

 Open access • Journal Article • DOI:10.1137/S0036142994269186

## Is the Pollution Effect of the FEM Avoidable for the Helmholtz Equation Considering High Wave Numbers — [Source link](#)

Ivo Babuška, Stefan A. Sauter

**Institutions:** University of Kiel

**Published on:** 01 Dec 1997 - SIAM Journal on Numerical Analysis (Society for Industrial and Applied Mathematics)

**Topics:** Helmholtz equation, Partial differential equation, Galerkin method, Finite element method and Wave equation

Related papers:

- [Finite element solution of the Helmholtz equation with high wave number Part I: The h-version of the FEM](#)☆
- [A Generalized Finite Element Method for solving the Helmholtz equation in two dimensions with minimal pollution](#)
- [Finite Element Solution of the Helmholtz Equation with High Wave Number Part II: The h - p Version of the FEM](#)
- [Finite Element Analysis of Acoustic Scattering](#)
- [Application of an Ultra Weak Variational Formulation of Elliptic PDEs to the Two-Dimensional Helmholtz Problem](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/is-the-pollution-effect-of-the-fem-avoidable-for-the-49yqsttap8>



University of Zurich  
Zurich Open Repository and Archive

Winterthurerstr. 190  
CH-8057 Zurich  
<http://www.zora.uzh.ch>

---

*Year: 1997*

---

## Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?

Babuska, I; Sauter, S

Babuska, I; Sauter, S (1997). Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? *SIAM Journal on Numerical Analysis*, 34(6):2392-2423.

Postprint available at:  
<http://www.zora.uzh.ch>

Posted at the Zurich Open Repository and Archive, University of Zurich.  
<http://www.zora.uzh.ch>

Originally published at:  
*SIAM Journal on Numerical Analysis* 1997, 34(6):2392-2423.

# IS THE POLLUTION EFFECT OF THE FEM AVOIDABLE FOR THE HELMHOLTZ EQUATION CONSIDERING HIGH WAVE NUMBERS?\*

IVO M. BABUŠKA<sup>†</sup> AND STEFAN A. SAUTER<sup>‡</sup>

**Abstract.** The development of numerical methods for solving the Helmholtz equation, which behaves robustly with respect to the wave number, is a topic of vivid research. It was observed that the solution of the Galerkin finite element method (FEM) differs significantly from the best approximation with increasing wave number. Many attempts have been presented in the literature to eliminate this lack of robustness by various modifications of the classical Galerkin FEM.

However, we will prove that, in two and more space dimensions, it is impossible to eliminate this so-called pollution effect. Furthermore, we will present a generalized FEM in one dimension which behaves robustly with respect to the wave number.

**Key words.** Helmholtz equation, high wave number, pollution effect, generalized FEM

**AMS subject classifications.** 65N12, 65N15, 65N30

**PII.** S0036142994269186

## 1. Introduction.

**1.1. Physical motivation.** Boundary value problems governed by the wave equation

$$\frac{\partial^2 w}{\partial t^2} - \Delta w = g$$

arise in many physical applications, e.g., electromagnetic wave propagation and acoustics. In applications like radar detection of moving bodies or acoustic scattering, a typical situation is that the inhomogeneity  $f$  is time periodic,

$$g(x, t) = f(x) e^{ikt}.$$

In this case, we know (cf. [9]) that the solution of the wave equation is of the form  $w(x, t) = u(x) e^{ikt}$ , where the amplitude  $u(x)$  satisfies the Helmholtz equation

$$-\Delta u - k^2 u = f.$$

In contrast to the Poisson equation in potential theory, the function  $u$  (and not its derivatives) is of main physical interest, denoting the amplitude of the electric or magnetic field or of the acoustic pressure, depending on the underlying application. This remark will be essential for the choice of an appropriate norm for measuring the accuracy of the discrete solution.

In many situations, e.g., scattering and transmission problems, the Helmholtz equation is defined in an unbounded exterior domain with Sommerfeld's radiation

---

\*Received by the editors June 6, 1994; accepted for publication (in revised form) July 1, 1996. This work was supported by ONR grant N00014-93-I-0131.

<http://www.siam.org/journals/sinum/34-6/26918.html>

<sup>†</sup>Texas Institute for Computational and Applied Mathematics, University of Texas at Austin, Taylor Hall 2.400, Austin, TX 78712.

<sup>‡</sup>Lehrstuhl Praktische Mathematik, Mathematisches Seminar, Universität Kiel, 24098 Kiel, Germany (sas@numerik.uni-kiel.de). This author's work was supported by the German Research Foundation (DFG) grant Sa 607/1-1.

conditions imposed at infinity (cf. [8]):

$$\begin{aligned} -\Delta u - k^2 u &= f && \text{in } \mathbf{R}^d \setminus \bar{D}, \\ u &= 0 && \text{on } \partial D, \\ \frac{\partial u}{\partial r} - iku &= o\left(\|x\|^{\frac{1-d}{2}}\right) && \|x\| \rightarrow \infty. \end{aligned}$$

Here,  $D$  denotes a bounded domain in  $\mathbf{R}^d$ , and  $\frac{\partial}{\partial r}$ , the derivative in the radial direction. If the so-called wave number  $k$  becomes large the solution of the Helmholtz equation becomes highly oscillating and the discretization very expensive. Hence, there is a growing interest in discretization methods where the computational complexity increases only moderately with increasing wave number.

**1.2. Transforming the Helmholtz equation on exterior domains onto a finite domain.** The fact that, in many situations, the Helmholtz equation is imposed on infinite domains rules out the straightforward use of finite element or finite difference discretizations. This difficulty can be avoided by introducing a sufficiently large ball  $B$  containing  $D \subset\subset B$ . The equation outside the artificial ball can be transformed to nonlocal boundary conditions on  $\partial B$  using the method of integral equations (see [13]). In this context, this technique is called the Dirichlet-to-Neumann map (see [17]). Thus, the Helmholtz equation has to be solved on the finite domain  $\Omega := B \cap \mathbf{R}^d \setminus \bar{D}$ :

$$(1.1) \quad \begin{aligned} -\Delta u - k^2 u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial D, \\ \frac{\partial u}{\partial n} - K[u] &= q && \text{on } \partial B, \end{aligned}$$

where  $\partial/\partial n$  denotes the normal derivative. For sufficiently large  $B$ , the integral operator  $K[u]$  can be approximated by the local term  $iku$ , i.e., the nonlocal boundary conditions are replaced by so-called Robin boundary conditions:

$$\frac{\partial u}{\partial n} - iku = q.$$

The size of the ball has to be adapted to the required accuracy and, additionally, has to be increased with increasing wave number  $k$ . A quantitative error analysis for this kind of approximation is given in [11, Thm. 3.1]. What is important for our purpose is the following. The domain of computation  $\Omega$  is, especially for high wave numbers  $k$ , much larger than the domain of physical interest. This observation shows that for large domains, weighted norms are appropriate to measure the accuracy of the numerical solution.

**1.3. The pollution effect of the Galerkin FEM for the Helmholtz equation.** It is well known that, for elliptic boundary value problems like (1.1), the Galerkin FEM leads to quasi-optimal error estimates with respect to the degrees of freedoms. This means that the accuracy of the Galerkin solution differs only by a constant factor from the best approximation in the finite element space. From numerical experiments and from theoretical analysis it is known (see [14], [15]) that this factor increases with increasing wave number; in other words, the Galerkin FEM does not behave robustly with respect to  $k$ . In [6, Thm. 2.6] it was shown that, for a model situation, the ratio of the error of the Galerkin solution and the error of the best approximation tends to infinity with increasing  $k$ . On the other hand, it was shown in [2] that the condition “ $k^2 h$  is small” would be sufficient to guarantee that the error

of the Galerkin solution is of the same magnitude as the error of the best approximation (independent of  $k$ ). In practice, however, this condition would rule out reliable wave computations in three dimensions for moderate and higher wave numbers for the following reasons. The condition  $k^2 h < 1$  would imply that the dimension  $N$  of the system matrix is of order  $N = O(h^{-3}) = O(k^6)$ . The arising system of linear equations has complex entries and is highly indefinite such that the solution process becomes too expensive for  $k > 10 \sim 20$ .

Recently, many attempts have been made in the mathematical and engineering literature to overcome this lack of robustness which, in this context, is called a pollution effect (see [14], [18], [6]). In many cases, one-dimensional model problems have been analyzed and then generalized to more space dimensions. Numerical experiments show that in some situations the pollution effect can be reduced, but in two-dimensions quantitative results about the size of the pollution appear to be very vague and a theoretical foundation is missing. These observations motivate us to study the following questions.

1. Is it possible to define a generalized finite element method (GFEM) for the Helmholtz equation that has no pollution—one in which the Galerkin solution converges at the same rate as the best approximation independent of the wave number  $k$ ?
2. How much insight can one get from the study of one-dimensional model problems? Are all higher-dimensional effects of the pollution visible in one dimension?

In this paper we will prove the following results.

- In one dimension, the pollution effect can be eliminated completely by a suitable modification of the discrete bilinear form.
- In two dimensions, the pollution effect can be reduced substantially, but cannot be avoided in principle. Hence, one-dimensional results are not fully representative for the two- and higher-dimensional cases.

The paper is organized as follows. After having introduced some notation we will specify what we mean by GFEMs. Then, we will explicitly define a pollution-free FEM in one dimension.

In two dimensions, we will prove that, for any modification of the FEM, we can define a family of domains and right-hand sides such that the ratio of the Galerkin error and the error of the best approximation tends to infinity. However, the proof of that theorem shows how a modified Galerkin method has to be designed such that the pollution effect is *minimal*. Numerical examples presented in [6] show that, by using these results, the classical Galerkin FEM can be improved substantially.

**2. Setting.** In this section, we will consider finite element discretizations for approximating the solution of the Helmholtz equation

$$(2.1) \quad \begin{aligned} -\Delta u - k^2 u &= f && \text{in } \Omega, \\ u &= 0 && \text{on } \partial D, \\ \frac{\partial u}{\partial n} - iku &= q && \text{on } \partial B. \end{aligned}$$

First, we will introduce some basic notation.

**2.1. Finite element spaces and Galerkin discretization.** The Galerkin discretization is applied to the variational formulation of the Helmholtz equation. For this, let  $H^1(\Omega)$  denote the usual Sobolev space as defined, e.g., in [1]. Incorporating the essential boundary condition on  $\partial D$  we obtain the space  $V := \{v \in H^1(\Omega) :$

$v|_{\partial D}=0\}$ . For given right-hand side  $f \in V'$  and  $q \in H^{-1/2}(\partial B)$ , the variational formulation of (2.1) is given by seeking  $u \in H^1(\Omega)$  such that

$$(2.2) \quad a(u, v) = F(v) \quad \forall v \in V,$$

with

$$\begin{aligned} a(u, v) &= \int_{\Omega} \nabla u \nabla \bar{v} - k^2 u \bar{v} dx - ik \int_{\partial B} u \bar{v} dx, \\ F(v) &= \int_{\Omega} f \bar{v} dx + \int_{\partial B} q \bar{v} dx. \end{aligned}$$

*Remark 1.* The bilinear form  $a(\cdot, \cdot)$  is symmetric, i.e.,  $a(u, v) = a(v, u)$ , but not Hermitian.

The Galerkin FEM is given by replacing in (2.2) the infinite-dimensional space  $V$  by finite element spaces. In this paper, we focus our attention on approximations of the Helmholtz equation of second order. For the FEM, this means that we are employing (bi)linear elements. Remarks on how the results can be generalized to higher-order discretizations will appear in various places. Let

$$\tau = \{\Delta_1, \Delta_2, \dots, \Delta_N\}$$

denote a finite element mesh consisting of simplicial or quadrilateral elements. The mesh width is denoted by

$$h := \max_{\Delta \in \tau} \text{diam } \Delta.$$

Let

$$\Theta := \{x_1, x_2, \dots, x_n\}$$

denote the set of vertices of  $\tau$  not lying on the essential boundary  $\partial D$ . For  $x_i \in \Theta$ , the usual local basis functions are given by

$$\begin{aligned} \phi_i(x) &= \delta_{i,j}, \quad 1 \leq i \leq n, \\ \phi_i|_{\Delta} &\text{ is (bi)linear for all } \Delta \in \tau. \end{aligned}$$

The finite element space corresponding to the grid  $\tau$  is given by

$$(2.3) \quad \mathcal{S}_h = \text{span } \{\phi_i \mid 1 \leq i \leq n\}.$$

**2.2. The Pollution effect of the FEM for the Helmholtz equation.** Let  $u_h \in \mathcal{S}_h$  denote the Galerkin finite element solution of the Helmholtz problem, while  $u$  denotes the exact solution in  $V$ . The error is given by  $e_h = u - u_h$ . The best approximation of  $u$  in the space  $\mathcal{S}_h$  with respect to an appropriate norm  $\|\cdot\|$  is defined by

$$\begin{aligned} u_h^{opt} &:= \arg \min_{v_h \in \mathcal{S}_h} \|u - v_h\|, \\ e_h^{opt} &:= \|u - u_h^{opt}\|. \end{aligned}$$

Obviously, we have

$$\|e_h^{opt}\| \leq \|e_h\|.$$

On the other hand, we know that the Galerkin method is quasi-optimal in the sense that, for sufficiently small  $h$ , the Galerkin solution satisfies

$$\|e_h\| \leq C \|e_h^{opt}\|.$$

Intuitively, we will say that the FEM has the pollution effect if  $e_h$  approaches zero increasingly slowly as  $e_h^{opt}$  has increasing  $k$ . This will be defined in a formal way now.

**DEFINITION 2.1.** *Here and in the following, we assume that the wave number is bounded away from zero,  $k \geq k_0 > 0$ . Let  $W$  denote a subspace of  $V' \times H^{-1/2}(\partial B)$ . For  $(f, q) \in W$ , let  $u_{f,q}$  denote the solution of the Helmholtz problem (2.2). We say that a FEM has the pollution effect if there are numbers  $r, s \in \mathbf{R}$  and  $t > 0$  such that the error of the best approximation satisfies*

$$\|e_h^{opt}\| \leq C_{f,q} h^r k^s \quad \forall (f, q) \in W,$$

and there exists a family of data, i.e., domains  $\Omega_k$ , right-hand sides  $(f_k, q_k) \in W$ , and meshes  $\tau_h$  characterized by the step size  $h = h(k)$  such that the error of the corresponding finite element solution can be estimated by

$$(2.4) \quad \frac{\|e_h\|}{h^r k^s} \geq C k^t.$$

In order to motivate a minimal requirement on the dependence of  $h$  on  $k$ , we consider the following model problem:

$$\begin{aligned} -u'' - k^2 u &= 0, & \text{in } \Omega = (0, 1), \\ u(0) &= 1, \\ u'(1) - iku(1) &= 0. \end{aligned}$$

The exact solution is given by  $u(x) = \exp(ikx)$ . The best approximation in the space of continuous, piecewise linear trial functions satisfies

$$\frac{\|u - u_h^{opt}\|_{H^s(\Omega)}}{\|u\|_{H^s(\Omega)}} \leq C h^{2-s} \frac{\|u\|_{H^2(\Omega)}}{\|u\|_{H^s(\Omega)}} \leq C (kh)^{2-s}.$$

Hence, a minimal requirement for the relative error of the best approximation to be small is that  $(kh)$  is small. In this light, the following assumption is very natural.

