\ln 4}{n}}+\varepsilon\right].$$

(ii) *For a degenerate kernel $K:X\times X\to\mathbb{R}$ such that $K(x,y)=\sum_{j=1}^m\xi_j(x)\eta_j(y)$ for all $x,y\in X$ and $n\geqslant(m+1)/2$*

$$\|(I-\lambda T_K)(\phi_{n,\varepsilon})-f\|_{\mathcal{L}^2(X)}\leqslant c_1'(K,\lambda)c_1(K,f,\lambda)\sqrt{\frac{(m+1)\ln(2en)+\ln 4}{n}}+\varepsilon,$$
$$\|\phi-\phi_{n,\varepsilon}\|_{\mathcal{L}^2(X)}\leqslant c_2'(K,\lambda)\left[c_1'(K,\lambda)c_1(K,f,\lambda)\sqrt{\frac{(m+1)\ln(2en)+\ln 4}{n}}+\varepsilon\right].$$

Proof. By compactness of T_K on $\mathcal{L}^2(X)$ (see, e.g., [4, Section 1.2]) and the Fredholm Alternative Theorem, both $(I-\lambda T_K)$ and $(I-\lambda T_K)^{-1}$ are bounded linear operators on $\mathcal{L}^2(X)$. So, $c_1'(K,\lambda)$ and $c_2'(K,\lambda)$ are well-defined. The remaining of the proof proceeds as the proof of Theorem 5, taking into account the relationship $\|\cdot\|_{\mathcal{L}^2(X)}\leqslant\sqrt{\mu(X)}\|\cdot\|_{\sup}$. □

Let us now investigate the relationship between the estimates from Theorems 5 and 6. The next theorem shows that the supremum norm of the error associated with the suboptimal solutions can be bounded from above in terms of its $\mathcal{L}^2(X)$-norm. For simplicity of exposition, we take X equal to the hypercube $\left[-\frac{1}{2},\frac{1}{2}\right]^d$ but a similar result holds when X is any given bounded domain (i.e., the closure of a bounded open connected set). We let $c_3(K,f,\lambda):=\frac{|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}$.

Theorem 7. *Let $X=\left[-\frac{1}{2},\frac{1}{2}\right]^d$, $K\in\mathcal{C}^1(X\times X)$, $\tau_K=\sup_{x,y\in X}|K(x,y)|$, $\rho_K=\int_X\sup_{y\in X}|K(x,y)|dx$, $\lambda\neq 0$ be such that $\frac{1}{\lambda}$ is not an eigenvalue of T_K, $|\lambda|<\frac{1}{\rho_K}$, and $f\in\mathcal{C}^1(X)$. Then there exists $c_4(K,f,\lambda,d)$, $c_5(K,f,\lambda,d)>0$ such that for every positive integer n and every $\phi_n\in\Upsilon_n$ one has*

$$\|(I-\lambda T_K)(\phi_n)-f\|_{\sup}\leqslant\max\left\{c_4(K,f,\lambda,d)\|(I-\lambda T_K)(\phi_n)-f\|_{\mathcal{L}^2(X)}^{\frac{2}{2+d}},c_5(K,f,\lambda,d)\|(I-\lambda T_K)(\phi_n)-f\|_{\mathcal{L}^2(X)}\right\}.\tag{24}$$

Proof. The solution ϕ of the Fredholm Eq. (1) is of class $\mathcal{C}^1(X)$ and its partial derivatives $\frac{\partial\phi}{\partial x_i}$, $i=1,\ldots,d$, satisfy the equation

$$\frac{\partial\phi(x)}{\partial x_i}-\lambda\int_X\frac{\partial K(x,y)}{\partial x_i}\phi(y)dy=\frac{\partial f(x)}{\partial x_i}.$$

So we get

$$\left\|\frac{\partial\phi}{\partial x_i}\right\|_{\sup}\leqslant|\lambda|\sup_{x,y\in X}\left|\frac{\partial K(x,y)}{\partial x_i}\right|\|\phi\|_{\mathcal{L}^1(X)}+\left\|\frac{\partial f}{\partial x_i}\right\|_{\sup}\leqslant\frac{|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}\sup_{x,y\in X}\left|\frac{\partial K(x,y)}{\partial x_i}\right|+\left\|\frac{\partial f}{\partial x_i}\right\|_{\sup}$$
$$=c_3(K,f,\lambda)\sup_{x,y\in X}\left|\frac{\partial K(x,y)}{\partial x_i}\right|+\left\|\frac{\partial f}{\partial x_i}\right\|_{\sup},$$

hence ϕ is Lipschitz continuous. Similarly, one can show that the functions $\phi_n \in \Upsilon_n$ are of class $\mathcal{C}^1(X)$ and that for all $i = 1,\ldots,d$

$$\left\|\frac{\partial \phi_n}{\partial x_i}\right\|_{\sup} \leqslant c_3(K,f,\lambda) \sup_{x,y\in X}\left|\frac{\partial K(x,y)}{\partial x_i}\right| + \left\|\frac{\partial f}{\partial x_i}\right\|_{\sup}. \tag{25}$$

Let $\bar{\phi}_n := (I - \lambda T_K)(\phi_n) + f$. By the definition of Υ_n and (25), it follows that $\bar{\phi}_n$ is Lipschitz continuous, with Lipschitz constant

$$L := L(K,f,\lambda,d) := 2\sqrt{\sum_{i=1}^{d}\left(c_3(K,f,\lambda)\sup_{x,y\in X}\left|\frac{\partial K(x,y)}{\partial x_i}\right| + \left\|\frac{\partial f}{\partial x_i}\right\|_{\sup}\right)^2}.$$

