An investigation of global radial basis function collocation methods applied to Helmholtz problems

Elisabeth Larsson[a] · Ulrika Sundin[a]

Communicated by E. Francomano

Abstract

Global radial basis function (RBF) collocation methods with infinitely smooth basis functions for partial differential equations (PDEs) work in general geometries, and can have exponential convergence properties for smooth solution functions. At the same time, the linear systems that arise are dense and severely ill-conditioned for large numbers of unknowns and small values of the shape parameter that determines how flat the basis functions are. We use Helmholtz equation as an application problem for the theoretical analysis and numerical experiments. We analyze and characterize the convergence properties as a function of the number of unknowns and for different shape parameter ranges. We provide theoretical results for the flat limit of the PDE solutions and investigate when the non-symmetric collocation matrices become singular. We also provide practical strategies for choosing the method parameters and evaluate the results on Helmholtz problems in a curved waveguide geometry.

1 Introduction

We started writing this paper in 2004. Some of the results can be found in the MSc thesis of the second author [39]. At that time, the first paper on the flat radial basis function (RBF) interpolation limit [7] had just been published, and most of the work on the paper about multivariate flat RBF limits [24] was done, but the paper was not published yet. The focus of research in RBF-based methods for partial differential equations (PDEs) was on global collocation methods, and we were interested in the limit behavior for RBF approximations to PDEs. Then the manuscript ended up 'in a drawer' due to various circumstances, and we came to pick it up again 15 years later. The current research focus has shifted to localized RBF-methods such as RBF-generated finite difference methods (RBF-FD) [12] and RBF partition of unity methods (RBF-PUM) [27]. However, we think that the results in this paper, even though they are on global RBF methods, provide insights that are generally useful also today. The objectives of the work are

- to investigate the approximation errors theoretically and numerically to gain understanding about the flat limit, the convergence properties, and the dependence on the shape parameter,
- to identify the gaps between theoretical results and numerical behavior,
- to provide practically useful strategies for choosing the method parameters and assessing the results.

Other articles that consider global RBF methods for Helmholtz problems are [31], where a Cauchy Helmholtz problem is solved using Bessel RBFs combined with regularization, and [53] where a global symmetric RBF collocation method is used for Helmholtz equation with a manufactured solution. For papers where Helmholtz wave propagation problems such as those considered in this paper are solved, we refer to [33] and [37] where localized RBF-FD methods are used.

The outline of the paper is as follows: In Section 2, we define three different Helmholtz test problems that are used throughout the paper. In Section 3 we derive the systems of equations for non-symmetric and symmetric collocation. Section 4 is devoted to cases where the non-symmetric collocation matrix is singular, and in Section 5, we discuss the limit properties. How to prove these properties is sketched in Appendix A. Section 6 contains a combination of theoretical error estimates, and more heuristic error approximations. Then in Section 7, we provide numerical results as well as practical strategies for method parameter selection. The paper ends with a discussion of the results in Section 8.

2 Generic and specific model problems

Throughout the paper, we consider time-independent, linear, partial differential equations (PDEs). We assume that the PDE, together with the different boundary conditions can be summarized as

$$\mathcal{L}^i u(\underline{x}) = f^i(\underline{x}), \quad \underline{x} \in \Omega^i, \quad i = 1, \dots, N_{\text{op}}, \tag{1}$$

[a]Scientific Computing, Department of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden (elisabeth.larsson@it.uu.se), (ulrika.sundin@it.uu.se).

where $\mathcal{L}^i$ is a linear operator, u is the solution function, f^i is a given function, $\underline{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$, and $\Omega^i \subseteq \bar{\Omega}$ is a region in the computational domain or a boundary segment.

To give examples and illustrate specific properties, we use a series of Helmholtz problems of increasing complexity. The Helmholtz equation models time-harmonic wave propagation, and in all cases, we consider waveguide problems with a wave originating from a source at the left boundary and propagating to the right. We allow reflected waves from the interior of the domain to propagate back to the left and out through the left boundary, but no waves may enter from outside the right boundary. The main reasons for our choice of model problems are the following:

- There is one problem parameter, the wavenumber κ, that can be varied to study its relation to the RBF method parameters.

- A Helmholtz problem is generally more difficult to solve than a Laplace or Poisson problem, especially for large wavenumbers, due to the indefiniteness of the operator, the wave nature of the solution, and the typically more complicated boundary conditions.

The Helmholtz PDE is in all examples given by

$$\mathcal{L}^1 u(\underline{x}) = -\Delta u(\underline{x}) - \kappa^2 u(\underline{x}) = 0, \quad \underline{x} \in \Omega^1 = \Omega. \tag{2}$$

The first and simplest model problem is one-dimensional, with $\Omega = (0, 1)$. The non-reflecting (or radiation) boundary conditions are given by

$$\mathcal{L}^2 u(x) = -\frac{du}{dx}(x) - i\kappa u(x) = -2i\kappa, \quad x = 0, \tag{3}$$

$$\mathcal{L}^3 u(x) = \frac{du}{dx}(x) - i\kappa u(x) = 0, \quad x = 1, \tag{4}$$

and the analytical solution is $u(x) = \exp(i\kappa x)$, if κ is constant.

The second problem is two-dimensional with a rectangular domain $\Omega = (0, L_1) \times (0, 1)$. At the top and bottom boundaries, we use the Dirichlet boundary condition

$$\mathcal{L}^4 u(\underline{x}) = u(\underline{x}) = 0, \quad \underline{x} = (0, x_2) \text{ or } \underline{x} = (L_1, x_2), \tag{5}$$

indicating that we consider a waveguide type of problem. The conditions at the left and right boundaries are

$$\mathcal{L}^2 u(\underline{x}) = -\frac{\partial u}{\partial x_2}(\underline{x}) - i\beta_m u(\underline{x}) = -2i\beta_m \sin(\alpha_m x_1), \quad \underline{x} = (x_1, 0), \tag{6}$$

$$\mathcal{L}^3 u(\underline{x}) = \frac{\partial u}{\partial x_2}(\underline{x}) - i\beta_m u(\underline{x}) = 0, \quad \underline{x} = (x_1, 1), \tag{7}$$

where $\alpha_m = \frac{m\pi}{L_1}$, $\beta_m = \sqrt{\kappa^2 - \alpha_m^2}$, and $m \geq 1$ is an integer. These conditions allow for just one propagating mode in the solution, which is given by $u(\underline{x}) = \exp(i\beta_m x_2)\sin(\alpha_m x_1)$, assuming a constant κ. It should be noted that if κ and m are chosen such that $\beta_m = 0$, the problem is not well-defined, and we avoid such combinations in the experiments.

The third and final problem is also two-dimensional, but the domain Ω is now enclosed between two curves $\gamma_1(x_2) < x_1 < \gamma_2(x_2)$, $x_2 \in (0, 1)$, see Figure 1. The Dirichlet condition (5) is modified to hold at γ_1 and γ_2.

$$\mathcal{L}^4 u(\underline{x}) = u(\underline{x}) = 0, \quad \underline{x} = (\gamma_j(x_2), x_2), \quad j = 1, 2. \tag{8}$$

At the left and right boundary, we use so called Dirichlet–to–Neumann map (DtN) radiation boundary conditions [22]

$$\begin{aligned}
\mathcal{L}^2 u(\underline{x}) &= -\frac{\partial u}{\partial x_2} - i\sum_{m=1}^{\infty} \beta_m \langle u(\cdot, 0), \psi_m^0 \rangle \psi_m^0(x_1) \\
&= -2i\sum_{m=1}^{\infty} A_m \beta_m \psi_m^0(x_1), \quad x_2 = 0, \\
\mathcal{L}^3 u(\underline{x}) &= \frac{\partial u}{\partial x_2} - i\sum_{m=1}^{\infty} \beta_m \langle u(\cdot, 1), \psi_m^1 \rangle \psi_m^1(x_1) = 0, \quad x_2 = 1,
\end{aligned} \tag{9}$$

where, for a fixed x_2, the modes $\psi_m^{x_2} = \sqrt{2}\sin(\alpha_m(x_1 - \gamma_1(x_2)))$, with $\alpha_m = \frac{m\pi}{\gamma_2(x_2) - \gamma_1(x_2)}$. The inner product is given by

$$\langle u(\cdot, x_2), \psi_m^{x_2} \rangle = \int_{\gamma_1(x_2)}^{\gamma_2(x_2)} u(x_1, x_2)\psi_m^{x_2}(x_1)\,dx_1, \tag{10}$$

and the amplitudes $A_m = \psi_m^0(x_s)$, where x_s is the position of the source in the vertical coordinate. The amplitudes are chosen to emulate a point source. The DtN conditions allow for any combination of modes to move transparently through the vertical boundaries. For practical and computational reasons, the infinite sum is truncated at $\mu_{x_2} = \lfloor \frac{\kappa(\gamma_2(x_2) - \gamma_1(x_2))}{\pi} \rfloor$. For a discussion of the assumptions behind this truncation and these particular DtN conditions, see [38].
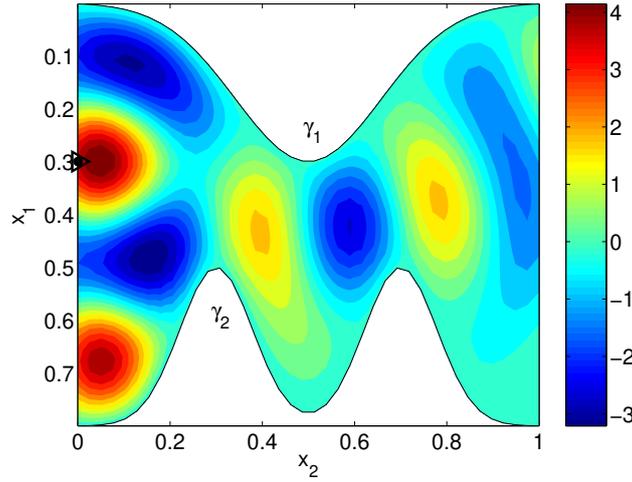
Figure 1: Wave propagation in an M-shaped duct. The source position is indicated by the marker at the left boundary and the wavenumber is $\kappa = 6\pi$. The real part of the solution is displayed.

3 The RBF approximations

In this section, we first describe Kansa's non-symmetric collocation method [20] for our model problems. The main advantage of the non-symmetric collocation method is its simplicity. It has been shown that the non-symmetric RBF discretization matrix, in rare cases [19], can become singular. This is discussed further in Section 4. It has further been shown that if a greedy method is used for selection of basis functions and collocation points, the non-symmetric method is convergent [32].

To ensure non-singularity, symmetric collocation [55, 8, 17] can instead be employed. This is slightly more involved, especially with non-trivial operators, which is why we include an example of how to do this for the one-dimensional model problem. Another approach to avoid singularity is to use non-symmetric collocation in a least-squares setting. See [40] for an example of global least squares RBF approximation and [50] for a general theoretical discussion of the stability of least squares approximation of linear PDE problems.

The non-symmetric and symmetric collocation methods were compared in [23] and were found to give almost identical approximation results. A similar comparison in [4], showed better performance with non-symmetric collocation in an adaptive setting.

3.1 Non-symmetric collocation

When we use non-symmetric collocation to discretize the problem (1), the RBF approximant is given by

$$s(\underline{x}) = \sum_{j=1}^{N} \lambda_j \phi(\varepsilon \|\underline{x} - \underline{x}_j\|) = \sum_{j=1}^{N} \lambda_j \phi_j(\underline{x}), \tag{11}$$

where $\underline{x}_j$, $j = 1, \ldots, N$ are the RBF center points and ε is the shape parameter. The collocation conditions are imposed at the N center points. Let $\underline{x}_j^k$, $j = 1, \ldots, N_k$ be the subset of center points that belong to the region or section Ω^k. The corresponding operator is used for collocation, and we get the equations

$$\mathcal{L}^k s(\underline{x}_i^k) = \sum_{j=1}^{N} \lambda_j \mathcal{L}^k \phi_j(\underline{x}_i^k) = f^k(\underline{x}_i^k), \quad i = 1, \ldots, N_k, \quad k = 1, \ldots, N_{\mathrm{op}}.$$

If the points are ordered according to the set affiliation, we get a system of equations, $M\underline{\lambda} = \underline{f}$, with the following general block structure

$$\begin{pmatrix} \mathcal{L}^1 \phi \\ \vdots \\ \mathcal{L}^{N_{\mathrm{op}}} \phi \end{pmatrix} \begin{pmatrix} \underline{\lambda} \end{pmatrix} = \begin{pmatrix} \underline{f}^1 \\ \vdots \\ \underline{f}^{N_{\mathrm{op}}} \end{pmatrix}, \tag{12}$$

where the block $\mathcal{L}^k \phi$ is of size $(N_k \times N)$.

Applying the operators in the specific model problems to the RBFs is straightforward, except for the DtN operators in (9). The left boundary condition applied to one of the RBFs and evaluated at the point $\underline{x} = (x_1, 0)$ takes the form

$$\mathcal{L}^2 \phi_j(\underline{x}) = -\frac{\partial \phi_j}{\partial x_2}(\underline{x}) - i \sum_{m=1}^{\mu_0} \beta_m \langle \phi_j(\cdot, 0), \psi_m^0 \rangle \psi_m^0(x_1).$$

To form the whole block $\mathcal{L}^2 \phi$, we need to evaluate $\mu_0 \cdot N$ inner products. This cannot in general be done analytically for infinitely smooth RBFs such as multiquadrics, inverse quadratics, or Gaussians.

One of our aims with choosing the Helmholtz model problems was to see if using RBFs would make it difficult to implement non-trivial boundary conditions. There are no fundamental issues preventing implementation of boundary conditions involving linear functionals applied to the basis functions. A practical issue is that the computational cost for the quadrature is quite large, although linear in N. In Section 7, we investigate how accurately we need to compute the inner products to not destroy the overall accuracy of the solution. The experiments show that we need to compute the inner products more accurately than the overall error tolerance, which increases the cost further.