**Assumption 2.2.** Throughout this paper we assume that  $(kh) \leq \alpha$  holds with a generic constant  $\alpha$  being independent of all parameters.

For a one-dimensional model problem the pollution effect was studied thoroughly in [15], [16], and [6]. It was shown for one-dimensional model problems that

- the Galerkin FEM contains the pollution effect in both the  $L^2$ - and  $H^1$ -norms;
- for the Galerkin FEM, estimate (2.4) holds for the  $L^2$ -norm with  $t = 1$  (cf. [6, Thm. 2.6]);
- the pollution effect is visible for all degrees of approximation, i.e., the  $hp$ -method for arbitrary but fixed  $p$ ;
- the pollution effect can be interpreted as a lack of stability since the discrete inf-sup constant behaves inverse-proportionally to the wave number.

It is clear that, in more space dimensions, one expects the pollution effect of at least the same magnitude as in the one-dimensional case.

**2.3. GFEMs.** In the following, we will characterize GFEMs. GFEMs were first introduced in [4] in a variational setting. For our purpose, an algebraic definition will turn out to be more appropriate.

We begin by introducing finite element interpolation and grid functionals.

DEFINITION 2.3. A vector  $\gamma \in \mathbf{C}^n$  is linked with a finite element function by the canonical prolongation

$$\mathcal{E}[\gamma](x) = \sum_{i=1}^n \gamma_i \phi_i(x), \quad x \in \Omega.$$

A grid functional  $\mathbf{Q}(f, q)$  is an operator which maps the right-hand side  $(f, q)$  of (2.1) onto a vector in  $\mathbf{C}^n$ .

Example 2.4. Using this notation, the Galerkin FEM is defined by seeking  $\mathbf{u}_h$  such that

$$\mathbf{G}^{Gal} \mathbf{u}_h = \mathbf{Q}^{Gal}(f, q),$$

with

$$\begin{aligned} \mathbf{G}_{i,j}^{Gal} &= a(\phi_j, \phi_i), \quad 1 \leq i, j \leq n, \\ \mathbf{Q}^{Gal}(f, q) &= \{F(\phi_i)\}_{1 \leq i \leq n}. \end{aligned}$$

The following remark concerns the sparsity of  $\mathbf{G}^{Gal}$ .

Remark 2. Two points  $x, y \in \Theta$  are called physically connected if there exists an element  $\Delta \in \tau$  having  $x$  and  $y$  as vertices.

If  $x_i$  and  $x_j$  are not physically connected, then  $\mathbf{G}_{i,j}^{Gal} = 0$ . The converse holds for almost all  $k > 0$ .

DEFINITION 2.5. Let  $\mathcal{S}_h$  be defined by (2.3). A GFEM is characterized by a (regular) matrix  $\mathbf{G}$  and a grid functional  $\mathbf{Q}$ . These operators have to have the same sparsity structure as the classical Galerkin FEM, which can be expressed (cf. Remark 2) by

$$\begin{aligned} \mathbf{G}_{i,j} &= 0 && \text{if } (x_i, x_j) \text{ are not physically connected,} \\ (\mathbf{Q}(f, q))_i &= 0, && \text{if } \text{supp } \phi_i \cap \text{supp } f = \emptyset \text{ and } \text{supp } \phi_i|_{\partial B} \cap \text{supp } q = \emptyset. \end{aligned} \quad ^1$$

Additionally, we require  $\mathbf{G}$  to be symmetric:  $\mathbf{G}_{i,j} = \mathbf{G}_{i,i}$  (cf. Remark 1).

The solution of

$$\mathbf{G}\mathbf{v} = \mathbf{Q}(f, q)$$

is then identified with the so-called generalized finite element approximation to (2.1) by  $v = \mathcal{E}[\mathbf{v}]$ .

In the next section we will prove that, in one dimension, it is possible to define  $\mathbf{G}$  and  $\mathbf{Q}$ , i.e., a GFEM which has no pollution.

**3. On the stabilization of the Helmholtz equation in one dimension with Robin boundary conditions.** In this chapter we will prove that, in one dimension, there exists a pollution-free GFEM for the Helmholtz problem. We have already mentioned that this is not possible in the higher-dimensional case. However,

<sup>1</sup>Here, the notation  $A \cap B = \emptyset$  means that  $A$  and  $B$  have disjoint interiors.

the one-dimensional analysis gives insights into how a GFEM has to be designed in higher dimensions such that the pollution effect is *minimal*.

To fix the ideas, let us consider the following one-dimensional model problem:

$$(3.1) \quad \begin{aligned} -u'' - k^2 u &= f && \text{in } \Omega := (0, 1), \\ u(0) &= 0, \\ u'(1) - iku(1) &= 0, \end{aligned}$$

and assume that  $k$  is bounded away from zero, i.e.,  $k \geq k_0 > 0$ . The variational formulation of problem (3.1) can be written in the form, seeking  $u \in V := \{v \in H^1(0, 1) \mid v(0) = 0\}$ , such that

$$(3.2) \quad a(u, v) = \int_{\Omega} f v dx \quad \forall v \in V$$

with

$$(3.3) \quad a(u, v) := \int_{\Omega} \langle \nabla u, \nabla \bar{v} \rangle - k^2 u \bar{v} dx - iku(1) \bar{v}(1).$$

Let  $\{x_i\}_{1 \leq i \leq n}$  denote a set of grid points  $0 < x_0 < x_1 < \dots < x_n = 1$ . The grid  $\tau_h$  consists of the intervals  $\Delta_i = [x_{i-1}, x_i]$ . The step size  $h$  and the corresponding finite element space

$$\mathcal{S}_h = \text{span } \{\phi_i : 1 \leq i \leq n\}$$

were already defined by (2.3). We will also need the space  $\mathcal{S}_h^0$  consisting of functions which are constant on each interval  $\Delta_i$ .

The Galerkin discretization of (3.2) leads to a system of difference equations of the following form:

$$G_{i-1,i}^{Gal} u_{i-1} + G_{i,i}^{Gal} u_i + G_{i,i+1}^{Gal} u_{i+1} = \int_{\Delta_i} f \phi_i dx + \int_{\Delta_{i+1}} f \phi_i dx, \quad 1 \leq i \leq n,$$

where we have already used the symmetry of  $\mathbf{G}^{Gal}$ . Terms in the equation above containing subscripts smaller than 1 or larger than  $n$  have to be skipped.

For the GFEM, we make the ansatz

$$(3.4) \quad G_{i-1,i} u_{i-1} + G_{i,i} u_i + G_{i,i+1} u_{i+1} = \mathbf{Q}(f|_{\Delta_i}, 0) + \mathbf{Q}(f|_{\Delta_{i+1}}, 0), \quad 1 \leq i \leq n.$$

In order to motivate how the coefficients  $G_{i,j}$  have to be chosen, we present the following consideration.

From the theory of ODEs, we know that, if  $f \equiv 0$  on an interval  $\Delta_m$ , the exact solution takes the form

$$u|_{\Delta_m} = A_m e^{ikx} + B_m e^{-ikx}.$$

In [6, Chap. 2] it was shown that on such intervals the solution of the Galerkin FEM is the interpolant of

$$u_h|_{\Delta_m} = A_{h,m} e^{ik_h x} + B_{h,m} e^{-ik_h x}$$

with suitable  $A_{h,m}$  and  $B_{h,m}$ . Furthermore, it was shown that the pollution effect is related to the *phase lag*  $k - k_h$ . The idea of constructing a pollution-free GFEM is to

eliminate the phase lag for as many right-hand sides as possible. The choice  $f = 0$  in  $\Delta_m \cup \Delta_{m+1}$  implies that the right-hand side in (3.4) vanishes. This leads to the condition that (3.4) has to be satisfied for the fundamental system  $u = \exp(\pm ikx)$  with right-hand side zero. Additionally, by choosing the operator  $\mathbf{Q}$  in a suitable way, it is possible to determine  $\mathbf{G}$  such that the GFEM solution interpolates the exact solution for piecewise constant right-hand sides, i.e.,  $f \in \mathcal{S}_h^0$ .

Working out these ideas properly, the following GFEM results.

Let the system matrix  $\mathbf{G}^{stab}$  be defined by

$$(3.5) \quad \mathbf{G}_{i,j}^{stab} = \frac{k^2 h}{2 \tan \frac{kh}{2}} \begin{cases} \frac{\sin(k(x_{i+1}-x_{i-1}))}{\sin(k(x_{i+1}-x_i)) \sin(k(x_i-x_{i-1}))} & \text{if } i = j < n, \\ \frac{e^{-ik(x_n-x_{n-1})}}{\sin(k(x_n-x_{n-1}))} & \text{if } i = j = n, \\ -\frac{1}{\sin(k|x_i-x_j|)} & \text{if } |j-i| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

and the mapping  $\mathbf{Q}^{stab}$  by

$$(3.6) \quad (\mathbf{Q}^{stab} f)_i = \frac{h}{2 \tan \frac{kh}{2}} \sum_{m=i}^{\min(i+1, n)} \frac{\tan\left(k \frac{x_m - x_{m-1}}{2}\right)}{(x_m - x_{m-1})} \frac{\int_{x_{m-1}}^{x_m} f(x) dx}{(x_m - x_{m-1})}.$$

Note that this definition<sup>2</sup> can be interpreted as follows. First, one replaces the exact right-hand side  $f$  by the  $L^2$ -projection of  $f$  onto  $\mathcal{S}_h^0$ :

$$f^0(x) := P^0 f(x) := \sum_{m=1}^n \frac{\int_{x_{m-1}}^{x_m} f(x) dx}{(x_m - x_{m-1})} \chi_m(x),$$

whereas  $\chi_m$  denotes the characteristic function on the interval  $[x_{m-1}, x_m]$  and then computes as usual the right-hand side vector by applying some weighting which is related to the nonuniformity of the grid  $\tau$ . Note that in the case of a uniform grid and  $f \in \mathcal{S}_h^0$ , we have

$$(\mathbf{Q}^{stab} f)_i = \int_{\text{supp} \phi_i} f(x) \phi_i(x) dx \quad \forall i.$$

In the definition of  $\mathbf{Q}^{stab}$  we assume that  $\int_0^1 f(x) \chi_m(x) dx$  is well defined for all  $m$ , which is ensured for  $f \in L^2(\Omega)$ .

For the latter purpose we will need the following approximation properties of the  $L^2$ -projection onto  $\mathcal{S}_h^0$ :

$$\|w - P^0 w\|_s \leq ch^{t-s} \|w\|_t$$

for  $s \in \{-1, 0\}$  and  $t \in \{-1, 0, 1\}$  with  $t \geq s$ , and the pointwise error can be estimated by

$$(3.7) \quad |(w - P^0 w)(x)| \leq c\sqrt{h} \|w\|_{H^1(0,1)}.$$

<sup>2</sup>We have chosen the name  $\mathbf{G}^{stab}, \mathbf{Q}^{stab}$  for the following reason. In [15], it was shown that, for the classical Galerkin method, the pollution effect is caused by a lack of stability. The discrete inf-sup constant behaves inverse-proportionally to the wave number  $k$ .

In the following, we will show that, under certain assumptions, the finite element solution corresponding to  $\mathbf{G}^{stab}$  and  $\mathbf{Q}^{stab}$  is pollution-free. We first have to prove some estimates for the exact solution.

LEMMA 3.1. (a) *Let  $f \in H^1(0, 1)$  and  $u$  denote the corresponding solution of (3.1). Then, the following stability estimate holds:*

$$(3.8) \quad \left\| u^{(s)} \right\|_0 \leq C k^{s-2} \|f\|_{H^1(0,1)}$$

for  $s \in \{0, 1, 2\}$ .

(b) *Let  $\delta f := f - P_0 f$  denote the right-hand side of (3.1). Then, the corresponding solution  $u_0$  can be estimated by*

$$(3.9) \quad \left\| u_0^{(s)} \right\|_0 \leq C k^{s-2} \|f\|_{H^1(0,1)}$$

for  $s \in \{0, 1, 2\}$ .

*Proof.* The proof is a slight modification of the proof of Lemma 1 of [15]. We first prove (3.9).

Green's function of problem (3.1) can be written in the form

$$G(x, y) := \frac{1}{k} \begin{cases} \sin(kx) e^{iky} & \text{for } 0 \leq x \leq y, \\ \sin(ky) e^{ikx} & \text{for } y \leq x \leq 1. \end{cases}$$

Thus, the exact solution  $u_0$  can be expressed by

$$u_0(x) = \int_0^1 G(x, y) \delta f(y) dy.$$

In the next step we will estimate  $|u_0(x)|$ . First, let us assume that  $x = x_j$ , i.e., that  $x$  coincides with a grid point  $x_j$ . Using (3.7), we have

$$\begin{aligned} |u_0(x_j)| &= \left| \int_0^1 G(x_j, y) \delta f(y) dy \right| \\ &= \left| \frac{e^{ikx_j}}{k} \int_0^{x_j} \sin(ky) \delta f(y) dy + \frac{\sin(kx_j)}{k} \int_{x_j}^1 e^{iky} \delta f(y) dy \right| \\ &= \left| \frac{e^{ikx_j}}{k^2} \sum_{m=1}^j \left( -\cos(ky) \delta f(y) \Big|_{x_{m-1}}^{x_m} + \int_{x_{m-1}}^{x_m} \cos(ky) \delta f'(y) dy \right) \right. \\ &\quad \left. + \frac{\sin(kx_j)}{k^2} \sum_{m=j+1}^n \left( -ie^{iky} \delta f(y) \Big|_{x_{m-1}}^{x_m} + i \int_{x_{m-1}}^{x_m} e^{iky} \delta f'(y) dy \right) \right| \\ &\leq k^{-2} \left( 2\sqrt{h} \sum_{m=1}^n \|f\|_{H^1(x_{m-1}, x_m)} + 2\sqrt{\int_0^1 |f'(y)|^2 dy} \right) \\ &\leq k^{-2} \left( 2\sqrt{\sum_{m=1}^n \|f\|_{H^1(x_{m-1}, x_m)}^2} + 2\|f'\|_{L^2(0,1)} \right) \\ &\leq 4k^{-2} \|f\|_{H^1(0,1)}. \end{aligned}$$

The proof in the case of  $x \neq x_j$  is analogous.

The  $L^2$ -norm of  $u_0$  can therefore be estimated by

$$\|u_0\|_{L^2(0,1)} = \sqrt{\int_0^1 |u_0(x)|^2 dx} \leq \frac{c}{k^2} \|f\|_{H^1(0,1)}.$$

The estimate of  $\|u'_0\|_{L^2(0,1)}$  can be obtained in the same way by using  $\frac{d}{dx}e^{ikx} = ike^{ikx}$  and  $\frac{d}{dx}\sin(kx) = k\cos(kx)$ .

