Analysis of Double Sweep Optimized Schwarz Methods: the Positive Definite Case

Martin J. Gander and Hui Zhang

1 Introduction

Over the last decade, substantial research efforts have gone into developing preconditioners for time harmonic wave propagation problems, like the Helmholtz and the time harmonic Maxwell's equations. Such equations are much harder to solve than diffusive problems like Laplace's equation, because of two main reasons: first, the pollution effect [1] requires much finer meshes than would be necessary just to resolve the signal computed, and second, classical iterative methods all exhibit severe convergence problems when trying to solve the very large discrete linear systems obtained [9]. These research efforts have led to innovative new preconditioners, like optimized Schwarz methods (OSM) [5, 11], Analytic Incomplete LU (AILU) [12], the sweeping preconditioner [7, 8], the source transfer domain decomposition [3, 4], the method based on single layer potentials [14], and the method of polarized traces [15], for a more complete treatment, see [13] and references therein. In [13], it was shown that all these methods can be written as alternating optimized Schwarz methods called Double Sweep Optimized Schwarz Methods (DOSMs). We study here analytically the contraction properties of DOSMs for the model problem

$$\begin{aligned} &\eta u - u_{xx} - u_{yy} = f \quad \text{in } \Omega := (-\tfrac{L}{2}, 1+\tfrac{L}{2}) \times (0,\pi), \\ &u = 0 \quad \text{at } y \in \{0,\pi\}, \quad \mathcal{B}_1^l u = 0 \quad \text{at } x = -\tfrac{L}{2}, \quad \mathcal{B}_N^r u = 0 \quad \text{at } x = 1+\tfrac{L}{2}, \end{aligned} \tag{1}$$

Martin J. Gander
University of Geneva, Section of Mathematics, Rue du Lievre 2-4, CP 64, 1211 Geneva 4, e-mail: martin.gander@unige.ch

Hui Zhang (corresponding author)
Xi'an Jiaotong-Liverpool University, Department of Mathematical Sciences & Laboratory for Intelligent Computing and Financial Technology (KSF-P-02), Suzhou 215123, China; Zhejiang Ocean University, Key Laboratory of Oceanographic Big Data Mining & Application of Zhejiang Province, Zhoushan 316022, China, e-mail: mike.hui.zhang@hotmail.com

where $L \geq 0$ is a parameter which will be related to the overlap, $\mathcal{B}_1^l$ and $\mathcal{B}_N^r$ are linear trace operators, and $N \geq 2$ is an integer number to be defined below. While in our derivations in Section 2 and 3 we consider $\eta \in \mathbb{C}$, and we thus also include the Helmholtz case of interest, we will focus for our results then in Section 4 on the positive definite case $\eta > 0$.

2 Iteration Matrix of DOSM

Using a Fourier sine expansion of u solution to (1) in the y direction with Fourier parameter k, we obtain for the Fourier coefficients [1] for $k = 1, 2, \ldots$ the problem

$$\begin{aligned} &(k^2+\eta)u - u_{xx} = f \quad \text{in } (-\tfrac{L}{2}, 1+\tfrac{L}{2}),\ k \in \mathbb{N}^+, \\ &\mathcal{B}_1^l u = 0 \quad \text{at } x = -\tfrac{L}{2}, \quad \mathcal{B}_N^r u = 0 \quad \text{at } x = 1+\tfrac{L}{2}. \end{aligned} \tag{2}$$

We decompose the domain $(-\frac{L}{2}, 1+\frac{L}{2})$ into N overlapping subdomains of equal width $H + L := \frac{1}{N} + L$, denoted by $\Omega_j := ((j-1)H - \frac{L}{2}, jH + \frac{L}{2})$, and we denote the restricted solution by $u_j := u|_{\Omega_j}$, $j = 1, .., N$. In DOSM, (2) is reformulated as transmission problems on the Ω_j for $j = 1, .., N$,

$$\begin{aligned} &(k^2+\eta)u_j - (u_j)_{xx} = f \quad \text{in } \Omega_j, \\ &\mathcal{B}_j^l(u_j - u_{j-1}) = 0 \text{ at } x = (j-1)H - \tfrac{L}{2}, \mathcal{B}_j^r(u_j - u_{j+1}) = 0 \text{ at } x = jH + \tfrac{L}{2}, \end{aligned} \tag{3}$$

where $\mathcal{B}_j^l$ and $\mathcal{B}_j^r$ are linear trace operators and u_0, u_{N+1} are identically zero. Let $g_j^l := \mathcal{B}_j^l u_j$ at $x = x_j^l := (j-1)H - \frac{L}{2}$ and $g_j^r := \mathcal{B}_j^r u_j$ at $x = x_j^r := jH + \frac{L}{2}$. To rewrite (3) in terms of the interface data $[g_2^l; ..; g_N^l; g_1^r; ..; g_{N-1}^r]$, we define the trace-to-trace operators (see also Figure 1)

