Convergence analysis of GMRES for the Helmholtz equation via pseudospectrum *

Antti Hannukainen†

June 1, 2015

Abstract

Most finite element methods for solving time-harmonic wave-propagation problems lead to a linear system with a non-normal coefficient matrix. The non-normality is due to boundary conditions and losses. One way to solve these systems is to use a preconditioned iterative method. Detailed mathematical analysis of the convergence properties of these methods is important for developing new and understanding old preconditioners. Due to non-normality, there is currently very little existing literature in this direction. In this paper, we study the convergence of GMRES for such systems by deriving inclusion and exclusion regions for the pseudospectrum of the coefficient matrix. All analysis is done a priori by relating the properties of the weak problem to the coefficient matrix. The inclusion is derived from the stability properties of the problem and the exclusion is established via field of values and boundedness of the weak form. The derived tools are applied to estimate the pseudospectrum of time-harmonic Helmholtz equation with first-order absorbing boundary conditions, with and without a shifted-Laplace preconditioner.

1 Introduction

Several different strategies for discretizing time-harmonic wave propagation problems using finite elements have been proposed in the literature. For typical problems, most of these strategies lead to a linear system with a large, sparse, indefinite and non-normal coefficient matrix. The indefiniteness is due to the wave-nature of the problem and the non-normality arises either from losses or truncation of infinite domains to finite ones. The large size of the system is due to the number of degrees of freedom required to resolve an oscillating solution. Because of their properties, the linear systems related to time-harmonic wave

---

*This work was supported by the Alfred Kordelin’s Foundation and Academy of Finland projects 13267297 and 14341
†Aalto University, Department of Mathematics and Systems Analysis, P.O. Box 11100, FI-00076 Aalto, Finland, email:antti.hannukainen@aalto.fi
propagation problems are difficult to solve. Memory is an issue with direct solvers and lack of efficient preconditioners with iterative ones.

In order to develop new preconditioners and to understand old ones, it is important to know their effect on the convergence properties of the applied iterative method. Unfortunately, the convergence of iterative methods for linear systems with a non-normal coefficient matrix is a difficult subject of study. When the non-normality is significant, the iterative properties can be very different from the ones indicated by eigenvalues, \cite{21,10}. Similar difficulties are met with other properties related to the non-normal matrices, e.g., behavior of matrix exponentials cannot be predicted by eigenvalues, \cite{17}. Determining when the non-normality has a significant impact to iterative properties is complicated. The simplest way to estimate the impact is to compute one of the commonly used scalar measures of non-normality, e.g., \( \|AA^* - A^*A\|/\|A\|^{-1} \), the conditioning of eigenvectors or the conditioning of individual eigenvalues, \cite{24}. However, except for the first one, these measures are not computable for large matrices. In addition, they can vary considerably even for relatively small systems \cite{24}.

The convergence of preconditioned iterative methods has been extensively studied in the context of finite element methods, \cite{22}. However, majority of the research has focused on real valued symmetric positive definite problems. The finite element discretization of these problems also leads to symmetric positive definite linear systems, which are solved using the preconditioned conjugate gradient method (PCG). The aim in the analysis of these methods is to estimate the convergence rate before computations. Only few of the existing works deal with indefinite linear systems, \cite{28,9,2,3,18}, and even fewer with non-normal indefinite ones, \cite{26,12,26}.

Most preconditioners for finite element discretizations of elliptic weak problems have been analyzed by using the abstract framework of Schwarz methods, \cite{22}. This framework is based on studying the properties of the underlying weak problem instead of the linear system. The convergence of PCG is related to the weak form via Rayleigh quotients. Such analysis is done in the inner product induced by the bilinear form. As Rayleigh quotients are the first step in the existing analysis, it does not carry over to complex valued, indefinite, non-normal linear systems. Such systems require a different set of analytics tools.

There currently exists three different ways to analyze iterative properties of a non-normal matrix \cite{5}: to study the field of values (FOV), pseudospectrum, or to include conditioning of eigenvectors to the convergence estimates. For time-harmonic wave-equations, estimating eigenvector conditioning before the matrices are constructed seems to be complicated and thus this approach is not suitable for our purposes. FOV has been applied to analyze the preconditioned time-harmonic Helmholtz equation e.g. in \cite{12}. However, FOV is always a convex set containing all eigenvalues of the matrix. As we will see, the spectrum of the problems we are interested in curls around the origin making FOV based methods unsuitable for our purposes. In contrast, the pseudospectrum can be a non-convex set and as we will show it can be estimated a priori, making it the best option of the three for this work.
In this paper, we study pseudospectrum as a tool for relating the properties of the weak problem to the convergence of the GMRES method. We derive convergence estimates for GMRES by establishing inclusion and exclusion regions for the pseudospectrum. The exclusion region is derived from the stability estimates of the weak problem and the inclusion region is based on the relation between pseudospectrum and FOV. In several cases, an inclusion for FOV can be easily obtained based on continuity properties of the weak form. All analysis is done a priori, so that the regions can be obtained without constructing the actual matrices or performing computations with them. The derived bounds are explicit in the relevant parameters of the problem, e.g., mesh size, wave-number, and the losses. The presented analysis relies on general properties of the weak problem, stability and continuity so it is possible that it can be applied to other preconditioners and problems.

The paper is organized as follows. We begin with some preliminaries and proceed to give estimates relating pseudospectrum to a weak problem. After establishing these abstract results, we apply them to three example problems. We begin the examples by considering the Poisson problem, which is included for easy reference on what kind of information the derived estimates can deliver. Then we apply the presented tools to time-harmonic Helmholtz equation with and without a shifted-Laplace preconditioner. We end the paper with a discussion of the presented material.

2 Preliminaries

Our model problem is: Find \( u \in V \) such that
\[
a(u, v) = L(v) \quad \forall \ v \in V, \tag{1}
\]
where \( V \) is some finite element space, \( a(\cdot, \cdot) : V \times V \to \mathbb{C} \) is a sesquilinear form and \( L(\cdot) : V \to \mathbb{C} \) an antilinear functional. The finite element space \( V \) is spanned by basis \( \{ \varphi_i \}_{i=1}^N \) so that every function \( u \in V \) admits the representation
\[
u = \sum_{i=1}^{N} (\bar{x}_u)_i \varphi_i,
\]
in which the vector of coefficients \( \bar{x}_u \in \mathbb{C}^N \). Problem (1) leads to the linear system
\[
A\bar{x}_u = \bar{b}_i,
\]
where \( A \in \mathbb{C}^{N \times N} \), \( \bar{b} \in \mathbb{C}^N \), \( A_{ij} := a(\varphi_j, \varphi_i) \) and \( \bar{b}_i := L(\varphi_i) \). Hence, the sesquilinear form and the matrix \( A \) are related as
\[
a(u, v) := \bar{x}_u^* A \bar{x}_u, \tag{2}
\]
where \( * \) is the conjugate transpose. The properties of the matrix \( A \) will depend on the properties of the sesquilinear form and the basis functions via the above equation.
We will describe the actual problem and discretization in detail in Section 4. For now, let us note that when the sesquilinear form $a$ is related to the time-harmonic Helmholtz equation with absorbing boundary conditions, the matrix $A$ can be very large. This is due to the fact that the finite element mesh size has to be sufficiently fine before finite element method can produce accurate results, see [13, 14]. Typical engineering rule of thumb is to use ten degrees of freedom per one wave-length. For example, a cube for which each dimension is ten wave-lengths long requires one to use $10^6$ degrees of freedom, this is, $N = 10^6$ or larger.

In the following, we assume that problem (1) has a unique solution and admits some kind of a stability estimate. Stability estimates are typically derived under additional assumptions on the domain and the antilinear functional $L$. In general, the functional $L$ can be from the space $V' = \{ f: V \to \mathbb{C} \mid \bar{f} \in V^* \}$, where $V^*$ is the dual space of $V$. As such functionals can be quite badly behaving, stability estimates are often derived under the assumption $L \in W'$. In this spirit, we make the following assumption.