To estimate  $\|u''_0\|_{L^2(0,1)}$  we have

$$\|u''_0\|_0^2 = \int_0^1 |u''_0(x)|^2 dx = \int_0^1 |(\delta f + k^2 u_0)(x)|^2 dx \leq \|\delta f\|_0^2 + 2k^2 \|\delta f\|_0 \|u_0\|_0 + k^4 \|u_0\|_0^2.$$

Using the previous estimate and  $\|\delta f\|_0 \leq \|f\|_0$ , we conclude that

$$\|u''_0\|_0^2 \leq c \left( \|f\|_0^2 + 2\|f\|_0 \|f\|_1 + \|f\|_1^2 \right) \leq c \|f\|_1^2.$$

The estimate of  $\|u\|_s$  for  $s \in \{0, 1, 2\}$  can be obtained analogously.  $\square$

In the following we will show that the finite element solution corresponding to  $\mathbf{G}^{stab}$  and  $\mathbf{Q}^{stab}$  coincides with the interpolant of the exact solution, provided  $f \in \mathcal{S}_h^0$ . The details can be found in the following lemma.

LEMMA 3.2. (a) Let  $f^m = \chi_m$ , while  $\chi_m$  denotes the characteristic function on the interval  $[x_{m-1}, x_m]$ . Then the exact solution of the boundary value problem (3.1) is given by

$$u^m(x) := \frac{1}{k^2(x_m - x_{m-1})} \begin{cases} 2 \sin\left(k \frac{x_m - x_{m-1}}{2}\right) e^{ik \frac{x_m + x_{m-1}}{2}} \sin(kx) & \text{for } x \leq x_{m-1}, \\ -ie^{ikx_m} \sin(kx) + \cos(kx_{m-1}) e^{ikx} - 1 & \text{for } x_{m-1} < x < x_m, \\ 2 \sin\left(k \frac{x_m - x_{m-1}}{2}\right) \sin\left(k \frac{x_m + x_{m-1}}{2}\right) e^{ikx} & \text{for } x \geq x_m. \end{cases} \quad (3.10)$$

(b) Let  $f \in \mathcal{S}_h^0$  and  $u$  denote the corresponding solution of (3.1). Let  $hk < \pi$  and  $\mathbf{v} \in \mathbf{C}^n$  be defined as the solution of

$$\mathbf{G}^{stab} \mathbf{v} = \mathbf{Q}^{stab} f.$$

Let  $v := \mathcal{E} \mathbf{v}$  be the finite element solution. Then  $v$  coincides with the piecewise linear nodal interpolant of  $u$ .

*Proof.* Case (a). Statement (a) follows easily by explicitly computing

$$u^m(x) = \int_0^1 G(x, y) f^m(y) dy.$$

Case (b). By the linearity of the Helmholtz equations it is sufficient to prove the assertion only for  $f^m = \chi_m$ ,  $1 \leq m \leq n$ . Define the vector  $\mathbf{u}^m \in \mathbf{C}^n$  by

$$\mathbf{u}_i^m = u^m(x_i) \quad \forall 1 \leq i \leq n.$$

Now applying  $\mathbf{G}^{stab}$  to  $\mathbf{u}^m$  we get, after somewhat tedious algebra which is skipped here,

$$(3.11) \quad (\mathbf{G}^{stab} \mathbf{u}^m)_i = \begin{cases} 0 & \text{if } i \neq m \text{ and } i \neq m-1, \\ \frac{h}{2 \tan \frac{kh}{2}} \frac{\tan\left(k \frac{x_m - x_{m-1}}{2}\right)}{k(x_m - x_{m-1})} & \text{otherwise.} \end{cases}$$

Now computing  $\mathbf{Q}^{stab} f^m$ , we obtain

$$(\mathbf{Q}^{stab} f^m)_i = \frac{h}{2 \tan \frac{kh}{2}} \sum_{j=i}^{\min(i+1n)} \frac{\tan\left(k \frac{x_j - x_{j-1}}{2}\right)}{(x_j - x_{j-1})} \frac{\int_{x_{j-1}}^{x_j} f^m(x) dx}{(x_j - x_{j-1})}$$

$$= \begin{cases} 0 & \text{if } i \neq m \text{ and } i \neq m-1, \\ \frac{h}{2 \tan \frac{kh}{2}} \frac{\tan\left(k \frac{x_m - x_{m-1}}{2}\right)}{(x_m - x_{m-1})} & \text{otherwise.} \end{cases}$$

Thus, we conclude that  $\mathbf{G}^{stab} \mathbf{u}^m = \mathbf{Q}^{stab} f^m$  for all  $1 \leq m \leq n$ .

It remains to prove that  $\mathbf{G}^{stab}$  is regular. The matrix  $\mathbf{F}_{l,m} := (\mathbf{Q} f^m)_l$  has the form

$$\mathbf{F} = \begin{bmatrix} \star & \star & & & \\ & \star & \star & & \\ & & \star & \ddots & \\ & & & \ddots & \star \\ & & & & \star \end{bmatrix},$$

whereas the entries marked by a star are nonzero elements due to  $hk < \pi$ . All other entries of  $\mathbf{F}$  vanish. Thus,  $\mathbf{F}$  is regular, yielding that  $\mathbf{G}^{stab}$  has full rank and therefore that  $\mathbf{G}^{stab}$  is regular.  $\square$

We come now to the main result of this section, which shows that, under certain assumptions, the finite element solution corresponding to  $\mathbf{G}^{stab}$  and  $\mathbf{Q}^{stab}$  is pollution-free.

**THEOREM 3.3.** *Let the right-hand side  $f$  of (3.1) be in  $H^1(0,1)$ . Let  $\mathbf{G}^{stab}$  and  $\mathbf{Q}^{stab}$  be defined by (3.5) and (3.6) and let  $\mathbf{u}_{fe}$  denote the solution of*

$$\mathbf{G}^{stab} \mathbf{u}_{fe} = \mathbf{Q}^{stab} f.$$

*Then, the corresponding finite element solution  $u_{fe} = \mathcal{E} \mathbf{u}_{fe}$  satisfies*

$$\|(u_{fe} - u)'\|_0 \leq ch \|f\|_1,$$

*provided  $hk < \pi$ .*

*Remark 3.* Let  $f \in H^1(0,1)$  and  $u$  denote the solution of (3.1). Using (3.8) the error of the best approximation  $u_h^{opt}$  in  $\mathcal{S}_h$  with respect to the  $H^1$ -seminorm can be estimated by

$$\|(u_h^{opt} - u)'\|_0 \leq ch \|u''\|_0 \leq ch \|f\|_1.$$

According to Definition 2.1 the stabilized FEM is pollution-free.

*Proof.* Let the bilinear form  $a : V \times V \rightarrow \mathbf{C}$  corresponding to problem (3.1) be defined by (3.3) and let  $K : V \rightarrow V'$  denote the operator associated with  $a(\cdot, \cdot)$ . For  $f \in V'$ , let  $u \in V$  be the solution of

$$Ku = f.$$

Let  $P^0$  denote the  $L^2$ -projection of  $f$  onto  $\mathcal{S}_h^0$ , and  $u^0$ , the corresponding solution in  $V$ , i.e.,

$$Ku^0 = f^0$$

with  $f^0 := P^0 f$ . Using the definition of  $\mathbf{Q}^{stab}$  it follows that  $\mathbf{Q}^{stab} f = \mathbf{Q}^{stab} f^0$ . Using Lemma 3.2 we see that the finite element solution  $u_{fe}$  corresponding to  $\mathbf{G}^{stab}$  and  $\mathbf{Q}^{stab}$  coincides with the nodal interpolant  $u_{int}^0$  of  $u^0$ . Therefore, we have

$$u - u_{fe} = u - u^0 + u^0 - u_{int}^0,$$

and using the triangle inequality, it is sufficient to estimate the norms of the differences  $u - u^0$  and  $u^0 - u_{int}^0$  separately. We begin by estimating  $u - u^0$ .

We will need the following stability estimate of  $K$  which is proved in [15, Thm. 1], namely,

$$\left\| (K^{-1}g)' \right\|_0 \leq Ck \|g\|_{-1} \quad \forall g \in H^{-1}(\Omega).$$

Thus, using the approximation property of  $P_0$ , we conclude that

$$(3.12) \quad \left\| (u - u^0)' \right\|_0 \leq Ck \|f - P^0 f\|_{-1} \leq ck h^2 \|f\|_1 \leq c\pi h \|f\|_1.$$

To estimate the difference  $\left\| (u^0 - u_{int}^0)' \right\|_0$ , we proceed as follows:

$$(3.13) \quad \left\| (u^0 - u_{int}^0)' \right\|_0 \leq ch \left\| (u^0)'' \right\|_0 \leq ch \left( \left\| (u - u^0)'' \right\|_0 + \|u''\|_0 \right).$$

We have  $u - u^0 = K^{-1}(f - P^0 f)$ . Applying Lemma 3.1 for the problem  $K(u - u^0) = f - P^0 f$  yields

$$(3.14) \quad \left\| (u - u^0)'' \right\|_0 \leq C \|f\|_{H^1(0,1)}.$$

Using Lemma 3.1 again, we have

$$(3.15) \quad \|u''\|_0 \leq c \|f\|_{H^1(0,1)}.$$

Inserting (3.14) and (3.15) into (3.13) we get

$$(3.16) \quad \left\| (u^0 - u_{int}^0)' \right\|_0 \leq ch \|f\|_1.$$

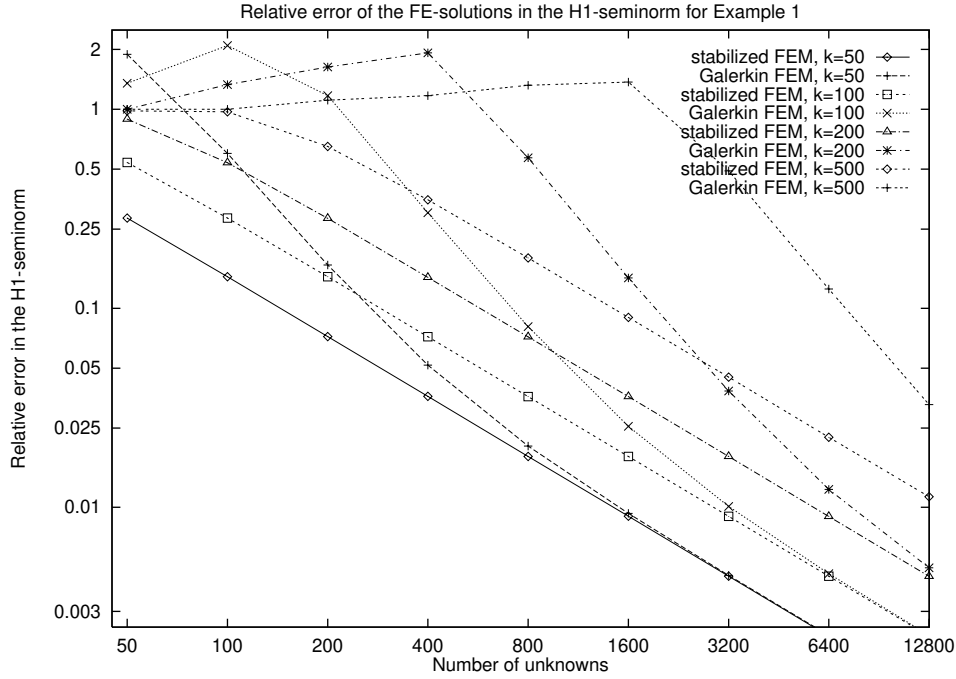
Estimating (3.12) together with (3.16) yields

$$\left\| (u - u_{fe})' \right\|_0 \leq ch \|f\|_1,$$

which completes the proof.  $\square$

In this chapter we have presented a pollution-free GFEM for a one-dimensional model problem. The principal underlying idea was to insure that if the right-hand side belongs to a sufficiently large subspace of  $V' \times H^{-1/2}(\partial B)$ , the GFEM interpolates the exact solution. Sufficiently large means that the continuous right-hand side can be approximated in this subspace consistently. Since the zero function will always belong to this subspace, the GFEM has to reproduce the fundamental system. We state that similar constructions can be made for boundary conditions possibly different from those chosen in (3.1) and also for higher-order approximations.

In the following we will show how significantly the stabilized FEM improves the classical FEM for the Helmholtz problem.

FIG. 4.1. Relative error in the  $H^1$ -seminorm for example 1.

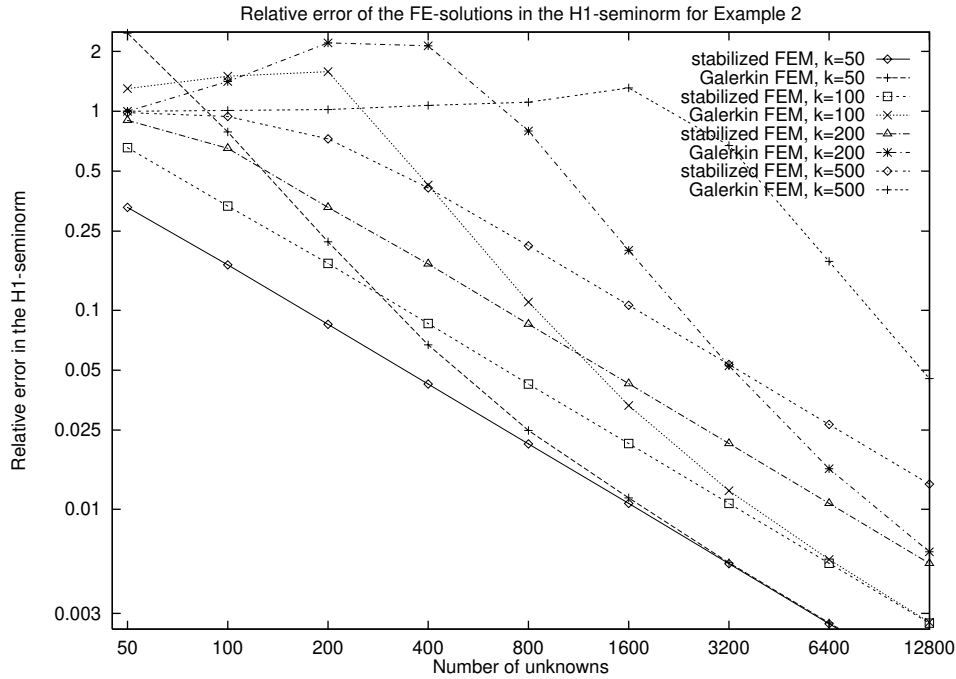
**4. Numerical examples.** In this section we will illustrate the pollution effect of the Galerkin FEM and the behavior of the stabilized FEM. We choose  $f = 1 + x^2$  as the right-hand side of (3.1). Our first example is characterized by choosing piecewise linear elements on a uniform mesh, while for example 2, we consider a highly nonuniform mesh in the following way. Let the step size  $h$  satisfy  $h^{-1} \in \mathbf{N}$ , and  $x_j = jh$  denote the grid points of the uniform mesh. We disturb these grid points randomly in the range

$$x_j^{dis} \in \left[ x_j - \frac{h}{2c}, x_j + \frac{h}{2c} \right] \cap \Omega$$

with  $c = 1.1$ . The grid of example 2 consists of the intervals  $[x_{j-1}^{dis}, x_j^{dis}]_{1 \leq j \leq h^{-1}}$ .