$$\begin{aligned} a_j &: \left(\ell_j \text{ at } x = x_j^l\right) \to \left(\mathcal{B}_{j+1}^l v_j \text{ at } x = x_{j+1}^l\right) \text{ with } v_j \text{ solving} \\ &(k^2+\eta)v_j - (v_j)_{xx} = 0 \quad \text{in } \Omega_j, \quad \mathcal{B}_j^l v_j = \ell_j \text{ at } x = x_j^l, \quad \mathcal{B}_j^r v_j = 0 \text{ at } x = x_j^r, \\ b_j &: \left(\gamma_j \text{ at } x = x_j^r\right) \to \left(\mathcal{B}_{j+1}^l v_j \text{ at } x = x_{j+1}^l\right) \text{ with } v_j \text{ solving} \\ &(k^2+\eta)v_j - (v_j)_{xx} = 0 \quad \text{in } \Omega_j, \quad \mathcal{B}_j^l v_j = 0 \text{ at } x = x_j^l, \quad \mathcal{B}_j^r v_j = \gamma_j \text{ at } x = x_j^r, \\ c_j &: \left(\gamma_j \text{ at } x = x_j^r\right) \to \left(\mathcal{B}_{j-1}^r v_j \text{ at } x = x_{j-1}^r\right) \text{ with } v_j \text{ solving} \\ &(k^2+\eta)v_j - (v_j)_{xx} = 0 \quad \text{in } \Omega_j, \quad \mathcal{B}_j^l v_j = 0 \text{ at } x = x_j^l, \quad \mathcal{B}_j^r v_j = \gamma_j \text{ at } x = x_j^r, \\ d_j &: \left(\ell_j \text{ at } x = x_j^l\right) \to \left(\mathcal{B}_{j-1}^r v_j \text{ at } x = x_{j-1}^r\right) \text{ with } v_j \text{ solving} \\ &(k^2+\eta)v_j - (v_j)_{xx} = 0 \quad \text{in } \Omega_j, \quad \mathcal{B}_j^l v_j = \ell_j \text{ at } x = x_j^l, \quad \mathcal{B}_j^r v_j = 0 \text{ at } x = x_j^r. \end{aligned}$$

[1] We still denote the Fourier transformed quantities for simplicity by the same symbols u, $\mathcal{B}_1^l$ and $\mathcal{B}_N^r$ to avoid a more complicated notation.

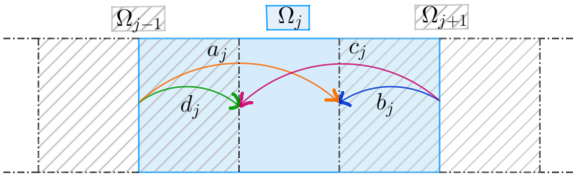

Fig. 1 Illustration of the interface-to-interface operators.

In the Fourier basis, the operators a_j, b_j, c_j, d_j reduce to scalars which we compute now explicitly. To simplify the notation, let $s := \sqrt{k^2+\eta}$ with $\operatorname{Re} s > 0$ if $k^2+\eta$ is not exactly on the negative real axis, $s := c_{\text{sgn}}\mathrm{i}\sqrt{-k^2-\eta}$ if $k^2+\eta < 0$ with $c_{\text{sgn}} \in \{1,-1\}$ a conventional sign-value from the time-dependence $\mathrm{e}^{c_{\text{sgn}}\mathrm{i}\sqrt{-\eta}t}$. For typical OSM transmission conditions (not just DOSM) of the form $\mathcal{B}_j^l = -q_j^l\partial_x + p_j^l$ and $\mathcal{B}_j^r = q_j^r\partial_x + p_j^r$, we define

$$R_j^l := \frac{p_j^l - q_j^l s}{p_j^l + q_j^l s}\mathrm{e}^{-Ls}, \qquad R_j^{lr} := \frac{p_{j-1}^r - q_{j-1}^r s}{p_j^l + q_j^l s}\mathrm{e}^{-Ls}, \qquad Q_j^{lr} := \frac{p_{j-1}^r + q_{j-1}^r s}{p_j^l + q_j^l s},$$
$$R_j^r := \frac{p_j^r - q_j^r s}{p_j^r + q_j^r s}\mathrm{e}^{-Ls}, \qquad R_j^{rl} := \frac{p_{j+1}^l - q_{j+1}^l s}{p_j^r + q_j^r s}\mathrm{e}^{-Ls}, \qquad Q_j^{rl} := \frac{p_{j+1}^l + q_{j+1}^l s}{p_j^r + q_j^r s},$$
$$R_j^{ll} := \frac{p_{j+1}^l - q_{j+1}^l s}{p_j^l + q_j^l s}\mathrm{e}^{-Ls}, \qquad R_j^{rr} := \frac{p_{j-1}^r - q_{j-1}^r s}{p_j^r + q_j^r s}\mathrm{e}^{-Ls}, \qquad Q_j^{ll(rr)} := \frac{p_{j\pm1}^{l(r)} + q_{j\pm1}^{l(r)} s}{p_j^{l(r)} + q_j^{l(r)} s},$$

and we have for $L \geq 0$

$$a_j = \frac{(Q_j^{ll} - R_j^r R_j^{ll})\mathrm{e}^{-Hs}}{1 - R_j^l R_j^r \mathrm{e}^{-2Hs}}, \qquad b_j = \frac{R_j^{rl} - R_j^l Q_j^{rl}\mathrm{e}^{-2Hs}}{1 - R_j^l R_j^r \mathrm{e}^{-2Hs}},$$
$$c_j = \frac{(Q_j^{rr} - R_j^l R_j^{rr})\mathrm{e}^{-Hs}}{1 - R_j^r R_j^l \mathrm{e}^{-2Hs}}, \qquad d_j = \frac{R_j^{lr} - R_j^r Q_j^{lr}\mathrm{e}^{-2Hs}}{1 - R_j^r R_j^l \mathrm{e}^{-2Hs}}.$$

