Analysis of the shifted Helmholtz expansion preconditioner for the Helmholtz equation

Pierre-Henri Cocquet[1], Martin J. Gander[2]

1 Introduction

Solving Helmholtz problem numerically is challenging [**?**] mainly because of the lack of coercivity of the continuous operator or highly oscillatory solutions. Krylov subspaces methods like GMRES are still used in regards of their robustness but they require a good preconditioner[1] to be fast enough. Among many proposed preconditioners like Incomplete LU, Analytic ILU or domain decomposition based preconditioner, the shifted Helmholtz preconditioner [**?**, **?**, **?**, **?**] has received a lot of attention over the last decade thanks to its simplicity and its relevance to heterogeneous media.

This paper focus on the recent idea of expansion preconditioner [**?**, **?**] which is based on the fact that the inverse of the discrete Helmholtz operator can be written as a superposition of inverse of discrete shifted Helmholtz operator only. This is achieved using the Taylor's expansion, around $\beta = 0$, of the matrix-valued function $f(\beta) = (-\Delta_h - (1+\mathrm{i}\beta)k^2)^{-1}$, where Δ_h corresponds to a finite difference discretization of the usual Laplace operator. The expansion-preconditioner is then defined as the truncation of the Taylor's series hence converging to the exact inverse of the discrete Helmholtz operator if the Taylor series actually converges. They also proposed to compute each inverse of shifted Helmholtz with some iteration of multigrid which is known to converge with a number of iterations independent of the wavenumber (see e.g. [**?**, **?**]). We emphasize that the rate of convergence of the expansion preconditioner toward $A_0^{-1} = f(0)$ is computed in [**?**] and is given to be a $O(\beta^n)$. However, the latter does not involves bounds on the higher derivative of f which can deteriorate the performance of the proposed preconditioner and no additional analysis is performed.

(1) Université de la Réunion, PIMENT, 2 rue Joseph Wetzell, 97490 Sainte-Clotilde.
(2) University of Geneva, 2-4 rue du Lièvre, CP 64, 1211 Genève, Switzerland,
{martin.gander@unige.ch}{pierre-henri.cocquet@univ-reunion.fr}

[1] For a linear system $C\mathbf{x} = \mathbf{y}$, a good preconditioner refer to a matrix B for which the spectrum of $B^{-1}C$ is clustered around 1 (see e.g Elman's estimate [**?**]).

The goal of this paper is to give a theoretical insight of the performances of the expansion preconditioner and to extend its definition to Finite Element discretization. We first build the expansion preconditioner using the generalized resolvent formula and study its performances. We next show, as proved in [**?**], that it is mandatory to have a shift of the order of the wavenumber to get wavenumber independant convergence of GMRES. This paper ends with some numerical simulations.

2 General analysis of the expansion preconditioner

Let Ω be a convex polygon of $\mathbb{R}^d$, with $d = 1,2,3$. The shifted Helmholtz equation with impedance boundary conditions is

$$\begin{cases} -\Delta u(x) - (k^2 + \mathrm{i}\varepsilon)u(x) = f(x),\ x \in \Omega, \\ \partial_{\mathbf{n}} u - \mathrm{i}\eta u = 0,\ \text{on } \partial\Omega, \end{cases} \tag{1}$$

where $\mathbf{n}$ is the unitary normal vector directed outward $\partial\Omega$, $\varepsilon > 0$ is the so-called shift, and $\eta > 0$ is the impedance parameter. The Helmholtz equation with approximate radiation condition is recovered from (**??**) by setting $\varepsilon = 0$ and $\eta = k$.