As explained in the first chapter, the standard Galerkin method is quasi-optimal if  $h$  is sufficiently small, but in the preasymptotic range the pollution effect influences the precision substantially. The error is always measured in the  $H^1$ -seminorm. The error of the best approximation  $u^{opt}$  differs from the error of the stabilized finite element solution  $u_{fe}^{stab}$  by less than 0.1%. Therefore, the plotted lines corresponding to  $u^{opt}$  and  $u_{fe}^{stab}$  coincide.

Figures 4.1 and 4.2 elucidate the effect of the pollution for the Galerkin method. As  $h$  becomes small, the corresponding error line approaches the error line of the best approximation, while for larger  $h$  the solution is spoiled substantially. The range of  $h$ , where the solution of the Galerkin FEM is polluted, increases with higher  $k$ . We state that even if  $kh$  is relatively large, i.e.,  $kh \in [1, 2]$ , the solution of the stabilized FEM already has the expected asymptotic behavior of  $O(kh)$ . As can be seen from Figure 4.2, the stabilized FEM works also on nonuniform grids, while the effect of the pollution is higher for the Galerkin FEM.

FIG. 4.2. Relative error in the  $H^1$ -seminorm for example 2.

The aim of our numerical investigation was to illustrate the improvement of the stabilized FEM compared with Galerkin FEM. A more thorough investigation of the pollution effect of the Galerkin FEM can be found in [15].

### 5. On the stabilization of the Helmholtz equation in two dimensions.

In this chapter we will show that in two dimensions it is impossible to eliminate the pollution effect completely. To be more concrete, we will consider a very simple model situation, by employing bilinear elements on a uniform grid. Furthermore, in order to avoid boundary effects we will consider a sequence of increasingly large domains and employ weighted norms. However, as pointed out in the introduction, this model situation is appropriate for drawing conclusions to practical applications as well.

In a mathematical setting this means that we focus on the question of whether, for such a simple model problem, it is possible to choose the coefficients of a GFEM for the *interior* difference equations such that the pollution is eliminated. If this is not possible for this simplified situation, we must expect that in more general situations it is not possible either.

However, the analysis will show that a GFEM which has *minimal* pollution must have the property that the difference equations for interior grid points satisfy

$$\sum_{j=1}^n G_{i,j} u_j = 0$$

for a maximal number of homogeneous solutions of the Helmholtz equation. Since, in contrast to the one-dimensional case, the number of homogeneous solutions is infinite, the above relation cannot be satisfied for all homogeneous solutions. It will turn out that this is the reason that the pollution effect is inevitable.

**5.1. The GFEM for the Helmholtz equation in two dimensions.** In this chapter,  $x$  always denotes a two-dimensional vector, i.e.,  $x = (x_1, x_2)^T$ . We will use two-dimensional multi-indices  $\nu \in \mathbf{Z}^2$ . The abbreviation  $|\nu|$  is defined by  $|\nu| = \nu_1 + \nu_2$ , and for a two-dimensional vector  $g$ , the notation  $g^\nu$  means  $g^\nu := g_1^{\nu_1} \cdot g_2^{\nu_2}$ .

Throughout this chapter we assume that the wave number  $k$  is bounded away from zero, i.e.,  $k \geq k_0 > 0$ .

Let the domains  $\Omega_n$  be defined by

$$(5.1) \quad \Omega_n := (-c_n, c_n)^2, \quad c_n \in \mathbf{N}, \quad c_n < c_{n+1}.$$

We consider the homogeneous Helmholtz equation on the domain  $\Omega_n$ .

$$(5.2) \quad \mathcal{L}_k u_n := -\Delta u_n - k^2 u_n = 0 \quad \text{in } \Omega_n,$$

with boundary conditions

$$(5.3) \quad B_n u_n = r_n \quad \text{on } \Gamma := \partial\Omega_n.$$

We will assume that problem (5.2), (5.3) has a unique solution  $u_n \in H^1(\Omega_n)$  for all  $r_n$ . To avoid notational technicalities we assume that the boundary conditions are not of Dirichlet type.

We consider GFEM discretizations to (5.2), (5.3) with piecewise bilinear elements. Let  $h$  denote a positive parameter denoting the step size which satisfies  $h^{-1} \in \mathbf{N}$ . The Cartesian grid points are given by  $\Theta_n = h\mathbf{Z}^2 \cap \bar{\Omega}_n$ . For convenience, we identify grid points  $x_\nu$  with the multi-index  $\nu$ . The notation  $\nu \in \Theta_n$  stands for  $\nu \in \{\mu \in \mathbf{Z}^2 \mid x_\mu \in \Theta_n\}$ . The corresponding Cartesian grid  $\tau_h$  consists of squares of side length  $h$ .

$$\begin{aligned} \square_\nu &:= h(\nu_1, \nu_1 + 1) \times h(\nu_2, \nu_2 + 1), \\ \tau_h &:= \{\square_\nu\}_{\nu \in \mathbf{Z}^2} \cap \Omega_n. \end{aligned}$$

The bilinear finite element space is denoted by  $\mathcal{S}_h$ , and the nodal basis, by  $\phi_\nu^h$ . The set of *interior* grid points is given by  $\Theta_n^{int} := h\mathbf{Z}^2 \cap \Omega_n$ . The system matrix of a GFEM method has to have the same sparsity pattern as the classical Galerkin discretization. For  $x_\nu \in \Theta_n^{int}$ , the difference equations of the corresponding Galerkin FEM can be written in the form

$$(5.4) \quad \begin{aligned} &G_2 u_{\nu_1-1, \nu_2+1} + G_1 u_{\nu_1, \nu_2+1} + G_2 u_{\nu_1+1, \nu_2+1} \\ &+ G_1 u_{\nu_1-1, \nu_2} + G_0 u_{\nu_1, \nu_2} + G_1 u_{\nu_1+1, \nu_2+1} \\ &+ G_2 u_{\nu_1-1, \nu_2-1} + G_1 u_{\nu_1, \nu_2-1} + G_2 u_{\nu_1+1, \nu_2-1} = 0. \end{aligned}$$

For convenience we have used two-dimensional indices for the vector  $u$ . Thus, the prolongation takes the form

$$\mathcal{E}[u](x) = \sum_{\nu \in \Theta} u_\nu \phi_\nu(x).$$

The interior matrix stencil

$$(5.5) \quad (\mathbf{G}_n^h)_\nu = \begin{bmatrix} G_2 & G_1 & G_2 \\ G_1 & G_0 & G_1 \\ G_2 & G_1 & G_2 \end{bmatrix}$$

expresses the sparsity of the matrix and has to be understood in the sense of (5.4). Since the matrix stencil contains nine coefficients, it is called a *nine-point stencil* (see [12]). For the Galerkin method, we have

$$(5.6) \quad \mathbf{G} = \begin{bmatrix} -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{8}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \end{bmatrix} - \alpha^2 \begin{bmatrix} \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \\ \frac{1}{9} & \frac{4}{9} & \frac{1}{9} \\ \frac{1}{36} & \frac{1}{9} & \frac{1}{36} \end{bmatrix}.$$

The GFEM is determined by defining a family of regular matrices  $\mathbf{G}_n^h$  and functionals  $\mathbf{Q}^h$  mapping the right-hand side of (5.3) onto the right-hand side vector of the linear system. According to Definition 2.5 the stencils of the GFEM have to satisfy the following conditions, A1–A2.

- A1. The matrix  $\mathbf{G}_n^h$  which depends on  $k$  is sparse in the sense that for every nodal point  $x_\nu \in \Theta^{int}$  the corresponding matrix row can be represented by a nine-point stencil.
- A2. The functional  $\mathbf{Q}^h$  is local in such a way that, for  $x_\nu \in \Theta^{int}$ , the corresponding entry of the right-hand side vector  $(\mathbf{Q}^h r_n)_\nu$  is zero.

We will impose further conditions which any reasonable GFEM should satisfy. Without the following assumptions, we would have to discuss much more pathological cases. Some detailed comments on these conditions are given in Remark 4.

- A3. The interior stencils have constant entries, i.e.,

$$(\mathbf{G}_n^h)_\nu = \begin{bmatrix} G_2 & G_1 & G_2 \\ G_1 & G_0 & G_1 \\ G_2 & G_1 & G_2 \end{bmatrix} \quad \forall x_\nu \in \Theta^{int}.$$

Furthermore, we require that the interior stencils are invariant of the domain  $\Omega$  but depend only on  $k$  and the step size  $h$ . For  $x_\nu \in \Theta_n^{int} \cap \Theta_{n'}^{int}$  and all  $h, k > 0$  we assume that the interior stencils coincide with

$$(\mathbf{G}_n^h)_\nu = (\mathbf{G}_{n'}^h)_\nu \quad \forall h, k > 0.$$

- A4. We assume that coefficients of the interior stencils of the finite element matrices  $\mathbf{G}$  (5.5) satisfy the following conditions:

- (i)  $G_0 = \sum_{m=0}^{\infty} (G_0)_m \alpha^{2m}$ ,
  - (ii)  $G_1 = \sum_{m=0}^{\infty} (G_1)_m \alpha^{2m}$ ,
  - (iii)  $G_2 = \sum_{m=0}^{\infty} (G_2)_m \alpha^{2m}$ ,
- with  $\alpha = kh$  and  $(G_t)_m$  independent of  $k$  and  $h$  for all  $t \in \{0, 1, 2\}$ ,  $m \in \mathbf{N}_0$ .

- A5. We assume that the principal part of  $\mathbf{G}$ , i.e.,

$$\mathbf{G}_{principal} := \begin{bmatrix} (G_2)_0 & (G_1)_0 & (G_2)_0 \\ (G_1)_0 & (G_0)_0 & (G_1)_0 \\ (G_2)_0 & (G_1)_0 & (G_2)_0 \end{bmatrix},$$

is an approximation of the principal part  $a_0(u, v) = \int_{\Omega} \langle \nabla u, \nabla v \rangle dx$  of order 2, implying

$$(5.7) \quad \begin{aligned} (G_0)_0 &> 0, \\ (G_0)_0 + 4((G_1)_0 + (G_2)_0) &= 0, \\ -(G_1)_0 - 2(G_2)_0 &= 1. \end{aligned}$$

These restrictions are very natural considering linear finite elements. Some comments are given in the following remark.

*Remark 4.* Condition A3 corresponds to the rotational symmetry and translation invariance of the Helmholtz equation and the mesh  $\tau_h$ .

Condition A4 reflects the fact that the Laplacian and the identity are operators of even order.

Condition A5 is the usual consistency condition if “ $-\Delta$ ” is discretized by (bi)linear elements.

Obviously, the Galerkin FEM satisfies conditions A1–A5.

**5.2. Weighted norms and the Fourier transform.** Now we will specify the norm in which the error of the approximation will be measured. In the introduction we have already explained why, for the Helmholtz problem, the  $L^2$ -norm is of main physical relevance and why the domain of computation might be much larger than the domain of interest. In this light, it is natural to introduce the following weighted  $L^2$ -norms. On a domain  $\Omega \subset \mathbf{R}^2$  the norms  $\|\cdot\|_-$  and  $\|\cdot\|_+$  are defined by

$$\|u\|_-^2 = \int_{\Omega} \frac{u(x) \bar{u}(x)}{1 + \|x\|^2} dx$$

and

$$\|w\|_+^2 = \int_{\Omega} w(x) \bar{w}(x) (1 + \|x\|^2) dx.$$

Let the space  $V_- := V_-(\Omega)$  be defined by the closure of  $C^\infty(\Omega)$  with respect to the norm  $\|\cdot\|_-$  and the space  $V_+$  correspondingly. If  $\Omega = \mathbf{R}^2$ , the norm  $\|v\|_-$  can be expressed by the Fourier transform of  $v$ . Before going into the details, we have to recall some facts about the theory of tempered distributions and the Fourier transform which will be needed below. The theory of the Fourier transform (integral) and the discrete Fourier transform could be found in [10, pp. 185 ff.], [3], [5].

Let us now define some function spaces and the integral and discrete Fourier transforms.

Let  $\mathcal{S}$  be the space of all arbitrary, often differentiable, complex-valued functions  $\phi$ , defined on  $\mathbf{R}^2$ , such that for all multi-indices  $q, k \in \mathbf{N}_0^2$ , there exist numbers  $C_{q,k}$  such that

$$\left| x^k \phi^{(q)}(x) \right| \leq C_{q,k}.$$

Let  $\mathcal{S}^*$  denote the space of tempered distributions on  $\mathcal{S}$ . If  $\phi \in \mathcal{S}$ , the Fourier transform  $\tilde{\phi}$  is defined by

$$\tilde{\phi}(\sigma) := \int_{\mathbf{R}^2} \phi(x) e^{i\langle \sigma, x \rangle} dx.$$

The Fourier transform of a distribution  $\psi$  is defined by the relation

$$(5.8) \quad \tilde{\psi}(\tilde{\phi}) = 4\pi^2 \psi(\phi) \quad \forall \phi \in \mathcal{S}.$$

To define the discrete Fourier transform, let  $\mathbf{S}^*$  denote the space of all infinite two-dimensional sequences  $\mathbf{a} = \{a_\nu\}_{\nu \in \mathbf{Z}^2}$  having the property that for every  $\mathbf{a} \in \mathbf{S}^*$  there exists a nonnegative integer  $q$  and a constant  $C$  such that

$$|a_\nu| \leq C (\|\mu\|^q + 1).$$

If  $\mathbf{a} \in \mathbf{S}^*$ , then the discrete Fourier transform of this element is defined by

$$\tilde{\mathbf{a}}(\sigma) := \sum_{\nu \in \mathbf{Z}^2} a_\nu e^{i\langle \nu, \sigma \rangle},$$

whereas  $\tilde{\mathbf{a}}(\sigma)$  is understood as a distribution over  $\mathcal{S}$ . Obviously,  $\tilde{\mathbf{a}}(\sigma)$  is a periodic function.

The relation between the norm  $\|v\|_{\pm}$  and the norm of  $\tilde{v}$  is explained in the following

LEMMA 5.1. *The norm  $\|v\|_{-}$  can be expressed by the Fourier transform  $\tilde{v}$  of  $v$ :*

$$(5.9) \quad \|v\|_{-} = \sup_{w \in H^1(\mathbf{R}^2)} \frac{\left| \int_{\mathbf{R}^2} \tilde{v}(\sigma) \bar{w}(\sigma) d\sigma \right|}{2\pi \|w\|_{H^1(\mathbf{R}^2)}}.$$

*Proof.* The relation

$$(5.10) \quad \|v\|_{-} = \sup_{w \in V_{+}} \frac{\left| \int_{\mathbf{R}^2} v(x) \bar{w}(x) dx \right|}{\|w\|_{+}}$$

is proved in the following two steps.