When $\mathcal{B}_j^l = \mathcal{B}_j^r = 1$, i.e. the classical alternating Schwarz case, we have

$$a_j = c_j = \frac{(1-\mathrm{e}^{-2Ls})\mathrm{e}^{-Hs}}{1-\mathrm{e}^{-2Hs-2Ls}}, \qquad b_j = d_j = \frac{(1-\mathrm{e}^{-2Hs})\mathrm{e}^{-Ls}}{1-\mathrm{e}^{-2Hs-2Ls}}.$$

Using the operators a_j, b_j, c_j and d_j, we can rewrite (3) as the linear system

$$\begin{pmatrix} I-A & -B \\ -D & I-C \end{pmatrix}\begin{pmatrix} g^l \\ g^r \end{pmatrix} = \begin{pmatrix} \tau^l \\ \tau^r \end{pmatrix}. \tag{4}$$

where A (C) has all its non-zero entries on the subdiagonal (superdiagonal) as $(a_2,\ldots,a_{N-1})$ $((c_2,\ldots,c_{N-1}))$, $B = \operatorname{diag}(b_1,\ldots,b_{N-1})$, $D = \operatorname{diag}(d_2,\ldots,d_N)$,

and $\tau_j^l := \mathcal{B}_j^l v_{j-1}$, $\tau_j^r := \mathcal{B}_j^r v_{j+1}$ with v_j satisfying

$$\begin{aligned}&(k^2+\eta)v_j - (v_j)_{xx} = f \quad \text{in } \Omega_j,\\ &\mathcal{B}_j^l v_j = 0 \quad \text{at } x = x_j^l, \quad \mathcal{B}_j^r v_j = 0 \quad \text{at } x = x_j^r.\end{aligned}$$

The DOSM, which comprises a class of the many recently invented preconditioners for time harmonic wave propagation [13], amounts to a block Gauss-Seidel iteration for (4): given an initial guess $g^{r,0}$ of g^r, we compute for iteration index $m = 0, 1, \ldots$

$$g^{r,m+1} := (I-C)^{-1}\left[\tau^r + D(I-A)^{-1}(\tau^l + Bg^{r,m})\right]. \tag{5}$$

We denote by $\epsilon^{r,m} := g^{r,m} - g^r$ the error, which then by (5) satisfies a recurrence relation with iteration matrix T,

$$\epsilon^{r,m+1} = T\epsilon^{r,m} := (I-C)^{-1}D(I-A)^{-1}B\epsilon^{r,m}. \tag{6}$$

3 Eigenvalues of the iteration matrix T

To understand the convergence properties of these methods, we need to study the spectral radius of T. We first compute the inverse of T if B and D are invertible. For simplicity, we assume from now on that $\mathcal{B}_j^l = \mathcal{B}^l$, $j = 2, .., N$ are the same and $\mathcal{B}_j^r = \mathcal{B}^r$, $j = 1, .., N-1$ are the same. Therefore, b_j, d_j, $j = 2, .., N$ have the same value and we denote them by b and d. In addition $a_j = c_j$, $j = 2, .., N-1$ which also have the same value denoted by a. We thus obtain

$$\begin{aligned}T^{-1} &= B^{-1}(I-A)D^{-1}(I-C)\\ &= b^{-1}d^{-1}\begin{pmatrix} b_1^{-1}b & -b_1^{-1}ba & & & \\ -a & a^2+1 & -a & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & a^2+1 & -a \\ & & & -a & a^2 + d_N^{-1}d \end{pmatrix} =: b^{-1}d^{-1}\tilde{T}.\end{aligned} \tag{7}$$

Let $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the largest and smallest eigenvalue in modulus, and $\rho := \lambda_{\max}(T)$. From (7), we have $\rho = bd\lambda_{\min}^{-1}(\tilde{T})$. Let λ be an eigenvalue of $\tilde{T}$ and $\mathbf{v} = (v_j)_{j=1}^{N-1} \in \mathbb{C}^{N-1}$ the associated eigenvector. It follows that

$$-av_{j-1} + (a^2+1-\lambda)v_j - av_{j+1} = 0, \tag{8a}$$

$$(b_1^{-1}b - \lambda)v_1 - b_1^{-1}bav_2 = 0, \tag{8b}$$

$$-av_{N-2} + \left(a^2 + d_N^{-1}d - \lambda\right)v_{N-1} = 0. \tag{8c}$$

Note that $v_j = \xi_1\mu^j + \xi_2\mu^{-j}$, $j \in \mathbb{Z}$ is the general solution of (8a) if $\mu \neq \pm 1$ satisfies

$$-a + (a^2 + 1 - \lambda)\mu - a\mu^2 = 0, \text{ or } \lambda = 1 + a^2 - a(\mu + \mu^{-1}). \tag{9}$$

Subtracting (8a) at $j = 1$ and $j = N - 1$ from (8b) and (8c) gives the equivalent boundary conditions

$$\begin{aligned} -av_0 + (a^2 + 1 - b_1^{-1}b)v_1 - a(1 - b_1^{-1}b)v_2 &= 0, \\ (1 - d_N^{-1}d)v_{N-1} - av_N &= 0. \end{aligned}$$