3.2 Symmetric collocation

Non-singularity of the RBF approximation matrix can be ensured through symmetric collocation [55, 8, 17]. The idea is to view the RBF $\phi(\varepsilon\|\underline{x}-\underline{\xi}\|)$ as a function of two variables $\psi(\underline{x},\underline{\xi})$. Then in the ansatz for the RBF approximation, for each basis function, the operator corresponding to its center location is applied to the second argument of the basis function. Since we consider complex operators, we also need to conjugate the operators in order to get a Hermitian matrix in the end. The approximation then takes the form

$$s(\underline{x}) = \sum_{k=1}^{N_{\mathrm{op}}} \sum_{j=1}^{N_k} \lambda_j^k \overline{\mathcal{L}_{\xi}^k} \psi(\underline{x},\underline{x}_j^k).$$

For the one-dimensional Helmholtz problem, collocation with this ansatz leads to a system of equations with the following structure

$$\begin{pmatrix} \mathcal{L}_x^1 \overline{\mathcal{L}_{\xi}^1} \psi & \mathcal{L}_x^1 \overline{\mathcal{L}_{\xi}^2} \psi & \mathcal{L}_x^1 \overline{\mathcal{L}_{\xi}^3} \psi \\ \mathcal{L}_x^2 \overline{\mathcal{L}_{\xi}^1} \psi & \mathcal{L}_x^2 \overline{\mathcal{L}_{\xi}^2} \psi & \mathcal{L}_x^2 \overline{\mathcal{L}_{\xi}^3} \psi \\ \mathcal{L}_x^3 \overline{\mathcal{L}_{\xi}^1} \psi & \mathcal{L}_x^3 \overline{\mathcal{L}_{\xi}^2} \psi & \mathcal{L}_x^3 \overline{\mathcal{L}_{\xi}^3} \psi \end{pmatrix} \begin{pmatrix} \underline{\lambda}^1 \\ \underline{\lambda}^2 \\ \underline{\lambda}^3 \end{pmatrix} = \begin{pmatrix} \underline{0} \\ -2i\kappa \\ 0 \end{pmatrix},$$

where the block $\mathcal{L}_x^j \overline{\mathcal{L}_{\xi}^k} \psi$ is of size $(N_j \times N_k)$. To see that the coefficient matrix M really is Hermitian, we can use the following differentiation rules for the RBFs

$$\frac{\partial^n}{\partial \xi^n} \psi(\underline{x}_j, \underline{x}_k) = (-1)^n \frac{\partial^n}{\partial x^n} \psi(\underline{x}_j, \underline{x}_k), \tag{13}$$

$$\frac{\partial^n}{\partial x^n} \psi(\underline{x}_k, \underline{x}_j) = (-1)^n \frac{\partial^n}{\partial x^n} \psi(\underline{x}_j, \underline{x}_k). \tag{14}$$

We can then show for the different blocks in the matrix that the matrix elements satisfy $m_{jk} = \overline{m}_{kj}$. As an example, for elements in the first two off-diagonal blocks we get

$$\mathcal{L}_x^1 \overline{\mathcal{L}_{\xi}^2} \psi(\underline{x}_j, \underline{x}_k) = (-\frac{\partial^2}{\partial x^2} - \kappa^2)(-\frac{\partial}{\partial \xi} + i\bar{\kappa})\psi(\underline{x}_j, \underline{x}_k) = (-\frac{\partial^2}{\partial x^2} - \kappa^2)(\frac{\partial}{\partial x} + i\bar{\kappa})\psi(\underline{x}_j, \underline{x}_k),$$

$$\overline{\mathcal{L}_x^2 \overline{\mathcal{L}_{\xi}^1}} \psi(\underline{x}_k, \underline{x}_j) = (-\frac{\partial}{\partial x} + i\bar{\kappa})(-\frac{\partial^2}{\partial \xi^2} - \kappa^2)\psi(\underline{x}_k, \underline{x}_j) = (\frac{\partial}{\partial x} + i\bar{\kappa})(-\frac{\partial^2}{\partial x^2} - \kappa^2)\psi(\underline{x}_j, \underline{x}_k).$$

Apart from the important non-singularity property, limited numerical experiments also show that the conditioning is slightly better (one order of magnitude) than for the non-symmetric method. However, the error curves, as functions of both x and ε, are close to identical.

It would be challenging to implement the symmetric collocation method for the two-dimensional problem with DtN boundary conditions. We would need to apply the non-local operators both to the basis functions and to their derivatives, and we would need to apply nested non-local operators leading to double integrals for all mode combinations. As the integrals form a large part of the computational cost, the symmetric collocation method is significantly more expensive than the non-symmetric collocation method for a given accuracy for the problem sizes we consider.

In the following, we choose to study the properties of the non-symmetric collocation method. We are not using the greedy method of [32] in our selection of center and collocation points. However, the results we present are in most cases valid both for coinciding and decoupled center and collocation points. The numerical results that we present are slightly worse than what we would get in the optimal case, but still represent the performance of the method well. In particular, we use well-distributed center points, which is a key factor in keeping the condition number of the rather badly conditioned matrices under control.

4 Singularity of the RBF collocation matrix

As already stated, the RBF collocation matrix may become singular with the non-symmetric collocation approach. This becomes particularly clear for problems with a parameter that can be varied freely as for our Helmholtz examples. For the one-dimensional Helmholtz model problem, we can in fact show that for any given node distribution (with distinct nodes) there are always wavenumbers κ that lead to a singular collocation matrix.

To get the equations in an appropriate form for eigenvalue analysis, we multiply the PDE (2) with -1 and the boundary conditions (3) and (4) with $i\kappa$. After collocation with the PDE at the interior points x_j^1, and the boundary conditions at the boundary points, we get a collocation matrix M with elements

$$m_{jk} = \begin{cases} \kappa^2 \phi_k(x_j^1) & + & \phi_k''(x_j^1), & j = 1,\dots,N-2, & k = 1,\dots,N, \\ \kappa^2 \phi_k(0) & - & i\kappa\phi_k'(0) & , & j = N-1, & k = 1,\dots,N, \\ \kappa^2 \phi_k(1) & + & i\kappa\phi_k'(1) & , & j = N, & k = 1,\dots,N. \end{cases}$$

We can express M as a matrix polynomial in κ,

$$M = \kappa^2 A + \kappa i B + C,$$

where A, B, and C are real matrices. Furthermore, A is the usual RBF interpolation matrix. The question of singularity of M can be posed as a quadratic eigenproblem

$$(\kappa^2 A + \kappa i B + C)\underline{v} = \underline{0}. \tag{15}$$

For standard RBFs and distinct points, A is non-singular. By introducing $\underline{w} = \kappa \underline{v}$ we can then reformulate (15) as a standard eigenvalue problem

$$\begin{pmatrix} 0 & I \\ -A^{-1}C & -iA^{-1}B \end{pmatrix} \begin{pmatrix} \underline{v} \\ \underline{w} \end{pmatrix} = \kappa \begin{pmatrix} \underline{v} \\ \underline{w} \end{pmatrix}.$$

Solving this problem leads to $2N$ eigenvalues. That is, values of κ for which the collocation matrix M is singular. Two of the eigenvalues have to be $\kappa = 0$ because of the scaling of the boundary conditions. By conjugating equation (15), we get

$$(\bar{\kappa}^2 A - \bar{\kappa} iB + C)\underline{\bar{v}} = ((-\bar{\kappa})^2 A + (-\bar{\kappa})iB + C)\underline{\bar{v}} = \underline{0}.$$

That is, if $(\kappa, \underline{v})$ is an eigenvalue–eigenvector pair, then $(-\bar{\kappa}, \underline{\bar{v}})$ also is. Hence, all eigenvalues with $\mathrm{Re}(\kappa) \neq 0$ must come in pairs $(\kappa, -\bar{\kappa})$. Then, there may also be a number of eigenvalues on the imaginary axis. The κ that are of interest in the Helmholtz problem are such that $\mathrm{Re}(\kappa) > 0$. We are then left with a maximum of $N - 1$ potentially interesting wavenumbers that lead to a singular problem.

In Figure 2, the eigenvalues that lead to a singular system are computed for different problem sizes using multiquadric and Gaussian RBFs. For multiquadrics, there are no eigenvalues in the region of interest, that is, real wavenumbers with solutions that are well resolved. The eigenvalues with the largest real part are closest to the real axis. These problems are resolved to $2\pi N/\kappa \approx 2$ points per wavelength, which is the theoretical lowest possible resolution to use for a wave propagation problem. The Gaussian approximation produces eigenvalues that are closer to the real axis, but also here the eigenvalues with large real part correspond to badly resolved problems. It should be noted that complex wavenumbers, typically with a significantly smaller imaginary part than real part, are used in practical applications to model damping within the media that the waves propagate through.
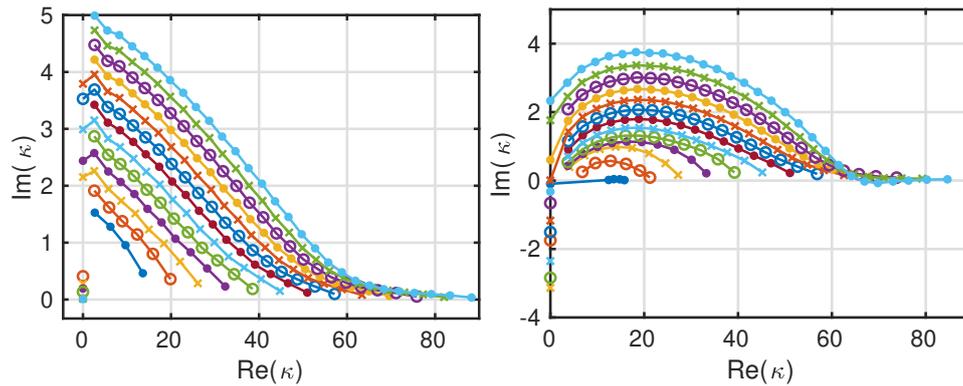
Figure 2: The wavenumbers that lead to a singular system for the one-dimensional problem using $N = 6, 8, \ldots, 30$ from bottom to top, for multiquadric RBFs with $\varepsilon = 5$ (left) and Gaussian RBFs with $\varepsilon = 10$ (right).

5   The flat RBF limit for PDE problems

The limits of multivariate RBF interpolants as the shape parameter ε goes to zero were analyzed in [24, 47, 30, 49, 29]. The same type of limits for finitely smooth RBFs were studied in [52, 28]. We also mention the recent article [21], where limits of linear functionals applied to Gaussian RBF approximations are investigated. The previous results show that the limit behavior is closely related to polynomial unisolvency [2] on the set of node points. We define

$$N_{K,d} = \begin{pmatrix} K + d \\ K \end{pmatrix}, \tag{16}$$

which is the dimension of the space of polynomials of degree K in $\mathbb{R}^d$. If $N = N_{K,d}$, and the node set is unisolvent, then the (infinitely smooth) flat limit RBF interpolant reproduces the multivariate polynomial interpolant of degree K on these nodes, and has the form

$$s(\underline{x}) = P_K(\underline{x}) + \varepsilon^2 P_{K+2}(\underline{x}) + \varepsilon^4 P_{K+4}(\underline{x}) + \cdots, \tag{17}$$

where $P_K(\underline{x})$ is the polynomial interpolant and $P_{K+2j}(\underline{x})$, $j \geq 1$ are zero at the node points.

When we apply the non-symmetric collocation method to a PDE problem, the RBF approximant has the same general form (11), and we can derive corresponding results for the limit. If the node set fulfills both unisolvency conditions described below, the limit RBF approximant has the form (17), but here $P_K(\underline{x})$ is the unique polynomial solution of degree K to the PDE problem, and P_{K+2j} have zero PDE residual at the node points. We note that for small shape parameters, the polynomial approximation is the dominating part of the RBF approximation.

In order to express the conditions for different limit results, we need to define two matrices, P and Q, from which we can determine polynomial unisolvency and unisolvency of the discrete PDE problem. Let $\{p_j(\underline{x})\}_{j=1}^N$ be N linearly independent monomials of minimal degree in d dimensions. For example, for $N = 7$ and $d = 2$, we can choose $\{1, x, y, x^2, xy, y^2, x^3\}$. If $N_{K-1,d} < N \le N_{K,d}$, then the degree of $p_N(\underline{x})$ is K.

The set of node points $\{x_i\}_{i=1}^N$ satisfies polynomial unisolvency if there, for any given data at the node points, is a unique linear combination $\sum_{j=1}^N \beta_j p_j(\underline{x})$ that interpolates the data. This is equivalent to non-singularity of the matrix

$$P = \begin{pmatrix} p_1(\underline{x}_1) & p_2(\underline{x}_1) & \cdots & p_N(\underline{x}_1) \\ p_1(\underline{x}_2) & p_2(\underline{x}_2) & \cdots & p_N(\underline{x}_2) \\ \vdots & \vdots & & \vdots \\ p_1(\underline{x}_N) & p_2(\underline{x}_N) & \cdots & p_N(\underline{x}_N) \end{pmatrix}. \tag{18}$$

In cases where P is singular, we instead construct a minimal non-degenerate basis [24]. Such a basis can be constructed by choosing N monomials of smallest possible degree under the constraint that they give linearly independent columns in the matrix P. The highest selected monomial degree M is then also the degree of $p_N(\underline{x})$. As an example, for $N = 5$ node points all located on the line $x = y$, a minimal non-degenerate basis is $\{1, x, x^2, x^3, x^4\}$ and $M = 4$.

Unisolvency of the discrete PDE problem on the set of node points $\{x_i\}_{i=1}^N$ with respect to $\{p_j(\underline{x})\}_{j=1}^N$ requires a unique linear combination $\sum_{j=1}^N \beta_j p_j(\underline{x})$ that satisfies the collocation conditions

$$\sum_{j=1}^N \beta_j \mathcal{L}^k p_j(\underline{x}_i^k) = f^k(\underline{x}_i^k), \quad i = 1, \ldots, N_k, \quad k = 1, \ldots, N_{\text{op}}.$$

This is equivalent to non-singularity of the matrix

$$Q = \begin{pmatrix} \mathcal{L}^1 p_1(\underline{x}_1^1) & \cdots & \mathcal{L}^1 p_N(\underline{x}_1^1) \\ \vdots & & \vdots \\ \mathcal{L}^{N_{\text{op}}} p_1(\underline{x}_{N_{N_{\text{op}}}}^{N_{\text{op}}}) & \cdots & \mathcal{L}^{N_{\text{op}}} p_N(\underline{x}_{N_{N_{\text{op}}}}^{N_{\text{op}}}) \end{pmatrix}. \tag{19}$$

As in [24], we need the RBFs to fulfill three conditions in order to get the results in the theorem given below. We repeat the conditions and discuss their validity briefly here, but for a full explanation, we refer the reader to [24].