(i) “ $\leq$ ”: Choosing  $w = \frac{v}{1+\|x\|^2}$  results in

$$\sup_{w \in V_{+}} \frac{\left| \int_{\mathbf{R}^2} v(x) \bar{w}(x) dx \right|}{\|w\|_{+}} \geq \frac{\|v\|_{-}^2}{\|v\|_{-}} = \|v\|_{-}.$$

(ii) “ $\geq$ ”:

$$\begin{aligned} \sup_{w \in V_{+}} \frac{\left| \int_{\mathbf{R}^2} v(x) \bar{w}(x) dx \right|}{\|w\|_{+}} &= \sup_{w \in V_{+}} \frac{\left| \int_{\mathbf{R}^2} \frac{v(x)}{\sqrt{1+\|x\|^2}} \bar{w}(x) \sqrt{1+\|x\|^2} dx \right|}{\|w\|_{+}} \\ &\leq \sup_{w \in V_{+}} \frac{\left\| \frac{v}{\sqrt{1+\|\cdot\|^2}} \right\|_{L^2(\mathbf{R}^2)} \left\| \bar{w} \sqrt{1+\|\cdot\|^2} \right\|_{L^2(\mathbf{R}^2)}}{\|w\|_{+}} = \sup_{w \in V_{+}} \frac{\|v\|_{-} \|w\|_{+}}{\|w\|_{+}} = \|v\|_{-}. \end{aligned}$$

Using the well-known relation  $\|\sqrt{1+\|\cdot\|^2} w\|_{L^2(\mathbf{R}^2)} = \frac{1}{2\pi} \|\tilde{w}\|_{H^1(\mathbf{R}^2)}$  and Parseval’s equality (5.8), one concludes that the right-hand side of (5.10) coincides with

$$\sup_{w \in H^1(\mathbf{R}^2)} \frac{\left| \int_{\mathbf{R}^2} \tilde{u}(\sigma) \bar{w}(\sigma) d\sigma \right|}{2\pi \|w\|_{H^1(\mathbf{R}^2)}},$$

which completes the proof.  $\square$

**5.3. On the pollution effect of the GFEM for the Helmholtz equation in two dimensions.** In this chapter we will show that, for each GFEM, there exists a family of right-hand sides  $r_n$  for problem (5.2), (5.3) dependent on  $k$ , such that the error of the finite element solution contains a pollution term. It will turn out that, for any GFEM, we can choose values  $\beta_0, \beta_1 \in [-\pi, \pi[$  such that the generalized

finite element approximation of the following homogeneous solution to the Helmholtz equation does not converge with the rate of the best approximation:

$$(5.11) \quad u_0(x) := u_{k, \beta_0, \beta_1}(x) := \frac{k}{4\pi^2} \int_{\beta_0}^{\beta_1} e^{-ik(x_1 \cos \beta + x_2 \sin \beta)} d\beta.$$

Some properties of  $u_0$  are stated in the following lemma.

LEMMA 5.2. (a) *The Fourier transform of  $u_0$  is given by*

$$(5.12) \quad \tilde{u}_0(r \cos \beta, r \sin \beta) = \delta(r - k) \chi_{[\beta_0, \beta_1]}(\beta),$$

whereas  $\delta$  denotes the Dirac point functional, and  $\chi_{[\beta_0, \beta_1]}$ , the characteristic function on the interval  $[\beta_0, \beta_1]$ .

(b) *The function  $u_0$  satisfies the homogeneous Helmholtz equation in the whole plane.*

(c) *The function  $u_0$  belongs to  $V_-(\mathbf{R}^2)$ .*

(d) *Let the family of right-hand sides  $r_n$  be defined by*

$$(5.13) \quad r_n := r_n(k, \beta_0, \beta_1) := B_n u_{k, \beta_0, \beta_1}.$$

Then, the restriction of  $u_0$  on  $\Omega_n$  is the unique solution of the Helmholtz problem (5.2), (5.3).

*Proof.* The first statement follows by computing the inverse Fourier transform of (5.12) (cf. [10, pp. 190 ff.]).

Transforming the homogeneous Helmholtz equation on the whole plane into the Fourier image results in

$$\left(\|\sigma\|^2 - k^2\right) \tilde{u} = 0.$$

Obviously, this equation is satisfied by (5.12), yielding statement (b).

The support of the Fourier transform of  $u_0$  is given by

$$\text{supp } \tilde{u}_0 = \mathcal{A} := \left\{ \sigma \in \mathbf{R}^2 \mid \exists \beta \in [\beta_0, \beta_1], \sigma = k(\cos \beta, \sin \beta)^T \right\}.$$

In combination with Lemma 5.1, we obtain

$$(5.14) \quad \begin{aligned} \|u_0\|_- &= \sup_{v \in H^1(\mathbf{R}^2)} \frac{\left| \int_{\mathbf{R}^2} \tilde{u}_0(\sigma) \bar{v}(\sigma) d\sigma \right|}{2\pi \|v\|_{H^1(\mathbf{R}^2)}} = \sup_{v \in H^1(\mathbf{R}^2)} \frac{\left| \int_{\mathcal{A}} \bar{v}(\sigma) d\sigma \right|}{2\pi \|v\|_{H^1(\mathbf{R}^2)}} \\ &\leq \sup_{v \in H^1(\mathbf{R}^2)} \frac{\|1\|_{L^2(\mathcal{A})} \|v\|_{L^2(\mathcal{A})}}{\|v\|_{H^1(\mathbf{R}^2)}} = \sqrt{k(\beta_1 - \beta_0)} \sup_{v \in H^1(\mathbf{R}^2)} \frac{\|v\|_{L^2(\mathcal{A})}}{\|v\|_{H^1(\mathbf{R}^2)}}. \end{aligned}$$

Using the trace theorem we know that

$$\|v\|_{L^2(\mathcal{A})} \leq \tilde{C} \|v\|_{H^1(\mathcal{A}_1)} \leq \tilde{C} \|v\|_{H^1(\mathbf{R}^2)}$$

with

$$\mathcal{A}_1 = \left\{ \sigma \in \mathbf{R}^2 \mid \exists \beta \in [\beta_0, \beta_1], \kappa \in [k, k+1], \sigma = \kappa(\cos \beta, \sin \beta)^T \right\}.$$

We had imposed the general condition that  $k \geq k_0 > 0$ ; therefore, the constant  $\tilde{C}$  depends only on the length of  $\mathcal{A}$ . Combining this estimate with (5.14) results in

$$(5.15) \quad \|u_0\|_- \leq C\sqrt{k(\beta_1 - \beta_0)},$$

where consequently,  $C$  depends only on the length of  $\mathcal{A}$ .

Assertion (d) is an immediate consequence of (a), (b), (c), and our assumption on  $B_n$ .  $\square$

In the following we will use infinite matrices having a sparse and constant stencil. For this, the space  $\mathcal{M}$  of infinite, sparse matrices is defined by

$$\mathcal{M} := \left\{ \mathbf{M} : \mathbf{Z} \times \mathbf{Z} \rightarrow \mathbf{C} \mid \forall \nu \in \mathbf{Z}^2 : \mathbf{M}_\nu = \begin{bmatrix} M_2 & M_1 & M_2 \\ M_1 & M_0 & M_1 \\ M_2 & M_1 & M_2 \end{bmatrix} \right\},$$

where  $\mathbf{M}_\nu$  denotes the stencil (cf. (5.5)) of the  $\nu$ th matrix row. In the following we will identify infinite matrices with its matrix stencil.

By assumption A3 the interior stencils of the matrices  $\mathbf{G}_n^h$  are independent of  $n$ . Let  $\mathbf{G}_\infty^h \in \mathcal{M}$  be the infinite matrix consisting of the interior stencils of  $\mathbf{G}_n^h$ , i.e., for  $x_\nu \in \Theta_n^{int}$ , we put

$$(\mathbf{G}_n^h)_\nu = \begin{bmatrix} G_2 & G_1 & G_1 \\ G_1 & G_0 & G_1 \\ G_2 & G_1 & G_2 \end{bmatrix} =: (\mathbf{G}_\infty^h)_\lambda \quad \forall \lambda \in \mathbf{Z}^2.$$

**THEOREM 5.3.** *Let  $\mathbf{G}_\infty^h \in \mathcal{M}$  denote an arbitrary but fixed infinite matrix, as explained above. Then, there exists constants  $c_{\mathcal{A}}$  and  $c_s$ , independent of  $k$  and  $h$ , but possibly depending on  $(G_t)_m$  (cf. A4) and constants  $\beta_0, \beta_1 \in [-\pi, \pi]$  having the property that  $\beta_1 - \beta_0 = \frac{c_{\mathcal{A}}}{k}$  such that every solution  $\mathbf{u} \in \mathbf{S}^*$  of*

$$\mathbf{G}_\infty^h \mathbf{u} = \mathbf{0}$$

*fulfills*

$$(5.16) \quad \|u_{k,\beta_0,\beta_1} - \mathcal{E}_h \mathbf{u}\|_- \geq c_s k^{3.5} h^3,$$

*with  $u_{k,\beta_0,\beta_1}$  defined by (5.11), provided  $k^{3.5} h^3$  is bounded.*

*Proof.* The proof of this theorem is somewhat sophisticated and needs some preparatory lemmas; therefore, it is postponed to the appendix.  $\square$

In the following, the constants  $\beta_0$  and  $\beta_1$  are fixed by Theorem 5.3.

In the following, we will show that the convergence rate of the best approximation is  $(kh)^2$ . In view of (5.16) it remains to generalize Theorem 5.3 to finite domains in order to prove that any GFEM contains the pollution effect. This is done by choosing the right-hand side in (5.2), (5.3) such that the restriction of  $u_0$  to  $\Omega_n$  denotes the exact solution.

We begin by estimating the error of the interpolant of  $u_0$  to obtain an error estimate of the best approximation by using

$$\|u_0 - u_n^{opt}\|_{V_-(\Omega_n)} \leq \|u_0 - u_n^{int}\|_{V_-(\mathbf{R}^2)}.$$

**THEOREM 5.4** (interpolation error). *Let the function  $u_0$  be defined by (5.11) with  $\beta_1 - \beta_0 = \frac{c_{\mathcal{A}}}{k}$ . Then we have*

$$\|u_0 - u_n^{h,int}\|_- \leq c_{int} (kh)^2,$$

*with a constant  $c_{int}$  independent of  $k$ ,  $h$ , and  $n$ .*

*Proof.* We have

$$\begin{aligned} \|u_0 - u_n^{h,int}\|_-^2 &= \int_{\mathbf{R}^2} \frac{|(u_0 - u_n^{h,int})(x)|^2}{1 + \|x\|^2} dx = \sum_{\square_\nu \subset \tau_h} \int_{\square_\nu} \frac{|(u_0 - u_n^{h,int})(x)|^2}{1 + \|x\|^2} dx \\ &\leq \sum_{\square_\nu \subset \tau_h} \frac{1}{1 + h^2 \|\nu\|^2} \|u_0 - u_n^{h,int}\|_{L^2(\square_\nu)}^2. \end{aligned}$$

Applying standard interpolation estimates (see [7, Rem. 15.2]) yields

$$\|u_0 - u_n^{h,int}\|_{L^2(\square_\nu)} \leq ch^2 |u_0|_{H^2(\square_\nu)}.$$

We conclude that

$$\begin{aligned} \|u_0 - u_n^{h,int}\|_-^2 &\leq ch^4 \sum_{\square_\nu \subset \tau_h} \frac{1}{1 + h^2 \|\nu\|^2} \sum_{|\lambda|=2} \int_{\square_\nu} \left| \frac{\partial^2 u_0(x)}{\partial x^\lambda} \right|^2 dx \\ &\leq ch^4 \left( \sup_{0 < h < h_0} \sup_{\nu \in \mathbf{Z}^2} \frac{1 + h^2 ((\nu_1 + 1)^2 + (\nu_2 + 1)^2)}{1 + h^2 (\nu_1^2 + \nu_2^2)} \right) \sum_{|\lambda|=2} \int_{\mathbf{R}^2} \frac{\left| \frac{\partial^2 u_0(x)}{\partial x^\lambda} \right|^2}{1 + \|x\|^2} dx. \end{aligned}$$

Using the estimate

$$\begin{aligned} \sup_{0 < h < h_0} \sup_{\nu \in \mathbf{Z}^2} \frac{1 + h^2 ((\nu_1 + 1)^2 + (\nu_2 + 1)^2)}{1 + h^2 (\nu_1^2 + \nu_2^2)} &\leq \sup_{\nu \in \mathbf{Z}^2} \frac{1 + h_0^2 ((\nu_1 + 1)^2 + (\nu_2 + 1)^2)}{1 + h_0^2 (\nu_1^2 + \nu_2^2)} \\ &\leq \sup_{\nu \in \mathbf{Z}^2} \left( 1 + \frac{2h_0^2 (\nu_1 + \nu_2) + h_0^2}{1 + h_0^2 (\nu_1^2 + \nu_2^2)} \right) \\ &\leq \sup_{\nu \in \mathbf{Z}^2} \left( 1 + 2 \frac{h_0^2 (\nu_1^2 + \nu_2^2)}{1 + h_0^2 (\nu_1^2 + \nu_2^2)} + h_0^2 \right) \leq \sup_{\nu \in \mathbf{Z}^2} \left( 1 + 2 \frac{1}{h_0^{-2} (\nu_1^2 + \nu_2^2)^{-1} + 1} + h_0^2 \right) \\ &\leq 3 + h_0^2 \leq c \end{aligned}$$

results in

$$\|u_0 - u_n^{h,int}\|_- \leq ch^2 \sum_{|\lambda|=2} \left\| \frac{\partial^2 u_0}{\partial x^\lambda} \right\|_-.$$

The proof is completed by showing that

$$\left\| \frac{\partial^{|\nu|} u_0}{\partial x^\nu} \right\|_- \leq ck^{|\nu|}$$

for  $\nu \in \mathbf{N}_0^2$ ,  $|\nu| = \nu_1 + \nu_2 \leq 2$ .

In view of (5.11) we get

$$\left\| \frac{\partial^{|\nu|} u_0}{\partial x^\nu} \right\|_- = |k_1^{\nu_1} k_2^{\nu_2}| \|u_0\|_-$$

with  $(k_1, k_2)^T := k(\cos \beta, \sin \beta)^T$ . Using the relation  $\beta_1 - \beta_0 = \frac{c_A}{k}$  and (5.15), we obtain

$$\|u_0\|_- \leq \check{C}.$$

The length of the arc  $\mathcal{A}$  which appeared in the proof of Lemma 5.2 is now of order 1, and hence  $\tilde{C}$  is independent of  $k$ . The estimate  $|k_j| \leq k$  for  $j = 1, 2$  completes the proof.  $\square$

The proof of Theorem 5.6 is done by contradiction. Thus, in the following lemma, some conclusions are drawn under the assumption that there exists a pollution-free GFEM. These conclusions concern the relation of the Helmholtz problem on the finite domain  $\Omega_n$  with the one on the whole plane.