**Assumption 2.1.** Let $W$ be a Hilbert space, $V \subset W$, $L \in W'$ and $u \in V$ be the unique solution to problem (1). Then there exists a constant $C_S > 0$ independent of $u$ and $L$ such that

$$\|u\| \leq C_S \|L\|_{W'}$$

(3)

where $\|\cdot\|$ is a norm on $V$ and $\|\cdot\|_{W'} := \sup\{ |L(w)| \mid w \in W \text{ and } \|w\|_W = 1 \}$.

The pseudospectrum of a matrix $A \in \mathbb{C}^{N \times N}$, $\Lambda_\epsilon(A)$, is a family of sets depending on a parameter $\epsilon > 0$. The sets in the family are defined as

$$\Lambda_\epsilon(A) := \{ z \in \mathbb{C} \mid |(zI - A)^{-1}| \geq \epsilon^{-1} \},$$

in which $|\cdot|$ is the standard spectral norm. When the matrix $(zI - A)$ is singular, we define $|(zI - A)^{-1}| = \infty$. The notation $|\cdot|$ is also used for the Euclidean norm of a vector. Clearly, the pseudospectrum can also be characterized as

$$\Lambda_\epsilon(A) := \{ z \in \mathbb{C} \mid \sigma_{\text{min}}(zI - A) \leq \epsilon \},$$

in which we denote the smallest singular value of a matrix $B \in \mathbb{C}^{N \times N}$ as $\sigma_{\text{min}}(B)$.

The pseudospectrum was independently proposed by several authors as an extension of the spectrum, suitable to study the properties non-normal matrices, [17]. The pseudospectrum has been extensively studied in the literature, see e.g. [24, 23, 17]. In the following, we write $\Lambda_\epsilon(A) = \Lambda_\epsilon$, when the matrix $A$ is clear from the context.

In the derivation of the inclusion region, we take advantage on the relation between FOV and pseudospectrum. The FOV of a matrix $A \in \mathbb{C}^{N \times N}$ is defined as the set

$$\text{FOV}(A) := \left\{ \frac{\bar{x}^* A\bar{x}}{\bar{x}^* \bar{x}} \mid \bar{x} \in \mathbb{C}^N \text{ and } \bar{x} \neq 0 \right\}.$$
The set $\text{FOV}(A)$ is convex, compact and contains all eigenvalues of $A$. As we will see, coarse inclusion for $\text{FOV}(A)$ can be obtained by using it’s close relation with the sesquilinear form. We postpone stating the relation between pseudospectrum and FOV to Section 3, where we have introduced sufficient notation for proving it.

Both pseudospectrum and FOV can be related to convergence of the GMRES method, [5 10 21]. The approximation error for the solution $\bar{x}_i$ generated by GMRES on step $i$ is measured as $|\bar{r}_i|$, where $\bar{r}_i$ is the residual, $\bar{r}_i = A\bar{x}_i - \bar{b}$. There holds that

$$|\bar{r}_i| = \inf_{p \in \tilde{P}_i \atop p(0)=1} |p(A)\bar{r}_0|, \quad (5)$$

in which $\tilde{P}_i$ is the space of monic polynomials of degree $i$. The matrix valued polynomial in the above minimization problem can be evaluated using Dunford integral [27, 5]. Let the open set $U \subset \mathbb{C}$ be such that $\sigma(A) \subset U$ and $\partial U$ is the union of rectifiable positively oriented Jordan curves. The set $\sigma(A)$ is the spectrum of $A$. Application of the Dunford integral gives

$$p(A) = \frac{1}{2\pi i} \int_{\partial U} p(z)(zI - A)^{-1}dz. \quad (6)$$

This integral can be used to derive estimates for equation (5). Let $\tilde{\Lambda}_\epsilon$ satisfy the assumptions made on the set $U$ and in addition let

$$\Lambda_\epsilon \subset \tilde{\Lambda}_\epsilon.$$

This is, $|zI - A| \leq \epsilon^{-1} \forall \ z \in \partial \tilde{\Lambda}_\epsilon$. In our case, $\tilde{\Lambda}_\epsilon$ is an approximation for the pseudospectral set. Estimating the integral gives

$$|p(A)\bar{r}_0| \leq |p(A)||\bar{r}_0| \leq \frac{|\bar{r}_0||\partial \tilde{\Lambda}_\epsilon|}{2\pi \epsilon} \sup_{z \in \Lambda_\epsilon} |p(z)|. \quad (7)$$

Combining equations (5) and (7) leads to the GMRES convergence estimate

$$\frac{|\bar{r}_i|}{|\bar{r}_0|} = \frac{|\partial \tilde{\Lambda}_\epsilon|}{2\pi \epsilon} \inf_{p \in \tilde{P}_i \atop p(0)=1} \sup_{z \in \tilde{\Lambda}_\epsilon} |p(z)|. \quad (8)$$

As we will illustrate in Section 4, this bound is useful for deriving worst case behavior of the GMRES convergence rate.

The convergence bound (8) is meaningful only if one can solve the complex polynomial minimization problem. Typically, the set $\tilde{\Lambda}_\epsilon$ is replaced with a larger set on which the minimization problem can be solved analytically. In simple cases, $\tilde{\Lambda}_\epsilon$ can replaced with a circular or an elliptical domain, [21]. Due to the constraint $p(0) = 1$, this approach gives useful information only when the circle or ellipsoid containing $\tilde{\Lambda}_\epsilon$ does not contain the origin. When this is the case, one can try to apply so-called bratwurst shaped domains [15].
the name suggests, a bratwurst shaped domain can curl around the origin and it can be used to derive convergence estimates for the minimization problem. The bratwurst shaped domains can be applied with the inclusion and exclusion regions derived in this paper. However, the construction given in [15] is not simple, and cannot yield easy to use a priori bounds.

In order to verify the analytically derived inclusion and exclusion results, we compute examples of the pseudospectral sets. Several different strategies for computing pseudospectrum have been proposed, see [24] and references therein. Several software packages, such as EigTool, are also freely available.

To have full control over the computation of the pseudospectrum, we have chosen to use our own implementation of GRID - approach to compute pseudospectral sets. In the GRID-approach, a mesh is placed in the complex plane and the norm of the resolvent is computed for each grid point. The computed data is used to isolines describing the set. In the simplest case, the norm is computed as the largest singular value of the matrix \((z - A)^{-1}\). Clearly, such an approach is very expensive for large number of points and large matrices. The process can be sped up by adapting the computational grid to the resolvent norm or by using a suitable matrix factorization to speed up the evaluation of the largest singular value. We have opted to speed up the computation by using an adaptive strategy to refine the computational grid. An initial triangular grid is placed in the complex plane. The grid is iteratively refined to conform to the shape of the resolvent norm. We use a refinement strategy based on splitting triangles intersecting with pre-specified level sets of the resolvent norm. This guarantees higher resolution at interesting regions of the complex plane.

\section{Abstract Framework}

In this section, we derive inclusion and exclusion regions for the pseudospectral set. For this purpose, it is easier to bound the complement of \(\Lambda_\epsilon\), i.e.

\[ \Lambda_\epsilon^c := \{ |(zI - A)^{-1}| < \epsilon^{-1} \}. \tag{9} \]

The inclusion and exclusion regions will lead to a set containing the pseudospectrum. If the boundary of this set is a rectifiable Jordan curve, it can be used in connection with equation (8) to compute convergence estimates for the GM-RES method. The exclusion will be a disc around the origin. For the results to be meaningful, the exclusion should not be fully contained in the inclusion. If this is the case, the polynomial minimization problem in equation (8) does not tend to zero and the bound does not provide useful information. This has to be studied separately for each problem.

