Analysis of the shifted Helmholtz expansion preconditioner for the Helmholtz equation

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

1 Introduction

Solving discretized Helmholtz problems by iterative methods is challenging [8], mainly because of the lack of coercivity of the continuous operator and the highly oscillatory nature of the solutions. Krylov subspaces methods like GMRES are the methods of choice because of their robustness, but they require a good preconditioner to be effective. Among many proposed preconditioners like Incomplete LU, Analytic ILU or domain decomposition based preconditioners, the shifted Helmholtz preconditioner has received a lot of attention over the last decade, because of its simplicity and its relevance to heterogeneous media, see [7, 2, 3, 4] and references therein.

We focus here on the recent idea of a generalization, the expansion preconditioner [5, 6], which is based on the fact that the inverse of the discrete Helmholtz operator can be written as a superposition of inverses of discrete shifted Helmholtz operators only. This is achieved with 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 Laplace operator, using a Taylor expansion to evaluate the function at $\beta = 0$,

$$f(0) = \sum_{j\geq 0} \frac{f^{(j)}(\beta)}{n!}(-\beta)^j = \sum_{j\geq 0} (-\mathrm{i}\beta k^2)^j f(\beta)^{-(j+1)}. \tag{1}$$

The expansion preconditioner is then defined as the truncation of the Taylor series, and converges to the exact inverse of the discrete Helmholtz operator if the Taylor series actually converges. The authors in [5, 6] also propose to compute each inverse of the shifted Helmholtz operator in the expansion preconditioner (1) approximately using multigrid, which can converge with a number of iterations independent of the

(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}

wavenumber for large enough shifts, see e.g. [2, 3]. The rate of convergence of the expansion preconditioner toward $A_0^{-1} = f(0)$ is computed in [5] to be $O(\beta^n)$, but this result is obtained without bounds on the higher derivatives of f which can deteriorate the performance of the proposed preconditioner.

The goal of this paper is to give theoretical and numerical insight for the performance of the expansion preconditioner, and to extend its definition to finite element discretizations. We first build the expansion preconditioner using the generalized resolvent formula and study its performance. We next show, as proved in [4] for the shifted Helmholtz operator, that a shift of the order of at most the wavenumber ensures wavenumber independant convergence of GMRES preconditioned with the expansion preconditioner. We then illustrate our results with numerical experiments, which indicate that even a larger shift might be tolerable.

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

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

where $\mathbf{n}$ is the unit outward normal on $\partial\Omega$, $\varepsilon > 0$ is the so-called shift, and $\eta > 0$ is the impedance parameter. The Helmholtz equation with approximate radiation condition is obtained from (2) by setting $\varepsilon = 0$ and $\eta = k$. The variational form of (2) is

$$\begin{cases} \text{Find } u \in H^1(\Omega) \text{ such that for all } v \in H^1(\Omega) : \\ a_\varepsilon(u,v) := \displaystyle\int_\Omega \nabla u \cdot \overline{\nabla v} - (k^2 + \mathrm{i}\varepsilon) u\bar{v} dx - \mathrm{i}\eta \int_{\partial\Omega} u\bar{v} d\sigma = \int_\Omega f\bar{v} dx. \end{cases} \tag{3}$$

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

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

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

$$\begin{cases} \text{Find } u_h \in \mathscr{V}_h \text{ such that :} \\ a_\varepsilon(u_h, v_h) = \displaystyle\int_\Omega f\overline{v_h} dx,\ \forall v_h \in \mathscr{V}_h. \end{cases} \tag{4}$$

This is equivalent to the linear system $A_\varepsilon \mathbf{w}_h = \mathbf{b}_h$ where $u_h = F_h \mathbf{w}_h$ 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 K the stiffness, M the mass and N the boundary mass matrix, we get

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

We denote by A_0 the discrete Helmholtz operator obtained with $\varepsilon = 0$ and $\eta = k$. The matrix A_0 is invertible because of the impedance boundary condition in (2). We now give a generalized resolvent formula which can be obtain by a direct computation.