LEMMA 5.5. *Let us assume that there exists a pollution-free GFEM satisfying conditions A1–A5, i.e.,*

$$(5.17) \quad \|u_n^{ex} - u_n^h\|_{V_-(\Omega_n)} \leq c(kh)^2,$$

with a constant  $c$  independent of  $k$ ,  $h$ , and  $n$ .

(a) *Let  $r_n = B_n u_0$  be the right-hand side of the Helmholtz problem (5.2), (5.3) and  $u_n^h$  the finite element solution. Let  $U_n^h \in V_-(\mathbf{R}^2)$  be the extension of  $u_n^h$  onto the whole plane by zero, which means that*

$$U_n^h(x) := \begin{cases} u_n^h(x) & \text{if } x \in \Omega_n, \\ 0 & \text{otherwise.} \end{cases}$$

*Then, there exists a subsequence  $U_{n_j}^h$  which for  $n_j \rightarrow \infty$  converges weakly to a function  $u_\infty^h \in V_-(\mathbf{R}^2)$ .*

(b) *Furthermore, the subsequence  $u_{n_j}^h$  converges to  $u_\infty^h$  on every compact domain, implying that, for every domain  $\Omega_{n_0}$ , the following holds:*

*$\forall \epsilon > 0$ , and  $\exists j_0, \forall j \geq j_0$  we have*

$$\|u_\infty^h - u_{n_j}^h\|_{V_-(\Omega_{n_0})} \leq \epsilon.$$

(c) *Let  $\Omega_n$  be an arbitrary but fixed domain. For any  $m > n$ , the relation*

$$\sum_{\nu \in \Theta_m} (\mathbf{G}_m^h)_{\lambda, \nu} (\mathbf{u}_m^h)_\nu = 0 \quad \forall \lambda \in \Theta_n$$

*holds, with  $\mathbf{u}_m^h$  denoting the nodal values of  $u_m^h$ .*

(d) *The limit function  $u_\infty^h$  has the property that the corresponding nodal vector  $\mathbf{u}_\infty^h$  satisfies*

$$\sum_{\nu \in \mathbf{Z}^2} (\mathbf{G}_\infty^h)_{\lambda, \nu} (\mathbf{u}_\infty^h)_\nu = 0 \quad \forall \lambda \in \mathbf{Z}^2.$$

*Proof.* (a) The space  $V_-(\mathbf{R}^2)$  is a Hilbert space with scalar product

$$(u, v)_- := \int_{\mathbf{R}^2} \frac{u(x) \bar{v}(x)}{1 + \|x\|^2} dx.$$

Since the constant  $c$  of assumption (5.17) is independent of  $k$ ,  $h$ , and  $n$ , it follows that the sequence  $\|u_0 - U_n^h\|_{V_-(\mathbf{R}^2)}$  is bounded independent of  $n$ . In view of Lemma 5.2 we know that  $u_0 \in V_-(\mathbf{R}^2)$ , and consequently, the sequence  $\|U_n^h\|_{V_-(\mathbf{R}^2)}$  is bounded independent of  $n$ . Hence, we can choose a subsequence  $U_{n_j}^h$ , which for  $n_j \rightarrow \infty$  converges weakly to a function  $u_\infty^h \in V_-(\mathbf{R}^2)$ .

For simplicity we will skip the index  $j$  in the following.

(b) Let  $n_0$  be fixed and  $n > n_0$ . Using assumption (5.17), we get

$$\|u^0 - u_n^h\|_{V_-(\Omega_{n_0})} \leq C,$$

with  $C$  independent of  $n$ . The function  $u_n^h$  on  $\Omega_{n_0}$  is characterized by the (countable) number of nodal values in  $\Omega_{n_0}$ . Therefore we can choose a subsequence  $u_{n_j}^h$  which converges to  $u_\infty^h$  on  $\Omega_{n_0}$ .

(c) From assumption A2 it follows that, for all  $x_\lambda \in \Theta_n$ , the right-hand side  $(\mathbf{r}_m^h)_\lambda$  is zero, yielding statement (c).

(d) By contradiction. Assume  $\sum_{\nu \in \mathbf{Z}^2} (\mathbf{G}_\infty^h)_{\lambda,\nu} (\mathbf{u}_\infty^h)_\nu \neq 0$  for an index  $\lambda \in \mathbf{Z}^2$ . Choose  $n \in \mathbf{N}$  such that  $x_\lambda \in \Theta_n^{\text{int}}$ . Thus, by using A3, we get

$$(5.18) \quad \sum_{\nu \in \mathbf{Z}^2} (\mathbf{G}_\infty^h)_{\lambda,\nu} (\mathbf{u}_\infty^h)_\nu = \sum_{\nu \in \Theta_n} (\mathbf{G}_n^h)_{\lambda,\nu} (\mathbf{u}_\infty^h)_\nu \neq 0.$$

On the other hand, from statement (c) it follows that

$$(5.19) \quad \sum_{\nu \in \Theta_n} (\mathbf{G}_m^h)_{\lambda,\nu} (\mathbf{u}_m^h)_\nu = 0$$

for all sufficiently large indices  $m > n$ . By assumption A3 we have for sufficiently large  $m > n$  that

$$(5.20) \quad (\mathbf{G}_\infty^h)_{\lambda,\nu} = (\mathbf{G}_m^h)_{\lambda,\nu} \quad \forall \nu \in \Theta_n,$$

and therefore, (5.19) together with (5.20) imply that

$$\sum_{\nu \in \Theta_n} (\mathbf{G}_\infty^h)_{\lambda,\nu} (\mathbf{u}_m^h)_\nu = 0 \quad \forall m \text{ sufficiently large.}$$

Passing to the limit  $m \rightarrow \infty$  yields

$$\sum_{\nu \in \Omega_n} (\mathbf{G}_\infty^h)_{\lambda,\nu} (\mathbf{u}_\infty^h)_\nu = 0,$$

which contradicts (5.18).  $\square$

We are now able to prove that, for every GFEM, there exists a sequence of domains  $\Omega_n$  with  $n$  dependent on  $k$  and  $h$  and a family of right-hand sides for the Helmholtz equation (5.2), (5.3) such that the error of the corresponding finite element solution contains a pollution term.

**THEOREM 5.6.** *For every GFEM which satisfies conditions A1–A5, there exists a family of domains  $\Omega_n$  and right-hand sides  $r_n$  with  $n = n(k, h)$  for (5.2), (5.3) such that the error of the finite element solution  $u_n^h$  compared with the exact solution  $u_n^{\text{ex}}$  can be estimated from below by*

$$(5.21) \quad \|u_n^{\text{ex}} - u_n^h\|_{V_-(\Omega_n)} \geq C k^{3.5} h^3,$$

provided  $k^{3.5} h^3 \leq C$ .

The error of the best approximation  $u_n^{h,\text{opt}} \in \mathcal{S}_h$  of  $u_n^{\text{ex}}$  with respect to the  $\|\cdot\|_-$ -norm can be estimated by

$$\|u_n^{\text{ex}} - u_n^{h,\text{opt}}\|_{V_-(\Omega_n)} \leq C (kh)^2.$$

Hence, for  $k \rightarrow \infty$  and  $h$  chosen such that  $k^{3.5}h^3 = 1$ , the error of the best approximation tends to zero, while the error of the finite element solution is larger than  $C$ . Therefore, the pollution effect is unavoidable in two dimensions.

*Proof.* The proof is given by contradiction. Let us assume that there exists regular matrices  $\mathbf{G}_n^h$  corresponding to the domains  $\Omega_n$  with  $n = n(k, h)$ , such that for every right-hand side  $r_n$  and corresponding exact solution  $u_n^{ex}$  the error of the finite element solution  $u_n^h$  can be estimated by

$$(5.22) \quad \|u_n^{ex} - u_n^h\|_- \leq c(kh)^2,$$

with a constant  $c$  independent of  $k$ ,  $h$ , and  $n$ .

Let the right-hand side of the Helmholtz equation be given by (5.13) with  $\beta_1 - \beta_0 = \frac{c_A}{k}$ . Then,  $u_0$  defined by (5.11) denotes the exact solution.

Let  $u_\infty^h := \lim_{n \rightarrow \infty} u_n^h$  denote the limit of the finite element solutions, as explained in Lemma 5.5. Therefore, for every domain  $\Omega_{n_0}$ , there exists  $n > n_0$  such that

$$(5.23) \quad \|u_\infty^h - u_n^h\|_{V_-(\Omega_{n_0})} \leq \frac{c_s}{3} k^{3.5} h^3.$$

From assumption (5.22) and Lemma 5.5 it follows that  $u_0 - u_\infty^h \in V_-(\mathbf{R}^2)$ . Hence, for every  $\epsilon > 0$  there exists  $n_0$  such that

$$\|u_0 - u_\infty^h\|_{V_-(\mathbf{R}^2 \setminus \Omega_{n_0})} \leq \epsilon.$$

Choosing  $\epsilon = \frac{c_s}{3} k^{3.5} h^3$  with  $c_s$  defined in Theorem 5.3 and sufficiently large  $n_0$ , possibly dependent on  $k$  and  $h$ , we obtain that, for all sufficiently large  $n \geq n_0$ , estimate (5.23) and

$$\|u_0 - u_\infty^h\|_{V_-(\mathbf{R}^2 \setminus \Omega_{n_0})} \leq \frac{c_s}{3} k^{3.5} h^3$$

are satisfied. Using Theorem 5.3 and the estimate above we conclude that

$$\begin{aligned} \|u_0 - u_n^h\|_{V_-(\Omega_n)} &\geq \|u_0 - u_n^h\|_{V_-(\Omega_{n_0})} \geq \|u_0 - u_\infty^h\|_{V_-(\Omega_{n_0})} - \|u_\infty^h - u_n^h\|_{V_-(\Omega_{n_0})} \\ &\geq \|u_0 - u_\infty^h\|_{V_-(\mathbf{R}^2)} - \|u_0 - u_\infty^h\|_{V_-(\mathbf{R}^2 \setminus \Omega_{n_0})} - \frac{c_s}{3} k^{3.5} h^3 \geq \frac{c_s}{3} k^{3.5} h^3 \end{aligned} \quad (5.24)$$

is satisfied. Combining assumptions (5.22) and (5.24), we obtain

$$(5.25) \quad \frac{c_s}{3} k^{3.5} h^3 \leq c(hk)^2.$$

Let  $k \rightarrow \infty$  and  $h$  be chosen such that  $k^{3.5}h^3 = 1$ , i.e.,  $h = k^{-3.5/3}$ . Therefore,  $(hk)^2 = k^{-1/3}$  tends to zero for  $k \rightarrow \infty$ . Hence, for  $k \rightarrow \infty$ , the left-hand side of (5.25) is  $c_s/3$ , while the right-hand side tends to zero, which contradicts our assumption.  $\square$

This theorem shows that the pollution effect is unavoidable in two dimensions. However, this theorem does not make any assertion about the *size* of the pollution for a *fixed* domain. From this theorem it is clear that a GFEM which satisfies

$$\|u_{k,\beta_0,\beta_1} - u_{fe}^h\|_- \leq Ck^{3.5}h^3$$

has “optimal” interior stencils. In the following appendix it will be explained how such stencils can be constructed. These insights have been used in [6] to design a GFEM with minimal pollution.

**6. Appendix.** In this appendix we will prove Theorem 5.3.

We use the notations introduced in section 3. We will consider an infinite matrix  $\mathbf{G} \in \mathcal{M}$  that has a constant nine-point stencil,

$$(6.1) \quad \mathbf{G} = \begin{bmatrix} G_1 & G_2 & G_2 \\ G_1 & G_0 & G_1 \\ G_2 & G_1 & G_2 \end{bmatrix},$$

and that satisfies conditions A1–A5. Let  $\mathbf{u}_h \in \mathbf{S}^*$  denote a solution of

$$(6.2) \quad \mathbf{G}\mathbf{u}_h = \mathbf{0},$$

which is identified with a finite element function  $u_h \in \mathcal{S}_h$  via  $u_h := \mathcal{E}_h \mathbf{u}_h$ . We will discuss here the question of which precision functions of the type

$$u_0(x) = \frac{k}{4\pi} \int_{\beta_0}^{\beta_1} e^{-ik(x_1 \cos \beta + x_2 \sin \beta)} d\beta$$

can be approximated by solutions of (6.2). We recall the definition of the norms  $\|\cdot\|_{\pm}$  (see section 5.2). Using Lemma 5.1, the error  $u_0 - u_h$  can be expressed by

$$(6.3) \quad \|u_0 - u_h\|_- = \sup_{w \in H^1(\mathbf{R}^2)} \frac{|\int_{\mathbf{R}^2} (\tilde{u}_0 - \tilde{u}_h)(\sigma) \bar{w}(\sigma) d\sigma|}{2\pi \|w\|_{H^1(\mathbf{R}^2)}}.$$

In view of (6.3) we will now compute the Fourier transformation of a finite element function  $u_h$  corresponding to a solution  $\mathbf{u}_h$  of (6.2), while  $\tilde{u}_0$  is given by (5.12).

LEMMA 6.1. *The discrete Fourier transform of any solution  $\tilde{\mathbf{u}}_h \in \mathbf{S}^*$  of (6.2) satisfies*

$$(6.4) \quad \hat{g}(\sigma) \tilde{\mathbf{u}}_h(\sigma) = 0,$$

with

$$\hat{g}(\sigma) := G_0 + 2G_1(\cos \sigma_1 + \cos \sigma_2) + 4G_2 \cos \sigma_1 \cos \sigma_2$$

in the distributional sense.