Further substituting $v_j = \xi_1\mu^j + \xi_2\mu^{-j}$ into the above equations leads to

$$\begin{aligned} \left[-a + (a^2 + 1 - b_1^{-1}b)\mu - a(1 - b_1^{-1}b)\mu^2\right]\xi_1 + \\ \left[-a + (a^2 + 1 - b_1^{-1}b)\mu^{-1} - a(1 - b_1^{-1}b)\mu^{-2}\right]\xi_2 = 0, \\ \left[(1 - d_N^{-1}d) - a\mu\right]\mu^{N-1}\xi_1 + \left[(1 - d_N^{-1}d) - a\mu^{-1}\right]\mu^{1-N}\xi_2 = 0. \end{aligned}$$

Since $\mathbf{v} \neq 0$, the determinant of the above linear system for $[\xi_1; \xi_2]$ must vanish, i.e.

$$\begin{aligned} &\left[-a + (a^2 + 1 - b_1^{-1}b)\mu - a(1 - b_1^{-1}b)\mu^2\right]\left[(1 - d_N^{-1}d) - a\mu^{-1}\right]\mu^{1-N} \\ &= \left[-a + (a^2 + 1 - b_1^{-1}b)\mu^{-1} - a(1 - b_1^{-1}b)\mu^{-2}\right]\left[(1 - d_N^{-1}d) - a\mu\right]\mu^{N-1}. \end{aligned} \tag{10}$$

Assume $a \neq 0$ and let $\beta_1 := a^{-1}(1 - b_1^{-1}b)$, $\beta_N := a^{-1}(1 - d_N^{-1}d)$. We can rewrite (10) as

$$\mu^{2N} = \frac{(1 - a\mu)(1 - \beta_1\mu)(1 - \beta_N\mu)}{(1 - a\mu^{-1})(1 - \beta_1\mu^{-1})(1 - \beta_N\mu^{-1})}. \tag{11}$$

In the special case when $\mathcal{B}_1^l = \mathcal{B}^l$ and $\mathcal{B}_N^r = \mathcal{B}^r$, we have $\beta_1 = \beta_N = 0$ so that

$$\mu^{2N} = \frac{1 - a\mu}{1 - a\mu^{-1}}. \tag{12}$$

Remark 1 The value $\lambda = (1 \mp a)^2$ corresponding to $\mu = \pm 1$ in (9) is an eigenvalue of T^{-1} if and only if $v_j = (\xi_1 + \xi_2 j)(\pm 1)^j$ is a non-zero solution of (8) or equivalently

$$[(1\mp a)(\pm 1\mp b_1^{-1}b - a)][(1 - d_N^{-1}d)(N-1)\mp aN] = [\pm a^2 + (\pm 1 - 2a)(1 - b_1^{-1}b)][1 - d_N^{-1}d \mp a].$$

In the special case of $\mathcal{B}_1^l = \mathcal{B}^l$ and $\mathcal{B}_N^r = \mathcal{B}^r$, the above condition becomes $\pm a^2 N(1 \mp a) = -a^3$, that is, $a = 0$ or $\pm\frac{N}{N-1}$.

4 Roots of the Polynomial Equation for μ

We first observe the following facts: $\mu = \pm 1$ are two roots of (11), and the other roots appear in pairs as μ, μ^{-1}. Our goal in this section is to locate all the roots in the complex plane. We assume from now on that $\eta \geq 0$ and thus $a, \beta_1, \beta_N \in \mathbb{R}$. Hence complex roots of (11) appear in conjugate pairs. We begin with the simplest

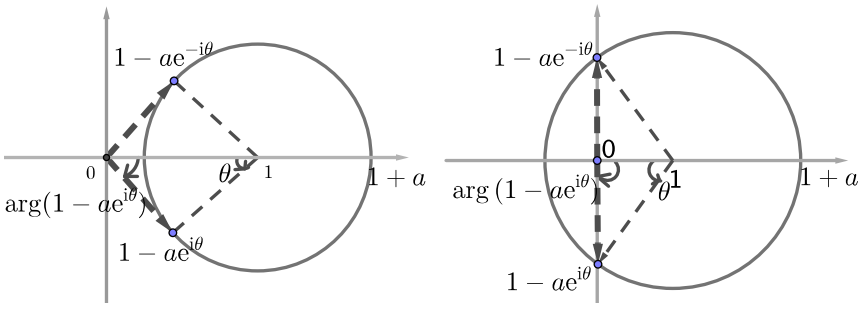

Fig. 2 Image of $1-ae^{i\theta}$. Left: $0<a<1$. Right: $a>1$.

case (12). We assume the argument $\arg z$ of a complex number z to take values in $(-\pi,\pi]$.

Lemma 1 *If $a\in[-1,1]\backslash\{0\}$, then the roots of* (12) *are* ±1 *and* $(\operatorname{sign} a)\,e^{\pm i\,\theta_j}$ *for some* $\theta_j\in\left[(j-\frac{1}{2})\pi/N, j\pi/N\right)$, $j=1,..,N-1$.