Lemma 1. *Let $A, B \in \mathbb{C}^{n\times n}$ with B invertible, and let $p, z \in \mathbb{C}$ be two complex numbers in the resolvent set of AB^{-1}. Let $R(z) := (A - zB)^{-1}$ be the generalized resolvent of A. Then the relation $R(p) - R(z) = (z-p)R(z)BR(p)$ holds.*

Using the Neumann series, Lemma 1 allows us to rewrite the inverse of the discrete Helmholtz operator as a superposition of discrete shifted Helmholtz operators:

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

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

and the series converges in the norm $\|x\|_M = \sqrt{\langle Mx, \overline{x}\rangle} = \|F_h x\|_{L^2(\Omega)}$.

Proof. Lemma 1 applied with $A = A_0$, $B = M$, $p = 0$ and $z = \mathrm{i}\varepsilon$ yields

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

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

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

Then using the Cauchy-Schwarz inequality and the lower bound

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

we obtain that $\|\mathbf{w}\|_M < \|\mathbf{b}\|_M \varepsilon^{-1}$, and thus $\left\|\varepsilon A_\varepsilon^{-1} M\right\|_M < 1$. Finally, $(\mathbb{I}_d + \mathrm{i}\varepsilon A_\varepsilon^{-1} M)^{-1}$ can be expanded as a Neumann series, which completes the proof.

Remark 2 *The mass matrix is symmetric and positive definite so it admits a square root $M^{1/2}$. For any $B \in \mathbb{C}^{N\times N}$, 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 in Theorem 1 converges also in the 2-norm.

Following [5], the expansion preconditioner of order $n \in \mathbb{N}^+$ is defined as a truncation of the Neumann series given in Theorem 1,

$$EX(n) = \left(\sum_{j=0}^{n-1} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M\right)^{j+1} \right) M^{-1} = \left(\sum_{j=0}^{n-1} (-\mathrm{i}\varepsilon)^j \left(A_\varepsilon^{-1} M\right)^{j} \right) A_\varepsilon^{-1}. \quad (5)$$

The preconditioned Helmholtz problem is thus given by

$$EX(n) A_0 \mathbf{w}_l = EX(n) \mathbf{b}_l. \quad (6)$$

From Elman's estimate (see e.g. Theorem 1.8 in [4]), the rate of convergence of GMRES used for solving any $A\mathbf{w} = \mathbf{b}$ can be estimated by an upper bound of $\|\mathbb{I}_d - A\|_2$. Denoting by $\mathbf{r}_m$ the GMRES residual and assuming that $\|\mathbb{I}_d - A\|_2 \leq \sigma < 1$, this reads

$$\frac{\|\mathbf{r}_m\|_2}{\|\mathbf{r}_0\|_2} \leq \left(\frac{2\sqrt{\sigma}}{(1+\sigma)^2} \right)^m. \quad (7)$$

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 bounds*

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

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

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

$$\mathbb{I}_d - EX(n)A_0 = (A_0^{-1} - EX(n))A_0 = \left(\sum_{j \geq n} (-\mathrm{i}\varepsilon)^j (A_\varepsilon^{-1} M)^j \right) A_\varepsilon^{-1} A_0. \quad (8)$$

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. Taking norms in (8) and summing the geometric series then gives

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

Remark 4 *The construction of the expansion preconditioner as well as Theorem 3 hold without any changes for high order finite element discretizations.*

The upper bound from Theorem 3 involves only $\varepsilon \mathscr{N}\left(A_\varepsilon^{-1} M\right)$. If this quantity Is bounded away from 1, the expansion preconditioner can reduce the number of GMRES iterations when using a large enough n, i.e. enough terms in the expansion.

3 Wavenumber-independent convergence of GMRES

As it was proved for the shifted Helmholtz preconditioner in [4], we now show that taking $\varepsilon \lesssim k$ is sufficient in the expansion preconditioner to ensure wavenumber-independant convergence of GMRES. We do this for two types of meshes: for $k^3h^2 \leq C_0$, for which one should have no pollution error according to [9], and the even higher resolution $h \sim k^{-2}$.