The variational form of (**??**) is given below

$$\begin{cases} \text{Find } u \in H^1(\Omega) \text{ such that for all } v \in H^1(\Omega) : \\ a_\eta(u,v) := \displaystyle\int_\Omega \nabla u \cdot \overline{\nabla v} - (k^2 + \mathrm{i}\varepsilon) u\bar{v} dx - \mathrm{i}\eta \int_{\partial\Omega} u\bar{v} d\sigma = \int_\Omega f\bar{v} dx. \end{cases} \tag{2}$$

Let $\mathscr{V}_l$ be the finite element space obtained with piecewise linear polynomials

$$\mathscr{V}_l = \left\{ v \in \mathscr{C}(\overline{\Omega}) \mid v|_T \in \mathbb{P}_1 \text{ for all } T \in \mathscr{T}_l \right\} = \mathrm{Span}(\phi_1, \cdots, \phi_N),$$

where $\left\{\phi_j\right\}_{j=1}^N$ is the finite element nodal basis. The discrete problem is then

$$\begin{cases} \text{Find } u_l \in \mathscr{V}_l \text{ such that :} \\ a_\varepsilon(u_l, v_l) = \displaystyle\int_\Omega f\overline{v_l} dx,\ \forall v_l \in \mathscr{V}_l. \end{cases} \tag{3}$$

The latter is equivalent to the linear system $A_\varepsilon \mathbf{z}_l = \mathbf{b}_l$ where $u_l = F_h \mathbf{z}_l$ is the Galerkin solution and

$$F_h : x = (x_1, \cdots, x_N) \in \mathbb{C}^N \mapsto \sum_{j=1}^N x_j \phi_j \in \mathscr{V}_h.$$

Denoting by S, M, N respectively the stiffness, mass and boundary mass matrix, one gets

$$A_\varepsilon = S - (k^2 + \mathrm{i}\varepsilon)M - \mathrm{i}\eta N.$$

We denote by A_0 the discrete Helmholtz operator obtained with $\varepsilon = 0$ and $\eta = k$. We emphasize that this matrix is invertible thanks to the impedance boundary

condition. Also, if Dirichlet or Neumann's boundary conditions are used, we assume throughout this paper that A_0 is invertible.

We now give a generalized resolvent formula whose proof can be done by routine computations.

Lemma 1. *Let $A,B \in \mathrm{Hom}(\mathbb{C}^n)$ with B invertible and $p,z \in \mathbb{C}$ be two complex numbers in the resolvent set of AB^{-1}. Let $R(z) = (A - zB)^{-1}$ be the generalized resolvent of A. The following formula then holds*

$$R(p) - R(z) = (z-p)R(z)BR(p).$$

Using Neumann's series, Lemma **??** allow to rewrite the inverse of the discrete Helmholtz operator as a superposition of discrete shifted Helmholtz operator.

Theorem 1. *The inverse of the discrete Helmholtz operator is given as follows*

$$A_0^{-1} = \left(\sum_{j \geq 0} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M \right)^j \right) A_\varepsilon^{-1},$$

where the serie converges with respect to the norm $\|x\|_M = \sqrt{\langle Mx, \bar{x} \rangle} = \|F_h x\|_{L^2(\Omega)}$.

Proof. Lemma **??** applied with $A = A_0$, $B = M$, $p = 0$ and $z = \mathrm{i}\varepsilon$ yields

$$A_0^{-1} = (I_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M)^{-1} A_\varepsilon^{-1}.$$

Note that $A_\varepsilon^{-1} M = (M^{-1} A_\varepsilon)^{-1}$. Let $\mathbf{z} \in \mathbb{C}^N$ such that $A_\varepsilon \mathbf{z} = M\mathbf{b}$ for some $\mathbf{b} \in \mathbb{C}^N$. From the definition of the mass matrix M, the operator F_h and A_ε, one gets

$$a_\eta(F_h\mathbf{z}, F_h\mathbf{z}) = \langle M\mathbf{b}, \bar{\mathbf{z}} \rangle = \left(F_h\mathbf{b}, \overline{F_h\mathbf{z}} \right)_{L^2(\Omega)}.$$

Cauchy-Schwartz inequality and the next lower bound

$$|a_\eta(F_h\mathbf{z}, F_h\mathbf{z})| > |\mathscr{I}\, a_\eta(F_h\mathbf{z}, F_h\mathbf{z})| = \varepsilon \|F_h\mathbf{z}\|^2_{L^2(\Omega)} + \eta \|F_h\mathbf{z}\|^2_{L^2(\partial\Omega)},$$

show that $\|\mathbf{z}\|_M < \|\mathbf{b}\|_M \varepsilon^{-1}$, and thus $\left\|\varepsilon A_\varepsilon^{-1} M\right\|_M < 1$. Finally, $(I_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M)^{-1}$ can be expanded as a Neumann's serie and the proof is finished.