Proof Since (12) is invariant under the transform $a\to -a, \mu\to-\mu$, we can assume $a>0$. Substituting the ansatz $\mu=e^{i\theta}$, $\theta\in(-\pi,\pi]$ into (12), we obtain for θ the equation

$$w(\theta):=e^{i2N\theta}=\frac{1-ae^{i\theta}}{1-ae^{-i\theta}}=:z(\theta). \tag{13}$$

Since (13) is invariant under the transform $\theta\to-\theta$, we need only to show that $w(\theta)=z(\theta)$ has $N-1$ roots for $\theta\in(0,\pi)$. On the one hand, we note that $1-ae^{i\theta}$ turns around 1 with radius a, see Figure 2. It follows that $z(\theta)$ moves on the unit circle: first from $z(0)=1$ clockwise to the extremal point $z(\arccos a)$ with $\arg z=-2\arcsin a$, and then back counter-clockwise to $z(\pi)=1$. On the other hand, $w(\theta)$ starts from $w(0)=1$ and turns counter-clockwise along the unit circle N times. Hence, in each lower semi-cycle $\theta\in\left[(j-\frac{1}{2})\pi/N, j\pi/N\right)$, $j=1,..,N-1$ there must exist a value of θ such that $w(\theta)=z(\theta)$. □

5 Numerical Study of the Convergence Factor

As before, we focus on the regime $k^2+\eta>0$, and therefore $s=\sqrt{k^2+\eta}$ varies in $[s_{\min},s_{\max}]$. Typically, $s_{\max}$ is linked to N, for example, if H is proportional to the mesh size and a second-order discretization is used, we have $s_{\max}=O(N)$. On the other hand, $s_{\min}$ is in this case a constant, which for our sine expansion has the value $s_{\min}=\sqrt{1+\eta}$ stemming from the lowest Fourier mode $k=1$.

In the special case of the classical alternating Schwarz methods, $\mathcal{B}_j^l=\mathcal{B}_j^r=1$, we have $a\in(0,1)$, $b\in(0,1)$. By (9) and Lemma 1, we get $\lambda_{\min}=1+a^2-2a\cos\theta_1>0$ for some $\theta_1\in\left[\frac{\pi}{2N},\frac{\pi}{N}\right)$. Therefore, the convergence factor $\rho=b^2\lambda_{\min}^{-1}$ becomes

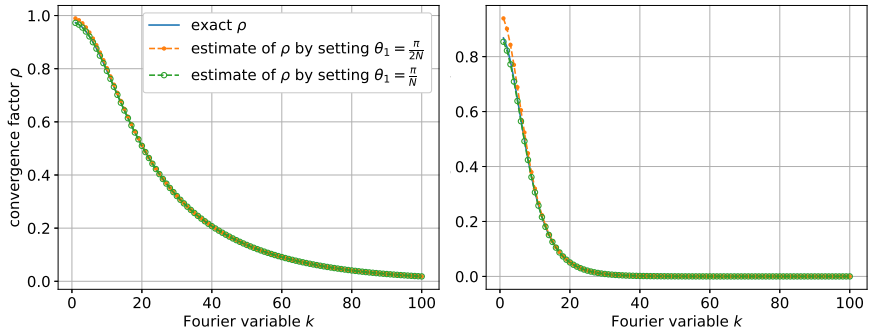

Fig. 3 Convergence factor of alternating Schwarz with $N = 10$, $L = \frac{1}{5N}$ (left), $\frac{4}{5N}$ (right) and $s \in [\sqrt{2}, \sqrt{10^4+1}]$.

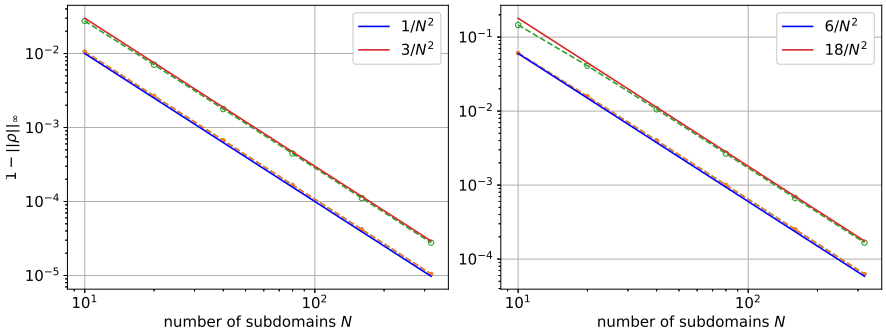

Fig. 4 Scaling of alternating Schwarz with $L = \frac{1}{5N}$ (left), $\frac{4}{5N}$ (right) and $s \in [\sqrt{2}, \sqrt{100N^2+1}]$. The dashed lines correspond to the upper and lower bounds of $1 - \|\rho\|_\infty$.

$$0 < \rho = \frac{e^{-2Ls}(1-e^{-2Hs})^2}{r_1^2 + r_2^2 - 2r_1 r_2 \cos\theta_1}, \quad r_1 := 1 - e^{-2Hs-2Ls}, \quad r_2 := (1-e^{-2Ls})e^{-Hs}.$$

Substituting θ_1 with its lower (upper) bound into the above expression yields an upper (lower) bound of ρ. In Figure 3, we compare these bounds with the exact value of ρ computed numerically. We see that the bounds are quite sharp. Then using these bounds we get the scaling of $\|\rho\|_\infty := \max_s |\rho(s)|$ with the number of subdomains N, and the convergence deteriorates, see Figure 4.