Then, for any point $\bar{x} \in X$ such that $|\bar{\phi}_n(\bar{x})| = \|\bar{\phi}_n\|_{\sup}$ one has $|\bar{\phi}_n(x)| \geqslant \frac{\|\bar{\phi}_n\|_{\sup}}{2}$ on the set $B_{2,\bar{x}}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right) \cap X$, where $B_{2,\bar{x}}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)$ is the closed ball of radius $\frac{\|\bar{\phi}_n\|_{\sup}}{2L}$ in the l^2-norm, centered on $\bar{x}$. So,

$$\|\bar{\phi}_n\|_{\mathcal{L}^2(X)} \geqslant \frac{\|\bar{\phi}_n\|_{\sup}}{2}\sqrt{\mu\left(B_{2,\bar{x}}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)\cap X\right)}. \tag{26}$$

For $r > 0$, let $\mathrm{vol}_2^{(d)}(r) := \pi^{\frac{d}{2}} r^d \Gamma^{-1}\left(\frac{d}{2}+1\right)$ denote the volume of the d-dimensional ball of radius r in the l^2-norm [36, p. 304]. Simple geometric arguments concerning the intersection with minimum overlap between a hypercube and a sphere of variable center and radius give the following estimates. We distinguish two cases:

1. $\frac{\|\bar{\phi}_n\|_{\sup}}{2L} \leqslant \frac{1}{2}$. In this case, $\mu\left(B_{2,\bar{x}}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)\cap X\right) \geqslant \mathrm{vol}_2^{(d)}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)$ (the equality holds if and only if $\bar{x}$ is any vertex of $\left[-\frac{1}{2},\frac{1}{2}\right]^d$);
2. $\frac{\|\bar{\phi}_n\|_{\sup}}{2L} \geqslant \frac{1}{2}$. In this case, $\mu\left(B_{2,\bar{x}}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)\cap X\right) \geqslant \mathrm{vol}_2^{(d)}\left(\frac{1}{2}\right)$ (the equality holds if and only if $\frac{\|\bar{\phi}_n\|_{\sup}}{2L} = \frac{1}{2}$ and $\bar{x}$ is any vertex of $\left[-\frac{1}{2},\frac{1}{2}\right]^d$).

Hence

$$\|\bar{\phi}_n\|_{\mathcal{L}^2(X)} \geqslant \frac{\|\bar{\phi}_n\|_{\sup}}{2}\min\left\{\sqrt{\mathrm{vol}_2^{(d)}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)}, \sqrt{\mathrm{vol}_2^{(d)}\left(\frac{1}{2}\right)}\right\} = \frac{\|\bar{\phi}_n\|_{\sup}}{2}\min\left\{\sqrt{\frac{\pi^{\frac{d}{2}}\left(\frac{\|\bar{\phi}_n\|_{\sup}}{2L}\right)^d}{\Gamma\left(\frac{d}{2}+1\right)}}, \sqrt{\frac{\pi^{\frac{d}{2}}2^{-d}}{\Gamma\left(\frac{d}{2}+1\right)}}\right\}, \tag{27}$$

or, equivalently,

$$\|\bar{\phi}_n\|_{\sup} \leqslant \max\left\{2L^{\frac{d}{2+d}}\frac{\left(\Gamma\left(\frac{d}{2}+1\right)\right)^{\frac{1}{2+d}}}{\pi^{\frac{d}{4+2d}}}\|\bar{\phi}_n\|_{\mathcal{L}^2(X)}^{\frac{2}{2+d}}, 2^{\frac{d+2}{2}}\frac{\left(\Gamma\left(\frac{d}{2}+1\right)\right)^{\frac{1}{2}}}{\pi^{\frac{d}{4}}}\|\bar{\phi}_n\|_{\mathcal{L}^2(X)}\right\}.$$

So, the bound (24) holds with

$$c_4(K,f,\lambda,d) := 2L^{\frac{d}{2+d}}\frac{\left(\Gamma\left(\frac{d}{2}+1\right)\right)^{\frac{1}{2+d}}}{\pi^{\frac{d}{4+2d}}}$$

and

$$c_5(K,f,\lambda,d) := 2^{\frac{d+2}{2}}\frac{\left(\Gamma\left(\frac{d}{2}+1\right)\right)^{\frac{1}{2}}}{\pi^{\frac{d}{4}}}. \qquad \square$$

According to Theorem 7, for every $\phi_n \in \Upsilon_n$ an upper bound on $\|(I - \lambda T_K)(\phi_n) - f\|_{\sup}$ can be obtained via an upper bound on $\|(I - \lambda T_K)(\phi_n) - f\|_{\mathcal{L}^2(X)}$..

8. Accuracy estimates for network training

In the collocation method [6, Chapter 11], approximations of values of solutions to Fredholm integral equations are calculated numerically merely in finitely many points from a subset of the domain X, typically a uniform grid

$$X_l := \{x^{(1)},\ldots,x^{(l)}\}.$$

Let $y^{(k)} \simeq \phi(x^{(k)})$ denote the approximation of the value of the solution in the point $x^{(k)}$ obtained by some numerical method of calculation. Then the set $z := \{(x^{(k)},y^{(k)})|k = 1,\ldots,l\}$ can be used as a *training set of input–output pairs* for a suitable neural network. Various neural-network learning algorithms aim to minimize the *empirical error*

$$\mathcal{E}_z(\psi) := \frac{1}{l}\sum_{k=1}^{l}(\psi(x^{(k)}) - y^{(k)})^2 \tag{28}$$

over the set $f + \mathrm{span}_n G$, where G is a dictionary and n is the number of computational units. Another method consists in minimizing directly the *residual empirical error*

$$\mathcal{E}_{r,z}(\psi) := \frac{1}{l}\sum_{j=1}^{l}((I-\lambda T_K)(\psi)(x^{(j)}) - f(x^{(j)}))^2 \tag{29}$$

over the set $f + \mathrm{span}_n G$, without requiring the preliminary approximate computation of $\phi(x^{(k)})$ by other numerical methods. Typical algorithms to minimize (28) and (29) are steepest descent, conjugate directions, quasi-Newton methods, etc. [37]. For example, in [7] the quasi-Newton BFGS (Broyden–Fletcher–Goldfarb–Shanno) method was implemented in the Matlab optimization toolbox [38].