Remark 2 *The mass matrix is symmetric and positive definite so it admits a square root $M^{1/2}$. For any $B \in \mathrm{Hom}(\mathbb{C}^N)$, the matrix norm induced by $\|\cdot\|_M$ is then defined by $\|B\|_M = \left\|M^{1/2} B M^{-1/2}\right\|_2$. This yields*

$$\left\|\varepsilon A_\varepsilon^{-1} M\right\|_M = \varepsilon \left\| M^{1/2} A_\varepsilon^{-1} M^{1/2} \right\|_2 = \varepsilon \left\| A_\varepsilon^{-1} M \right\|_2 < 1,$$

and thus the series from Theorem **??** *converges with respect to the* 2*-norm as well.*

Following [**?**], the expansion preconditioner of order $n \in \mathbb{N}^*$ is defined as a truncation of the Neumann's serie given in Theorem **??**

$$EX(n) = \left(\sum_{j=0}^{n-1} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M\right)^{j+1} \right) M^{-1} = \left(\sum_{j=0}^{n-1} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M\right)^{j} \right) A_\varepsilon^{-1}. \quad (4)$$

The preconditioned problem is thus given as follow

$$EX(n) A_0 \mathbf{z}_l = EX(n) \mathbf{b}_l. \quad (5)$$

From Elman's estimate (see e.g. Theorem 1.8 [**?**]), the rate of convergence of GMRES used for solving $C\mathbf{x} = \mathbf{y}$ only depend on the upper bound of $\|\mathbb{I} - C\|_2$. We now compute this term for the expansion preconditioner.

Theorem 3. *For any shift $\varepsilon > 0$, impedance parameter $\eta > 0$, meshsize h and $n \in \mathbb{N}$, the expansion preconditioner satisfies the following bounds*

$$\mathscr{N}\left(\mathbb{I}_d - EX(1)A_0\right) \leq \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right),$$
$$\forall n \geq 1,\ \mathscr{N}\left(\mathbb{I}_d - EX(n)A_0\right) \leq \frac{1 + \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)}{1 - \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)} (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^n,$$

where $\mathscr{N}(B)$ denotes any matrix norm or $\rho(B)$.

Proof. The first item follows from $\mathbb{I} - EX(1)A_0 = \mathbb{I} - A_\varepsilon^{-1} A_0 = \mathrm{i}\varepsilon A_\varepsilon^{-1} M$. For the second one, we compute

$$\mathbb{I} - EX(n)A_0 = (A_0^{-1} - EX(n))A_0 = \left(\sum_{j \geq n} (-\mathrm{i}\varepsilon)^j (A_\varepsilon^{-1} M)^j \right) A_\varepsilon^{-1} A_0.$$

Note that $A_\varepsilon^{-1} A_0 = \mathbb{I}_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M$ and thus $A_\varepsilon^{-1} A_0$ and $A_\varepsilon^{-1} M$ commute. Now, using that $\varepsilon \rho(A_\varepsilon^{-1} M) \leq \varepsilon \left\| A_\varepsilon^{-1} M \right\|_2 < 1$, we can use Gelfand's formula to get the convergence of the Neumann series with respect to any matrix norm. Majoring and expanding using geometric serie then give

$$\begin{aligned} \mathscr{N}\left(\mathbb{I} - EX(n)A_0\right) &\leq \mathscr{N}\left(\mathbb{I}_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M\right) (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^n \sum_{j \geq 0} (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^j \\ &\leq \frac{1 + \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)}{1 - \varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)} (\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right))^n. \end{aligned}$$

Remark 4 *The construction of the expansion preconditioner as well as Theorem* **??** *hold without any changes for high order Finite Element discretization.*

The upper bound from Theorem **??** involves only $\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)$. If the latter is bounded away from 1, the expansion preconditioner can greatly reduce the number of GMRES iterations by considering a large enough n.