(I) The RBF $\phi(r)$ can be Taylor expanded as $\phi(r) = \sum_{j=0}^\infty a_j r^{2j}$.

(II) The PDE collocation matrix M in system (12) is non-singular in the interval $0 < \varepsilon \le R$, for some $R > 0$.

(III) Certain matrices $A_{p,J}$, built from the coefficients a_j in the Taylor expansion of $\phi(r)$, are non-singular for $0 \le p \le d$ and $0 \le J \le K$.

Condition (I) is true for all infinitely smooth RBFs that are commonly used. Condition (II) is likely to hold for some value of R, but the previous section shows that M can become singular at any ε, given a specific combination of PDE problem and node points. Condition (III) was shown to hold for these RBFs in [30].

The following theorem gives the different possibilities for the limiting RBF approximant as the shape parameter $\varepsilon \to 0$.

Theorem 5.1. *Assume that the RBF $\phi(r)$ fulfills conditions (I)–(III) and that the number of node points satisfies $N_{K-1,d} < N \le N_{K,d}$. The degree of a minimal non-degenerate basis for the point set is either K for a unisolvent set or M for a non-unisolvent set. If*

(i) *P and Q are non-singular, the limit exists and is a polynomial of deg K. If $N = N_{K,d}$ it is the unique degree K polynomial solution to the discrete PDE problem, otherwise the final polynomial depends on the choice of RBF.*

(ii) *P is singular, but Q is non-singular, the limit exists and is an RBF-dependent polynomial of degree M.*

(iii) *P is non-singular, but Q is singular, divergence will occur unless the right hand side $\underline{f}$ of system (12) happens to be in the range of Q. If there is just a single null-space polynomial $n(\underline{x})$ of degree K, the divergent term is proportional to $\varepsilon^{-2} n(\underline{x})$.*

(iv) *P has a nullspace of dimension $m > 0$ and Q has a nullspace of dimension $p > 0$, then if $m \ge p$ the limit is likely, but not certain to exist. If it exists it is of degree M. If $m < p$ divergence is likely, but not certain.*

The proof builds on the results for RBF interpolation in [24]. The key arguments and differences are pointed out in Appendix A.

The uncomplicated case (i) is of course the most common and the other types are deviations stemming from degenerate node point configurations or specific combinations of PDE problem parameters and node points that lead to degeneracy. Below, we give an example of each type of degeneracy for the two-dimensional Helmholtz problem given by (2) and (5)–(7) with $m = 1$.

**Example (ii): The node set is not polynomially unisolvent**
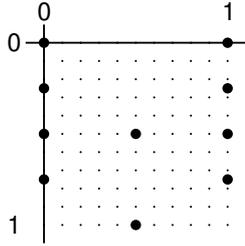
The $N = 10$ points are $(1/2, 1/2)$, $(1, 1/2)$, and $(k/4, 0)$, $(k/4, 1)$, $k = 0, \ldots, 3$. The matrix P has a nullspace defined by $n(\underline{x}) = x_2(x_2 - \frac{1}{2})(x_2 - 1)$. A non-degenerate basis is given by $\{1, x_1, x_2, x_1^2, x_1 x_2, x_2^2, x_1^3, x_1^2 x_2, x_1 x_2^2, x_1^4\}$ with $M = 4$. The limit is hence a polynomial of degree four whose coefficients depend on the choice of RBF. To illustrate what this dependence can look like, we give the general form of the limit polynomial $p(\underline{x})$.

$$
\begin{aligned}
p(\underline{x}) = \ & \beta_0 + \beta_1 x_2 + \beta_2 x_1 + \beta_3 x_2^2 + \beta_4 x_2 x_1 + \beta_5 x_1^2 \\
& + \beta_6 \left[-12 a_3 (x_2^3 + 3 x_2 x_1^2) + \frac{8 a_2^2}{a_1} (x_2^3 + x_2 x_1^2) \right] \\
& + \beta_7 \left[-12 a_3 (x_1^3 + 3 x_2^2 x_1) + \frac{8 a_2^2}{a_1} (x_1^3 + x_2^2 x_1) \right] \\
& + \beta_8 \left[-4 a_3 (5 x_1^3 + 3 x_2^2 x_1) + \frac{8 a_2^2}{a_1} (x_1^3 + x_2^2 x_1) \right] \\
& + \beta_9 \Big[-4 a_4 (9 x_2^3 + 36 x_2^2 x_1 + 45 x_2 x_1^2 + 20 x_1^3 - 24 x_2^3 x_1 - 40 x_2 x_1^3) \\
& \quad + \frac{6 a_2 a_3}{a_1} (3 x_2^3 + 4 x_2^2 x_1 + 3 x_2 x_1^2 + 4 x_1^3) - \frac{72 a_3^2}{a_2} (x_2^3 x_1 + x_2 x_1^3) \Big],
\end{aligned}
$$

where a_j are the Taylor expansion coefficients of the RBF, and β_j are the ten unknown coefficients that are determined by the ten discrete PDE collocation conditions. The result is κ-dependent as well as RBF-dependent.

Example (iii): The node set is not PDE-unisolvent
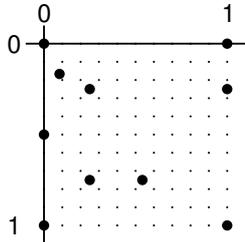
The $N = 10$ points are $(0, 0)$, $(1/2, 0)$, $(1, 0)$, $(0, 1)$, $(1/4, 1)$, $(1, 1)$, $(1/6, (2545 - 23\sqrt{9233})/3936)$, $(1/4, 1/4)$, $(3/4, 1/4)$, and $(3/4, 969/1804)$. For $\kappa = 4\sqrt{246}/9$ the matrix Q has a nullspace defined by $q(\underline{x}) = -\frac{5}{32} x_2(x_2 + 1) + \frac{x_1}{16}(8 - 24 x_1 + 3 x_2 + 16 x_1^2 + 4 x_1 x_2 - 7 x_2^2)$.

In this case, we get divergence of order ε^{-2} as $\varepsilon \to 0$ for all RBFs that obey conditions (I)–(III). This can be observed not only in exact arithmetic, but also in for example a double precision numerical simulation. However, if we move just one of the points or change κ slightly, there is no longer a nullspace. This kind of degeneracy is very rare, since it requires very special combinations of wavenumber and node points.

Example (iv): Both P and Q have a nullspace
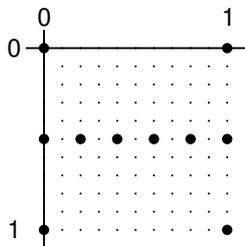
The $N = 10$ points are $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$ and $(1/2, k/5)$, $k = 0, \ldots, 5$. The matrices P and Q have a common nullspace of dimension two defined by $q_1(\underline{x}) = x_1(x_1 - \frac{1}{2})(x_1 - 1)$ and $q_2(\underline{x}) = x_2(x_1 - \frac{1}{2})(x_2 - 1)$. The limit exists and is an RBF- and κ-dependent polynomial of degree $M = 5$.

A practical implication of this result is that if we use node sets that are not unisolvent, e.g., Cartesian nodes, both PDE approximation and interpolation are expected to behave poorly in the small shape parameter regime. The condition number of the linear system is larger than in the unisolvent case, and the result contains a term that diverges as $\varepsilon \to 0$.

An important property of interpolation with Gaussian RBFs is that it never diverges [47]. In the PDE case, this property also holds as long as Q is non-singular. However, it can still be difficult to compute the limit numerically using a stable evaluation method. The RBF-QR method derived in [15, 13], and further explored for solving PDEs [26] uses pivoting to handle non-unisolvent cases. This means that a limit can always be computed, but it may be different from the Gaussian limit. The RBF-GA

method [14] always reproduces the correct Gaussian limit, but instead cannot handle large values of N. To compute limit approximations for other types of RBFs, the Contour-Padé method [16], see also the further developed version called the RBF-RA method [54], can be used. It works for any small shape parameter value, but is limited to intermediate values of N.

6 Convergence properties and error estimates

In this section, we investigate errors and convergence properties from different perspectives, as well as quantify how the choice of shape parameter affects the results. We start by formulating general residual-based error estimates in the following subsection.

6.1 General error estimates using Green's functions

We define the error function as the difference between the RBF approximant and the exact solution to the PDE problem (1)

$$e(\underline{x}) = s(\underline{x}) - u(\underline{x}). \tag{20}$$

In the interpolation case, the error and the residual are the same, and if the function $u(\underline{x})$ is known at a point, the corresponding error can be explicitly computed. In the PDE case, we can compute the residual for each operator, while the error is governed by the same type of PDE as the solution

$$\mathcal{L}^i e(\underline{x}) = \mathcal{L}^i s(\underline{x}) - f^i(\underline{x}) \equiv r^i(\underline{x}), \quad \underline{x} \in \Omega^i, \quad i = 1, \dots, N_{\text{op}}, \tag{21}$$

where r^i are residuals. One way to find the error is to solve the above PDE problem. However, because the residuals are zero at the collocation points, they are highly oscillatory and more node points are required to approximate the error than to solve the original PDE.

Instead of solving the error equation, we can formulate *a posteriori* error estimates in terms of the residual. Writing out the error PDE for the one-dimensional problem we get

$$\begin{cases} -\Delta e(x) - \kappa^2 e(x) &=& r(x), \\ -e'(0) - i\kappa e(0) &=& 0, \\ e'(1) - i\kappa e(1) &=& 0. \end{cases} \tag{22}$$

A Green's function satisfying the boundary conditions is given by

$$G(x, \xi) = \frac{i}{2\kappa} e^{i\kappa|x-\xi|}, \tag{23}$$

with

$$\frac{\partial G}{\partial \xi} = \begin{cases} \frac{1}{2} e^{i\kappa|x-\xi|}, & x \geq \xi, \\ -\frac{1}{2} e^{i\kappa|x-\xi|}, & x < \xi, \end{cases} \quad \text{and} \quad \Delta_\xi G = -\frac{i\kappa}{2} e^{i\kappa|x-\xi|} - \delta(x - \xi), \tag{24}$$

such that $-\Delta_\xi G - \kappa^2 G = \delta(x - \xi)$, which allows us express the error as

$$e(\xi) = \int_0^1 G(x, \xi) r(x) \, dx. \tag{25}$$

We can use this form of the error to formulate an error estimate as

$$\|e\|_\infty \leq \int_0^1 |G(x, \xi)| |r(x)| \, dx = \frac{1}{2\kappa} \int_0^1 |r(x)| dx. \tag{26}$$

For the two-dimensional Helmholtz problem in a rectangular domain, the corresponding Green's function satisfying the boundary conditions is given by

$$G(\underline{x}, \underline{\xi}) = \sum_{m=1}^\infty \frac{i}{2\beta_m} e^{i\beta_m|x_2-\xi_2|} \psi_m(x_1) \psi_m(\xi_1), \tag{27}$$

with $-\Delta_\xi G - \kappa^2 G = \delta(x_2 - \xi_2) \sum_{m=1}^\infty \psi_m(x_1) \psi_m(\xi_1)$, and with the vertical eigenfunctions $\psi_m(x_1) = \sqrt{2} \sin(m\pi x_1)$. Similarly as for the one-dimensional problem we define the error as

$$e(\underline{\xi}) = \int_0^1 \int_0^{L_1} G(\underline{x}, \underline{\xi}) r(\underline{x}) \, dx_1 \, dx_2. \tag{28}$$

To see how this works, we note that the vertical eigenfunctions form an orthonormal basis, and we can express the residual as

$$r(x_1, x_2) = \sum_{m=1}^\infty \langle r(\cdot, x_2), \psi_m^{x_2} \rangle \psi_m^{x_2} \equiv \sum_{m=1}^\infty r_m(x_2) \psi_m(x_1). \tag{29}$$

This allows us to simplify the error expression

$$\begin{aligned} e(\underline{\xi}) &=& \sum_{m=1}^\infty \frac{i}{2\beta_m} \int_0^1 e^{i\beta_m|x_2-\xi_2|} \int_0^{L_1} \psi_m(x_1) \psi_m(\xi_1) \sum_{n=1}^\infty r_n(x_2) \psi_n(x_1) \, dx_1 \, dx_2 \\ &=& \sum_{m=1}^\infty \frac{i}{2\beta_m} \psi_m(\xi_1) \int_0^1 e^{i\beta_m|x_2-\xi_2|} r_m(x_2) \, dx_2 \end{aligned} \tag{30}$$

To convert this into an error estimate, we first note that for $m \le \mu_0 = \lfloor \kappa L_1/\pi \rfloor$, the horizontal wavenumber β_m is real, and $|e^{i\beta_m |x_2 - \xi_2|}| = 1$, while for $m > \mu_0$, β_m is purely imaginary and $|\int_0^1 e^{i\beta_m |x_2 - \xi_2|} dx_2| \le |\int_0^1 e^{i\beta_m |x_2 - \frac{1}{2}|} dx_2| = \frac{2}{|\beta_m|}(1 - e^{-\frac{1}{2}|\beta_m|})$. Then we get

$$
\begin{aligned}
\|e\|_\infty &\le \sum_{m=1}^{\mu_0} \frac{1}{\sqrt{2}\beta_m} \int_0^1 |r_m(x_2)| \, dx_2 + \sum_{m=\mu_0+1}^{\infty} \frac{\sqrt{2}}{|\beta_m|^2}(1 - e^{-\frac{1}{2}|\beta_m|}) \int_0^1 |r_m(x_2)| \, dx_2 \\
&\le \sum_{m=1}^{\mu_0} \frac{1}{\sqrt{2}\beta_m} \|r_m\|_\infty + \sum_{m=\mu_0+1}^{\infty} \frac{\sqrt{2}}{|\beta_m|^2}(1 - e^{-\frac{1}{2}|\beta_m|})\|r_m\|_\infty.
\end{aligned} \tag{31}
$$

For the two-dimensional problem in a domain with curved boundaries, we cannot provide an explicit Green's function. The horizontal wavenumber and the modal expansion in this case varies with x_2. The residual in the error PDE should be understood as a source function that sends out waves in each direction from each point. For the rectangular duct, the projected residual in an infinitesimal slice, located at x_2 and with width dx_2, provides the contribution $\frac{r_m(x_2)}{\sqrt{2}\beta_m(x_2)} dx_2$ to the total error for a propagating mode. In the curved case, we use this quantity as a heuristic approximation of the local error contribution, and then integrate over the horizontal direction to get the global error approximation

$$
\begin{aligned}
\|e\|_\infty &\approx \sum_{m=1}^{\infty} \int_{X_m^{\Re e}} \frac{|r_m(x_2)|}{\sqrt{2}\beta_m(x_2)} \, dx_2 \\
&\quad + \sum_{m=1}^{\infty} \int_{X_m^{\Im m}} \frac{\sqrt{2}}{|\beta_m(x_2)|^2}(1 - e^{-\frac{1}{2}|\beta_m(x_2)|})|r_m(x_2)| \, dx_2,
\end{aligned} \tag{32}
$$

where $X_m^{\Re e} = \{x_2 \in [0,1] \mid \Re e(\beta_m(x_2)) > 0\}$ and $X_m^{\Im m} = \{x_2 \in [0,1] \mid \Im m(\beta_m(x_2)) > 0\}$. We evaluate this error approximation numerically in Section 7 and find that we get surprisingly good results.