The next theorem provides theoretical insights into simulation results obtained in recent experimental works [7,9] on approximations of solutions to Fredholm equations by neural networks. We consider $X = \left[-\frac{1}{2},\frac{1}{2}\right]^d$ and X_l a uniform grid of cardinality $l := (m+1)^d$; so, m controls the distance $\frac{1}{m}$ between adjacent points of the grid. To simplify some formulas, we assume that l is even.

Theorem 8. *Let $X = \left[-\frac{1}{2},\frac{1}{2}\right]^d, K \in \mathcal{C}^1(X\times X), \tau_K := \sup_{x,y\in X}|K(x,y)|, \rho_K := \int_X \sup_{y\in X}|K(x,y)|dx, \lambda \neq 0$ be such that $\frac{1}{\lambda}$ is not an eigenvalue of $T_K : \mathcal{L}^2(X) \to \mathcal{L}^2(X), |\lambda| < \frac{1}{\rho_K}, f \in \mathcal{C}^1(X)$, and $c_1(K,f,\lambda) := \frac{4\tau_K|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}$. Let l be an even integer and $X_l \subset X$ a uniform grid of size $l := (m+1)^d$. Then there exist a positive integer m^* and three positive constants $c_6(K,f,\lambda,d)$, $c_7(K,f,\lambda,d)$, $c_8(K,f,\lambda,d)$ such that, for every $m \geqslant m^*$, every positive integer n, and every $\phi_n \in \Upsilon_n$*

$$\|(I-\lambda T_K)(\phi_n) - f\|_{\sup} \leqslant \max\left\{c_6(K,f,\lambda,d)\Delta(K,\lambda,l,d,n)^{\frac{2}{2+d}}, c_7(K,f,\lambda,d)\Delta(K,\lambda,M,d,n)\right\} + \frac{c_8(K,f,\lambda,d)}{m}, \tag{30}$$

where $\Delta(K,\lambda,l,d,n) := \sqrt{\frac{1}{l}\sum_{j=1}^{l}((I-\lambda T_K)(\phi_n)(x^{(j)}) - f(x^{(j)}))^2}$.

Proof. Let $\bar{\phi}_n := (I-\lambda T_K)(\phi_n) + f$. By the definition of Υ_n and (25), it follows that $\bar{\phi}_n$ is Lipschitz continuous, with Lipschitz constant

$$L := L(K,f,\lambda,d) := 2\sqrt{\sum_{i=1}^{d}\left(c_3(K,f,\lambda)\sup_{x,y\in X}\left|\frac{\partial K(x,y)}{\partial x_i}\right| + \left\|\frac{\partial f}{\partial x_i}\right\|_{\sup}\right)^2},$$

where $c_3(K,f,\lambda) := \frac{|\lambda|\|f\|_{\mathcal{L}^1}}{1-|\lambda|\rho_K}$. By the Lipschitz continuity of $\bar{\phi}_n$ and the regularity of the grid X_l, we get

$$\|\bar{\phi}_n\|_{\sup} \leqslant \max_{j=1,\ldots,l}\left|\bar{\phi}_n(x^{(j)})\right| + L\sup_{x\in X}\min_{j=1,\ldots,l}\|x-x^{(j)}\|_2 = \max_{j=1,\ldots,l}\left|\bar{\phi}_n(x^{(j)})\right| + \frac{L}{m}.$$

For any point $\bar{x} \in X_l$ such that $|\bar{\phi}_n(\bar{x})| = \max_{j=1,\ldots,l}|\bar{\phi}_n(x^{(j)})|$, one has $|\bar{\phi}_n(x)| \geqslant \frac{|\bar{\phi}_n(\bar{x})|}{2}$ on the set $B_{1,\bar{x}}\left(\frac{|\bar{\phi}_n(\bar{x})|}{2L}\right)\cap X_l$, where $B_{1,\bar{x}}\left(\frac{|\bar{\phi}_n(\bar{x})|}{2L}\right)$ is the closed ball of radius $\frac{|\bar{\phi}_n(\bar{x})|}{2L}$ in the l^1-norm, centered on $\bar{x}$.

So,

$$\sqrt{\frac{1}{l}\sum_{j=1}^{l}|\bar{\phi}_n(x^{(j)})|^2} \geqslant \sqrt{\frac{1}{l}}\frac{|\bar{\phi}_n(\bar{x})|}{2}\sqrt{\mathrm{card}\left(B_{1,\bar{x}}\left(\frac{|\bar{\phi}_n(\bar{x})|}{2L}\right)\cap X_l\right)}. \tag{31}$$

For $r > 0$, let $\mathrm{vol}_1^{(d)}(r) := \frac{2^d r^d}{d!}$ denote the volume of the d-dimensional ball of radius r in the l^1-norm [39, p. 3]. It is easy to see by simple geometric arguments that the following two cases are possible (in the second one we exploit the assumption that m is even):

1. if $\frac{|\bar{\phi}_n(\bar{x})|}{2L} \leqslant \frac{1}{2}$, then $\mathrm{card}\left(B_{1,\bar{x}}\left(\frac{|\bar{\phi}_n(\bar{x})|}{2L}\right)\cap X_l\right) \geqslant m^d\mathrm{vol}_1^{(d)}\left(\frac{1}{m}\left\lfloor\frac{m|\bar{\phi}_n(\bar{x})|}{2L}\right\rfloor\right) + 1$;
2. if $\frac{|\bar{\phi}_n(\bar{x})|}{2L} \geqslant \frac{1}{2}$, then $\mathrm{card}\left(B_{1,\bar{x}}\left(\frac{|\bar{\phi}_n(\bar{x})|}{2L}\right)\cap X_l\right) \geqslant m^d\mathrm{vol}_1^{(d)}\left(\frac{1}{2}\right) + 1$.