*Proof.* Let  $e_1 = (1, 0)^T$  and  $e_2 = (0, 1)^T$ . The proof follows from

$$\begin{aligned} \widetilde{\mathbf{G}\mathbf{u}_h}(\sigma) &= \sum_{\nu \in \mathbf{Z}^2} (\mathbf{G}\mathbf{u}_h)_{\nu} e^{i\langle \sigma, \nu \rangle} \\ &= \sum_{\nu \in \mathbf{Z}^2} (G_0(\mathbf{u}_h)_{\nu} + G_1((\mathbf{u}_h)_{\nu+e_1} + (\mathbf{u}_h)_{\nu-e_1} + (\mathbf{u}_h)_{\nu+e_2} + (\mathbf{u}_h)_{\nu-e_2}) \\ &\quad + G_2((\mathbf{u}_h)_{\nu-e_1-e_2} + (\mathbf{u}_h)_{\nu+e_1-e_2} + (\mathbf{u}_h)_{\nu-e_1+e_2} + (\mathbf{u}_h)_{\nu+e_1+e_2})) e^{i\langle \sigma, \nu \rangle} \\ &= G_0 \tilde{\mathbf{u}}_h(\sigma) + G_1(e^{-i\langle \sigma, e_1 \rangle} \tilde{\mathbf{u}}_h(\sigma) + e^{i\langle \sigma, e_1 \rangle} \tilde{\mathbf{u}}_h(\sigma) + e^{-i\langle \sigma, e_2 \rangle} \tilde{\mathbf{u}}_h(\sigma) \\ &\quad + e^{i\langle \sigma, e_2 \rangle} \tilde{\mathbf{u}}_h(\sigma)) \\ &\quad + G_2(e^{i\langle \sigma, e_1+e_2 \rangle} \tilde{\mathbf{u}}_h(\sigma) + e^{i\langle \sigma, -e_1+e_2 \rangle} \tilde{\mathbf{u}}_h(\sigma) \\ &\quad + e^{i\langle \sigma, e_1-e_2 \rangle} \tilde{\mathbf{u}}_h(\sigma) + e^{i\langle \sigma, -e_1-e_2 \rangle} \tilde{\mathbf{u}}_h(\sigma)) \\ &= \hat{g}(\sigma) \tilde{\mathbf{u}}_h(\sigma). \quad \square \end{aligned}$$

From assumption A5, it follows that  $G_0 \neq 0$  is fulfilled for sufficiently small  $kh$ ; thus, the function  $g$  is well defined:

$$g(\sigma) = 4 + 2g_1(\cos \sigma_1 + \cos \sigma_2) + 4g_2 \cos \sigma_1 \cos \sigma_2,$$

with

$$g_1 = 4 \frac{G_1}{G_0}, \quad g_2 = 4 \frac{G_2}{G_0}.$$

Using A5, the coefficients  $g_1$  and  $g_2$  can be expanded accordingly:

$$\begin{aligned} \text{(a)} \quad g_1 &= \sum_{m=0}^{\infty} b_m \alpha^{2m}, \\ \text{(b)} \quad g_2 &= \sum_{m=0}^{\infty} c_m \alpha^{2m}, \end{aligned}$$

with  $\alpha = kh$ .

Note that conditions (5.7) imply that

$$\begin{aligned} 1 + b_0 + c_0 &= 0, \\ b_0 + 2c_0 &\neq 0. \end{aligned} \quad (6.5)$$

In view of (6.4) we conclude that

$$\text{supp } \tilde{\mathbf{u}}_h \subset \mathcal{N}_{\mathbf{G}}^1 := \{\sigma \in \mathbf{R}^2 \mid g(\sigma) = 0\}.$$

For later use we define the scaled set  $\mathcal{N}_{\mathbf{G}}^h$  by

$$\mathcal{N}_{\mathbf{G}}^h := \{\sigma \in \mathbf{R}^2 \mid g(h\sigma) = 0\}.$$

The relation of the discrete Fourier transform of a solution  $\mathbf{u}_h$  of (6.2) and the (integral) Fourier transform of the corresponding finite element function is discussed in the following lemma.

LEMMA 6.2. *Let  $u_h = \mathcal{E}_h \mathbf{u}_h$  denote the finite element function corresponding to a solution  $\mathbf{u}_h$  of (6.2). Then, we have*

$$\text{supp } \tilde{u}_h \subset \mathcal{N}_{\mathbf{G}}^h.$$

*Proof.* The inverse of the discrete Fourier transform is given by

$$(\mathbf{u}_h)_\nu = \frac{1}{4\pi^2} \int_{[-\pi, \pi]^2} \tilde{\mathbf{u}}_h(\sigma) e^{-i\langle \nu, \sigma \rangle} d\sigma.$$

Therefore, the Fourier transform of the corresponding finite element function can be written in the following form:

$$\tilde{u}_h(\sigma) = \sum_{\nu \in \mathbf{Z}^2} \tilde{\phi}_\nu^h(\sigma) \frac{1}{4\pi^2} \int_{[-\pi, \pi]^2} \tilde{\mathbf{u}}_h(s) e^{-i\langle \nu, s \rangle} ds, \quad (6.6)$$

with  $\phi_\nu^h$  denoting the bilinear basis functions. Explicit calculations yield that

$$\tilde{\phi}_\nu^h(\sigma) = \frac{16 \sin^2 \frac{h\sigma_1}{2} \sin^2 \frac{h\sigma_2}{2}}{h^2 \sigma_1^2 \sigma_2^2} e^{ih\langle \sigma, \nu \rangle}. \quad (6.7)$$

Inserting (6.7) into (6.6) results in

$$\tilde{u}_h(\sigma) = \frac{16 \sin^2 \frac{h\sigma_1}{2} \sin^2 \frac{h\sigma_2}{2}}{h^2 \sigma_1^2 \sigma_2^2} \int_{[-\pi, \pi]^2} \tilde{\mathbf{u}}_h(s) \left( \frac{1}{4\pi^2} \sum_{\nu \in \mathbf{Z}^2} e^{i\langle \nu, h\sigma - s \rangle} \right) ds.$$

Using the well-known relation

$$\frac{1}{4\pi^2} \sum_{\nu \in \mathbf{Z}^2} e^{i\langle \nu, \sigma - s \rangle} = \sum_{\nu \in \mathbf{Z}^2} \delta(\sigma - s + 2\pi\nu),$$

whereas  $\delta$  denotes the Dirac point functional, one obtains

$$\begin{aligned}\tilde{u}_h(\sigma) &= \frac{16 \sin^2 \frac{h\sigma_1}{2} \sin^2 \frac{h\sigma_2}{2}}{h^2 \sigma_1^2 \sigma_2^2} \sum_{\nu \in \mathbf{Z}^2} \int_{[-\pi, \pi]^2} \tilde{\mathbf{u}}_h(s) \delta(h\sigma - s + 2\pi\nu) ds \\ &= \frac{16 \sin^2 \frac{h\sigma_1}{2} \sin^2 \frac{h\sigma_2}{2}}{h^2 \sigma_1^2 \sigma_2^2} \sum_{\nu \in \mathbf{Z}^2} \tilde{\mathbf{u}}_h(h\sigma + 2\pi\nu).\end{aligned}$$

If  $\sigma \notin \mathcal{N}_{\mathbf{G}}^h$  then  $h\sigma \notin \mathcal{N}_{\mathbf{G}}^1$ . Using the periodicity of  $g(\sigma)$  it is obvious that  $h\sigma \notin \mathcal{N}_{\mathbf{G}}^1$  implies  $h\sigma + 2\pi\nu \notin \mathcal{N}_{\mathbf{G}}^1$  resulting in  $\tilde{\mathbf{u}}_h(h\sigma + 2\pi\nu) = 0$ . Consequently, we conclude that if  $\sigma \notin \mathcal{N}_{\mathbf{G}}^h$  then  $\tilde{u}_h(\sigma) = 0$ , which completes the proof.  $\square$

The norm  $\|u_0 - u_h\|_-$  will be estimated as follows. For given GFEM we will choose  $u_0$ , i.e.,  $\beta_0, \beta_1$ , such that  $\text{supp } \tilde{u}_0 \cap \text{supp } \tilde{u}_h = \emptyset$  and a function  $\eta \in H^1(\mathbf{R}^2)$ , having the property that

$$\mathcal{D} := \text{supp } \eta \cap \mathcal{N}_G^h = \emptyset.$$

Under these assumption and taking into account (6.3) and (5.12), the norm  $\|u_0 - u_h\|_-$  can be estimated from below by

$$\|u_0 - u_h\|_- \geq \frac{\left| \int_{\text{supp } \tilde{u}_0} \bar{\eta}(\sigma) d\sigma \right|}{2\pi \|\eta\|_{H^1(\mathbf{R}^2)}}.$$

To determine the domain  $\mathcal{D}$  we will use the following lemma.

LEMMA 6.3. *Let  $\mathbf{G} \in \mathcal{M}$  be an arbitrary but fixed matrix, fulfilling A1–A5. We assume that  $kh$  is sufficiently small. Then there exist positive constants  $c_0$  and  $c_s$ , independent of  $h$  and  $k$ , but possibly dependent on the stencils  $(\mathbf{G}_t)_m$  (cf. A4) and constants  $\tilde{\beta}_0, \tilde{\beta}_1 \in [-\pi, \pi]$  having the property that*

$$\tilde{\beta}_0 - \tilde{\beta}_1 \geq c_0$$

such that

$$\mathcal{D}_1 := \left\{ \sigma = r \begin{pmatrix} \cos \beta \\ \sin \beta \end{pmatrix} \mid \forall r \in [kh - c_s k^7 h^7, kh + c_s k^7 h^7], \beta \in [\tilde{\beta}_0, \tilde{\beta}_1] \right\} \cap \mathcal{N}_{\mathbf{G}}^1 = \emptyset.$$

*Proof.* We have to show that  $\sigma \in \mathcal{N}_{\mathbf{G}}^1$  implies that  $\sigma \notin \mathcal{D}_1$ . Therefore, we investigate the roots of  $g(\sigma)$ . The zeros of  $g$  are  $2\pi$ -periodic; i.e., if  $g(\sigma) = 0$ , then  $g(\sigma + 2\pi\nu) = 0$  for all  $\nu \in \mathbf{Z}^2$ . In view of the definition of  $\mathcal{D}_1$ , we are interested only in the zeros of  $g$  which are of order  $hk$ , i.e., are small. We make the following ansatz using the abbreviation  $\alpha := kh$ :

$$(6.8) \quad \sigma = r(\beta, \alpha) \begin{pmatrix} \cos \beta \\ \sin \beta \end{pmatrix}$$

with  $r : [-\pi, \pi[ \times \mathbf{R}_+ \rightarrow \mathbf{R}_+$ ,

$$(6.9) \quad r(\beta, \alpha) = \alpha + \sum_{m=1}^{\infty} r_m(\beta) \alpha^{2m+1}.$$

To simplify the notation we write  $r = r(\beta, \alpha)$ ,  $r_m = r_m(\beta)$ . Formally, we set  $r_0 = 1$ . We make use of the abbreviations

$$\kappa_n = \frac{(-\cos^2 \beta)^n}{(2n)!}, \quad \lambda_n = \frac{(-\sin^2 \beta)^n}{(2n)!}, \quad (\kappa \star \lambda)_n = \sum_{m=0}^n \kappa_m \lambda_{n-m},$$

and

$$\rho_{2n,m} = \begin{cases} \delta_{0,m} & \text{if } n = 0, \\ \left( \underbrace{r \star r \star \cdots \star r}_{2n\text{-fold convolution}} \right)_m & \text{otherwise,} \end{cases}$$

with  $\delta_{n,m}$  denoting the Kronecker delta and  $r = (r_0, r_1, \dots)^T$ . We state that

$$\left( \alpha + \sum_{m=1}^{\infty} r_m \alpha^{2m+1} \right)^{2n} = \alpha^{2n} \sum_{m=0}^{\infty} \rho_{2n,m} \alpha^{2m},$$

which will be used later.

Using ansatz (6.8), the condition  $g(\sigma) = 0$  is equivalent to

$$g(r \cos \beta, r \sin \beta) = 4 + 2g_1(\cos(r \cos \beta) + \cos(r \sin \beta)) + 4g_2 \cos(r \cos \beta) \cos(r \sin \beta) = 0.$$

Replacing  $\cos(r \cos \varphi)$  and  $\cos(r \sin \varphi)$  by the corresponding Taylor series about  $r = 0$  and inserting expansions (a), (b), and (6.9) results in

$$\begin{aligned} & 4 + 2g_1 \sum_{n=0}^{\infty} (\kappa_n + \lambda_n) r^{2n} + 4g_2 \sum_{n=0}^{\infty} \kappa_n r^{2n} \sum_{n=0}^{\infty} \lambda_n r^{2n} \\ &= 4 + 2 \sum_{l=0}^{\infty} b_l \alpha^{2l} \sum_{n=0}^{\infty} (\kappa_n + \lambda_n) \alpha^{2n} \sum_{m=0}^{\infty} \rho_{2n,m} \alpha^{2m} \\ &\quad + 4 \sum_{l=0}^{\infty} c_l \alpha^{2l} \sum_{n=0}^{\infty} (\kappa \star \lambda)_n \alpha^{2n} \sum_{m=0}^{\infty} \rho_{2n,m} \alpha^{2m} \\ &= 4 + 2 \sum_{l=0}^{\infty} b_l \alpha^{2l} \sum_{n=0}^{\infty} \alpha^{2n} \sum_{m=0}^n (\kappa_m + \lambda_m) \rho_{2m,n-m} \\ &\quad + 4 \sum_{l=0}^{\infty} c_l \alpha^{2l} \sum_{n=0}^{\infty} \alpha^{2n} \sum_{m=0}^n (\kappa \star \lambda)_m \rho_{2m,n-m} \\ &= 4 + 4 \sum_{l=0}^{\infty} \alpha^{2l} \left( \sum_{m=0}^l b_{l-m} \sum_{n=0}^m \frac{\kappa_n + \lambda_n}{2} \rho_{2n,m-n} \right. \\ &\quad \left. + \sum_{m=0}^l c_{l-m} \sum_{n=0}^m (\kappa \star \lambda)_n \rho_{2n,m-n} \right) \\ &= 4 + 4 \sum_{l=0}^{\infty} \alpha^{2l} \left( \sum_{n=0}^l \sum_{m=n}^l b_{l-m} \frac{\kappa_n + \lambda_n}{2} \rho_{2n,m-n} + c_{l-m} (\kappa \star \lambda)_n \rho_{2n,m-n} \right) \\ &= 4 + 4 \sum_{l=0}^{\infty} \alpha^{2l} \sum_{n=0}^l \sum_{m=0}^{l-n} \rho_{2n,m} \left( b_{l-n-m} \frac{\kappa_n + \lambda_n}{2} + c_{l-n-m} (\kappa \star \lambda)_n \right) \stackrel{!}{=} 0. \end{aligned}$$

We conclude that the condition “ $\stackrel{!}{=}$ ” is equivalent to the conditions

$$(6.10) \quad \gamma_0 := 4 + 4(b_0 + c_0) = 0$$

and

$$(6.11) \quad \gamma_l := \sum_{n=0}^l \sum_{m=0}^{l-n} \rho_{2n,m} \left( b_{l-n-m} \frac{\kappa_n + \lambda_n}{2} + c_{l-n-m} (\kappa \star \lambda)_n \right) = 0 \quad \forall l \geq 1.$$

Condition (6.10) is fulfilled by using (6.5). Condition (6.11) can be rewritten in the form

$$(6.12) \quad -\frac{r_{l-1}}{2} (b_0 + 2c_0) + l.o.t = 0.$$

(The abbreviation *l.o.t.* denotes the remaining sum of (6.11) containing only functions  $r_j$  with  $j < l - 1$ .) In view of (6.5), relation (6.12) serves as a recursion formula for the functions  $r_j$ .

In the following, we will show that it is impossible to choose the stencil coefficients  $(\mathbf{G}_t)_m$  of A4 (or equivalently the coefficients  $b_m, c_m$  of  $g_{1,2}$ ) such that all coefficients  $r_j(\beta)$  of expansion (6.9) vanish.

For  $l = 0$  and  $l = 2$ , conditions (6.10) and (6.11) can be written in the explicit form.