6.2 Convergence properties for small ε

As discussed in the previous section, we approach the polynomial limit, $s(\underline{x}) = p(\underline{x})$, as $\varepsilon \to 0$. For polynomial *interpolation* in one dimension, the interpolation error $e_I(x)$ takes the form

$$
e_I(x) = u(x) - p(x) = \frac{\prod_{j=1}^N (x - x_j)}{N!} u^{(N)}(\xi),
$$

where $\xi \in (x_1, x_N)$. For equispaced points, $x_{j+1} - x_j = h$, this can be estimated by

$$
|e_I(x)| \le \frac{h^N}{N} \max_{\xi \in (x_1, x_N)} |u^{(N)}(\xi)|,
$$

see [18, pp. 39–40]. In the PDE case, the residual plays the role of the error. By following the steps for the proof of the polynomial error [18, pp. 43–44], we can get a similar estimate for the residual.

Theorem 6.1. *For a one-dimensional linear PDE problem*

$$
\begin{cases}
\mathcal{L}^1 u(x) = f^1(x), & x_1 < x < x_N, \\
\mathcal{L}^2 u(x) = f^2(x), & x = x_1, \\
\mathcal{L}^3 u(x) = f^3(x), & x = x_N,
\end{cases}
$$

with a polynomial solution $p(x)$ determined through collocation at the nodes x_i, $i = 1, \dots, N$ the residual $r(x) = \mathcal{L}^1 p(x) - f(x)$ has the form

$$
r(x) = \frac{\prod_{j=2}^{N-1}(x - x_j)}{(N-2)!} r^{(N-2)}(\xi),
$$

where $\xi \in (x_1, x_N)$. For equispaced points, $x_{j+1} - x_j = h$, this can be estimated by

$$
|r(x)| \le h^{N-2} \max_{\xi \in (x_1, x_N)} |r^{(N-2)}(\xi)| \begin{cases} 1 & x < x_1 \text{ or } x > x_{N-1}, \\ 1/(N-2) & x_1 \le x \le x_{N-1}. \end{cases}
$$

Proof. Let $\Psi(s) = r(s) - \frac{r(x)}{\chi(x)}\chi(s)$, where $\chi(x) = \prod_{j=2}^{N-2}(x - x_j)$. Then $\Psi(x) = 0$ and $\Psi(x_j) = 0$, $j = 2, \dots, N-1$, since $r(x_j) = 0$ at all interior node points where the equation is enforced. This means that $\Psi(s)$ has at least $N - 1$ zeros. By repeated application of Rolle's theorem, we find that $\Psi^{(N-2)}(s)$ has at least one zero. That is,

$$
0 = \Psi^{(N-2)}(\xi) = r^{(N-2)}(\xi) - \frac{r(x)}{\chi(x)}(N-2)!.
$$

Rearranging gives the expression for $r(x)$.

For the interior part of the interval $x_1 \le x \le x_{N-1}$ the bound on the residual follows from the interpolation bound. For the end intervals, the absolute value of the product term is largest at the end points, and is given by $\prod_{i=2}^{N-1}(i-1)h = h^{N-2}(N-2)!$, providing the second part of the bound. $\qquad\square$

We note that $N - 2 = h^{-1} - 1$ and use the residual bound in the integral over the absolute value of the residual to get

$$\int_0^1 |r(x)|dx \le h^{N-2}\left(h + \frac{1-2h}{h^{-1}-1} + h\right)\|r^{(N-2)}\|_\infty \le 3h^{N-1}\|r^{(N-2)}\|_\infty. \tag{33}$$

This bound is not tight. By numerically integrating $\prod_{j=2}^{N-1}(x - x_j)/(N-2)!$ for different values of N, we find that the result is always smaller than h^{N-1}, and that for $N \ge 10$ it is smaller than $0.5h^{N-1}$.

To see how this can help us in understanding the behavior of the error for small ε, we insert the numerical estimate of the integral, h^{N-1}, in the error estimate (26) for the one-dimensional Helmholtz problem to get

$$\|e\|_\infty \le \frac{h^{N-1}}{2\kappa}\|r^{(N-2)}\|_\infty. \tag{34}$$

In the flat limit, the residual is $r(x) = -p''(x) - \kappa^2 p(x)$, where $p(x)$ is the limit polynomial of degree $N - 1$. Then $r^{(N-2)}(x) = -\kappa^2 p^{(N-2)}(x)$. We know that $p(x) \approx u(x) = \exp(i\kappa x)$, but even if $p(x)$ is a very good approximation of $u(x)$, $p^{(N-2)}(x)$ (which is a line) is a rather poor approximation of $u^{(N-2)}(x)$. However, numerical tests indicate that the order of magnitude is right. That is, $|p^{(N-2)}| \approx |\frac{d^{N-2}\exp(i\kappa x)}{dx^{N-2}}| = \kappa^{N-2}$. We cannot use this as a bound, but we get an approximate expression for the error in the limit

$$\|e\|_\infty \approx \frac{h^{N-1}}{2\kappa}\kappa^2\kappa^{N-2} = \frac{1}{2}(\kappa h)^{N-1}. \tag{35}$$

Note that the quantity κh is small only if the problem is adequately resolved.

For the two-dimensional problem, the limit polynomial has degree K if $N_{K-1,d} < N \le N_{K,d}$ and it is zero at the interior node points. To get an estimate for the residual in terms of its derivatives, we could potentially use a sampling inequality such as [35, Theorem 3.5], which says that for all $h \le h_0$

$$\|r\|_\infty \le C_k h^k \sum_{|\sigma|=k} \|D^\sigma r\|_\infty, \tag{36}$$

where h_0 depends on the geometry of Ω. For the unit square, which we are using here, $h_0 = \frac{1}{2kc_2}$ with $c_2 = 12$. In the discretizations that we use $h = 1/(\sqrt{N} - 1)$. Requiring $h \le h_0$ leads to the following condition $k \le (\sqrt{N} - 1)/24$. We want to use the theorem for $k = K - 1$, where inverting the expression for $N_{K,d}$ yields that $K = \lfloor \sqrt{2}\sqrt{N + 1/8} - 1.5 \rfloor$. That is, the theorem does not hold in this case. From practical experience, the result holds also for larger k (larger h), and we will therefore use it to approximate the residual.

In this case, using that $r(\underline{x}) = -\Delta p(\underline{x}) - \kappa^2 p(\underline{x})$, and, for $|\sigma| = K - 1$, $D^\sigma r(\underline{x}) = -\kappa^2 D^\sigma p(\underline{x}) \approx -\kappa^2 D^\sigma u(\underline{x})$, we get

$$\|r\|_\infty \le C_{K-1}h^{K-1}\kappa^2 \sum_{|\sigma|=K-1} \beta_1^{\sigma_2}\alpha_1^{\sigma_1} \le C_{K-1}\kappa^2 K(\kappa h)^{K-1},$$

Combining the approximate expression for the residual with the error estimate (31) restricted to the first mode (scaled by $1/\sqrt{2}$) gives

$$\|e\|_\infty \approx \frac{1}{2|\beta_1|}C_{K-1}\kappa^2 K(\kappa h)^{K-1}$$

Numerical experiments show that $KC_{K-1} = C/(K-1)$, provides the appropriate behavior with respect to N (both K and h are coupled with N). This leads to

$$\|e\|_\infty \approx \frac{C\kappa^2(\kappa h)^{K-1}}{2|\beta_1|(K-1)} \approx \tilde{C}(\kappa h)^K,$$

where the final expression is just to show that the dimension is similar to that of the one-dimensional error approximation.
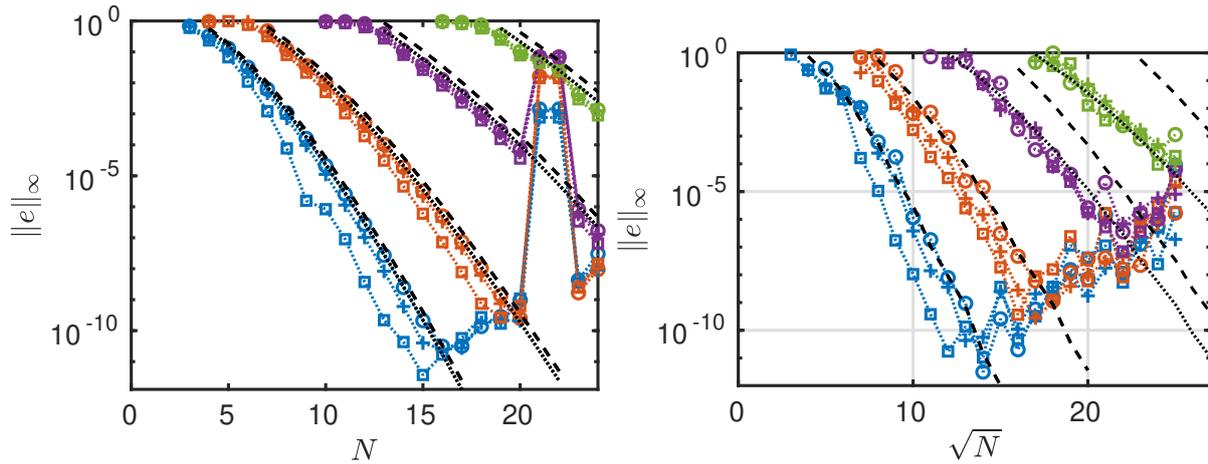
Figure 3 shows the computed errors of the one-dimensional and two-dimensional problems for small values of the shape parameter. The error behavior agrees well with the derived error approximations. For the two-dimensional problem, we also show that the error expression can be multiplied by a constant to get a very good fit to the actual error. This means that we can use $\|e\|_\infty \approx C(\kappa h)^K$ *a priori* with $C = 1$ to determine the necessary resolution for a given tolerance. Given at least two numerical solutions, we can also estimate the constant C. The error approximation is most likely to be valid for problems that are almost rectangular or with mildly varying coefficients, but only for small shape parameter values.

6.3 Convergence properties for larger ε

As shown in [48, 50, 27], the convergence of a PDE approximation can be expressed in terms of the approximation properties of the interpolant (consistency error) and a stability term. The consistency error of the PDE operator can for example be expressed as

$$\mathcal{E}_\mathcal{L} = \mathcal{L}(I_h(u) - u),$$

where $I_h(u)$ interpolates u using a node set with fill distance h. Several authors have derived exponentially converging error results for RBF interpolation [43, 36, 34, 3, 56, 44]. The first papers are focused on interpolation errors, while [44] also provides

Figure 3: The computed errors using the Gaussian RBF together with the RBF-QR method for $\varepsilon = 0.5$ ($\square$), $\varepsilon = 0.25$ (+), and $\varepsilon = 0.01$ ($\circ$) for $\kappa = \pi, 2\pi, 4\pi, 6\pi$, from left to right, together with the approximation $\|e\|_\infty \approx \frac{1}{2}(\kappa h)^{N-1}$ for the one-dimensional problem (left), and for $\kappa = 1.2\pi, 2.4\pi, 4.8\pi, 7.2\pi$, from left to right, together with the approximation $\|e\|_\infty \approx (\kappa h)^K$ for the two-dimensional problem (right) (dashed curves). In the one-dimensional case, we also show the error approximation using the numerically computed integral of the product term (dotted curves), and in the two-dimensional case for $\kappa = 4.8\pi$ and 7.2π, we also show the error approximation using $C = 1/40$ and $C = 1/800$, respectively (dotted curves).

estimates for derivatives of functions with many zeros, such as the interpolation error. We use the optimality property of RBF interpolants in the native space (reproducing kernel Hilbert space) [9]

$$\|I_h(u)\|_{\mathcal{N}(\Omega)} \le \|u\|_{\mathcal{N}(\Omega)},$$

to replace the interpolation error norm with the function norm, since $\|I_h(u) - u\|_{\mathcal{N}(\Omega)} = \|\mathcal{E}_{\mathcal{I}}\|_{\mathcal{N}(\Omega)} \le 2\|u\|_{\mathcal{N}(\Omega)}$. We get the following estimates for RBF interpolants in compact cube domains using [44, Corollary 5.1] for Gaussians

$$\|\mathcal{E}_{\mathcal{I}}\|_\infty \le e^{C_G \log(h)/h}\|\mathcal{E}_{\mathcal{I}}\|_{\mathcal{N}_{\mathcal{G}}(\Omega)} \le 2e^{C_G \log(h)/h}\|u\|_{\mathcal{N}_{\mathcal{G}}(\Omega)},$$

where $C_G > 0$, and for inverse multiquadrics

$$\|\mathcal{E}_{\mathcal{I}}\|_\infty \le e^{-C_Q/h}\|\mathcal{E}_{\mathcal{I}}\|_{\mathcal{N}_{\mathcal{Q}}(\Omega)} \le 2e^{-C_Q/h}\|u\|_{\mathcal{N}_{\mathcal{Q}}(\Omega)},$$

when $h \le h_0$, and with $C_Q > 0$. The constants C_G and C_Q depend on the number of dimensions d and properties of the domain Ω.