Then

$$\begin{aligned}\sqrt{\frac{1}{l}\sum_{j=1}^{l}|\bar{\phi}_n(x^{(j)})|^2} &\geqslant \sqrt{\frac{1}{l}}\frac{|\bar{\phi}_n(\bar{x})|}{2}\min\left\{\sqrt{m^d\mathrm{vol}_1^{(d)}\left(\frac{1}{m}\left\lfloor\frac{m|\bar{\phi}_n(\bar{x})|}{2L}\right\rfloor\right)+1}, \sqrt{m^d\mathrm{vol}_1^{(d)}\left(\frac{1}{2}\right)+1}\right\}\\ &= \sqrt{\frac{1}{l}}\frac{|\bar{\phi}_n(\bar{x})|}{2}\min\left\{\sqrt{\left(2\left\lfloor\frac{m|\bar{\phi}_n(\bar{x})|}{2L}\right\rfloor\right)^d+1}, \sqrt{\frac{m^d}{d!}+1}\right\}.\end{aligned} \tag{32}$$

Let m be sufficiently large, so that $|\bar{\phi}_n(\bar{x})| \geqslant \frac{4L}{m}$, and consequently $\left\lfloor \frac{m|\bar{\phi}_n(\bar{x})|}{2L} \right\rfloor \geqslant \frac{m|\bar{\phi}_n(\bar{x})|}{4L}$. By (32) we get

$$|\bar{\phi}_n(\bar{x})| \leqslant \max\left\{ 2^{\frac{2+d}{2}} L^{\frac{2d}{4+2d}} \left(\sqrt{\frac{1}{l}\sum_{j=1}^{l} |\bar{\phi}_n(x^{(j)})|^2} \right)^{\frac{2}{2+d}}, 2^{\frac{2+d}{2}} \sqrt{d!} \sqrt{\frac{1}{l}\sum_{j=1}^{l} |\bar{\phi}_n(x^{(j)})|^2} \right\}.$$

Then, the estimate (30) holds with $c_6(K,f,\lambda,d) := 2^{\frac{2+d}{2}} L^{\frac{2d}{4+2d}}$, $c_7(K,f,\lambda,d) := 2^{\frac{2+d}{2}}\sqrt{d!}$, and $c_8(K,f,\lambda,d) := L(K,f,\lambda,d)$. □

9. Numerical results

In this section we present some simulation results of numerical solutions of Fredholm integral equations of the second kind. The simulations were performed using Matlab 7.7 on a personal computer with a 2.40 GHz Core2 Quad Q6600 CPU and 2 GB of RAM.

Example 1. In the first example, we compared several methods for the approximate solution of the integral Eq. (1) with $X = [0,1]$, $\lambda = -1$, $K(x,y) = e^{-2xy}$, and $f(x) = e^{-2x} + \frac{e^{-(2+2x)}-1}{2(1+x)}$. The example considered was the same as in [27, Table 2, Example I] and its exact solution is $\phi(x) = e^{-2x}$. The methods considered were:

1. the simplified Fredholm integral equation solver from [27,40] (see the Appendix for a short description);
2. the minimization of the residual empirical error

$$\mathcal{E}_{r,z}(\psi) := \frac{1}{l}\sum_{j=1}^{l} ((I - \lambda T_K)(\psi)(x^{(j)}) - f(x^{(j)}))^2 \tag{33}$$

over the set $f + \mathrm{span}_n G$, investigated theoretically in the previous sections.

The dictionaries G considered in item (2) were $G = G_K$ and $G = G_{F_d(X,\mathbb{R}^d)}$ induced by the kernel of the integral equation and the Gaussian kernel, resp.

In our simulations, likewise in [27], integrals of the form $\int_0^1 K(x,y)\psi(y)dy$ (resp., $\int_0^1 R(x,y)f(y)dy$) were discretized and approximated by finite summations of the form $\frac{1}{l}\sum_{j=1}^{l} K(x,y_j)\psi(y_j)$ (resp. $\frac{1}{l}\sum_{j=1}^{l} R(x,y_j)f(y_j)$), where the y_j's are the points of a uniform grid on $X = [0,1]$ made up of l points (the cases $l = 100$ and $l = 1000$ were considered). For the minimization of the objectives in item (2), a multistart procedure was used, based on the quasi-Newton BFGS method, as implemented by the function *fminunc* in the Matlab optimization toolbox [38].

The approximate solution $\tilde{\psi}$ obtained by the simplified Fredholm integral equation solver has the form

$$\tilde{\psi}(x) = f(x) - \lambda\frac{1}{l}\sum_{j=1}^{l} \widetilde{R}_K^{\lambda}(x,y_j)f(y_j), \tag{34}$$

where $\widetilde{R}_K^{\lambda}(x,y)$ is an approximation of the resolvent kernel which is obtained by truncating the expansions (37) and (38) to $j = 13$, likewise in [27]. In particular, formula (34) shows that for the uniform grid with $l = 1000$ (the one considered in the simulations made in [27]), one has $\tilde{\psi} \in f + \mathrm{span}_n G_{\widetilde{R}_K^{\lambda}}$, where $n = 1000$. So, we consider the method in item (2) efficient when it provides an accuracy similar to the one of the the simplified Fredholm integral equation solver, but with a significantly smaller number of terms n. For instance, sparseness of the approximate solution is useful when evaluating it outside the grid.