When \((zI - A)\) is non-singular, the matrix norm in equation (9) is defined as

\[ |(zI - A)^{-1}| := \sup_{u \in V} \frac{|(zI - A)^{-1} \tilde{x}_u|}{|\tilde{x}_u|}. \tag{10} \]

\footnote{\textsuperscript{1}see the Pseudospectral Gateway, \url{http://www.cs.ox.ac.uk/pseudospectra/}}
To eliminate the inverse and to establish a connection to the weak problem, we define an auxiliary vector $\vec{x}_v \in \mathbb{C}^N$ such that

$$(zI - A)\vec{x}_v = \vec{x}_u. \tag{11}$$

Estimates for the resolvent norm are derived using the auxiliary variable. First, we establish the stability bound $|\vec{x}_v| \leq f(z)|\vec{x}_u|$. When $f(z)$ is bounded from above, this implies that $(zI - A)$ is non-singular. In this case, the auxiliary vector is uniquely defined and

$$\vec{x}_v = (zI - A)^{-1}\vec{x}_u.$$

The norm (10) can be estimated using the stability estimate for $\vec{x}_v$ as

$$|(zI - A)^{-1}| = \sup_{u \in V} \frac{|\vec{x}_u|}{|\vec{x}_u|} \leq f(z). \tag{12}$$

We begin by taking advantage of the stability of the weak problem, Assumption 2.1. Due to the duality between coefficient vectors and functions, stability of the weak problem implies stability of the linear system. As all finite dimensional norms are equal, there exists positive constants $\alpha, \alpha_W > 0$ independent of $u$ such that

$$\alpha|\vec{x}_u| \leq \|u\| \quad \text{and} \quad \alpha_W|\vec{x}_u| \leq \|u\|_W \quad \forall u \in V. \tag{13}$$

When the derived framework is applied to a specific problem, $\alpha$ and $\alpha_W$ are typically dependent on the mesh size. The dependency of these constants on relevant problem parameters are discussed in Section 4. Combining these norm equivalences with Assumption 2.1 leads to the following corollary.

**Corollary 3.1.** Let $\vec{b} \in \mathbb{C}^n$ and $\vec{x}_u$ be such that $A\vec{x}_u = \vec{b}$. Then there holds that

$$|\vec{x}_u| \leq C_{2S}|\vec{b}|.$$

Where $C_{2S} := C_S(\alpha_W \alpha)^{-1}$.

**Proof.** Let $q \in V$ be such that

$$(q, v)_W = \vec{x}_v^* \vec{b} \quad \forall v \in V,$$

where $(\cdot, \cdot)_W$ is inner product on $W$. Using Cauchy-Schwarz inequality and the norm equivalence given in equation (13) there holds that $\|q\|_W \leq \alpha_W^{-1}|\vec{b}|$. Via this construction, vector $\vec{b}$ defines an antilinear functional on $W'$ as $L(v) := (q, v)_W$. By the definition of the dual norm and Cauchy-Schwarz inequality

$$\|L\|_{W'} = \sup_{w \in W} \frac{|(w, q)_W|}{\|w\|_W} \leq \sup_{w \in W} \frac{\|w\|_W \|q\|_W}{\|w\|_W} = \|q\|_W. \tag{14}$$

It follows that

$$\|L\|_{W'} \leq \alpha_W^{-1}|\vec{b}|.$$

Combining the above equation with Assumption 2.1 and equation (13), we obtain

$$\alpha_W \alpha|\vec{x}_u| \leq C_S|\vec{b}|.$$

□
The above Corollary essentially gives a lower bound for the smallest singular value of $A$. There holds that

$$\sigma_{\text{min}}(A)^{-1} = \min_{\vec{x} \in \mathbb{C}^N} \frac{|A^{-1}\vec{x}|}{|\vec{x}|}$$

so, that $C_{2S}^{-1} \leq \sigma_{\text{min}}(A)$. Corollary 3.1 can be used to derive exclusion region near the origin. We give here a direct proof that fits well to the framework of the paper. Same result can be established from the lower bound for the smallest singular value by using Theorem 3 from [16].

**Theorem 3.1.** Let Assumption 2.1 hold and let $C_{2S}$ be as defined in Corollary 3.1. Then there holds that

$$B(0, \frac{1}{C_{2S}} - \epsilon) \subset \Lambda^\epsilon,$$

in which $B(z_0, r) := \{ z \in \mathbb{C} \mid |z - z_0| < r \}$. 

**Proof.** From the definition of the auxiliary variable (11) it follows that

$$A\vec{x}_v = z\vec{x}_v - \vec{x}_u$$

Application of Corollary 3.1 gives

$$|\vec{x}_v| \leq C_{2S} (|z||\vec{x}_v| + |\vec{x}_u|)$$

i.e.

$$|\vec{x}_v| \leq \frac{C_{2S}}{1 - C_{2S}|z|} |\vec{x}_u|. \quad (15)$$

When $|z| < C_{2S}^{-1}$, the above bound implies that $(zI - A)$ is non-singular. In this case, combining equations (15) and (12) gives

$$|(zI - A)^{-1}| \leq \frac{C_{2S}}{1 - C_{2S}|z|}.$$ 

To obtain the exclusion region, we set

$$\frac{C_{2S}}{1 - C_{2S}|z|} < \epsilon^{-1},$$

which gives the bound

$$|z| < \frac{1}{C_{2S}} - \epsilon.$$

The inclusion is obtained by relating pseudospectrum to FOV. The following Theorem is proven e.g. in, [23]. For completeness, we give a proof using the notation used in this Section.
**Theorem 3.2.** Let $S_\epsilon := \{ z \in \mathbb{C} \mid \text{dist}(z, \text{FOV}(A)) \leq \epsilon \}$ in which

$$\text{dist}(z, Q) := \inf_{q \in Q} |z - q|.$$ 

Then there holds that $\Lambda_\epsilon \subset S_\epsilon$.

**Proof.** The auxiliary variable is defined as

$$(A - zI)\bar{x}_v = \bar{x}_u.$$ 

Testing the above equation with $\bar{x}_v$ gives

$$\bar{x}_v^* A\bar{x}_v - z\bar{x}_v^*\bar{x}_v = \bar{x}_v^*\bar{x}_u.$$ 

Using Cauchy-Schwarz inequality gives

$$|\bar{x}_v| |\bar{x}_u| \geq |\bar{x}_v^* A\bar{x}_v - z\bar{x}_v^*\bar{x}_v| = \bar{x}_v^*\bar{x}_v \frac{\bar{x}_u^* A\bar{x}_u}{\bar{x}_v^*\bar{x}_v} - z.$$ 

This is,

$$|\bar{x}_u| \left| \frac{\bar{x}_u^* A\bar{x}_u}{\bar{x}_v^*\bar{x}_v} - z \right| \leq |\bar{x}_v|.$$ 

By the definition of FOV(A) in equation (4) there holds that

$$|\bar{x}_u| \geq \text{dist}(z, \text{FOV}(A)) |\bar{x}_v|.$$ 

\[\Box\]

Theorem 3.2 gives tools for deriving an inclusion for the pseudospectrum. The FOV is directly related to the boundedness properties of the sesquilinear form of the original problem. This relation arises from the connection $\bar{x}_v^* A\bar{x}_v = a(v, v)$. The simplest estimate follows from boundedness of the sesquilinear form. Assume that there exists $C > 0$ such that

$$|a(u, u)| < C\|u\|_V^2 \quad \forall u \in V.$$ 

Then there holds that

$$\text{FOV}(A) \subset B(0, C).$$ 

This is a very crude estimate, but it demonstrates how FOV can be bounded in simple cases. However, as we will see, more refined estimates are required to avoid inclusion of zero to the approximate pseudospectrum.

### 4 Examples

In this section, we demonstrate the presented theory with three examples. In all examples, we assume that $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ is a bounded domain with Lipschitz continuous boundary. We use standard notation for Sobolev spaces, see \[1\].
The finite element space \( V \) is defined as
\[
V := \{ u \in H^1(\Omega) \mid u \in P^1(K) \quad \forall K \in \mathcal{T} \},
\] (16)
where \( \mathcal{T} \) is a shape regular triangular or tetrahedral partition of \( \Omega \). This is, \( V \) is the space of first order Lagrange finite elements. The space of first order polynomials over set \( K \) is denoted by \( P^1(K) \) and the mesh-size by \( h \), respectively.

The presented theoretical results are independent of the domain, but the actual numerical examples are computed on \( \Omega = (-1,1)^2 \setminus (0,1)^2 \). The meshes used in the tests are generated from a coarse mesh with approximately 100 nodes using uniform refinement. The coarse mesh is called level one mesh, once refined coarse mesh as a level two mesh and so on.