The results for derivatives of the interpolation error are given for Lipschitz domains, which are more general than compact cubes, but the results are instead weaker in terms of the convergence rate. From [44, Theorem 3.5], we get

$$\|\mathcal{E}_{\mathcal{L}}\|_\infty \le 2e^{\tilde{C}_G \log(h)/\sqrt{h}}\|u\|_{\mathcal{N}_{\mathcal{G}}(\Omega)},$$

$$\|\mathcal{E}_{\mathcal{L}}\|_\infty \le 2e^{-\tilde{C}_Q/\sqrt{h}}\|u\|_{\mathcal{N}_{\mathcal{Q}}(\Omega)},$$

for Gaussians and inverse multiquadrics respectively. The higher rate of Gaussian RBFs is related to the behavior of embedding constants for the native space in relation to Sobolev spaces of increasing order. The constants $\tilde{C}_G$ and $\tilde{C}_Q$ depend on properties of the domain Ω, and $\tilde{C}_Q$ also depends on $\mathcal{L}$ and d. In [45], it is shown that the better convergence rates are obtained also for derivatives of the interpolation error if the nodes are clustered in a layer close to the boundary.

In order to investigate numerically what the actual behavior of the error is for the Helmholtz problem, we solve the one-dimensional problem for a range of shape parameter values and different numbers of node points. In this test, we have used multiquadric RBFs. We assume that the error for multiquadric RBFs has the form

$$\|e\|_\infty = A_M \exp(-C_M f(h)),$$

where $C_M > 0$, $f(h) = 1/h$ or $f(h) = 1/\sqrt{h}$, and the native space norm has been absorbed into the constant. Then a plot of the logarithm of the error against $f(h)$ should result in a straight line. From Figure 4, it is clear that $f(h) = 1/h$ is a better fit. The dashed lines correspond to a fit of the model with $f(h) = 1/h$ to the actual errors, where the results suffering from ill-conditioning effects have been ignored.

Figure 5 shows the fitted model parameters A_M and C_M for different shape parameter values. The different curves correspond to different wavenumbers, and it should be noted that the exponential rate C_M becomes independent of the wavenumber when $\varepsilon \gtrsim 0.5$. The rate C_M also decreases with increasing shape parameter values. The optimal rate is attained for a small positive shape parameter value, and for even smaller shape parameters, the asymptotic (polynomial) rate is dominating. The coefficient A_M instead seems to be largest where the rate is optimal, and smallest where the rate is lowest, which makes it harder to determine the best shape parameter value. We discuss this further in the following subsection.
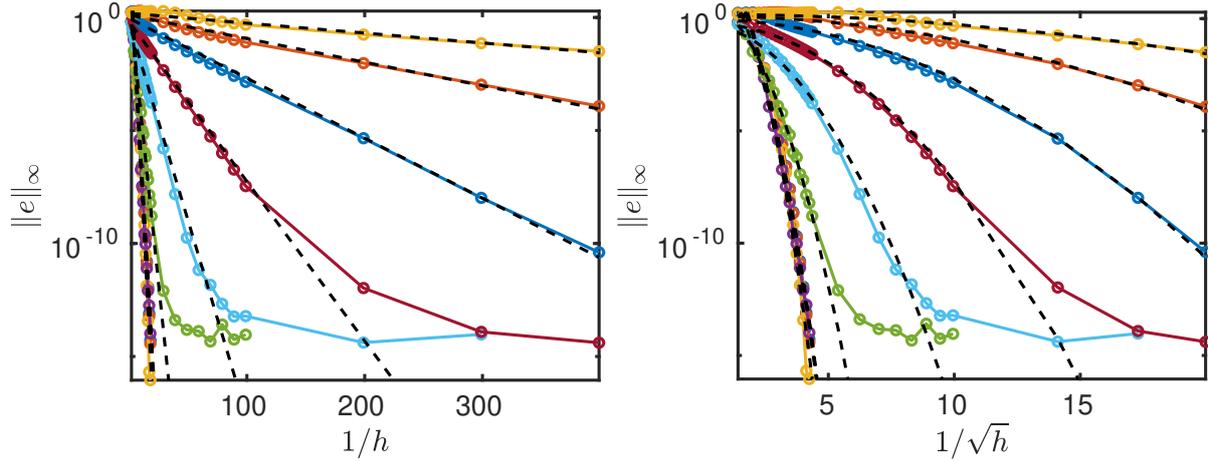
Figure 4: The error in the one-dimensional Helmholtz solution when multiquadric RBFs are used as a function of $1/h$ (left) and $1/\sqrt{h}$ (right) for shape parameters $\varepsilon = 10^{-2+\frac{4}{9}q}$, $q = 1, \dots, 9$ (left to right). The dashed black lines/curves correspond to a fit of $\|e\|_\infty = A_M \exp(-C_M/h)$ to the error data (in both cases). The computations were performed in quad precision and the Contour-Padé method [16] was used for the three smallest shape parameters.
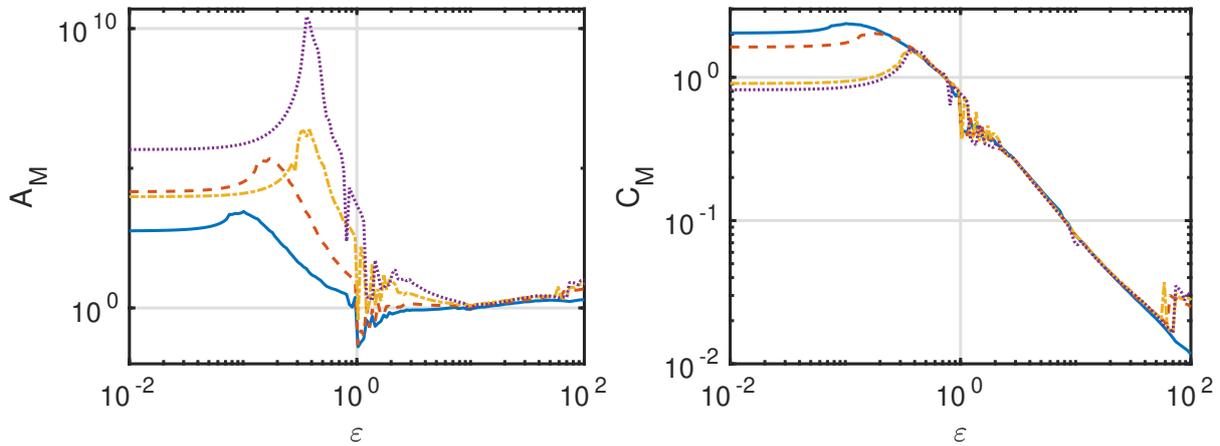
Figure 5: The result of fitting the model parameters A_M and C_M to the computed errors for the one-dimensional Helmholtz problem using multiquadric RBFs and different values of the shape parameter ε, and for $\kappa = \pi$ (solid line), $\kappa = 2\pi$ (dashed line), $\kappa = 4\pi$ (dash-dot line), and $\kappa = 6\pi$ (dotted line).

6.4 Convergence as a function of the shape parameter

Dependence on the shape parameter is not discussed in [44], and the results reported in the previous subsection hold for a fixed value of ε. However, using a shape parameter $\varepsilon_0 \neq 1$ for an interpolation problem defined in the domain Ω with fill distance h is equivalent to using a shape parameter $\varepsilon_1 = 1$ for a problem in the scaled domain $\varepsilon_0 \Omega$ with fill distance $\varepsilon_0 h$. This can be understood by noting that $\phi(\varepsilon_0 \|\underline{x}_i - \underline{x}_j\|) = \phi(1 \cdot \|\varepsilon_0 \underline{x}_i - \varepsilon_0 \underline{x}_j\|)$. Hence, the native space norm is the same in both cases, and the errors are the same in both cases.

If we let the constants A_M and C_M in the error estimate for a specific domain Ω and shape parameter ε be denoted by $A_M(\Omega, \varepsilon)$ and $C_M(\Omega, \varepsilon)$, this means that

$$A_M(\Omega, \varepsilon) e^{-C_M(\Omega, \varepsilon)/h} = A_M(\varepsilon\Omega, 1) e^{-C_M(\varepsilon\Omega, 1)/(\varepsilon h)}. \tag{37}$$

That is, the convergence rate in terms of h for a fixed value of ε is larger when ε is small. This can also be seen in Figure 5, where the slope in the logarithmic plot of C_M against ε is approximately equal to -1. It should be stressed that this does not hold in the flat limit regime, only for $\varepsilon \gtrsim 0.5$ (in our case). This also corresponds to the theoretical result given in [34], where an explicit constraint on the smallest shape parameter for which the results hold is given as $\varepsilon \geq 1/D$, where for a cube domain, D is the side. This coincides well with the numerical results. However, there is also an upper bound $\varepsilon \leq 1$, which is harder to reconcile with what we observe.

Figures 6 and 7 show the error as a function of ε for two one-dimensional problems, and one two-dimensional problem, respectively. The error curves represent a common behavior for smooth solution functions. Starting from a large shape parameter and moving towards smaller values, the error first decreases rapidly then reaches an optimal region, and finally levels out at the polynomial approximation error, see [24] for a more detailed discussion about the error curve and the optimal shape parameter.

Due to the conditioning problems for decreasing values of ε and increasing values of N, a common approach in the literature is to scale the shape parameter such that, e.g., $\varepsilon h = C$, which is called stationary interpolation. A problem is that stationary interpolation does not converge as h goes to zero. This can be understood by again looking at analogous problems. If we start from a problem on the domain Ω with shape parameter ε and fill distance h, and we refine to get fill distance h/q and shape parameter $q\varepsilon$, then the equivalent problem is $(q\Omega, \varepsilon, h)$. That is, the refinement corresponds to stretching out the domain, while keeping the fill distance and shape parameter constant. This makes the apparent solution function become increasingly smooth, and approaching a constant. Since constants are only reproduced for $\varepsilon = 0$ for the commonly used infinitely smooth RBFs, there is no convergence for a fixed non-zero ε. By augmenting the RBF approximation with polynomial terms, convergence corresponding to the polynomial order can be recovered also in the stationary case [11].

The convergence curves when choosing the shape parameter as $\varepsilon = Ch^\beta$ for different exponents β are shown as dashed lines in Figures 6 and 7. As expected, the stationary choice, $\beta = -1$ levels out as N increases. For $\beta > -1$ we get convergence along different paths. The choice $\beta = 0$ corresponds to the exponential convergence case for fixed shape parameter values. For these Helmholtz problems, the curve with $\varepsilon = Ch^{3/2}$ captures the optimal shape parameter values well. For other types of problems, the relation would look different.
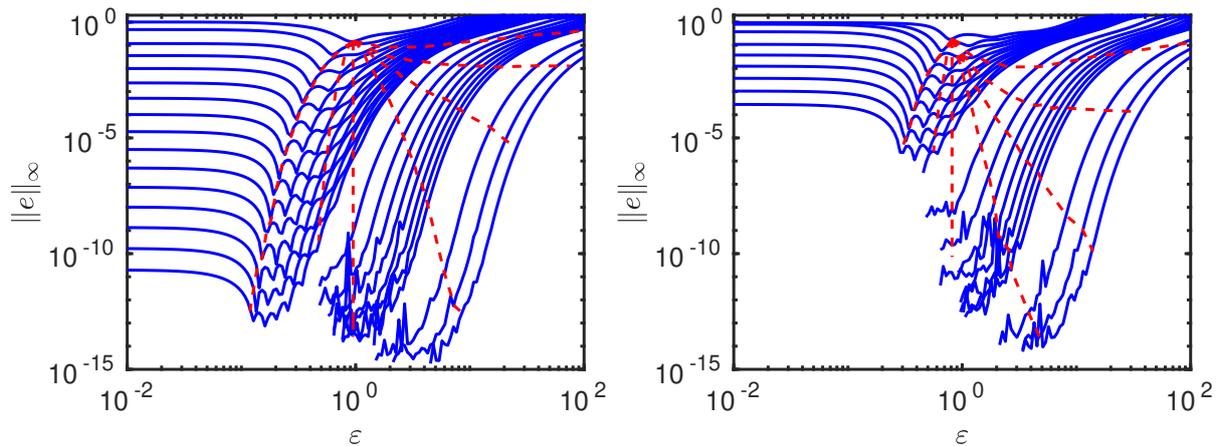
Figure 6: The maximum error as a function of ε for $\kappa = 2\pi$ (left) and $\kappa = 4\pi$ (right) using multiquadric RBFs. The number of node points is from top to bottom $N = 6, 7, \ldots, 21, 30, 40, \ldots, 100, 200, 300, 400$ in the left subfigure, and $N = 10, 11, \ldots, 20, 30, \ldots, 100, 200, 300, 400$ in the right subfigure. The dashed lines show how the error curves are traversed if the shape parameter is chosen as $\varepsilon = Ch^\beta$, with $\beta = \frac{3}{2}, \frac{1}{2}, 0, -\frac{1}{2}, -\frac{3}{4}, -1, -\frac{3}{2}$ from left to right. The computations were performed in quad precision and the Contour-Padé method [16] was used for small shape parameter values.

For the two-dimensional problem, the curves are more irregular due to several interacting terms in the error [24]. However, the overall behavior for the different ways to choose the shape parameter is very similar to the one-dimensional case.

Assuming that $C_M(\varepsilon\Omega, 1)$ in (37) does not vary strongly with ε, something that can be verified by noting that the slope the line in Figure 5 for C_M is approximately equal to -1, we can finally provide a convergence rate for the scaled ε convergence case. If we have exponential convergence as $1/\varepsilon h$ and $\varepsilon = Ch^\beta$ we end up with

$$\|e\|_\infty = A_M^\varepsilon e^{-C_M^\varepsilon / h^{\beta+1}}, \quad -1 < \beta \leq 0, \tag{38}$$

where $C_M^\varepsilon > 0$ and the superscript indicates the potential ε-dependence. If $\beta > 0$, the convergence curves may enter the polynomial region, and we cannot in general get increasing convergence rates for increasing β. The validity of this expression is further investigated numerically in Section 7.