For an approximation ψ of the solution ϕ, the absolute error is defined as $\max_{j=1}^{l}|\psi(x_j) - \phi(x_j)|$, whereas the l_2 error and the l_2 relative error are defined, resp., as $\sqrt{\frac{1}{l}\sum_{j=1}^{l}(\psi(x_j) - \phi(x_j))^2}$ and $\sqrt{\frac{1}{l}\sum_{j=1}^{l}\frac{(\psi(x_j)-\phi(x_j))^2}{(\phi(x_j))^2}}$. Tables 1 and 2 show the results obtained for the two methods described above. Note that in the numerical results from [27] only the l_2 relative error was provided.

We can see from the Table 1 that, for $l = 100$, method 2 with $G = G_K$ and $n = 5$ computational units was as good as the method 1 (for this value of l, the two simulations were of similar time length). The difference is that the method 2 required only 5 computational units, while the approximate solution obtained by the method 1 is expressed in terms of 100 computational units. For the same value of l, method 2 with $G = G_{F_d(X,\mathbb{R}^d)}$ performed even better than the method 1 in terms of the l_2 relative error. This required only 10 computational units but simulations of longer durations with respect to the case

Table 1
Simulation results for the Example 1 with $l = 100$.

$l = 100$	abs. err.	l_2 err.	l_2 rel. err.
Method 1	$4.3 \cdot 10^{-5}$	$3.6 \cdot 10^{-5}$	$1.2 \cdot 10^{-4}$
Method 2	$4.2 \cdot 10^{-5}(G_K, n = 5)$	$3.6 \cdot 10^{-5}(G_K, n = 5)$	$1.2 \cdot 10^{-4}(G_K, n = 5)$
	$4.7 \cdot 10^{-5}(G_{F_d(X,\mathbb{R}^d)}, n = 10)$	$3.5 \cdot 10^{-5}(G_{F_d(X,\mathbb{R}^d)}, n = 10)$	$1.1 \cdot 10^{-4}(G_{F_d(X,\mathbb{R}^d)}, n = 10)$

Table 2
Simulation results for the Example 1 with $l = 1000$.

$l = 1000$	abs. err.	l_2 err.	l_2 rel. err.
Method 1	$4.3 \cdot 10^{-7}$	$3.6 \cdot 10^{-7}$	$1.2 \cdot 10^{-6}$
Method 2	$5.9 \cdot 10^{-7}(G_K, n = 20)$	$3.9 \cdot 10^{-7}(G_K, n = 20)$	$1.2 \cdot 10^{-6}(G_K, n = 20)$
	$1.3 \cdot 10^{-5}(G_{F_d(X,\mathbb{R}^d)}, n = 25)$	$6.7 \cdot 10^{-6}(G_{F_d(X,\mathbb{R}^d)}, n = 25)$	$2.2 \cdot 10^{-5}(G_{F_d(X,\mathbb{R}^d)}, n = 25)$

$G = G_K$ because of the larger number of parameters to be optimized (30 instead of 10). One may expect a reduction of simulation times by adopting an incremental training technique similar as in [9]. For $l = 1000$, the method 2 with $G = G_K$ provided similar results as the method 1 with only 20 computational units (compared with 1000 computational units used by method 1), but this sparseness was obtained at the cost of a significantly longer simulation time. For $l = 100$ the two methods obtained similar results. Moreover, for $l = 1000$ the choice $G = G_K$ provided better results than the one $G = G_{F_d(X,\mathbb{R}^d)}$, even with a smaller number of computational units. This may be caused by the fact that $\phi(x) - f(x)$ has the integral representation $\lambda \int_X \phi(y) K(x,y) dy$, which is expressed in terms of the kernel K with the possibility of knowing a priori an upper bound on $\sup_{y \in X} |\phi(y)|$, using the techniques of Sections 4 and 5 (the knowledge of this upper bound allows one to restrict the search from $f + \mathrm{span}_n G$ to $f + |\lambda| \sup_{y \in X} |\phi(y)| \mathrm{conv}_n(G \cup -G)$).

Example 2. In the second example, we considered the integral Eq. (1) with $X = [0,1]$, $\lambda = -1$,

$$K(x,y) = \begin{cases} e^{2x}, & 0 \leqslant y < 0.5, \\ e^{-2x}, & 0.5 \leqslant y \leqslant 1 \end{cases}$$

and $f(x) = \gamma_1 e^{-2x} - \gamma_2 e^{2x}$, where $\gamma_1 = 1 - \frac{1}{2}(e^{-1} - e^{-2})$, and $\gamma_2 = \frac{1}{2}(1 - e^{-1})$. The example considered is the same as in [27, Table 2, Example III] and its exact solution is $\phi(x) = e^{-2x}$. Thanks to the form of this kernel K, one has $\mathrm{span}_n G_K = \mathrm{span}_2 G_K$ for every $n \geqslant 2$ and $\phi - f \in \mathrm{span}_2 G_K$. For this reason, we expected a good performance of the method 2 with $G = G_K$ and $n = 2$, which was confirmed by the simulation results reported in Tables 3 and 4. As shown in these tables, the results for the method 2 were better than the ones obtained by the method 1; the simulations made for the two methods required about the same running time.

Example 3. In the third example, we considered the integral Eq. (1) with $X = [0, \sqrt{\pi}] \times [0, \sqrt{\pi}]$, $\lambda = -1/5$, $K(x,y) = \cos(x_1 y_1) \cos(x_2 y_2)$, and $f(x) = 1 - \frac{\sin(x_1)\sin(x_2)}{5 x_1 x_2}$. The example considered is the same as in [8, Example 2] and its exact solution is $\phi(x) = 1$. The simplified Fredholm integral equation solver cannot be applied here (at least in its original formulation), since its extension to the case $d > 1$ has not yet been developed (see [27,40]). The Tables 7 and 8 show the results of the simulations in terms of the absolute error and the l_2 error (for this case, the l_2 relative error is equal to the l_2 error, as $\phi(x) = 1$). Our results in terms of the absolute error were better than the ones reported in [8, Table 3] for the same problem (note that [8, Table 3] reported only the absolute errors, not the l_2 errors). Indeed, for the same choices of the number of Gaussian computational units (4 and 9, resp.), the smallest absolute errors obtained therein were much larger than the smallest ones shown in the Tables 5 and 6 ($6.0 \cdot 10^{-2}$ versus $6.3 \cdot 10^{-4}$, and $1.2 \cdot 10^{-2}$ versus $2.7 \cdot 10^{-4}$, resp.). This may be ascribed to the fact that the simulations reported in [8] employed fixed centers and widths for the Gaussian computational units, whereas in our simulations the centers and widths of the Gaussians were among the tunable parameters.