Throughout this Section, \( c, C > 0 \) are generic positive constants independent of mesh size \( h \), solution, load, and parameters of the weak problem, if not otherwise stated. They may depend on the shape regularity constant of the partition \( \mathcal{T} \) and the domain \( \Omega \).

### 4.1 Poisson equation

We begin by considering the finite element discretization of the Poisson equation: Find \( u \in V_0 \) such that
\[
(\nabla u, \nabla v) = (f, v) \quad \forall v \in V_0.
\] (17)
In which \( V_0 = V \cap H^1_0(\Omega) \) and \( f \in L^2(\Omega) \). This is
\[
a(u,v) := (\nabla u, \nabla v) \quad \text{and} \quad L(v) := (f, v)
\]
so that \( L \in (L^2(\Omega))' \). We use the standard \( H^1 \)-norm
\[
\|u\|_1^2 := (\nabla u, \nabla u) + (u, u)
\]
for the space \( V_0 \).

It is straightforward to see that the matrix \( A \) related to problem (17) is symmetric and positive definite. The convergence of iterative methods for such linear systems can be analyzed using much easier techniques than pseudospectrum. However, such a simple example is useful for demonstrating what kind of information on GMRES convergence can be obtained based on the inclusion and exclusion results.

Pseudospectrum of a normal matrix can be easily computed from it’s eigenvalues. All normal matrices are unitary diagonalizable, hence there exists a diagonal \( D \in \mathbb{C}^{N \times N} \) and a unitary \( Q \in \mathbb{C}^{N \times N} \) such that \( A = Q^* D Q \). Based on this expansion, there holds that
\[
| (zI - A)^{-1} | = | (z - D)^{-1} | = \max_{\lambda \in \sigma(A)} |(z - \lambda)^{-1}|.
\]
Thus, pseudospectrum of any normal matrix is a union of discs centered around its eigenvalues $\lambda_i$,

$$\Lambda_\epsilon = \bigcup_{i=1}^{N} B(\lambda_i, \epsilon).$$

The pseudospectrum for level one mesh is visualized in Figure 4.1 for different values of $\epsilon$.

Figure 1: Pseudospectral set for the Poisson problem on level one mesh. For sufficiently small $\epsilon$, the set is composed of disjoint disks with radius $\epsilon$.

Next, we derive inclusion and exclusion regions using Theorem 3.1 and 3.2. First, we need to establish a stability estimate satisfying Assumption 2.1. As we are interested in mesh size explicit bounds, we use $h$-explicit norm equivalences instead of equation (13). For the Poisson problem, stability estimate follows from the weak problem (17) by using Poincare-Friedrichs inequality. Let $u \in V_0$ be the solution to (17) then there exists a constant $C > 0$ such that

$$\|u\|_1 \leq C\|f\|_0.$$  

Following this stability estimate, we choose the space $W$ as $L^2(\Omega)$ and $\|\cdot\|_W = \|\cdot\|_0$. Norm equivalences between $H^1(\Omega)$-, $L^2(\Omega)$- and the Euclidian norm can be derived in the finite element space $V$ using the scaling argument and inverse inequality, [20]. There exists $c$ and $C$ such that

$$ch^{d/2} |\vec{x}_u| \leq \|u\|_0 \leq C \bar{h}^{d/2} |\vec{x}_u| \quad \forall u \in V$$  

and

$$ch^{d/2} |\vec{x}_u| \leq \|u\|_1 \leq \bar{C} \bar{h}^{d/2-1} |\vec{x}_u| \quad \forall u \in V.$$  

11
Now, we can apply Corollary 3.1 to derive a stability constant for the linear system arising from the weak problem (17). Let $\vec{x}_u$ be such that $A\vec{x}_u = \vec{b}$. Then by Corollary 3.1 and the $h$-explicit norm equivalences, there exists a constant $C$ such that

$$|\vec{x}_u| \leq C h^{-d} |\vec{b}|.$$ 

Application of Theorem 3.1 gives the following exclusion near the origin,

$$B(0, Ch^d - \epsilon) \subset \Lambda_\epsilon^c.$$ 

We proceed by deriving an inclusion for $FOV(A)$, which together with Theorem 3.2 gives inclusion for $\Lambda_\epsilon$. It is easy to derive the estimates

$$\Im a(u, u) = \Im \|\nabla u\|_0^2 = 0 \quad \forall u \in V_0$$

and

$$ch^d |\vec{x}_u|^2 \leq \Re a(u, u) < Ch^d - 2 |\vec{x}_u|^2 \quad \forall u \in V_0.$$ 

So that $FOV(A) \subset \{ x \in \mathbb{R} \mid ch^d < x < Ch^d - 2 \}$. An application of Theorem 3.2 gives the inclusion $\Lambda_\epsilon \subset \tilde{S}_\epsilon$, in which

$$\tilde{S}_\epsilon := \{ z \in \mathbb{C} \mid \text{dist}(z, \{ x \in \mathbb{R} \mid ch^d < x < Ch^d - 2 \}) \leq \epsilon \}.$$ 

The above inclusion and exclusion regions give us an approximation of pseudospectrum $\tilde{\Lambda}_\epsilon$,

$$\tilde{\Lambda}_\epsilon := \tilde{S}_\epsilon \setminus B(0, C_1 h^d - \epsilon).$$

Where the constant $C_1 > 0$ is independent of $h$ and $\epsilon$. To exclude the origin from this approximate pseudospectrum, we have to choose the parameter $\epsilon$ as $\epsilon \leq C_1 h^d$. In this case, the length of the boundary curve around the approximate pseudospectrum satisfies $|\partial \tilde{\Lambda}_\epsilon| \leq C_2 h^{d-2}$ for some $C_2 > 0$ independent of $h$ and $\epsilon$. When combined with equation (8) approximate pseudospectrum gives the GMRES convergence bound

$$|\vec{r}_i| \leq \frac{C_2 h^{d-2}}{2\pi \epsilon} \inf_{p \in P_i} \sup_{z \in \tilde{\Lambda}_\epsilon} |p(z)||\vec{r}_0| \quad \forall \epsilon \leq C_1 h^d \tag{20}$$

The set $\tilde{\Lambda}_\epsilon$ can be covered either with an ellipsoid or a circle and the minimization problem can be solved using estimates given in [10, 21]. There holds that

$$\inf_{p \in P_i} \sup_{z \in B(c, r)} |p(z)| \leq \left( \frac{r}{|c|} \right)^i$$

Although the estimate could be optimized with respect to parameter $\epsilon$, we have chosen $\epsilon = 0.5C_1 h^d$, which gives correct asymptotic behavior with respect to $h$. Using this $\epsilon$ and $c = C_2 h^{d-2}$, the circle based bound leads to the estimate

$$|\vec{r}_i| \leq \frac{C_2 h^{-2}}{\pi C_1} \left( \frac{1}{1 + \frac{1}{2C_2 h^2}} \right)^i |\vec{r}_0|$$

12
When the termination criteria for GMRES is chosen such that the relative residual satisfies $\frac{||r_i||}{||r_0||} \leq tol$, the above estimate gives the required number of iterations $N$ as

$$N \approx -\frac{2C_2}{C_1}h^{-2}\left(\log tol - \log \frac{C_2h^{-2}}{\pi C_1}\right).$$  \hspace{1cm} (21)$$

Our approximate pseudospectrum cannot capture the behavior of $\Lambda_\epsilon$ for very small values of $\epsilon$. For example in the current case, the exact pseudospectrum is composed of small discs with boundary length $2\pi\epsilon$. Let $\epsilon$ be such that the discs generating the pseudospectrum do not intersect. Any finite union of disjoint disks satisfies the conditions placed on the set $U$ in the Dunford integral. Using equation (6) we obtain the estimate

$$|p(A)| \leq \frac{1}{2\pi\epsilon} \sum_{i=1}^{N_0} |\partial B(\epsilon, \lambda_i)| \sup_{z \in B(\epsilon, \lambda_i)} |p(z)| \leq N_0 \sup_{z \in \Lambda_\epsilon} |p(z)|,$$

which is valid for sufficiently small $\epsilon$. Here $N_0$ is the number of disjoint eigenvalues of $A$. For quasi-uniform meshes, there exists $C$ such that $N_0 \leq Ch^{-d}$ so that

$$|p(A)| \leq Ch^{-d} \sup_{z \in \Lambda_\epsilon} |p(z)|.$$

Combining the above estimate with equation (5) gives

$$\frac{||r_i||}{||r_0||} \leq Ch^{-d} \inf_{p \in \tilde{P}, \ p(0)=1} \sup_{z \in \Lambda_\epsilon} |p(z)|. \hspace{1cm} (22)$$