7 Numerical experiments

In this section, we focus on the third test problem with curved boundaries, see Figure 1. We look at how to choose the method parameters and how we can use the theoretical estimates to interpret the results. Unless otherwise mentioned, the problem parameters are given by wavenumber $\kappa = 6\pi$, source location $x_s = 0.3$, and boundary curves

$$\begin{aligned} \gamma_1 &= 0.3\exp(-20(x_2 - 0.5)^2), \\ \gamma_2 &= 0.8 - 0.3\big(\exp(-80(x_2 - 0.3)^2) + \exp(-80(x_2 - 0.7)^2)\big). \end{aligned}$$

For global RBF approximations and shape parameters that are not in the flat limit a uniform node spacing is in general recommended [41]. However, when the problem size is large enough, there can instead be problems at the boundaries unless the nodes are clustered towards the boundaries [42, 13]. In our experiments, we do not reach the regime where this is an issue. Therefore, we use quasi uniform nodes. The nodes are constructed from the input parameters n_1 and n_2, that specify the number of nodes in the vertical direction at the left boundary, and the number of nodes in the horizontal direction. We define the step sizes $h_1 = 0.8/(n_1 - 1)$ and $h_2 = 1/(n_2 - 1)$. Based on these step sizes, the nodes are then placed uniformly along vertical lines with as similar node distance as possible. The nodes at the top and bottom boundaries are placed with uniform arc length. If the
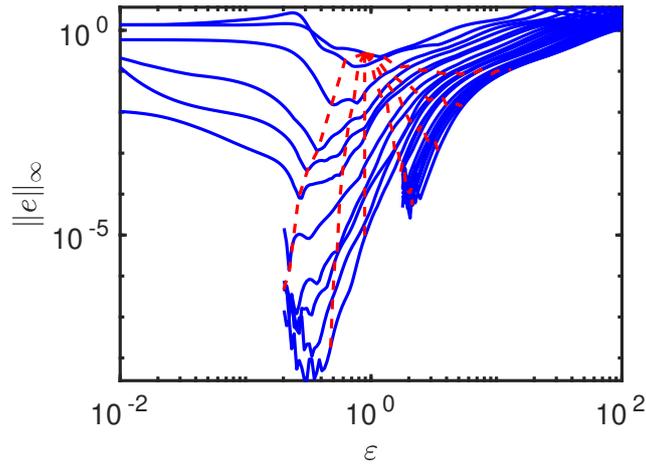
Figure 7: The maximum error as a function of ε for $\kappa = 2.2\pi$ for the two-dimensional problem using multiquadric RBFs. The number of node points is from top to bottom $N \approx n^2$, for $n = 3, \ldots, 25$. The dashed lines show how the error curves are traversed if the shape parameter is chosen as $\varepsilon = Ch^\beta$, with $\beta = \frac{3}{2}, \frac{1}{2}, 0, -\frac{1}{2}, -\frac{3}{4}, -1, -\frac{3}{2}$ from left to right. The computations were performed in quad precision and the Contour-Padé method [16] was used for small shape parameter values.
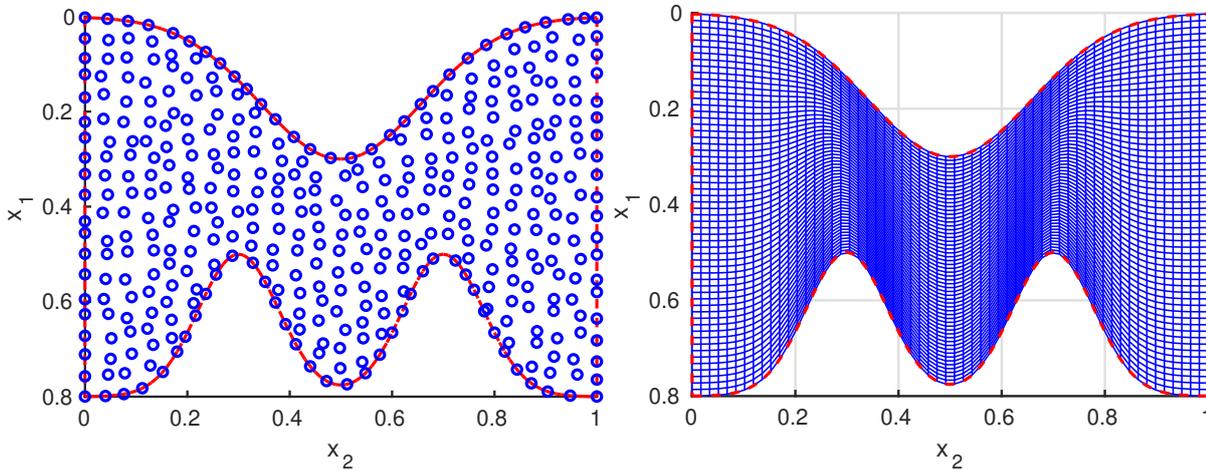
Figure 8: Node points with $n_1 = 20$ and $n_2 = 25$ (left) and the evaluation grid with 60×60 points used for the convergence experiments (right).

nodes are too regular, they are not unisolvent, and the conditioning gets higher at least for shape parameters that are small [24]. Therefore, we add a random perturbation to each node. In all experiments performed here, the size of the random perturbation is $0.25(h_1, h_2)$ for the interior nodes, while boundary nodes are only perturbed along the boundary. The solution, residual, and errors are evaluated on a grid. An example of both nodes and evaluation grid is given in Figure 8. The resulting numbers of node points for the grids we have used in the experiments are shown in Table 1.

Errors are measured against a reference solution computed using the largest node set with $n_1 \times n_2 = 100 \times 125$. This is the largest problem size that fits in the memory of the Dell Latitude E6230 laptop with an i5-3360 dual core CPU running at 2.8 GHz that was used for the experiments. When we refer to the maximum norm of the numerical errors or the solution, we evaluate them on the 60×60 evaluation grid, except for the solutions with higher wavenumbers, where we use 100×100 grid points. We use multiquadric RBFs in all numerical experiments. No stable methods are used here. The MATLAB implementations of the solvers that were used in the experiments are available at the first author's software page.

7.1 Selecting a tolerance for constructing the DtN boundary conditions

As mentioned in Section 3.1, we need to compute N inner products with each vertical eigenmode ψ_m present in the problem at the two vertical boundaries. Accurate numerical computation of these integrals is a significant computational cost, e.g., up to $n_f = 1700$ function evaluations per integral are needed for tolerance $1e-15$. The question is which tolerance to choose.

The sensitivity of the problem (ill-conditioning) depends strongly on the shape parameter ε with an exponentially increasing condition number as the shape parameter goes to zero. By using a stable evaluation method such as RBF-QR for Gaussian RBFs, the sensitivity is removed and the tolerance for the integrals does not need to be smaller than the desired error in the solution. However, for the test problem considered here, too small values of ε, leading to a global polynomial approximation is not an appropriate choice, and we are not able to use RBF-QR.

Table 1: The size N of the different node sets that are used in the experiments. The parameters n_1 and n_2 are chosen to make h_1 and h_2 as equal as possible.

$n_1 \times n_2$	10×12	11×14	12×15	13×16	14×17	15×19	16×20
N	104	131	152	174	194	235	261

$n_1 \times n_2$	17×21	18×22	19×24	20×25	22×27	24×30	26×32
N	287	317	362	396	462	563	639

$n_1 \times n_2$	28×35	30×37	32×40	34×42	36×45	38×47	40×50
N	747	844	971	1079	1219	1341	1493

$n_1 \times n_2$	50×62	60×75	70×87	80×100	90×112	100×125	
N	2294	3306	4434	5813	7300	9029	

Table 2 shows the average number of function evaluations needed by MATLAB's `quadl` to approximate one integral to a prescribed absolute tolerance for different values of the shape parameter ε for a node set with $n_1 \times n_2 = 30 \times 38$. The bold faced entries in the table show the largest tolerance that can be used before the approximation changes significantly. The tolerance is much smaller than the absolute error in the solution, which is about 0.5 compared with the reference solution. The condition numbers computed by MATLAB are between $1 \cdot 10^{17}$ for $\varepsilon = 5$ and $1 \cdot 10^{11}$ for $\varepsilon = 12$. For larger N, the ill-conditioning also increases, so we expect that even smaller tolerances are needed in this case.

Table 2: The average number of function evaluations for approximating one integral of the type in (9) using `quadl` for multiquadric RBFs. Bold faced numbers show the largest tolerance that does not significantly alter the result. The relative error against the reference solution is also given. A × indicates that the approximation had an error of the same order as the size of the solution.

| Tolerance	1e−4		1e−6		1e−8		1e−10	
$\varepsilon = 5$	33	×	52	×	97	×	**164**	2.5e−1
$\varepsilon = 6$	34	×	54	×	**102**	1.4e−1	168	1.4e−1
$\varepsilon = 7$	35	×	56	6.3e−1	**106**	6.1e−2	173	5.9e−2
$\varepsilon = 8$	36	×	58	6.6e−2	**109**	5.2e−2	180	5.2e−2
$\varepsilon = 9$	36	×	**59**	6.0e−2	112	5.0e−2	187	5.0e−2
$\varepsilon = 10$	36	×	**60**	5.0e−2	114	5.0e−2	193	5.0e−2
$\varepsilon = 11$	37	3.5e−1	**62**	5.1e−2	115	5.1e−2	198	5.1e−2
$\varepsilon = 12$	**37**	7.9e−2	63	5.2e−2	117	5.2e−2	202	5.2e−2

7.2   Choosing the starting value for the shape parameter

To solve a large scale problem efficiently it pays off to choose the shape parameter carefully, since it does not affect the cost, only the accuracy. As was discussed in Section 6.4, a practical way to achieve convergence in spite of the ill-conditioning is to choose the shape parameter as $\varepsilon = Ch^\beta$, with $\beta > -1$. We are going to use $\beta = -1/2$, which provides a trade-off between convergence rate and conditioning problems. Then we need to decide which C to use.

Compared with the full solution, it is not so expensive to solve a much less resolved problem a few times for different shape parameters. A related approach was used for estimating the shape parameter in [6]. We want to test if the residual-based error approximation (32) can help us find the best shape parameter value for such a problem, and from there the C to use. We also try the ℓ_2-norm of the residual as an indicator, since the residual should be small when the error is small. The maximum norm of the residual was also tested, but did not correlate strongly with the error. Figure 9 shows the relative error approximation as well as the relative ℓ_2-norm of the residual together with the actual relative error against the reference solution. In the first example, the shape parameter values corresponding to the smallest error approximation, ε^{app}, and the smallest residual norm, ε^{res}, are both close to the actual minimum ε^*. In the second example, the minimum for the error approximation is a bit higher than the true value. Note that in the left subfigure, for large shape parameters, the error approximation underestimates the actual error. In our experiments, this only happened for small N and large actual errors.

Table 3 gives the minimal shape parameter values for ten different (small to medium) problem sizes. In most of the cases the error approximation, the residual estimate, or both are close to the true value. We have also computed the C-values corresponding to the average of the two estimates. If we had solved only the first problem, we would have chosen $C = 1.5$. This is what we have used for the convergence experiments in the following subsection. We also tried $C = 1$, but then the ill-conditioning prevented us from solving the largest problems.

An alternative method to find a good shape parameter value is to use the leave-one-out cross validation method. It was first introduced for RBF interpolation methods [46], and a cost effective version of the method was derived in [57]. It was suggested to use LOOCV for PDE problems using the residual as error indicator in [5], and this was implemented in [10]. We tried to use residual-based LOOCV on the Helmholtz problems in this paper, but the preliminary results were not close enough to the optimal values, and we therefore decided to use the error approximation instead.
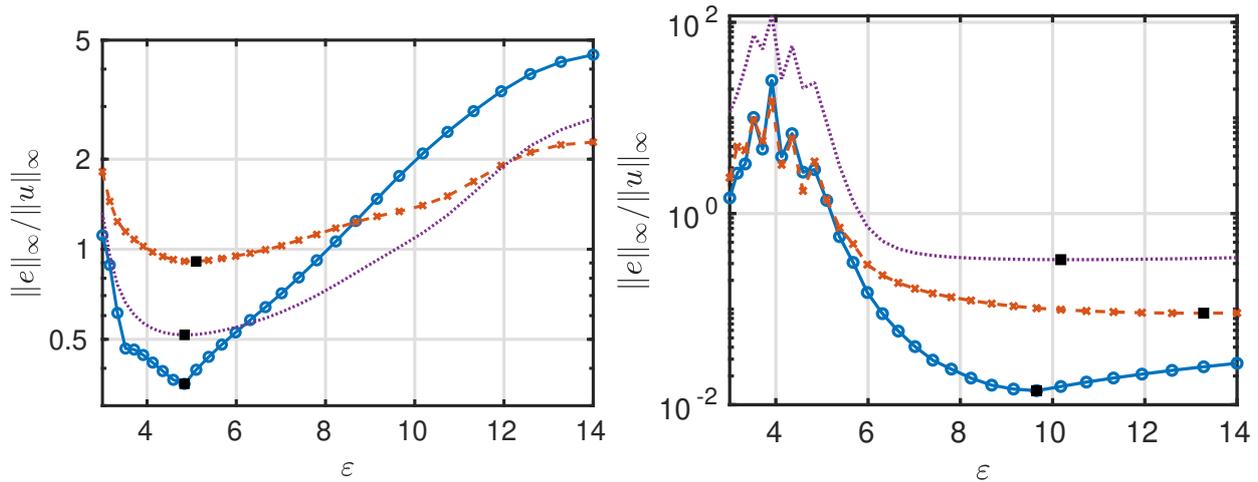
Figure 9: The error approximation (32) (×), the ℓ_2-norm of the residual (dotted line) and the error against a highly resolved reference solution (○) for the 10×12 (left) and 40×50 (right) node sets. The minima are indicated by black squares.

Table 3: The optimal shape parameter for the error against the reference solution ε^*, the error approximation ε^{app}, and the ℓ_2-norm of the residual ε^{res} and the constant $\tilde{C}$ implied by the average of the two estimates for different problem sizes.