Example 4. In the fourth example, we considered a nonlinear Fredholm integral equation of the second kind of the form (23), and we searched for its approximate solution by minimining the residual empirical error (33) on $f + \mathrm{span}_n G_{F_d(X,\mathbb{R}^d)}$. It has to be remarked that, since here T_K is a nonlinear operator, the simplified Fredholm integral equation solver cannot be applied (it refers to linear integral equations). The equation considered is similar to the one of [7, Example 6.3] and is given by

$$\phi(x) - 2 \int_0^{\pi} \phi^2(y) dy = f(x), \tag{35}$$

which is of the form (23) with $K(x,y,\phi(y)) = \phi^2(y)$; we chose $f(x) = \sin(x) - \pi$. This equation has the two solutions $\phi_1(x) = \sin(x)$ and $\phi_2(x) = \sin(x) + \frac{7}{2\pi}$. For an approximate solution ψ, the performance indices are the absolute error

$$\min \left\{ \max_{j=1}^{l} |\psi(x_j) - \phi_1(x_j)|, \max_{j=1}^{l} |\psi(x_j) - \phi_2(x_j)| \right\},$$

Table 3
Simulation results for the Example 2 with $l = 100$.

$l = 100$	abs. err.	l_2 err.	l_2 rel. err.
Method 1	$6.5 \cdot 10^{-5}$	$3.7 \cdot 10^{-5}$	$1.8 \cdot 10^{-4}$
Method 2	$6.5 \cdot 10^{-5}(G_K, n = 2)$	$3.7 \cdot 10^{-5}(G_K, n = 2)$	$1.8 \cdot 10^{-4}(G_K, n = 2)$

Table 4
Simulation results for the Example 2 with $l = 1000$.

$l = 1000$	abs. err.	l_2 err.	l_2 rel. err.
Method 1	$6.6 \cdot 10^{-7}$	$3.7 \cdot 10^{-7}$	$1.8 \cdot 10^{-6}$
Method 2	$6.5 \cdot 10^{-7}(G_K, n = 2)$	$3.7 \cdot 10^{-7}(G_K, n = 2)$	$1.7 \cdot 10^{-6}(G_K, n = 2)$

Table 5
Simulation results for the Example 3 with $l = 100$.

$l = 100$	abs. err.	l_2 err.
method 2	$6.3 \cdot 10^{-4}(G_{F_d(X,\mathbb{R}^d)}, n = 4)$	$2.4 \cdot 10^{-3}(G_{F_d(X,\mathbb{R}^d)}, n = 4)$

Table 6
Simulation results for the Example 3 with $l = 225$.

$l = 225$	abs. err.	l_2 err.
Method 2	$2.7 \cdot 10^{-4}(G_{F_d(X,\mathbb{R}^d)}, n = 9)$	$1.3 \cdot 10^{-3}(G_{F_d(X,\mathbb{R}^d)}, n = 9)$

Table 7
Simulation results for the Example 4 with $l = 100$.

$l = 100$	abs. err.	l_2 err.
Method 2	$5.2 \cdot 10^{-5}(G_{F_d(X,\mathbb{R}^d)}, n = 5)$	$9.2 \cdot 10^{-5}(G_{F_d(X,\mathbb{R}^d)}, n = 5)$

Table 8
Simulation results for the Example 4 with $l = 1000$.

$l = 1000$	abs. err.	l_2 err.
Method 2	$8.2 \cdot 10^{-6}(G_{F_d(X,\mathbb{R}^d)}, n = 10)$	$2.7 \cdot 10^{-6}(G_{F_d(X,\mathbb{R}^d)}, n = 10)$

and the l_2 error

$$\min\left\{\sqrt{\frac{1}{l}\sum_{j=1}^{l}(\psi(x_j)-\phi_1(x_j))^2}, \sqrt{\frac{1}{l}\sum_{j=1}^{l}(\psi(x_j)-\phi_2(x_j))^2}\right\}.$$

We did not consider the l_2 relative error, as one of the two solutions ϕ_1 and ϕ_2 has zeros on $X = [0, \pi]$. The Tables 7 and 8 show the obtained results. Although this is not shown in these tables, good approximations of both the solutions ϕ_1 and ϕ_2 were obtained via method 2, starting from different initial choices for the parameters.

10. Discussion

We have applied tools from nonlinear approximation theory to obtain a theoretical background for recent experimental studies [7–9] of approximation of solutions to Fredholm integral equations by neural networks. We have derived estimates of speeds of decrease of errors in approximation of solutions by kernel networks with increasing numbers of computational units. The upper bounds are formulated in terms of the properties of dictionaries defined by kernels. In the estimates, both the supremum norm and the $\mathcal{L}^1$-norm of the function f defining the integral equation $\phi(x) - \lambda \int_X K(x,y)\phi(y)dy = f(x)$ play crucial roles. Our theoretical results provide some guidelines in the choice of radial and kernel-based approximators, for which performance guarantees can be proved.