This estimate based on the exact set $\Lambda_\epsilon$ has a different multiplicative term in comparison to equation (20). Interestingly, for $d = 1$, multiplicative term is smaller, for $d = 2$ it is equivalent and for $d = 3$ bigger. Regardless of the multiplicative constant, the estimate (22) can deliver improved convergence number estimates. The best possible bound can be obtained at the limit $\epsilon = 0$, when the minimization problem can be solved using Chebychev polynomials, see e.g. [21]. Based on the FOV, the condition number $\kappa(A) \leq Ch^{-2}$. We obtain an estimate for the number of iterations

$$N \approx -Ch^{-1}(\log(tol) - \log(Ch^{-d})) \hspace{1cm} (23)$$

The main difference between the estimates (21) and (23) is in in the power of the mesh size $h$. For the particular problem, this difference is due to the fact, that the set $\tilde{\Lambda}_\epsilon$ cannot capture the behaviour of the pseudospectrum for small $\epsilon$. For complicated problems, such knowledge is very difficult to come by and one has to be satisfied with worst case estimates, such as equation (21). The second difference between the two estimates is in the additive terms. These additive terms are relevant only when tolerance is of the same order of magnitude with $Ch^{-d}$, which requires usage of very fine mesh sizes.
4.2 Helmholtz equation with absorbing boundary conditions

The Helmholtz equation with first-order absorbing boundary conditions is a more realistic example for the analysis presented in this paper. The weak problem reads: Find \( u \in H^1(\Omega) \) such that
\[
a(u, v) = L(v) \quad \forall v \in H^1(\Omega).
\]
(24)
in which
\[
a(u, v) := (\nabla u, \nabla v) + i\kappa (u, v)_{\partial\Omega} - \kappa^2 (u, v) \quad \text{and} \quad L(v) := (f, v) + (g, v)_{\partial\Omega}.
\]
(25)
The parameter \( \kappa \in \mathbb{R}, \kappa > 0 \), \( f \in L^2(\Omega) \) and \( g \in L^2(\partial\Omega) \). The inner product \((\cdot, \cdot)_{\partial\Omega}\) is the standard \(L^2\)-inner product over \( \partial\Omega \). The stability of this problem has been analyzed in domains excluding any resonant behavior, [19].

**Theorem 4.1.** Let \( \Omega \) be a bounded, star shaped domain with a smooth boundary and let \( u \in H^1(\Omega) \) be the solution to problem (24). Then there exists a constant \( C_S > 0 \) independent of \( u, f, g \) and \( \kappa \) such that
\[
\|u\|_{\kappa} \leq C_S (\|f\|_0 + \|g\|_{0,\partial\Omega}),
\]
in which the norm \( \| \cdot \|_{\kappa} \) is defined as
\[
\|u\|_{\kappa}^2 := \|\nabla u\|_0^2 + \kappa^2 \|u\|_0^2.
\]
(26)
The finite element approximation \( u_h \) is defined as: Find \( u_h \in V \) such that
\[
a(u_h, v) = (f, v) + (g, v)_{\partial\Omega} \quad \forall v \in V.
\]
When the solution has \( H^2(\Omega)\)-regularity, the existence of a unique solution to this problem can be guaranteed, when the mesh size requirement \( \kappa^2 h << 1 \) is satisfied, [13, 14, 19]. In this case, there exists a constant \( C \) such that the a priori error estimate
\[
\|u - u_h\|_{\kappa} \leq C h (\|f\|_0 + \|g\|_{0,\partial\Omega}).
\]
(27)
holds.

Due to the boundary term \( i\kappa (u, v)_{\partial\Omega} \), problem (24) leads to a linear system with a non-normal coefficient matrix. As the boundary term depends on \( \kappa \), it is complicated to determine if the non-normality is meaningful or not. In addition, due to the relation between the wave-number and the mesh size it is difficult to study the asymptotic behaviour of GMRES, when \( \kappa \) tends to infinity.

When the mesh size is sufficiently small so that the a priori error estimate (27) holds, Theorem 4.1 implies stability of the discrete problem. We obtain
\[
\|u_h\|_{\kappa} \leq (1 + Ch) (\|f\|_0 + \|g\|_{0,\partial\Omega}) \leq C (\|f\|_0 + \|g\|_{0,\Omega}).
\]
(28)
This discrete stability estimate holds under the following assumptions.
Assumption 4.1. Assume that $\Omega$ is a bounded, star shaped domain with a smooth boundary, the mesh size $h$ is such that $\kappa^2 h << 1$ and the solution $u$ to problem (24) has $H^2(\Omega)$-regularity.

Following the discrete stability result (28) we choose the space $W$ as $L^2(\Omega)$ with the norm $\| \cdot \|_W = \| \cdot \|_0$. The $h$- explicit norm equivalences given in equation (18) can be used for this space. As we are interested in wavenumber and the mesh size explicit estimates, we use the $\kappa$-dependent norm given in equation (26) for the space $V$. Norm equivalences for this $\kappa$-dependent norm are easily established using equation (18) and (19) as

$$c \kappa h^{d/2} |\vec{x}_u| \leq \| u \|_C \leq C (h^{d/2-1} + \kappa h^{d/2}) |\vec{x}_u| \quad \forall u \in V,$$

for some $c, C$. Application of Corollary 3.1 gives the stability estimate for the coefficient vector

$$|\vec{x}_{uh}| \leq C \frac{h^{-d}}{\kappa} |\vec{b}|.$$

Using Theorem 3.1 leads to the exclusion region

$$B(0, C\kappa h^d - \epsilon) \subset \Lambda^\epsilon_c.$$
Figure 3: Pseudospectral set for the Helmholtz equation with first order absorbing boundary conditions. The parameter $\kappa = 8\pi$. The mesh levels is three on left and four on right. One can observe the convergence of the set when mesh size tends to zero.

around the origin. To obtain an inclusion, we again derive an inclusion for $FOV(A)$ and apply Theorem 3.2. The sesquilinear form satisfies the boundedness estimates

$$|\Re a(u, u)| \leq C\|u\|_\kappa^2 \quad \forall u \in V$$

and

$$0 \leq \Im a(u, u) \leq C\kappa h^{d-1}|\vec{x}_u|^2 \quad \forall u \in V$$

for some $C$. The estimate between the $L^2(\partial\Omega)$- and Euclidian norm used in above is derived using identical techniques as used for proving inequality (18). When Assumption 4.1 is satisfied, combining the two boundedness estimates leads to the inclusion

$$FOV(A) \subset \{ z \in \mathbb{C} \mid |z| \leq C \text{ and } 0 \leq \Im z \leq C\kappa h^{d-1} \}.$$ 

This set contains the origin, so it cannot be used to derive GMRES convergence bounds. In this case, the presented theory is genuinely required to understand GMRES convergence.

To validate the derived inclusion and exclusion regions, we have computed examples from the exact set $\Lambda_\epsilon$ for $\kappa = 8\pi$ using mesh levels three and four. The results are visualized in Figures 2 and 3. Although, the $L$-shaped domain used in computations does not have smooth boundary nor $H^2(\Omega)$-regularity, the actual pseudospectral set is in good agreement with our theoretical results. Most importantly, when $\epsilon$ is sufficiently large, the pseudospectrum curls around the origin as predicted. Due to the solution having less that $H^2(\Omega)$-regularity, the requirement on the mesh size on $L$-shaped domain just takes the form $h^\alpha \kappa < < 1$, for some $\alpha < 2$, depending on regularity of the exact solution.