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

(A1) $\eta \sim k$ *and* $k^3h^2 \leq C_0$ *with* C_0 *small enough.*
(A2) $\eta \lesssim k$, $k \geq k_0$ *for a given* $k_0 > 0$ *and* $kh\sqrt{|k^2-\varepsilon|} \leq C_0$ *with* C_0 *small enough.*

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

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

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

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

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

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

$$\varepsilon\rho(A_\varepsilon^{-1}M) = \max_{\lambda_j \in \sigma(M^{-1}A_0)} \frac{\varepsilon}{|\lambda_j + \mathrm{i}\varepsilon|}. \tag{9}$$

Let $\mathbf{b} \in \mathbb{C}^N$ be fixed and $\mathbf{v}_h \in \mathbb{C}^N$ 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 (2) 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 (2) belongs to $H^2(\Omega)$. Since (A1) holds, one can apply [9, Corollary 4.4 p.12] to get

$$\|\nabla\varphi_h\|_{L^2(\Omega)} + k\|\varphi_h\|_{L^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)}. \tag{10}$$

Then (10) shows that

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

Using [4, Eq. (4.2) p. 24], we have $\|F_h\|_{\mathbb{C}^N \to \mathscr{V}_h} \sim h^{d/2}$, which gives

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

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

$$\left\|A_0^{-1}M\right\|_2 \lesssim \frac{1}{k}. \tag{11}$$

The upper bound (11) 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

$$\rho(A_\varepsilon^{-1}M) \leq \frac{C_2}{k}. \tag{12}$$

Assuming now that (A2) holds allows us to apply [4, 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, the authors proved in [4, Lemma 4.1 p. 598] that there exists a constant C_2 depending only on Ω such that

$$\left\|A_\varepsilon^{-1}M\right\|_2 \leq \frac{C_2}{k}. \tag{13}$$

Using now (12) and (13) together with the bound proved in Theorem 3 concludes the proof.

4 Numerical experiments

We discretize the Helmholtz equation on the unit square with classical Robin radiation boundary conditions, $\eta = k$, using P1 finite elements and the resolution $hk^{3/2} = 1$. In Table 1, we show the iteration numbers that GMRES preconditioned with the expansion preconditioner needed to reach a relative residual reduction of $1e-6$ for the right hand side $f = 1$. We see that for all expansion preconditioners, $n = 1,2,3$, iteration numbers are constant for the shift $\varepsilon = k$, and for larger n even a little better. For larger shifts, the expansion preconditioner with $n = 1$, which is identical to the shifted Helmholtz preconditioner, has growing iteration numbers, as shown in [4]. For the shift $\varepsilon = k^{3/2}$, the expansion preconditioners with $n = 2,3$ still seem to have constant iteration numbers, which is remarkable, and still increasing n lowers the iteration numbers a bit. When the shift is however $\varepsilon = k^2$, iteration numbers now also grow rapidly for the expansion preconditioner with $n = 2,3$, at

	$n=1$			$n=2$			$n=3$		
ε	k	$k^{3/2}$	k^2	k	$k^{3/2}$	k^2	k	$k^{3/2}$	k^2
$k=5$	5	6	8	5	6	9	5	7	9
$k=10$	6	8	13	5	8	12	5	7	13
$k=20$	6	11	24	5	8	21	4	7	20

Table 1 GMRES iteration numbers for Helmholtz with the expansion preconditioner.