Appendix A

A.1. The resolvent kernel

For a continuous kernel $K : X \times X \to \mathbb{R}$ on a compact set $X \subset \mathbb{R}^d$ and $\lambda \neq 0$ such that $\frac{1}{\lambda}$ is not an eigenvalue of T_K, the *resolvent kernel* R_K^λ is defined as

$$R_K^{\lambda}(x,y) = \frac{\mathcal{N}(x,y,\lambda)}{\mathcal{D}(\lambda)}, \tag{36}$$

where

$$\mathcal{N}(x,y,\lambda) := -\sum_{j=0}^{\infty} \frac{(-\lambda)^j}{j!} \int_X \int_X \dots \int_X \mathcal{K}\begin{pmatrix} x, \xi_1, \xi_2, \dots, \xi_j \\ y, \xi_1, \xi_2, \dots, \xi_j \end{pmatrix} d\xi_1 d\xi_2 \dots d\xi_j \tag{37}$$

and

$$\mathcal{D}(\lambda) := 1 + \sum_{j=1}^{\infty} \frac{(-\lambda)^j}{j!} \int_X \int_X \dots \int_X \mathcal{K}\begin{pmatrix} \xi_1, \xi_2, \dots, \xi_j \\ \xi_1, \xi_2, \dots, \xi_j \end{pmatrix} d\xi_1 d\xi_2 \dots d\xi_j \tag{38}$$

(see [41, Section 7.2] and [42]). For every positive integer j, the integrands in (37) and (38) have the expressions

$$\mathcal{K}\begin{pmatrix} x, \xi_1, \xi_2, \dots, \xi_j \\ y, \xi_1, \xi_2, \dots, \xi_j \end{pmatrix} := \det \begin{pmatrix} K(x,y) & K(x,\xi_1) & \dots & K(x,\xi_j) \\ K(\xi_1,y) & K(\xi_1,\xi_1) & \dots & K(\xi_1,\xi_j) \\ \dots & \dots & \dots & \dots \\ K(\xi_j,y) & K(\xi_j,\xi_1) & \dots & K(\xi_j,\xi_j) \end{pmatrix}$$

and

$$\mathcal{K}\begin{pmatrix} \xi_1, \xi_2, \dots, \xi_j \\ \xi_1, \xi_2, \dots, \xi_j \end{pmatrix} := \det \begin{pmatrix} K(\xi_1,\xi_1) & \dots & K(\xi_1,\xi_j) \\ \dots & \dots & \dots \\ K(\xi_l,\xi_1) & \dots & K(\xi_j,\xi_l) \end{pmatrix},$$

respectively. Both series defining $\mathcal{N}(x,y,\lambda)$ and $\mathcal{D}(\lambda)$ converge for every $\lambda \in \mathbb{C}$, and $R_K^{\lambda}(x,y)$ is a meromorphic function of λ (i.e., in any bounded region of the complex plane, the only singularities are poles) [24, Section 2.5].

For large values of j and d, in general evaluating the integrals in (37) and (38) is computationally demanding. Indeed, taking into account the definition of the determinant and focusing for instance on the computation of the term

$$\int_X \int_X \dots \int_X \mathcal{K}\begin{pmatrix} x, \xi_1, \xi_2, \dots, \xi_j \\ y, \xi_1, \xi_2, \dots, \xi_j \end{pmatrix} d\xi_1 d\xi_2 \dots d\xi_j \tag{39}$$

in formula (38), for every j this requires the evaluation of $j!$ integrals, each of which has jd variables of integration.

A.2. On the simplified Fredholm integral equation solver

For $d = 1$ it was shown in [27] that the computation of (39) can be simplified by reducing it to the evaluation of a summation depending on j one-dimensional integrals, where the number of terms in the summation is equal to the number $R(j)$ of solutions of the equation

$$1z_1 + 2z_2 + \dots + jz_j = j, \tag{40}$$

where $(z_1, z_2 \dots, z_j)$ is a vector of natural numbers including 0. The solver based on this method was called "simplified Fredholm integral equation solver" in [40]. The number $R(j)$ was computed in [27,40] for small values of j up to $j = 8$ but neither a general expression of $R(j)$ nor its asymptotic behaviour were investigated therein. However, one can see from the definition that $R(j)$ is equal to the number of distinct and order-independent ways in which j can be decomposed as the sum of natural numbers. Therefore, it coincides with the so-called "partition function" $p(j)$ in number theory [43], whose behavior is of the form $p(j) \sim \frac{1}{4j\sqrt{3}} e^{\pi\sqrt{\frac{2j}{3}}}$ for $j \to +\infty$ [44]. Just to give two examples of this behaviour, $p(100) = 190569292$ and $p(1000)$ is about $2.4 \cdot 10^{31}$. So, the computation of each term (39) is demanding for large j, even under the simplification made in [27,40] (for some integral equations arising from thermal engineering problems it was observed in [40] that a truncation of the series (37) and (38) to $j = 13$, for which one has $p(13) = 101$, may be satisfactory). An extension of the simplified Fredholm integral equation solver to the d-dimensional is under investigation (see [27,40]). For the computation of (39), this extension would likely require the evaluation of a summation depending on j integrals with d variables of integration, where the number of terms in the summation would be again $R(j) = p(j)$.

References

[1] Y. Lu, L. Shen, Y. Xu, Integral equation models for image restoration: high accuracy methods and fast algorithms, Inverse Problems 26, doi:10.1088/0266-5611/26/4/045006.
[2] W.V. Lovitt, Linear Integral Equations, Dover, New York, 1950.
[3] A.T. Lonseth, Sources and applications of integral equations, SIAM Rev. 19 (1977) 241–278.
[4] K. Atkinson, The Numerical Solution of Integral Equations of the Second Kind, Cambridge University Press, 1997.
[5] A. Forrester, A. Sobester, A. Keane, Engineering Design via Surrogate Modelling: A Practical Guide, Wiley, 2008.