For optimized Schwarz, since $\eta \geq 0$, it is natural to use positive p_j^l, p_j^r [10]. In the special case of $p_j^l = p_j^r = p(k^2) > 0$, we find that $R \in (-1,1), a \in (0,1), b = d \in (-1,1)$. Again, by (9) and Lemma 1, we have $\lambda_{\min} = 1 + a^2 - 2a\cos\theta_1 > 0$ for some $\theta_1 \in \left[\frac{\pi}{2N}, \frac{\pi}{N}\right)$. Therefore, the convergence factor $\rho = b^2\lambda_{\min}^{-1}$ becomes

$$0 < \rho = \frac{R^2(1-e^{-2Hs})^2}{r_1^2 + r_2^2 - 2r_1 r_2 \cos\theta_1}, \quad r_1 := 1 - R^2 e^{-2Hs}, \quad r_2 := (1-R^2)e^{-Hs}.$$

We first take $p > 0$ a constant. In Figure 5, we show how good the bounds of ρ are,

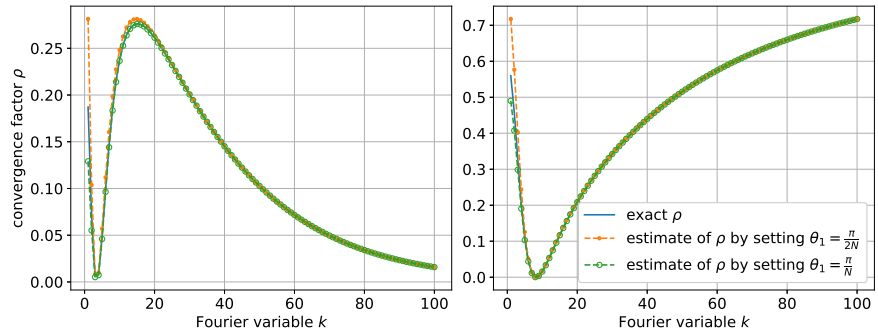

Fig. 5 Convergence factor of DOSM with p a constant obtained by numerically minimizing the upper bound of $\|\rho\|_\infty$, $N = 10$, $L = \frac{1}{5N}$ (left), $L = 0$ (right) and $s \in [\sqrt{2}, \sqrt{10^4+1}]$.

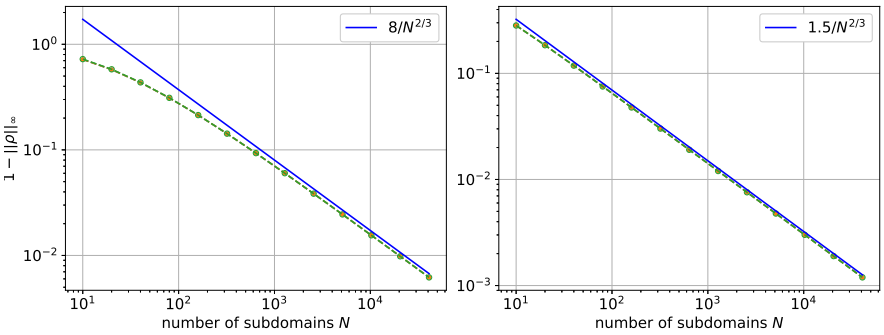

Fig. 6 Scaling of DOSM with p a constant obtained by numerically minimizing the upper bound of $\|\rho\|_\infty$, $L = \frac{1}{5N}$ (left), $L = 0$ (right) and $s \in [\sqrt{2}, \sqrt{100N^2+1}]$. The dashed lines correspond to the upper and lower bounds of $1 - \|\rho\|_\infty$ which are too close to be distinguished.

and also the results of minimizing the upper bound of $\|\rho\|_\infty$. Given p optimized in this way (dependent on N), we find that $\|\rho\|_\infty \approx 1 - O(N^{-2/3})$, both with minimal overlap and without overlap; see Figure 6. Next, we take $p = \tilde{p}_0 + \tilde{p}_2 k^2$ corresponding to the second-order boundary condition $\mathcal{B}_{l,r} = \mp\tilde{p}_0\partial_x - \tilde{p}_2\partial_{yy}$. We show the upper and lower bounds of ρ in Figure 7. Using numerically optimized parameters $\tilde{p}_0$ and $\tilde{p}_2$ (dependent on N), we find that $\|\rho\|_\infty \approx 1 - O(N^{-1/3})$ with minimal overlap and $\|\rho\|_\infty \approx 1 - O(N^{-2/5})$ without overlap; see Figure 8.

We can also choose $p(k^2)$ to be a more accurate approximation of $s = \sqrt{k^2+\eta}$ to obtain an even smaller reflection coefficient $R = \exp(-Ls)(p-s)/(p+s)$. In the recently invented methods [13], Perfectly Matched Layers (PMLs; see [2, 6]) are most commonly used. Starting from a boundary $x = x_0$, a PML $[x_0, x_0 + D]$ is added outside a domain, and a new variable

$$\tilde{x} := \begin{cases} x + \int_0^{x-x_0} \sigma(|t|)\,\mathrm{d}t, & x \in [x_0, x_0 + D], \\ x, & \text{inside the domain}, \end{cases}$$