$n_1 \times n_2$	10×12	11×14	12×15	13×16	14×17
ε^*	4.8	4.6	5.7	4.6	7.0
ε^{app}	5.1	6.7	8.7	6.7	3.2
ε^{res}	4.8	5.4	7.4	5.4	3.7
$\tilde{C}$	1.5	1.7	2.2	1.6	0.9
$n_1 \times n_2$	15×19	16×20	20×25	30×37	40×50
ε^*	7.4	4.8	9.7	9.2	9.7
ε^{app}	7.8	6.0	9.2	13.3	13.3
ε^{res}	6.3	5.4	7.0	9.7	10.2
$\tilde{C}$	1.7	1.3	1.7	1.9	1.7

7.3 Convergence experiments

Here, we use the relation $\varepsilon = C/\sqrt{h} = 1.5/\sqrt{h}$ to run a convergence experiment. We solve the test problem for different problem sizes and compute the error approximation and the error against the reference solution. According to equation (38), with this choice of shape parameter scaling, the error should be of the form

$$\|e\|_\infty = A_M \exp(-C_M/\sqrt{h}).$$

In Figure 10, we plot the relative error and the relative error approximation (32) against $1/\sqrt{h}$. A line has been fitted to the data set, and it is clear from the picture that it is a good fit of the convergence trend. The slopes C_M are 0.78 for the error and 0.75 for the error approximation, which means that the error approximation gives very good results for the ratio of errors at different resolutions, even if the constant is not precise. The constant A_M is 3.0 times larger for the error approximation than for the error. Based on the curves in Figure 9, we expect A_M to be problem and/or parameter dependent.

If we compare the error reduction from the smallest to the largest problem size with what we would get with an algebraically converging method where the error is $\mathcal{O}(h^p)$, a reduction in error with a factor 242 for a step size reduction of 10 corresponds to $p = 2.4$. That is, even if we have exponential convergence, the overall error reduction is not that impressive. However, the small numbers of points we can use, while still getting reasonable results are impressive. The smallest problem has 12 points in the horizontal direction, which corresponds to 4 points per wavelength. A rule of thumb for a finite difference method is that at least 15 points per wavelength, that is 45 for this problem, are needed for geometric resolution. For the largest problems, the tolerance for the quadrature had to be lowered. The small perturbations introduced by the inexact quadrature with tolerance $1 \cdot 10^{-10}$ are enough to prevent the convergence curve from following the straight line, and the convergence rate then seems to decrease. These experiments were run using tolerance $1 \cdot 10^{-14}$.

In the right subfigure of Figure 10, the ℓ_2-norm of the residual is plotted together with the same relative error results. Even though the residual norm gives reasonable estimates for the optimal shape parameter, it is clear that we cannot use it to follow the error trend.

7.4 Experiments with larger wavenumbers

We have also solved problems with larger wavenumbers as this usually is a challenge for wave propagation problems. For these problems $\kappa = 12\pi$ and 24π, corresponding to 6 and 12 wavelengths along the duct. The solution functions are shown in
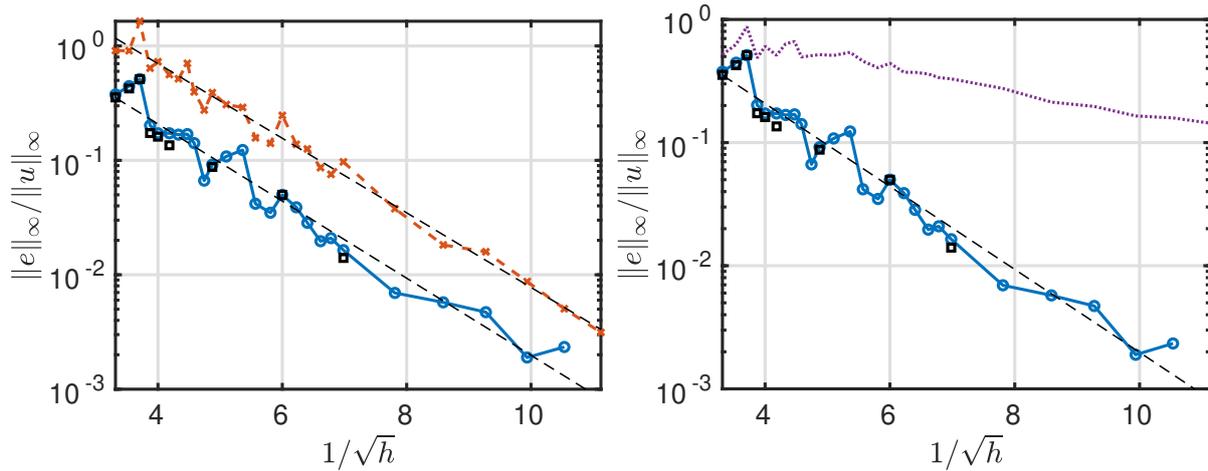
Figure 10: The relative error approximation (32) (×) and the error against the reference solution as a function of $1/\sqrt{h}$ are shown in the left subfigure. The ℓ_2-norm of the residual is shown together with the same error curve in the right subfigure. The black squares are the results for the optimal shape parameter values. The dashed lines represent lines fitted to the data points.

Figure 11.
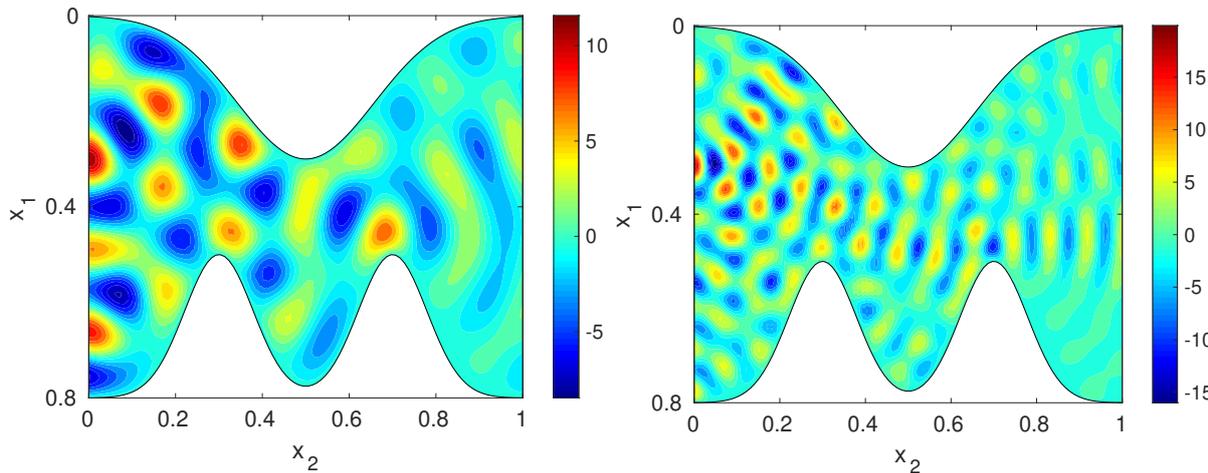
Figure 11: The solution function for $\kappa = 12\pi$ (left) and $\kappa = 24\pi$ (right). The solution is computed using nodes with $n_1 \times n_2 = 50 \times 62$. The source is located at the left boundary at $x_s = 0.3$.

These solutions have 9 and 19 propagating modes at the left boundary, respectively. A problem here was to compute the inner products with the eigenmodes to high enough accuracy. The accuracy of the boundary conditions is crucial to get the correct wave pattern. We were not able to run the simulations for $\kappa = 24\pi$ for a larger problem size than 50×62 (with good results). The same shape parameter scheme as for the convergence experiment was used.

For each problem, we ran three different problem sizes in order to get an estimate of the errors in the solutions. Then we computed the relative errors of the coarser solutions with respect to the finest solution. The computed errors are compared with the error estimate (32) to find the approximate ratio between real errors and estimate. Then we use the worst case ratio to project an error estimate also for the finest solution. The results are shown in Tables 4 and 5, indicating around 1% error for $\kappa = 12\pi$ and around 12% error for $\kappa = 24\pi$. The numbers of points per wavelength are 75/6=12.5 and 62/12 ≈ 5.2, respectively. For the problem with $\kappa = 6\pi$ and the same two node sets, we had 0.6–0.7% error and 21–25 points per wavelength. If we look at the error for 5 points per wavelength for $\kappa = 6\pi$, it is around 20%. That is, it seems that we do not need to resolve more with increasing frequency. For finite difference and finite element methods the error in a waveguide Helmholtz problem is typically proportional to $h^p \kappa^{p+1}$ [1, 38, 25]. This effect diminishes as the order of the method increases, and for a spectral method it disappears. This is consistent with the results for small ε in Section 6.2, where the error approximations are proportional to $(\kappa h)^K$.

Table 4: The relative error in relation to the finest solution, the relative error estimate, the ratio, the adjusted error estimate, and the local slope for the problem with $\kappa = 12\pi$.

$n_1 \times n_2$	$\|e\|_\infty/\|u\|_\infty$	$\|\tilde{e}\|_\infty/\|u\|_\infty$	$\|\tilde{e}\|_\infty/\|e\|_\infty$	$\frac{\|\tilde{e}\|_\infty}{\|u\|_\infty} / \min \frac{\|\tilde{e}\|_\infty}{\|e\|_\infty}$	C_M
40×50	0.0435	0.3083	7.1	0.0435	—
50×62	0.0240	0.1781	7.4	0.0251	0.66
60×75	—	0.0699	—	0.0099	1.2

Table 5: The relative error in relation to the finest solution, the relative error estimate, the ratio, the adjusted error estimate, and the local slope for the problem with $\kappa = 24\pi$.

$n_1 \times n_2$	$\|e\|_\infty/\|u\|_\infty$	$\|\tilde{e}\|_\infty/\|u\|_\infty$	$\|\tilde{e}\|_\infty/\|e\|_\infty$	$\frac{\|\tilde{e}\|_\infty}{\|u\|_\infty} / \min \frac{\|\tilde{e}\|_\infty}{\|e\|_\infty}$	C_M
30×37	0.3842	1.4909	3.9	0.3842	—
40×50	0.1292	0.7941	6.1	0.2047	0.64
50×62	—	0.4756	—	0.1226	0.62

8 Discussion

The main benefits of using global RBF methods for solving Helmholtz-type problems are that very few points per wavelength are needed to obtain a qualitatively correct solution, and that the number of points per wavelength does not need to increase with κ (the number of wavelengths). It is also relevant that non-trivial waveguide geometries can be managed easily, since there is no need for an orthogonal or even a structured grid. In [38, 25], we used orthogonal grids, which limits how much the boundaries can vary. It should be mentioned that the DtN boundary conditions assume a smooth continuation with horizontal boundaries outside of the domain. In our example, the derivative of the boundary curves is non-zero at $x_2 = 0, 1$, which introduces an error. However, since we got optimal convergence rates in the experiments, these errors are not large enough to influence the results at the level of errors that we could reach.

The main challenge of using a global RBF method for a PDE problem is the computational cost. In Helmholtz applications it is of interest to solve problems that are large in terms of wavelengths, and therefore require a certain resolution. With a dense linear system, both the storage requirements and the computational cost for a direct solver quickly become difficult to manage at least without using distributed computing. On top of that, the severe ill-conditioning of the linear systems makes them sensitive to numerical errors in the quadrature employed in DtN conditions as well as to rounding errors. An attractive alternative to using global RBF collocation methods is to use localized methods such as RBF-generated finite differences (RBF-FD) [12] and RBF partition of unity methods (RBF-PUM) [27]. In [51] it was shown that for an option pricing application, there was no significant difference in accuracy between the global method and RBF-PUM for a given problem size, while the computational cost is significantly lower for RBF-PUM due to sparsity of the linear systems.

We compared the non-symmetric and symmetric collocation approaches and found that the symmetric method, even though elegant, becomes cumbersome especially for non-trivial operators. The main benefit of the symmetric collocation is the guaranteed non-singularity of the interpolation matrix. However, for the non-symmetric method, singularity only occurred for wavenumbers that were physically uninteresting or for problems that were numerically unresolved. Furthermore, if a greedy method is used for the selection of center points and collocation points, then also the non-symmetric method is guaranteed to lead to non-singular matrices and convergence as h is refined [32].

We investigated the error behavior as a function of N and ε from different perspectives. Some of this can be explained by the limit behavior. We studied the limits for interpolation in [24], but here we looked at what is different for PDE problems. If the node set is unisolvent and PDE unisolvent, the limit RBF approximant has the form $s(\underline{x}) = P_K(\underline{x}) + \varepsilon^2 P_{K+2}(\underline{x}) + \ldots$, where $P_K(\underline{x})$ is the unique polynomial solution of degree K to the PDE problem, and P_{K+2j} have zero PDE residual at the node points. When ε is small, $P_K(\underline{x}) - u(\underline{x})$ dominates the error. This is the flat region in the error as a function of ε, see Figures 6 and 7. Then as ε starts to grow, there may be an optimal ε-range where the additional terms improve on the polynomial error, but eventually, the ε-terms dominate the error, and the exponential convergence rate depends mainly on ε and not on the problem, see Figure 5.

A contribution that we think is novel and of practical interest is the discussion about convergence for scaled shape parameters. We provide arguments for why $\varepsilon = Ch^\beta$ should lead to a convergence rate of the form $e^{-C_M/h^{\beta+1}}$, and show that this is what we also get numerically for $\beta = -1/2$.

Another practical contribution is that we have shown that given a reasonable error estimate, we can decide on a good choice for the shape parameter based on a small test problem. Then using a converging shape parameter strategy, we can solve the real problem, and also based on a comparison of error estimates and errors against the finest solution, we can get an improved error estimate for the solution of the most resolved problem.

Even though global collocation methods are not really practical for large scale problems, many of the things we have learned can be transferred also to localized methods, as these are based on 'local global collocation'.

Acknowledgements

A Proof sketches

In order to save space and not repeat already published material, we do not give the full proof for Theorem 5.1 here, instead we give instructions on how to carry out the proof using the machinery laid down in [24, pp. 122–127]. Because the RBF approximant in the PDE case has exactly the same form as the usual RBF interpolant, we get the exact same expansion [24, Eq. (28)] of the solution for small ε

$$s(\underline{x}, \varepsilon) = \varepsilon^{-2K}(\varepsilon^{-2q}P_{-q}(\underline{x}) + \cdots + \varepsilon^{2K}P_K(\underline{x}) + \cdots).\tag{39}$$

What differs from the interpolation case is the conditions that the polynomials must fulfill. In the PDE case we have that

$$\begin{array}{ll} P_K & \text{satisfies the inhomogeneous PDE and} \\ & \text{boundary conditions at the } N \text{ node points,} \\ P_j, \quad j \neq K & \text{satisfy the homogeneous PDE and} \\ & \text{boundary conditions at the } N \text{ node points.} \end{array}\tag{40}$$

The proof of part (i) is completely analogous to the proofs of Theorems 4.1 and 4.2 [24]. For part (iii), we follow the steps in the proof of Theorem 4.1. For simplicity, we first assume that the nullspace $n(\underline{x})$ of the matrix Q defined in (19) is of degree K. The steps are identical until the point were we are considering the conditions for P_{-q+K}. There are three possibilities

- If $q = 0$, then the polynomial is P_K and must satisfy the PDE. However, since the matrix Q is singular, this can only happen in the (unlikely) case that the right hand side $\underline{f}$ is in the range of Q.