3 Wavenumber-independance convergence of GMRES

We show in this section that, as proved in [**?**], taking $\varepsilon \sim k$ is mandatory to ensure wavenumber-independant convergence of GMRES when using an expansion preconditioner. This is done in the next result for two types of meshes: one for which one has pollution-free FEM[2] and one for $h \sim k^{-2}$.

Theorem 5. *Assume that one of the following assumptions holds*

(A1) $\eta \sim k$ *and* $k^3h^2 \leq C_0$ *holds with* C_0 *small enough.*

(A2) $\eta \lesssim k$, $k \geq k_0$ *for a given* $k_0 > 0$ *and* $kh\sqrt{|k^2-\varepsilon|} \leq C_0$ *holds with* C_0 *small enough.*

Then there exists a constant $C_2 > 0$ *depending only on* Ω *such that for any* $\varepsilon > 0$ *with* $\varepsilon C_2 < k$, *one has*

$$\forall\, n \geq 1,\ \mathscr{N}\left(\mathbb{I}_d - EX(n)A_0\right) \leq \left(\frac{C_2\varepsilon}{k}\right)^n \frac{k+C_2\varepsilon}{k-C_2\varepsilon},$$

where $\mathscr{N}(.) = \rho(.)$ *if (A1) hold and* $\mathscr{N}(.) = \|.\|_2$ *if (A2) hold.*

Proof. Assume that (A1) hold. Let $\lambda \in \mathbb{C}$ be an eigenvalue of $M^{-1}A_\varepsilon = (A_\varepsilon^{-1}M)^{-1}$ and $\mathbf{v} \in \mathbb{C}^N$ the associated eigenvector. One has

$$M^{-1}A_\varepsilon \mathbf{v} = \left(M^{-1}(S - \mathrm{i}\eta N) - (k^2 + \mathrm{i}\varepsilon)\mathbb{I}_N\right)\mathbf{v} = \lambda \mathbf{v}.$$

Therefore, the spectrum of $M^{-1}A_\varepsilon$ is given by

$$\sigma(M^{-1}A_\varepsilon) = \left\{\lambda_j + \mathrm{i}\varepsilon \mid \lambda_j \in \sigma(M^{-1}A_0)\right\},$$

from which we infer that

$$\varepsilon\rho(A_\varepsilon^{-1}M) = \max_{\lambda_j \in \sigma(M^{-1}A_0)} \frac{\varepsilon}{|\lambda_j + \mathrm{i}\varepsilon|} \tag{6}$$

Let $\mathbf{b} \in \mathbb{C}^N$ be fixed and $\varphi_h \in \mathbb{C}^N$ be the solution to $A_0\mathbf{v}_h = M\mathbf{b}$. Note that $\varphi_h = F_h\mathbf{v}_h \in \mathscr{V}_h$ corresponds to the FEM discretization of the solution to (**??**) with $f = F_h\mathbf{b}$. Since $f \in L^2(\Omega)$ and Ω is assumed to be convex, the solution to the Helmholtz equation (**??**) belongs to $H^2(\Omega)$. Since (A1) hold, one can apply [**?**, Corollary 4.4 p.12] to get

$$\|\nabla\varphi_h\|_{L^2(\Omega)} + k\|\varphi_h\|_{L^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)}. \tag{7}$$

Then (**??**) shows that

$$\|F_h\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \frac{1}{k}\|F_h\mathbf{b}\|_{L^2(\Omega)}.$$

Using [**?**, Eq. (4.2) p. 24], one has $\|F_h\|_{\mathbb{C}^N \to \mathscr{V}_h} \sim h^{d/2}$ which gives

$$\|\mathbf{v}_h\|_2 = \left\|A_0^{-1}M\mathbf{b}\right\|_2 \lesssim \frac{\|\mathbf{b}\|}{k}.$$

[2] According to [**?**] no pollution effect occurs if $k^3h^2 \leq C_0$ holds with C_0 small enough.