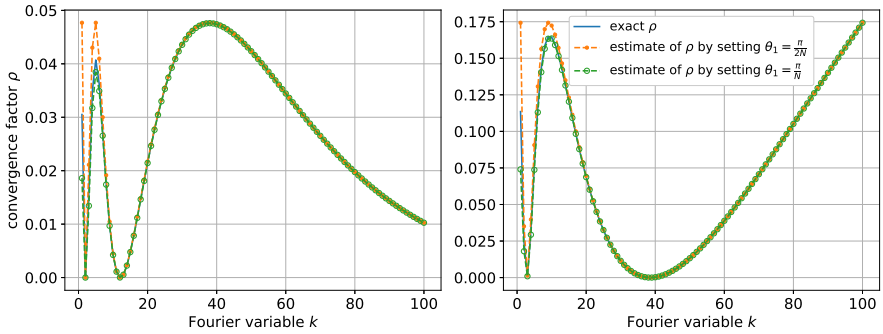

Fig. 7 Convergence factor of DOSM with $p = \tilde{p}_0 + \tilde{p}_2 k^2$ obtained by numerically minimizing the upper bound of $\|\rho\|_\infty$, $N = 10$, $L = \frac{1}{5N}$ (left), $L = 0$ (right) and $s \in [\sqrt{2}, \sqrt{10^4 + 1}]$.

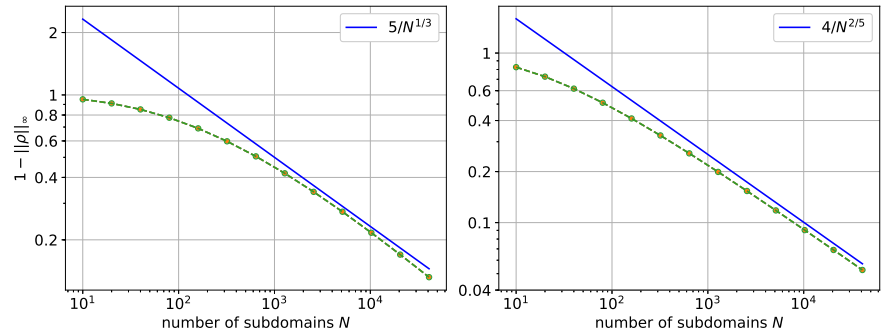

Fig. 8 Scaling of DOSM with $p = \tilde{p}_0 + \tilde{p}_2 k^2$ obtained by numerically minimizing the upper bound of $\|\rho\|_\infty$, $L = \frac{1}{5N}$ (left), $L = 0$ (right) and $s \in [\sqrt{2}, \sqrt{100N^2 + 1}]$. The dashed lines correspond to the upper and lower bounds of $1 - \|\rho\|_\infty$ which are too close to be distinguished.

is used for the model on the augmented domain $(k^2 + \eta)u - u_{\tilde{x}\tilde{x}} = \tilde{f}$, where $\tilde{f}$ is the zero extension of f and a homogeneous Dirichlet condition is put on the augmented boundary $x = x_0 + D$. This model amounts to imposing on $x = x_0$ the boundary condition $\mathrm{sign}(D)\partial_x u + \mathrm{DtN}_D u = 0$, where DtN_D is the Dirichlet-to-Neumann operator defined by

$$\begin{aligned} \mathrm{DtN}_D : (\gamma \text{ at } x = x_0) &\to (-\mathrm{sign}(D)\partial_x v \text{ at } x = x_0) \text{ with } v \text{ solving} \\ (k^2 + \eta)v - v_{\tilde{x}\tilde{x}} &= 0 \quad \text{for } x \in [x_0, x_0 + D], \\ v = 0 \quad \text{at } x = x_0 + D, &\quad v = \gamma \quad \text{at } x = x_0. \end{aligned}$$

In our case, DtN_D reduces to a scalar. Note that $\tilde{x}(x = x_0) = x_0$, $\tilde{x}(x = x_0 + D) = x_0 + D + \int_0^D \sigma(|t|)\,\mathrm{d}t =: x_0 + D + \bar{\sigma}$. From the above definition, we have

$$v = \xi_1 \mathrm{e}^{-s(\tilde{x} - x_0)} + \xi_2 \mathrm{e}^{s(\tilde{x} - x_0)}, \quad v(\tilde{x} = x_0 + D + \bar{\sigma}) = 0, \quad v(\tilde{x} = x_0) = \gamma.$$

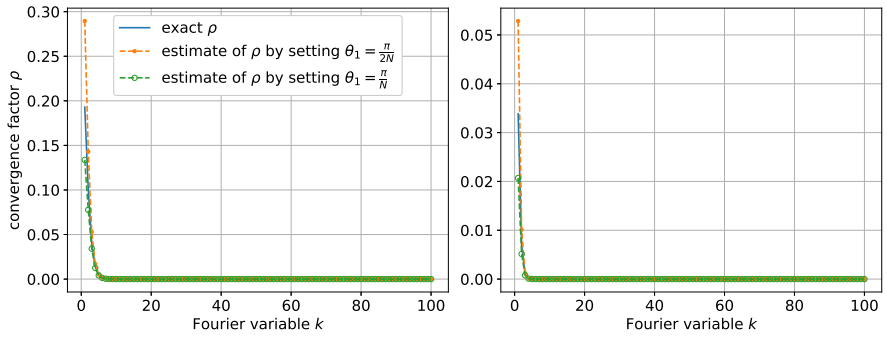

Fig. 9 Convergence factor of DOSM with $p = \mathrm{DtN}_D$ from PMLs, $N = 10$, $L = 0$, $\bar{\sigma} = 5D$, $D = 0.05$ (left), 0.1 (right) and $s \in [\sqrt{2}, \sqrt{10^4 + 1}]$.