- If $q > 0$ and P_{-q+K} is identically zero, then the moment vector $\underline{\sigma}_{-q}$ is zero, leading to $\underline{\lambda}_{-q}$, because of the non-singularity of P. This means that we could have omitted the $-q$ term in the expansion and we must have $q = 0$. This is in conflict with the previous case.

- Then we must have $q > 0$ and P_{-q+K} must contain a nullspace component $\alpha n(\underline{x})$. This means that we have at least one divergent term in the expansion of the solution.

If there is just a single nullspace component of degree K, and extending Q with an appropriate monomial of degree $K + 1$ leads to rank$(Q) = N$, then at the next step looking at P_{-q+1+K} we get the two possibilities $\alpha = 0$, which has been ruled out, or $P_{-q+1+K} = P_K$. Hence, we must have $q = 1$ and divergence of order ε^2.
    If the nullspace is of lower degree than K, we will also get divergence, but the negative power of ε could be higher.
    The argument behind part (ii) is that we need to go to the polynomial P_{-q+M} before we have enough degrees of freedom to satisfy the discrete PDE problem. Therefore, the limit must have degree M. However, because Q is non-singular, all previous polynomials must be identically zero and accordingly there can be no divergence. Compare with the proofs of Theorems 4.2 and 4.3.
    For part (iv) of the proof, we follow the proof of Theorem 4.3. The important difference is that the relation between the moments is determined by the nullspace of P, but the possible nullspace parts in the polynomials P_{-q+J} are determined by the nullspace of Q. In [24], we arrive at an equation $C^T B^{-1} C \underline{\alpha} = \underline{0}$. The corresponding equation here becomes

$$C^T B^{-1} D \underline{\alpha} = \underline{0},\tag{41}$$

where C is of size $n \times m$ and D has dimensions $n \times p$. To be precise, at step J of the proof, m is the dimension of the J-degree part of the nullspace of P and p is the corresponding dimension for Q.
    If $m = p$, the system (41) is square, but non-singularity cannot be guaranteed when C and D are different. If $m > p$, the system is over-determined and it is likely that the only solution is $\underline{\alpha} = 0$. If on the other hand, $m < p$ the system is under-determined, allowing for non-zero nullspace components in the expansion polynomials.
    If $n(x)$ defines a nullspace component for P, then $p(\underline{x})n(\underline{x})$ defines a higher degree nullspace component using any polynomial $p(\underline{x})$. Therefore, the dimension m typically grows with J. However, there is no similar mechanism for the nullspace of Q (since $Ln(\underline{x}) = 0$ does not generally imply $L(p(\underline{x})n(\underline{x})) = 0$). Accordingly, the dimension p is likely to stay the same or decrease with J.
    These facts taken together lead to the statements in part (iv). We use the formulation *likely*, since it should be theoretically possible to construct counter examples in both the convergent and the divergent case.

References

[1] A. Bayliss, C. Goldstein, and E. Turkel. The numerical solution of the Helmholtz equation for wave propagation problems in underwater acoustics. *Comput. Math. Applic.*, 11(7):655–665, 1985.

[2] L. Bos. On certain configurations of points in $\mathbb{R}^n$ which are unisolvent for polynomial interpolation. *Journal of approximation theory*, 64(3):271–280, 1991.

[3] M. Buhmann and N. Dyn. Spectral convergence of multiquadric interpolation. *Proc. Edinburgh Math. Soc. (2)*, 36(2):319–333, 1993.

[4] R. Cavoretto and A. De Rossi. An adaptive LOOCV-based refinement scheme for RBF collocation methods over irregular domains. *Appl. Math. Lett.*, 103:106178, 10, 2020.

[5] A. H.-D. Cheng, M. A. Golberg, E. J. Kansa, and G. Zammito. Exponential convergence and h-c multiquadric collocation method for partial differential equations. *Numer. Methods Partial Differential Equations*, 19(5):571–594, 2003.

[6] O. Davydov and D. T. Oanh. On the optimal shape parameter for Gaussian radial basis function finite difference approximation of the Poisson equation. *Comput. Math. Appl.*, 62:2143–2161, 2011.

[7] T. A. Driscoll and B. Fornberg. Interpolation in the limit of increasingly flat radial basis functions. *Comput. Math. Appl.*, 43(3–5):413–422, 2002.

[8] G. Fasshauer. Solving partial differential equations by collocation with radial basis functions. In A. LeMéhauté, C. Rabut, and L. Schumaker, editors, *Surface Fitting and Multiresolution Methods, Volume 2 of the Proceedings of the 3rd International Conference on Curves and Surfaces, Chamonix-Mont-Blanc*, pages 131–138, Nashville, TN, 1997. Vanderbilt University Press.

[9] G. E. Fasshauer. *Meshfree approximation methods with MATLAB*, volume 6 of *Interdisciplinary Mathematical Sciences*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2007.

[10] A. J. M. Ferreira, C. M. C. Roque, R. M. N. Jorge, G. Fasshauer, and R. Batra. Analysis of functionally graded plates by a robust meshless method. *J. Mech. Adv. Mater. Struct.*, 14(8):577–587, 2007.

[11] N. Flyer, B. Fornberg, V. Bayona, and G. A. Barnett. On the role of polynomials in RBF-FD approximations: I. Interpolation and accuracy. *J. Comput. Phys.*, 321:21–38, 2016.

[12] B. Fornberg and N. Flyer. *A primer on radial basis functions with applications to the geosciences*, volume 87 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015.

[13] B. Fornberg, E. Larsson, and N. Flyer. Stable computations with Gaussian radial basis functions. *SIAM J. Sci. Comput.*, 33(2):869–892, 2011.

[14] B. Fornberg, E. Lehto, and C. Powell. Stable calculation of Gaussian-based RBF-FD stencils. *Comput. Math. Appl.*, 65(4):627–637, 2013.

[15] B. Fornberg and C. Piret. A stable algorithm for flat radial basis functions on a sphere. *SIAM J. Sci. Comput.*, 30(1):60–80, 2007.

[16] B. Fornberg and G. Wright. Stable computation of multiquadric interpolants for all values of the shape parameter. *Comput. Math. Appl.*, 48(5):853–867, 2004.

[17] C. Franke and R. Schaback. Solving partial differential equations by collocation using radial basis functions. *Appl. Math. Comput.*, 93(1):73–82, 1998.

[18] G. H. Golub and J. M. Ortega. *Scientific computing and differential equations*. Academic Press, Inc., Boston, MA, 1992.

[19] Y. C. Hon and R. Schaback. On unsymmetric collocation by radial basis functions. *Appl. Math. Comput.*, 119(2-3):177–186, 2001.

[20] E. J. Kansa. Multiquadrics—a scattered data approximation scheme with applications to computational fluid-dynamics. II. Solutions to parabolic, hyperbolic and elliptic partial differential equations. *Comput. Math. Appl.*, 19(8–9):147–161, 1990.

[21] T. Karvonen and S. Särkkä. Worst-case optimal approximation with increasingly flat Gaussian kernels. *Adv. Comput. Math.*, 46(2):No. 21, 17 pp., 2020.

[22] J. B. Keller and D. Givoli. Exact non-reflecting boundary conditions. *J. Comp. Phys.*, 82(1):172–192, 1989.

[23] E. Larsson and B. Fornberg. A numerical study of some radial basis function based solution methods for elliptic PDEs. *Comput. Math. Appl.*, 46(5-6):891–902, 2003.

[24] E. Larsson and B. Fornberg. Theoretical and computational aspects of multivariate interpolation with increasingly flat radial basis functions. *Comput. Math. Appl.*, 49(1):103–130, 2005.

[25] E. Larsson. A domain decomposition method for the Helmholtz equation in a multilayer domain. *SIAM J. Sci. Comp.*, 20(5):1713–1731, 1999.

[26] E. Larsson, E. Lehto, A. Heryudono, and B. Fornberg. Stable computation of differentiation matrices and scattered node stencils based on Gaussian radial basis functions. *SIAM J. Sci. Comput.*, 35(4):A2096–A2119, 2013.

[27] E. Larsson, V. Shcherbakov, and A. Heryudono. A least squares radial basis function partition of unity method for solving PDEs. *SIAM J. Sci. Comput.*, 39(6):A2538–A2563, 2017.

[28] Y. J. Lee, C. A. Micchelli, and J. Yoon. On convergence of flat multivariate interpolation by translation kernels with finite smoothness. *Constr. Approx.*, 40(1):37–60, 2014.

[29] Y. J. Lee, C. A. Micchelli, and J. Yoon. A study on multivariate interpolation by increasingly flat kernel functions. *J. Math. Anal. Appl.*, 427(1):74–87, 2015.

[30] Y. J. Lee, G. J. Yoon, and J. Yoon. Convergence of increasingly flat radial basis interpolants to polynomial interpolants. *SIAM J. Math. Anal.*, 39(2):537–553, 2007.

[31] J. Lin, W. Chen, and K. Y. Sze. A new radial basis function for Helmholtz problems. *Eng. Anal. Bound. Elem.*, 36(12):1923–1930, 2012.

[32] L. Ling and R. Schaback. Stable and convergent unsymmetric meshless collocation methods. *SIAM Journal on Numerical Analysis*, 46(3):1097–1115, 2008.

[33] M.-A. Londoño and H. Montegranario. Optimal shape parameter for meshless solution of the 2D Helmholtz equation. *CT&F - Cienc. Tecn. Fut.*, 9:15–35, 2019.

[34] W. R. Madych. Miscellaneous error bounds for multiquadric and related interpolators. *Comput. Math. Appl.*, 24(12):121–138, 1992.

[35] W. R. Madych. An estimate for multivariate interpolation. II. *J. Approx. Theory*, 142(2):116–128, 2006.

[36] W. R. Madych and S. A. Nelson. Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation. *J. Approx. Theory*, 70(1):94–114, 1992.

[37] P. K. Mishra, G. E. Fasshauer, M. K. Sen, and L. Ling. A stabilized radial basis-finite difference (RBF-FD) method with hybrid kernels. *Comput. Math. Appl.*, 77(9):2354–2368, 2019.

[38] K. Otto and E. Larsson. Iterative solution of the Helmholtz equation by a second-order method. *SIAM J. Matrix. Anal. Appl.*, 21(1):209–229, 1999.

[39] U. Pettersson. Radial basis function approximations for the Helmholtz equation. M.Sc. thesis, UPTEC Report F 03 082, School of Engineering, Uppsala Univ., Uppsala, Sweden, 2003.

[40] R. B. Platte and T. A. Driscoll. Eigenvalue stability of radial basis function discretizations for time-dependent problems. *Comput. Math. Appl.*, 51(8):1251–1268, 2006.

[41] R. B. Platte and T. A. Driscoll. Polynomials and potential theory for Gaussian radial basis function interpolation. *SIAM J. Numer. Anal.*, 43(2):750–766, 2005.

[42] R. B. Platte, L. N. Trefethen, and A. B. J. Kuijlaars. Impossibility of fast stable approximation of analytic functions from equispaced samples. *SIAM Rev.*, 53(2):308–318, 2011.

[43] M. J. D. Powell. Univariate multiquadric interpolation: some recent results. In *Curves and surfaces (Chamonix-Mont-Blanc, 1990)*, pages 371–382. Academic Press, Boston, MA, 1991.

[44] C. Rieger and B. Zwicknagl. Sampling inequalities for infinitely smooth functions, with applications to interpolation and machine learning. *Adv. Comput. Math.*, 32(1):103–129, 2010.

[45] C. Rieger and B. Zwicknagl. Improved exponential convergence rates by oversampling near the boundary. *Constr. Approx.*, 39(2):323–341, 2014.

[46] S. Rippa. An algorithm for selecting a good value for the parameter c in radial basis function interpolation. *Adv. Comput. Math.*, 11(2–3):193–210, 1999.

[47] R. Schaback. Multivariate interpolation by polynomials and radial basis functions. *Constr. Approx.*, 21(3):293–317, 2005.

[48] R. Schaback. Convergence of unsymmetric kernel-based meshless collocation methods. *SIAM J. Numer. Anal.*, 45(1):333–351, 2007.

[49] R. Schaback. Limit problems for interpolation by analytic radial basis functions. *J. Comput. Appl. Math.*, 212(2):127–149, 2008.

[50] R. Schaback. All well-posed problems have uniformly stable and convergent discretizations. *Numer. Math.*, 132(3):597–630, 2016.

[51] V. Shcherbakov and E. Larsson. Radial basis function partition of unity methods for pricing vanilla basket options. *Comput. Math. Appl.*, 71(1):185–200, 2016.

[52] G. Song, J. Riddle, G. E. Fasshauer, and F. J. Hickernell. Multivariate interpolation with increasingly flat radial basis functions of finite smoothness. *Adv. Comput. Math.*, 36(3):485–501, 2012.

[53] P. Thounthong, M. N. Khan, I. Hussain, I. Ahmad, and P. Kumam. Symmetric radial basis function method for simulation of elliptic partial differential equations. *Mathematics*, 6(12), 2018.

[54] G. B. Wright and B. Fornberg. Stable computations with flat radial basis functions using vector-valued rational approximations. *J. Comp. Phys.*, 331:137–156, 2017.

[55] Z. M. Wu. Hermite-Birkhoff interpolation of scattered data by radial basis functions. *Approx. Theory Appl.*, 8(2):1–10, 1992.

[56] Z. M. Wu and R. Schaback. Local error estimates for radial basis function interpolation of scattered data. *IMA J. Numer. Anal.*, 13(1):13–27, 1993.

[57] F. Yang, L. Yan, and L. Ling. Doubly stochastic radial basis function methods. *J. Comput. Phys.*, 363:87–97, 2018.