The approximate pseudospectrum could also be used to to derive convergence estimate for GMRES method using Bratwurst shaped domains to solve...
the minimization problem. However, as a preconditioner would always be applied, the current case is not very interesting hence we do not proceed further with it.

### 4.3 Shifted-Laplace preconditioned Helmholtz equation

The analysis of inclusion and exclusion regions is more complicated, when a preconditioner is applied to speed up the convergence of the GMRES method. Several different preconditioners have been proposed for problem (24), see e.g. [6]. We consider here the shifted-Laplace preconditioner [8]. This preconditioner is based on solving an auxiliary problem on each step of the iteration. The auxiliary problem is defined as: For a given $u \in V$ find $Pu \in V$ such that

$$b(Pu, v) = \bar{x}^*_v \bar{x}_u \quad \forall v \in V. \quad (31)$$

The sesquilinear form $b$ in the above equation is given as

$$b(u, v) = (\nabla u, \nabla v) + i\kappa (u, v)_{\partial \Omega} - \kappa^2 (u, v) + i\sigma (u, v),$$

in which $\kappa$ is as defined in Section 4.2 and $\sigma \in \mathbb{R}, \sigma > 0$. This is, a loss term $i\sigma (u, v)$ is added to the sesquilinear from defined in equation (25). The addition of the loss term leads to a stability estimate on the finite element space $V$ independent of the mesh size. Choosing $v = Pu$ in equation (31) gives

$$b(Pu, Pu) = \bar{x}^*_u \bar{x}_u \quad \forall v \in V.$$

Taking imaginary part leads to

$$\kappa \|Pu\|_{\partial \Omega} + \sigma \|Pu\|^2_0 = \Im \bar{x}^*_Pu \bar{x}_u.$$

This is,

$$\sigma \|Pu\|^2_0 \leq \Im \bar{x}^*_Pu \bar{x}_u.$$

Now, using Cauchy-Schwarz inequality and norm equivalence [12] gives

$$\|Pu\|_0 \leq C\sigma^{-1} ||u||_0 \quad \forall u \in V \quad (32)$$

for some $C$. The matrix form of the preconditioner is denoted as $B^{-1}$, where $B_{ij} = b(\varphi_j, \varphi_i)$. Hence, the problem to be solved by the GMRES method is

$$AB^{-1} \bar{x} = \bar{b}, \quad \bar{x} = B^{-1}\bar{\bar{x}}.$$

The rationale behind using shifted-Laplace preconditioners is that when a sufficiently large loss term is added, the action of the preconditioner can be efficiently evaluated using a multigrid method [7]. When applied directly to solve the original problem (24), multigrid methods face two challenges, [4]. The standard smoothing iteration is not stable and the coarse grid correction has to be made on a sufficiently fine mesh. The introduction of a loss term has been analyzed in [11] for a problem with zero Dirichlet boundary conditions. In this
case, additional losses improve the multigrid solver by allowing the coarse grid correction to be made on a coarser mesh. The coarse grid depends on the loss term, hence there is a tradeoff between the number of GMRES iterations and the cost of applying the preconditioner. Typically, the loss parameter is chosen as $\sigma = 0.5\kappa^2$. For simplicity, we will consider here only the exact preconditioner. This gives good insight on what one can expect from the inexact case.

As we will see, a shifted-Laplace preconditioner can eliminate the mesh size dependency from the pseudospectral set. This is, the inclusion and exclusion regions are independent of the applied mesh size. This is a desired property, as the mesh size dependency in the non-preconditioned case leads quickly to an unbearably large number of iterations. The exclusion regions will, however depend on the ratio of $\kappa$ and $\sigma$.

The shifted-Laplace preconditioner has been previously analyzed in [25] by estimating the location of the eigenvalues. The existing analysis is not explicit in $\sigma$ and does not take the non-normality into account. In addition, the previous work does not include the exclusion region around the origin, which we can obtain using Theorem 3.1. and the stability result given in equation (28).

To study the shifted-Laplace preconditioner, we interpret the matrix $AB^{-1}$ as the matrix form of the sesquilinear form $a(Pu, u)$, where $a(u, v)$ is as defined in equation (25) and $P$ in equation (31). A suitable stability estimate for this sesquilinear form is established by the following Corollary.

**Corollary 4.1.** Let $u \in V$ be such that

$$a(Pu, v) = (f, v) \quad \forall v \in V. \quad (33)$$

In addition, let Assumption [4.7] be satisfied. Then there exists a constant $C > 0$ independent of $u, f, \kappa, h$ and $\sigma$ such that

$$|\vec{x}_u| \leq C h^{d/2} \left( 1 + \frac{\sigma}{\kappa} \right) \|f\|_0. \quad (34)$$

**Proof.** Application of equation (28) gives

$$\|Pu\|_\kappa \leq C \|f\|_0. \quad (35)$$

It follows from definition (31) that

$$a(Pu, u) = |\vec{x}_u|^2 - i\sigma(Pu, u).$$

Combining above with equation (33) gives

$$|\vec{x}_u|^2 = (f, u) + i\sigma(Pu, u). \quad (35)$$

Using Cauchy-Schwarz inequality, estimate (34) and norm equivalence (18) gives

$$|\vec{x}_u| \leq C h^{d/2} \left( \|f\|_0 + C_S \frac{\sigma}{\kappa} \|f\|_0 \right).$$

$\square$
The above stability estimate is given in the norm \( \| u \| = |\mathbf{x}_u| \). Hence, we choose this as the norm of the space \( V \). The above Corollary also suggest to choose the space \( W = L^2(\Omega) \) as previously. With these choices, a direct application of Theorem 3.1 gives the exclusion

\[
B(0, C \frac{\kappa}{\kappa + \sigma} - \epsilon) \subset \Lambda_\epsilon^c. \tag{36}
\]

When \( \sigma = 0 \), the preconditioner solves the problem exactly and \( \Lambda_\epsilon = B(1, \epsilon) \).

As the constant in above is \( C \) is independent of \( \sigma \) and \( \kappa \), setting \( \sigma = 0 \), leads to \( C \leq 1 \). A field of values based inclusion can be obtained as follows. There holds that

\[
\bar{\mathbf{x}}_u^a A B^{-1} \bar{\mathbf{x}}_u^a = a(Pu, u) = \bar{\mathbf{x}}_u^a \bar{\mathbf{x}}_u - i \sigma(Pu, u).
\]

An inclusion for FOV follows by estimating the last term. By the stability result given in equation (32) and norm equivalence (18), there holds that

\[
\sigma(Pu, u) \leq \sigma \| Pu \|_0 \| u \|_0 \leq C \bar{\mathbf{x}}_u^a \bar{\mathbf{x}}_u.
\]

This is, the FOV is located inside the set \(| 1 - z | \leq C_1 \).

The polynomial minimization problem in the GMRES convergence bound (8) does not give any information on the convergence, when the approximate pseudospectrum is an annulus surrounding the origin. To apply the FOV based estimate, one has to explicitly know the constants in derived inclusion and exclusion regions to guarantee that this cannot happen. The constant \( C_1 \) in the inclusion for FOV is related to the norm equivalence between \( L^2(\Omega) \) and Euclidian norm. It is easy to see, that \( C_1 = \sqrt{\text{cond}(M)} \), where \( M_{ij} = (\varphi_i, \varphi_j) \) is the mass matrix. In typical cases \( C_1 \approx 4 \), so that derived inclusion is not useful when \( \sigma = 0.5 \kappa^2 \) and the dimension of the exclusion tends to zero when \( \kappa \) grows.

Due to the close relation between the preconditioner and the original problem, we can estimate the pseudospectrum using a problem specific technique.