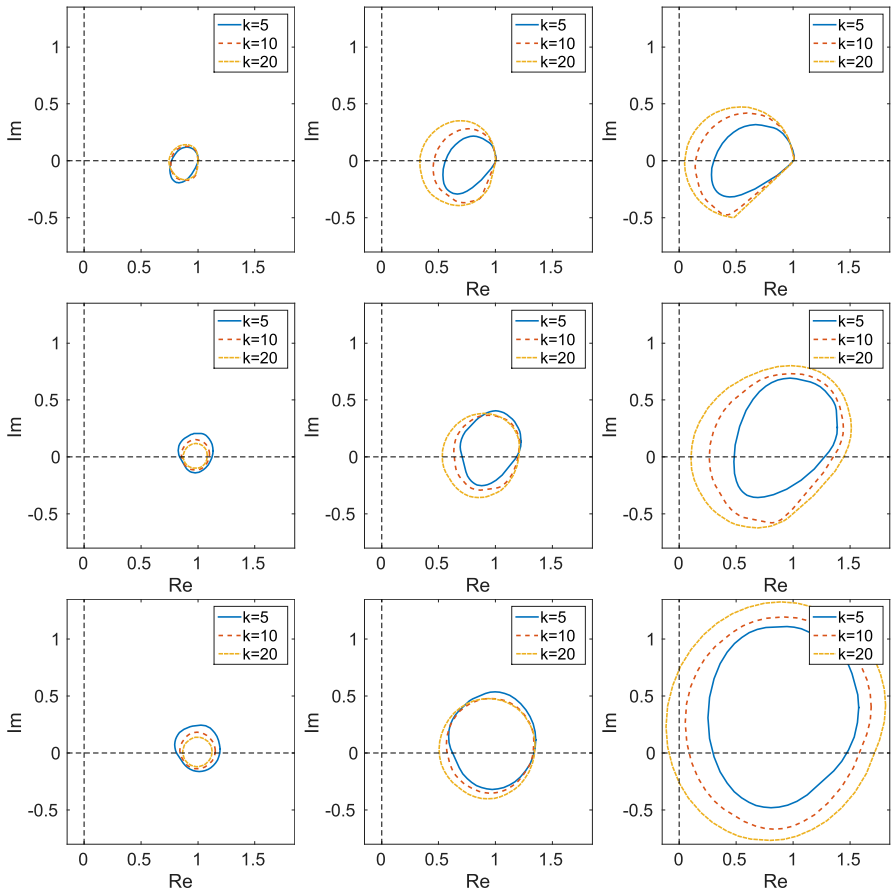

Fig. 1 Numerical range for shifts $\varepsilon = k, k^{3/2}, k^2$ (from left to right). Top row: shifted Helmholtz which is identical to the expansion preconditioner with $n = 1$. Middle row: expansion preconditioner $n = 2$. Bottom row: expansion preconditioner $n = 3$.

a linear rate in the wave number k, which was also observed in [4] for the shifted Helmholtz preconditioner with Robin boundary conditions[1]. The numerical ranges for the preconditioned operators of Table 1 are shown in Figure 1. We can see that the expansion preconditioner is robust for the shifts $\varepsilon = k$ and $\varepsilon = k^{3/2}$ in the range $k \in \{5, 10, 20\}$ tested, since the corresponding numerical range does not approach zero. It thus seems to deteriorate only for larger shifts than the shifted Helmholtz preconditioner, but is also substantially more costly, since one has to invert the shifted Helmholtz operator n times.

[1] Quadratic growth was even observed in the wave guide configuration.

5 Conclusions

We presented a convergence analysis of the expansion preconditioner for discretized Helmholtz problems. We showed that like for the shifted Helmholtz preconditioner, which coincides with the expansion preconditioner for $n = 1$, wave number independent convergence of GMRES can be guaranteed for shift $\varepsilon \lesssim k$. For larger shifts, we tested the expansion preconditioner numerically and found that in the range of wave numbers tested, the expansion preconditioner seems still to be robust for shifts $\varepsilon = k^{3/2}$, which is quite remarkable. Unfortunately for shifts of $O(k^2)$, which is required for the effective solution of the shifted problems by multigrid [2, 3], the expansion preconditioner is also not robust any more, like the shifted Helmholtz preconditioner. The effort for pushing this approach to larger shifts is thus still ongoing, see also the important recent work in domain decomposition [10] presented as a plenary lecture in the present conference [11].

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.
10. Graham, I.G., Spence, E.A. & Vainikko, E. (2015), Domain decomposition preconditioning for high-frequency Helmholtz problems using absorption, ArXiv e-prints.
11. Graham, I.G. (2017), Domain Decomposition for high frequency Helmholtz problems, this proceeding volume.