$l = 0 :$

$$(6.13) \quad 1 + (b_0 + c_0) = 0.$$

$l = 2 :$

$$(6.14) \quad b_2 + c_2 - \frac{b_1 + 2c_1}{4} + \frac{3(b_0 + 4c_0) + (b_0 - 4c_0) \cos 4\beta}{192} = \frac{r_1(\beta)}{2} (b_0 + 2c_0).$$

A necessary condition for  $r_1(\beta) \equiv 0$  is that (6.13) and  $b_0 - 4c_0 = 0$  are satisfied, yielding

$$(6.15) \quad b_0 = -\frac{4}{5}, \quad c_0 = -\frac{1}{5}.$$

Inserting (6.15), condition (6.11) for  $l = 1$  and  $l = 3$  takes the following form.

$l = 1 :$

$$(6.16) \quad b_1 + c_1 + \frac{3}{10} = 0.$$

$l = 3 :$

$$b_3 + c_3 - \frac{b_2 + 2c_2}{4} + \frac{3(b_1 + 4c_1) + (b_1 - 4c_1) \cos(4\beta)}{192} + \frac{5 - \cos(4\beta)}{4800} = -\frac{3}{5} r_2(\beta).$$

Again, a necessary condition for  $r_2(\beta) \equiv 0$  is that (6.16) and

$$\frac{(b_1 - 4c_1) \cos(4\beta)}{192} - \frac{\cos(4\beta)}{4800} = 0$$

hold, resulting in

$$(6.17) \quad b_1 = -\frac{29}{125}, \quad c_1 = -\frac{17}{250}.$$

Using (6.15) and (6.17), condition (6.11) for  $l = 2$  and  $l = 4$  can be written in the following form.

$l = 2 :$

$$b_2 + c_2 + \frac{67}{1000} = 0.$$

$l = 4 :$

$$(6.18) \quad \begin{aligned} \frac{3}{5}r_3(\beta) = & \frac{1427}{4608000} + b_4 + c_4 - \frac{b_3 + 2c_3}{4} + \frac{b_2 + 4c_2}{64} \\ & - \left( \frac{131}{1920000} - \frac{b_2 - 4c_2}{192} \right) \cos(4\beta) - \frac{\cos(8\beta)}{1290240}. \end{aligned}$$

Now, it is impossible to choose  $b_2$  and  $c_2$  in such a way that  $r_3(\beta)$  vanishes identically. That means for any stencil  $(\mathbf{G}_t)_m$  there exist values of  $\beta$  such that  $r_3(\beta) \neq 0$ . By easy analysis one obtains that the number of extrema of  $r_3(\beta)$  is bounded by 16. Therefore, it is possible to choose  $\hat{\beta}_0$  and  $\hat{\beta}_1$  with  $\hat{\beta}_1 - \hat{\beta}_0 \geq \frac{2\pi}{32}$  such that

$$\sup_{\beta \in [\hat{\beta}_0, \hat{\beta}_1]} |r_3(\beta)| \geq \hat{c},$$

while  $\hat{c}$  is independent of  $h$  and  $k$ , but possibly depends on  $(\mathbf{G}_t)_m$ . Consequently, if  $\alpha = hk$  is sufficiently small, the function  $r$  of (6.9) can be estimated by

$$|r(\beta, \alpha) - \alpha| \geq c_s \alpha^7$$

for all  $\beta \in [\tilde{\beta}_0, \tilde{\beta}_1]$  with  $\tilde{\beta}_0, \tilde{\beta}_1 \in [\hat{\beta}_0, \hat{\beta}_1]$  and  $\tilde{\beta}_1 - \tilde{\beta}_0 \geq c$ , while  $c$  and  $c_s$  do not depend on  $k$  and  $h$ .  $\square$

The set  $\mathcal{N}_{\mathbf{G}}^h$  is defined by a suitable scaling of  $\mathcal{N}_{\mathbf{G}}^1$ ; thus, by using the previous lemma, it follows that the scaled domain

$$\mathcal{D}_h := \left\{ \sigma = r \begin{pmatrix} \cos \beta \\ \sin \beta \end{pmatrix} \middle| \forall r \in [k - c_s k^7 h^6, kh + c_s k^7 h^6], \beta \in [\beta_0, \beta_1] \right\}$$

satisfies

$$\mathcal{D}_h \cap \mathcal{N}_{\mathbf{G}}^h = \emptyset.$$

We are now able to prove Theorem 5.3.

*Proof of Theorem 5.3.* Let  $\tilde{\beta}_0, \tilde{\beta}_1$  be defined as in the proof of Lemma 6.3. We had assumed that the wave number  $k \geq k_0 > 0$  is bounded from below, and hence that there exists  $\beta_0, \beta_1 \in [\tilde{\beta}_0, \tilde{\beta}_1]$  with

$$\beta_1 - \beta_0 = \frac{c_A}{k}.$$

Let the function  $\eta$  be defined by

$$\eta(\sigma) = \eta(r \cos \beta, r \sin \beta) := \rho(r) \chi(\beta)$$

with

$$\rho(r) := \begin{cases} \frac{r-k+\delta}{\delta} & \text{if } r \in [k-\delta, k], \\ \frac{\delta+k-r}{\delta} & \text{if } r \in [k, k+\delta], \\ 0 & \text{otherwise,} \end{cases}$$

whereas  $\delta := c_s k^7 h^6$  with  $c_s$  from Lemma 6.3 and

$$\chi(\beta) := \begin{cases} \sin\left(\frac{\beta-\beta_1}{\beta_0-\beta_1}\pi\right) & \text{if } \beta \in [\beta_0, \beta_1], \\ 0 & \text{otherwise.} \end{cases}$$

The function  $\eta$  has the property that  $\text{supp } \eta \subset \mathcal{D}_h$ . We will use the function  $\eta$  to estimate the right-hand side of

$$\begin{aligned} \|u_0 - u_h\|_- &= \sup_{w \in H^1(\mathbf{R}^2)} \frac{\left| \int_{\mathcal{A}_k} \tilde{u}_0(\sigma) \bar{w}(\sigma) d\sigma - \int_{\text{supp} \tilde{u}_h} \tilde{u}_h(\sigma) \bar{w}(\sigma) d\sigma \right|}{2\pi \|w\|_{H^1(\mathbf{R}^2)}} \\ &\geq \frac{\left| \int_{\mathcal{A}_k} \tilde{u}_0(\sigma) \eta(\sigma) d\sigma - \int_{\text{supp} \mathcal{N}_{\mathbf{G}}^h} \tilde{u}_h(\sigma) \eta(\sigma) d\sigma \right|}{2\pi \|\eta\|_{H^1(\mathbf{R}^2)}}, \end{aligned}$$

while the set  $\mathcal{A}_k$  is defined by

$$\mathcal{A}_k := \left\{ \sigma = k \begin{pmatrix} \cos \beta \\ \sin \beta \end{pmatrix}, \forall \beta \in [\beta_0, \beta_1] \right\}.$$

We have that  $\text{supp } \tilde{u}_h \subset \mathcal{N}_{\mathbf{G}}^h$ , and by construction,  $\text{supp } \eta \cap \mathcal{N}_{\mathbf{G}}^h = \emptyset$ . Using (5.12), one obtains

$$\|u_0 - u_h\|_- \geq \frac{\left| \int_{\mathcal{A}_k} \eta(\sigma) d\sigma \right|}{2\pi \|\eta\|_{H^1(\mathbf{R}^2)}}.$$

The proof of Theorem 5.3 is given by showing that the function  $\eta$  satisfies

$$(6.19) \quad \left| \int_{\mathcal{A}_k} \eta(\sigma) d\sigma \right| = \frac{2c_{\mathcal{A}}}{\pi}$$

and

$$\begin{aligned} \|\eta\|_{H^1(\mathbf{R}^2)} &= \sqrt{\left( \frac{\delta}{3} + \frac{1}{\delta} \right) c_{\mathcal{A}} + \frac{k}{c_{\mathcal{A}}} \cdot \frac{(k+\delta)^2 \log \frac{k+\delta}{k} - (k-\delta)^2 \log \frac{k-\delta}{k} - 2k\delta}{\delta^2} \frac{\pi^2}{2}} \\ (6.20) \quad &= \sqrt{\frac{c_{\mathcal{A}}}{\delta} + \frac{c_{\mathcal{A}}^2 + \pi^2}{3c_{\mathcal{A}}} \delta + \delta O\left(\left(\frac{\delta}{k}\right)^2\right)} \leq \frac{c}{\sqrt{\delta}} \end{aligned}$$

for sufficiently small  $hk$ .

Statement (6.19) follows from

$$\begin{aligned} \left| \int_{\mathcal{A}} \eta d\mu \right| &= k \left| \int_{\beta_0}^{\beta_1} \eta(k \cos \beta, k \sin \beta) d\beta \right| = k \left| \int_{\beta_0}^{\beta_1} \chi(\beta) d\beta \right| = k \left| \int_{\beta_0}^{\beta_1} \sin \left( \frac{\beta - \beta_1}{\beta_0 - \beta_1} \pi \right) d\beta \right| \\ &= \frac{k(\beta_1 - \beta_0)}{\pi} \cos \left( \frac{\beta - \beta_1}{\beta_0 - \beta_1} \pi \right) \Big|_{\beta_0}^{\beta_1} = \frac{2c_{\mathcal{A}}}{\pi}. \end{aligned}$$

To prove estimate (6.20), we proceed as follows:

$$\begin{aligned} \|\eta\|_{H^1(\mathbf{R}^2)}^2 &= \int_{k-\delta}^{k+\delta} \int_{\beta_0}^{\beta_1} r \rho(r)^2 \sin^2 \left( \frac{\beta - \beta_1}{\beta_0 - \beta_1} \pi \right) d\beta dr \\ &\quad + \int_{k-\delta}^{k+\delta} \int_{\beta_0}^{\beta_1} r \left( \frac{\partial \rho(r)}{\partial r} \right)^2 \sin^2 \left( \frac{\beta - \beta_1}{\beta_0 - \beta_1} \pi \right) d\beta dr \\ &\quad + \int_{k-\delta}^{k+\delta} \int_{\beta_0}^{\beta_1} \frac{1}{r} \rho(r)^2 \left( \frac{\partial}{\partial \beta} \sin \left( \frac{\beta - \beta_1}{\beta_0 - \beta_1} \pi \right) \right)^2 d\beta dr. \end{aligned}$$

Explicit calculations yield

$$\begin{aligned}\|\eta\|_{H^1(\mathbf{R}^2)}^2 &= k(\beta_1 - \beta_0) \left( \frac{\delta}{3} + \frac{1}{\delta} \right) + \frac{\pi^2}{2(\beta_1 - \beta_0)} \frac{(k + \delta)^2 \log \frac{k+\delta}{k} - (k - \delta)^2 \log \frac{k-\delta}{k} - 2k\delta}{\delta^2} \\ &= \left( \frac{\delta}{3} + \frac{1}{\delta} \right) c_{\mathcal{A}} + \frac{k\pi^2}{2c_{\mathcal{A}}} \frac{\left(1 + \frac{\delta}{k}\right)^2 \log \left(1 + \frac{\delta}{k}\right) - \left(1 - \frac{\delta}{k}\right)^2 \log \left(1 - \frac{\delta}{k}\right) - 2\frac{\delta}{k}}{\left(\frac{\delta}{k}\right)^2}.\end{aligned}$$

Using the Taylor expansion about  $\frac{\delta}{k} = 0$ , we conclude that

$$\|\eta\|_{H^1(\mathbf{R}^2)}^2 = \left( \frac{c_{\mathcal{A}}}{\delta} + \frac{(c_{\mathcal{A}}^2 + \pi^2)\delta}{3c_{\mathcal{A}}} + \delta O\left(\left(\frac{\delta}{k}\right)^2\right) \right).$$

In the theorem, we assumed that  $k^{3.5}h^3$  is bounded; therefore,  $\delta = c_s k^7 h^6$  is also bounded. Hence, one obtains from the equation above that

$$\|\eta\|_{H^1(\mathbf{R}^2)}^2 \leq \frac{c}{\delta}. \quad \square$$

#### REFERENCES

- [1] R.A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
- [2] A.K. AZIZ, R.B. KELLOGG, AND A.B. STEPHENS, *A two point boundary value problem with a rapidly oscillating solution*, Numer. Math., 53 (1988), pp. 107–121.
- [3] I. BABUŠKA, *The Fourier transform in the theory of difference equations and its applications*, Arch. Mech. Stos., 11 (1959), pp. 349–381.
- [4] I. BABUŠKA AND J.E. OSBORN, *Generalized finite element methods: Their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.
- [5] I. BABUŠKA, M. PRÁGER, AND E. VITÁSEK, *Numerical Processes in Differential Equations*, Wiley, New York, 1966.
- [6] I.M. BABUŠKA, F. IHLENBURG, E.T. PAIK, AND S.A. SAUTER, *A generalized finite element method for solving the Helmholtz equation in two dimensions with minimal pollution*, Comput. Meth. Appl. Mech. Engrg., 128 (1995), pp. 325–359.
- [7] P.G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II: Finite Element Methods, (Part 1), P.G. Ciarlet and J.L. Lions, eds., North-Holland, Amsterdam, 1991.
- [8] D. COLTON AND R. KREB, *Integral Equation Methods in Scattering Theory*, Wiley, New York, 1983.
- [9] R. DAUTRAY AND J.-L. LIONS, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 1, Springer-Verlag, New York, 1990.
- [10] I.M. GELFAND AND G.E. SHILOV, *Verallgemeinerte Funktionen (Distributionen)*, VEB Deutscher Verlag der Wissenschaften, Berlin, Germany, 1960.
- [11] C.I. GOLDSTEIN, *The finite element method with non-uniform mesh sizes applied to the exterior Helmholtz problem*, Numer. Math., 38 (1982), pp. 61–82.
- [12] W. HACKBUSCH, *Iterative Solution of Large Sparse Systems*, Springer-Verlag, Berlin, 1994.
- [13] W. HACKBUSCH, *Integral Equations*, Internat. Ser. Numer. Math., Birkhäuser, Basel, 1995.
- [14] I. HARARI AND T.J.R. HUGHES, *Finite element methods for the Helmholtz equation in an exterior domain: Model problems*, Comput. Meth. Appl. Mech. Engrg., 87 (1991), pp. 59–96.
- [15] F. IHLENBURG AND I. BABUŠKA, *Finite element solution to the Helmholtz equation with high wave number. Part I: The h-version of the FEM*, Comput. Math. Appl., 39 (1995), pp. 9–37.
- [16] F. IHLENBURG AND I. BABUŠKA, *Finite element solution to the Helmholtz equation with high wave number. Part II: The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.
- [17] J.B. KELLER AND D. GIVOLI, *Exact non-reflecting boundary conditions*, J. Comput. Phys., 82 (1989), pp. 172–192.
- [18] L.L. THOMPSON AND P.M. PINSKY, *A Galerkin least squares finite element method for the two-dimensional Helmholtz equation*, Internat. J. Numer. Methods Engrg., 38 (1995), pp. 9–37.