**Lemma 4.1.** There exists a positive constant \( C > 0 \) such that

\[
\left\{ z \in \mathbb{C} \mid C \left( \left| \frac{1}{|z|^2 - \Re z} \right| + \frac{1}{|1 - z|} \right) < \frac{1}{\epsilon} \right\} \cup B \left( \frac{1}{2}, \frac{1}{2} \right) \subset \Lambda_\epsilon^c
\]

**Proof.** There holds that \( A = A^T \) and \( B = B^T \). Using the identity \(|C| = |C^*| \) for any \( C \in \mathbb{C}^{N \times N} \), it follows that

\[
\sup_{\mathbf{x}_u \in \mathbb{C}^n} \left| \frac{(zI - AB^{-1})^{-1} \bar{\mathbf{x}}_u}{|\bar{\mathbf{x}}_u|} \right| = \sup_{\mathbf{x}_u \in \mathbb{C}^n} \left| \frac{(zB - A)^{-1} B \bar{\mathbf{x}}_u}{|\bar{\mathbf{x}}_u|} \right|
\]

Let \( \bar{\mathbf{x}}_v \in \mathbb{C}^N \) be such that

\[
(A - zB) \bar{\mathbf{x}}_v = B \bar{\mathbf{x}}_u.
\]
As in Section 3, we establish the stability estimate $|\vec{x}_v| \leq f(z)|\vec{x}_u|$. When $f(z)$ is finite, this estimate yields the desired bound. Testing with any $\vec{x}_w \in \mathbb{R}^N$ gives

$$(1 - z)a(v, w) - i\sigma z(v, w) = \vec{x}_w^r\vec{B}\vec{x}_u.$$ 

Assuming that $z \neq 1$ and dividing by $1 - z$ yields

$$a(v, w) - \frac{i\sigma z}{1 - z}(v, w) = a(\frac{u}{1 - z}, w) + \frac{i\sigma}{1 - z}(u, w).$$

By adding an subtracting a suitable term, the above can be written as

$$a(v - (1 - z)^{-1}u, w) - \frac{i\sigma z}{1 - z}(v - (1 - z)^{-1}u, w) = \frac{i\sigma}{(1 - z)^2}(u, w)$$

Choosing $w = v - (1 - z)^{-1}u$, using the identity $\frac{1}{1 - z} = \frac{z - |z|^2}{1 - |z|^2}$ and taking imaginary part gives

$$\kappa\|v - (1 - z)^{-1}u\|_0^2 + \sigma|\frac{|z|^2 - \Re z}{1 - |z|^2}\|v - (1 - z)^{-1}u\|_0^2 = \Im \frac{i\sigma}{(1 - z)^2}(u, v - (1 - z)^{-1}u)$$

When $z \neq 1$ and $\Re z - |z|^2 > 0$, this is

$$\left|\frac{1}{2} - z\right| \geq \frac{1}{2},$$

the coefficient of the $L^2(\Omega)$ - term is positive and we obtain the estimate

$$\|v - (1 - z)^{-1}u\|_0 \leq \frac{1}{|z|^2 - \Re z}\|u\|_0.$$ 

Using the norm equivalence given in equation (18) yields

$$|\vec{x}_v - (1 - z)^{-1}\vec{x}_u| \leq C\frac{1}{|z|^2 - \Re z}|\vec{x}_u|.$$ 

The stability estimate follows from the above equation and triangle inequality as

$$|\vec{x}_v| \leq |\vec{x}_v - (1 - z)^{-1}\vec{x}_u| + \frac{|\vec{x}_u|}{1 - z} \leq C\left(\frac{1}{|z|^2 - \Re z} + \frac{1}{1 - z}\right)|\vec{x}_u|.$$ 

To obtain an overview of the derived bounds we have computed the pseudospectrum for $\kappa = 16\pi$ and $\sigma = 0.5\kappa, 0.5\kappa^2$ using the level three mesh. The results are presented in Figure 4. Based on these results, analysis given in this Section seems to capture the behavior of the pseudospectrum rather well. In both cases, when $\epsilon$ is sufficiently small, pseudospectrum is located inside the...
disc \( B\left(\frac{1}{2}, \frac{1}{2} \right) \), as predicted by Lemma 4.1. When the loss term is small, the pseudospectrum has a rather small diameter and is located close to 1. For large values of \( \sigma \), the set moves closer to the origin. These results are in good agreement with the exclusion given in equation (39).

The GMRES convergence bound gives usable information only if the origin is outside the approximate pseudospectrum. In the current case, this requirement limits the value of \( \epsilon \) and thus determines the GMRES convergence rate. We have studied the pseudospectrum close to the origin in more detail by using a bisection search to find \( x \in \mathbb{R} \) closest to the origin such that \( \| (xI - AB^{-1})^{-1} \| = 2 \cdot 10^{-2} \) for different \( \kappa \) between \( 4\pi \) and \( 64\pi \) for \( \sigma = 0.5\kappa \) and \( \sigma = 0.5\kappa^2 \). The results are visualized in Figure 5. These results indicate, that the exclusion given in equation (39) corresponds well with the real behavior of the set.

![Figure 4: Pseudospectrum for Example 4.3 with \( \epsilon = 1, 10, 100, 1000 \). The parameter \( \kappa = 16\pi \) and level three mesh was used. On left the loss term is chose as \( \sigma = 0.5\kappa \) and on right as \( \sigma = 0.5\kappa^2 \). The circle \( B\left(\frac{1}{2}, \frac{1}{2} \right) \) is visualized with a dashed line.](image)

A rigorous derivation of convergence estimate based on bratwurst shaped domains would require us to relate the parameters of these domains to \( \tilde{\Lambda}_\epsilon \), which is out of the scope of this paper. Our computations indicate that the pseudospectrum for sufficiently large \( \epsilon \) can be contained inside a circle, hence we will instead use the bound for circles to derive an approximate convergence rate. Based on the numerical and theoretical results, it seems to be reasonable to choose

\[
\tilde{\Lambda}_\epsilon = B(1, 1 - \frac{\kappa}{\kappa + \sigma}) \oplus B(0, \epsilon).
\]

When \( \sigma = 0.5\kappa^2 \), there holds that \( \tilde{\Lambda}_\epsilon \subset B(1, 1 - 0.5\kappa^{-1}) \oplus B(0, \epsilon) \). To exclude the origin, we choose \( \epsilon = 0.25\kappa^{-1} \). Using equation (8) and polynomial minimization over circles [10], this leads to the estimate

\[
\frac{|\tilde{r}_i|}{|\tilde{r}_0|} \leq 4\kappa \left( \frac{1}{1 + 0.25\kappa^{-1}} \right)^i.
\]
Figure 5: The point $x \in \mathbb{R}$ closest to the origin such that $|(xI - A)^{-1}| = 2 \cdot 10^{-2}$ for Example 4.3. The dependency is as predicted by the exclusion. Mesh level five was used in this computation.

Figure 6: The number of GMRES iterations required to solve the problem for different values of $\kappa$. The stopping criteria was set as $tol = 10^{-6}$. Level seven mesh was used in the computation.

This is the required number of iterations $N$ to reach tolerance $tol$ is

$$N \approx -4\kappa \log tol + 4\kappa \log \kappa. \quad (37)$$

So, asymptotically, the dominating term is $\kappa \log \kappa$. We cannot observe this effect in our numerical examples as it would require us to use extremely large values of $\kappa$. For instance, when $\epsilon = 10^{-6}$, $\kappa$ would need to be of the order $10^6$, before it has an impact on the required number of GMRES iterations. This means, that the non-normality is not practically relevant in this case.

Estimate (37) rises the question, how should the stopping tolerance $tol$ be chosen. Using the tools derived in this paper, the size of relative residual can be related to $\kappa$-dependent norm. As we have studied right preconditioning, the solution obtained from GMRES $\vec{x}_i = B^{-1}\tilde{\vec{x}}_i$ satisfies $\vec{r}_i = A\vec{x}_i - \vec{b} = AB^{-1}\tilde{\vec{x}}_i - \vec{b}$. Hence, we can derive the estimate for the system without a preconditioner. The derived result holds for all left preconditioned systems.