The above estimate holds for any $\mathbf{b} \in \mathbb{C}^N$ and thus

$$\left\|A_0^{-1}M\right\|_2 \lesssim \frac{1}{k}. \tag{8}$$

The upper bound (**??**) proves that, for any $\mu \in \sigma(A_0^{-1}M)$, $|\mu| \lesssim k^{-1}$. Since any $\lambda \in \sigma(M^{-1}A_0)$ can be written as $\lambda = 1/\mu$, one gets $k \lesssim |\lambda|$. We finally infer that there exists $C_2 > 0$ depending only on Ω such that

$$\rho(A_\varepsilon^{-1}M) \leq \frac{C_2}{k}. \tag{9}$$

Assuming now that (A2) hold allow to apply [**?**, Lemma 3.5 p.595] that gives the quasi-optimality of the bilinear form a_ε on $\mathscr{V}_h$ with respect to the weighted norm $\|u\|_{1,k}^2 = \|\nabla u\|_{L^2(\Omega)}^2 + k^2\|u\|_{L^2(\Omega)}^2$. Using this, they proved [**?**, Lemma 4.1 p. 598] that there exists a constant C_2 depending only on Ω such that

$$\left\|A_\varepsilon^{-1}M\right\|_2 \leq \frac{C_2}{k}. \tag{10}$$

Using now (**??**) and (**??**) together with the bound proved in Theorem **??** ends the proof.

4 Numerical simulations

5 Conclusions

References

1. Cai, X. C., & Widlund, O. B. (1992). Domain decomposition algorithms for indefinite elliptic problems. SIAM Journal on Scientific and Statistical Computing, 13(1), 243-258.
2. Cocquet, P. H., & Gander, M. J. (2016). On the minimal shift in the shifted Laplacian preconditioner for multigrid to work. In Domain Decomposition Methods in Science and Engineering XXII (pp. 137-145). Springer International Publishing.
3. Cocquet, P. H., & Gander, M. J. (2017). How Large a Shift is Needed in the Shifted Helmholtz Preconditioner for its Effective Inversion by Multigrid?. SIAM Journal on Scientific Computing, 39(2), A438-A478.
4. Gander, M.J., Graham, I.G., & Spence, E.A. (2015). Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: what is the largest shift for which wavenumber-independent convergence is guaranteed?. Numerische Mathematik, 131(3), 567-614.
5. Cools, S., & Vanroose, W. (2015). Generalization of the complex shifted Laplacian: on the class of expansion preconditioners for Helmholtz problems. ArXiv e-prints.
6. Cools, S., & Vanroose, W. (2017). On the Optimality of Shifted Laplacian in a Class of Polynomial Preconditioners for the Helmholtz Equation. In Modern Solvers for Helmholtz Problems (pp. 53-81). Springer International Publishing.
7. Y.A Erlangga, C. Vuik, C.W. Oosterlee. On a class of preconditioners for solving the discrete Helmholtz equation, *Applied Numerical Mathematics*, p. 409-425, 2004.

8. Ernst, O. G., & Gander, M. J. (2012). Why it is difficult to solve Helmholtz problems with classical iterative methods. In Numerical analysis of multiscale problems (pp. 325-363). Springer Berlin Heidelberg.
9. Du, Y., & Wu, H. (2015). Preasymptotic error analysis of higher order FEM and CIP-FEM for Helmholtz equation with high wave number. SIAM Journal on Numerical Analysis, 53(2), 782-804.