[6] K. Atkinson, W. Han, Theoretical Numerical Analysis: A Functional Analysis Framework, Springer, 2005.
[7] S. Effati, R. Buzhabadi, A neural network approach for solving Fredholm integral equations of the second kind, Neural Computing and Applications doi:10.1007/s00521-010-0489-y.
[8] A. Alipanah, S. Esmaeili, Numerical solution of the two-dimensional Fredholm integral equations by Gaussian radial basis function, J. Comput. Appl. Math. 235 (2011) 5342–5347.
[9] A. Golbabai, S. Seifollahi, Numerical solution of the second kind integral equations using radial basis function networks, Appl. Math. Comput. 174 (2006) 877–883.
[10] A.R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inform. Theory 39 (1993) 930–945.
[11] G. Gnecco, V. Kůrková, M. Sanguineti, Some comparisons of complexity in dictionary-based and linear computational models, Neural Networks 24 (2011) 171–182.
[12] G. Gnecco, V. Kůrková, M. Sanguineti, Can dictionary-based computational models outperform the best linear ones?, Neural Networks 24 (2011) 881–887
[13] V. Kůrková, M. Sanguineti, Comparison of worst-case errors in linear and neural network approximation, IEEE Trans. Inform. Theory 48 (2002) 264–275.
[14] R. Gribonval, P. Vandergheynst, On the exponential convergence of matching pursuits in quasi-incoherent dictionaries, IEEE Trans. Inform. Theory 52 (2006) 255–261.
[15] G. Gnecco, V. Kůrková, M. Sanguineti, Bounds for approximate solutions of Fredholm integral equations using kernel networks, in: T. Honkela, W. Duch, M. Girolami, S. Kaski (Eds.), Lecture Notes in Computer Science, vol. 6791, Springer, Heidelberg, 2011, pp. 126–133. Proc. ICANN 2011.
[16] A. Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.
[17] H.N. Mhaskar, Versatile Gaussian networks, in: Proceedings of IEEE Workshop of Nonlinear Image Processing, 1995, pp. 70–73.
[18] L.K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Stat. 20 (1992) 608–613.
[19] V. Kůrková, M. Sanguineti, Geometric upper bounds on rates of variable-basis approximation, IEEE Trans. Inform. Theory 54 (2008) 5681–5688.
[20] C. Darken, M. Donahue, L. Gurvits, E. Sontag, Rate of approximation results motivated by robust neural network learning, in: Proceedings of 6th Annual ACM Conference on Comput Learning Theory, The Association for Computing Machinery, New York, 1993, pp. 303–309.
[21] F. Girosi, Approximation error bounds that use VC- bounds, in: Proceedings of International Conference on Artificial Neural Networks, Paris, 1995, pp. 295–302.
[22] L. Gurvits, P. Koiran, Approximation and learning of convex superpositions, J. Comput. Syst. Sci. 55 (1997) 161–170.
[23] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[24] F.G. Tricomi, Integral Equations, Interscience, New York, 1957.
[25] R.M. Dudley, Balls in $\mathbb{R}^k$ do not cut all subsets of k+2 points, Adv. Math. 31 (1979) 306–308.
[26] E. Sontag, VC dimension of neural networks, in: C. Bishop (Ed.), Neural Networks and Machine Learning, Springer, Berlin, 1998, pp. 69–95.
[27] K.G. Terry Hollands, A simplification to Fredholm's solution to the Fredholm integral equation of the second kind, Appl. Math. Comput. 189 (2007) 1078–1086.
[28] H. Hochstadt, Integral Equations, Wiley-Interscience, 1989.
[29] B. Schölkopf, A.J. Smola, Learning With Kernels – Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.
[30] P.C. Kainen, V. Kůrková, M. Sanguineti, Complexity of Gaussian radial-basis networks approximating smooth functions, J. Complex. 25 (2009) 63–74.
[31] E.M. Stein, Singular Integrals and Differentiability Properties of Functions, Princeton University Press, Princeton, NJ, 1970.
[32] C. Martínez, M. Sanz, The Theory of Fractional Powers of Operators, Elsevier, Amsterdam, 2001.
[33] P.C. Kainen, V. Kůrková, M. Sanguineti, Dependence of computational models on input dimension: Tractability of approximation and optimization tasks, in: IEEE Transactions on Information Theory 58, doi:10.1109/TIT.2011.2169531.
[34] H.N. Mhaskar, On the tractability of multivariate integration and approximation by neural networks, J. Complex. 20 (2004) 561–590.
[35] M. Krasnov, A. Kiselev, G. Makarenko, Problems and Exercises in Integral Equations, Mir Publishers, Moscow, 1971.
[36] R. Courant, Differential and Integral Calculus, vol. II, Wiley-Interscience, 1988.
[37] D.P. Bertsekas, Nonlinear Programming, Athena Scientific, Belmont, MA, 1999.
[38] MathWorks, Matlab Optimization Toolbox. URL <http://www.mathworks.com/products/optimization/>.
[39] K. Ball, An elementary introduction to modern convex geometry, in: Flavors of Geometry, Cambridge University Press, 1997, pp. 1–58.
[40] K.G. Terry Hollands, The simplified-Fredholm integral equation solver and its use in thermal radiation, J. Heat. Trans. 132 (2010) 023401-1–023401-6.
[41] I. Gohberg, S. Goldberg, N. Krupnik, Traces and Determinants of Linear Operators, Birkhäuser Verlag, 2000.
[42] J. Feinberg, Fredholm's minors of arbitrary order: their representations as a determinant of resolvents and in terms of free fermions and an explicit formula for their functional derivative, J. Phys. A: Math. General 37 (2004) 6299–6310.
[43] G.E. Andrews, The Theory of Partitions, Cambridge University Press, 1998.
[44] G.H. Hardy, S. Ramanujan, Asymptotic formula in combinatory analysis, Proc. Lond. Math. Soc. 17 (1918) 17–75.