**Lemma 4.2.** Consider the problem $A\vec{x}_u = \vec{b}$, in which $A \in \mathbb{C}^{N \times N}$ and $\vec{b} \in \mathbb{C}^N$ are related to the finite element discretization of problem (24). Let $\vec{x}_u$ be such that $|A\vec{x}_u - \vec{b}| \leq tol |\tilde{\vec{b}}|$. In addition, let Assumption 4.1 hold. Then there exists a constant $C > 0$ independent of $tol, u, \tilde{u}, \kappa$ and $h$ such that

$$||u - \tilde{u}||_{\kappa} \leq C tol \left(||f||_0 + h^{-1/2}||g||_{0, \partial\Omega}\right).$$

*Proof.* Denote $\vec{r} = A\vec{x}_u - \vec{b}$. There holds that $A\vec{x}_u - \vec{b} = A(\vec{x}_u - \vec{x}_u)$, hence,
error \[ \bar{e} = \bar{x}_u - \bar{x}_w \] is a solution to the equation,

\[ A\bar{e} = \bar{r}. \]

Let the space \[ W = L^2(\Omega) \]. Using the same construction as in the proof of Corollary 3.1, we define \( q \in V \) such that \( (q, v)_W = \bar{r}^\ast \bar{x}_v \) \( \forall v \in V \) and the linear functional \( L(w) = (q, w)_W \). Using standard tools and the norm equivalence gives \( \| L \|_{W^\prime} \leq C h^{-d/2} |\bar{r}| \). The stability estimate given in equation (30) leads to

\[ \| u - \tilde{u} \|_\kappa \leq C h^{-d/2} |\bar{r}|. \]

Now, this can be written as

\[ \| u - \tilde{u} \|_\kappa \leq C h^{-d/2} tol |\bar{b}|. \]

As there holds that

\[ |\bar{b}| = \max_{x_v \in \mathbb{R}^N} \bar{b}^\ast \bar{x}_v = \frac{(f, v) + (g, v)_{\partial\Omega}}{|\bar{x}_v|}, \]

Cauchy-Schwarz inequality and norm equivalence gives

\[ |\bar{b}| \leq C h^{d/2} \left( \| f \|_0 + h^{-1/2} \| g \|_0 \right) \]

One should note that identical techniques that were used to prove the above Lemma can be used to derive a relation between the \( V \)-norm and \( tol \) for any finite dimensional variational problem satisfying Assumption 2.1.

We conclude by solving the shifted-Laplace preconditioned problem for right-hand side

\[ f = \exp \left( -10^3 ((x - 0.5)^2 + (y + 0.5)^2) \right) \]

and different values of \( \kappa \). The loss term for the preconditioner was chosen as \( \sigma = 0.5 \kappa^2 \) and \( 0.5 \kappa \) and the level five mesh was used in the computations. The number of GMRES iterations is plotted in Figure 6. In this case, we observe a linear relationship between \( \kappa \) and the number of iterations for \( \sigma = 0.5 \kappa^2 \). The number of iterations stays constant for \( \sigma = 0.5 \kappa \). These results are in good agreement with the estimate (37).

5 Conclusions

The main result of the paper is the derivation of exclusion region for pseudospectral set near the origin, Theorem 3.1. The derivation was made under Assumption 2.1 stability of the weak problem. All analysis was done a priori, without constructing any matrices. Theorem 3.1 was applied in all three tests, and the derived results were in good agreement with the true behavior of the pseudospectral set. In addition, an inclusion region was derived using
the connection between FOV and the pseudospectrum. Boundedness estimates for FOV were derived based on the properties of the weak problem. All given analysis is applicable to a wide range of different problems.

As demonstrated by the examples, the proposed inclusion and exclusion regions led to a worst case convergence estimate for the GMRES method. However, the effect of this overestimation was significant only for extreme parameter values. As illustrated by the first example, more refined convergence estimate would require knowledge from behavior of pseudospectrum for $\epsilon \to 0$. Such analysis is one direction for continuing this work.

The aim of the paper was to investigate, if pseudospectrum based convergence estimate can be used for relating properties of weak form to convergence of GMRES. This was proven to be possible. As in Example 4.3, one needs to establish stability and boundedness of the preconditioned problem. The application of the derived theory will lead to inclusion and exclusion regions for pseudospectrum. Second possible direction for future work is to study different preconditioners and problems using the derived tools. Natural extension would be to investigate convergence of GMRES for time-harmonic Maxwell’s equations.

References

[1] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, third edition, 2007.

[2] X-C. Cai and O. Widlund. Domain decomposition algorithms for indefinite elliptic problems. *SIAM J. Sci. Statist. Comput.*, 12(1):243–258, 1992.

[3] X-C. Cai and O. Widlund. Multiplicative Schwarz algorithms for some nonsymmetric and indefinite problems. *SIAM J. Numer. Anal.*, 30(4):936–952, 1993.

[4] H. C. Elman, O. G. Ernst, and D. P. O’Leary. A multigrid method enhanced by Krylov subspace iteration for discrete Helmholtz equations. *SIAM J. Sci. Comput.*, 23:1291–1315, April 2001.

[5] M. Embree. How descriptive are GMRES convergence bounds? Technical report, Oxford University Computing Laboratory, 1999.

[6] Y. A. Erlangga. Advances in iterative methods and preconditioners for the Helmholtz equation. *Arch. Comput. Methods Eng.*, 15(1):37–66, 2008.

[7] Y. A. Erlangga, C. W. Oosterlee, and C. Vuik. A novel multigrid based preconditioner for heterogeneous Helmholtz problems. *SIAM J. Sci. Comput.*, 27(4):1471–1492, 2006.

[8] Y. A. Erlangga, C. Vuik, and C. W. Oosterlee. On a class of preconditioners for solving the Helmholtz equation. *Appl. Numer. Math.*, 50(3-4):409–425, 2004.
[9] J. Gopalakrishnan and J. Pasciak. Overlapping Schwarz preconditioners for indefinite time harmonic Maxwell equations. *Math. Comp.*, 72:1–15, 2003.

[10] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, 1997.

[11] A. Hannukainen. Analysis of shifted-Laplace preconditioners for the Helmholtz equation. Submitted to BIT. Num. Math, 2013.

[12] A. Hannukainen. Field of values analysis of a two-level preconditioner for the Helmholtz equation. *SIAM J. Numer. Anal.*, 51(3):1567–1584, 2013.

[13] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number. I. The $h$-version of the FEM. *Comput. Math. Appl.*, 30(9):9–37, 1995.

[14] F. Ihlenburg and I. Babuška. Finite element solution of the Helmholtz equation with high wave number. II. The $h$-$p$ version of the FEM. *SIAM J. Numer. Anal.*, 34(1):315–358, 1997.

[15] T. Koch and J. Liesen. The conformal ”bratwurst” maps and associated Faber polynomials. *Numerische Mathematik*, 86(1):173–191, 2000.

[16] I. Koutis. Spectrum through pseudospectrum. arXiv:math/0701368, 2001.

[17] M. Embree L. N. Trefethen. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

[18] Z. Leyk, J. E. Pasciak, and J. Bramble. Iterative schemes for non-symmetric and indefinite elliptic boundary value problems. *Math. Comp.*, 60:1–22, 1993.

[19] J. Melenk. *On Generalized Finite Element Methods*. PhD thesis, The University of Maryland, 1995.

[20] A. Quarteroni and A. Valli. *Numerical Approximation of Partial Differential Equations*. Springer-Verlag, 1994.

[21] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2003.

[22] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.

[23] L. N. Trefethen. Pseudospectra of linear operators. *SIAM Rev.*, 39(3):383–406, September 1997.

[24] L. N. Trefethen. Computation of pseudospectra. *Acta Numerica*, 8:247–295, 1999.

[25] M. B. van Gijzen, Y. A. Erlangga, and C. Vuik. Spectral analysis of the discrete Helmholtz operator preconditioned with a shifted Laplacian. *SIAM J. Sci. Comput.*, 29(5):1942–1958, 2007.
[26] M.B van Gijzen and Y.A. Erlangga. Convergence bounds for preconditioned GMRES using element-by-element estimates of the field of values. In J. Périaux P. Wesseling, E. Oñate, editor, ECCOMAS CFD, 2006.

[27] K. Yoshida. Functional Analysis. Springer-Verlag, 1995.

[28] H. Yserentant. Preconditioning indefinite discretization matrices. Numer. Math., 54(6):719–734, 1989.