Hence, we obtain in our case

$$\mathrm{DtN}_D = s \cdot \frac{1 + \mathrm{e}^{-2(D+\bar{\sigma})s}}{1 - \mathrm{e}^{-2(D+\bar{\sigma})s}}.$$

Typically, one chooses $\bar{\sigma}$ linearly dependent on D. Using $p = \mathrm{DtN}_D$, we show in Figure 9 how good our upper and lower bounds of ρ are. It is impressive that doubling D decreases $\|\rho\|_\infty$ by a factor of about six. Then, for D proportional to the subdomain size $H = 1/N$, we look at their scaling with N in Figure 10. We see that

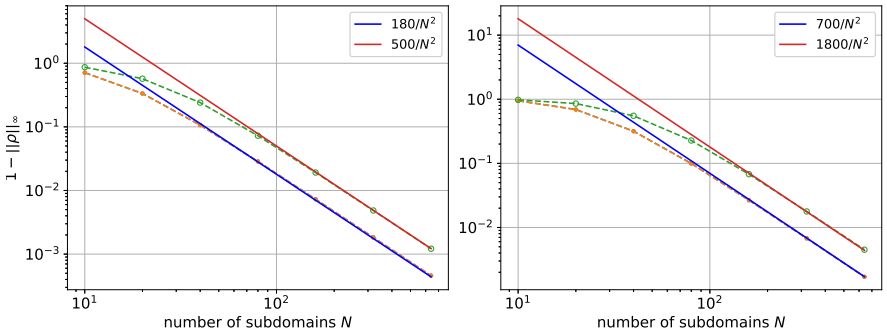

Fig. 10 Scaling of DOSM with $p = \mathrm{DtN}_D$ from PMLs, $L = 0$, $\bar{\sigma} = 5D$, $D = \frac{1}{2N}$ (left), $D = \frac{1}{N}$ (right) and $s \in [\sqrt{2}, \sqrt{100N^2 + 1}]$. The dashed lines correspond to the upper and lower bounds of $1 - \|\rho\|_\infty$.

the improvement by doubling D is only on the constant factor, and the deterioration $\|\rho\|_\infty \approx 1 - O(N^{-2})$ is the same as for the alternating Schwarz method. Hence, to have convergence independent of N, we must let the relative PML width D/H grow with N. To see how big a D is necessary, we test a range of D in Figure 11, where we can read for which size D and which N the bounds of $\|\rho\|_\infty$ equal to 0.2. We then plot

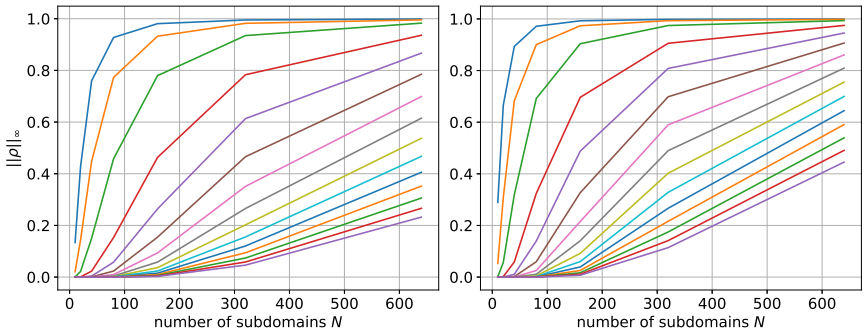

Fig. 11 Scaling of DOSM with different sizes D of PMLs, $L = 0$, $\bar{\sigma} = 5D$ and $s \in [\sqrt{2}, \sqrt{100N^2+1}]$. Left: lower bound. Right: upper bound. From top to bottom: $D/H = 0.5, 1, 2, 4, 6, .., 26$ where $H = 1/N$ is the subdomain width.

these pairs in Figure 12, which indicates that a constant PML size D, independent of

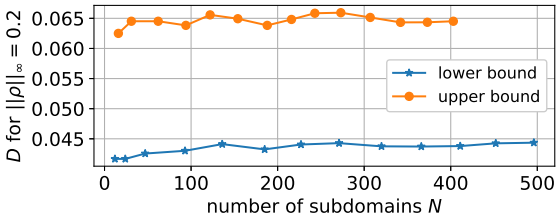

Fig. 12 Necessary PML size D to let DOSM converge independently of N, when $L = 0$, $\bar{\sigma} = 5D$ and $s \in [\sqrt{2}, \sqrt{100N^2+1}]$.

the number of subdomains N, is necessary and sufficient. The sufficiency is further shown in Figure 13. Note that in our setting a fixed physical PML size independent of the number of subdomains N means a linear growth of mesh points in the PMLs, not a logarithmic one.

References

1. Babuska, I.M., Sauter, S.A.: Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers? SIAM Journal on numerical analysis **34**(6), 2392–2423 (1997)
2. Berenger, J.P.: A perfectly matched layer for the absorption of electromagnetic waves. J. Comput. Phys. **114**, 185–200 (1994)
3. Chen, Z., Xiang, X.: A source transfer domain decomposition method for Helmholtz equations in unbounded domain. SIAM J. Numer. Anal. **51**, 2331–2356 (2013)
4. Chen, Z., Xiang, X.: A source transfer domain decomposition method for Helmholtz equations in unbounded domain Part II: Extensions. Numer. Math. Theor. Meth. Appl. **6**, 538–555 (2013)

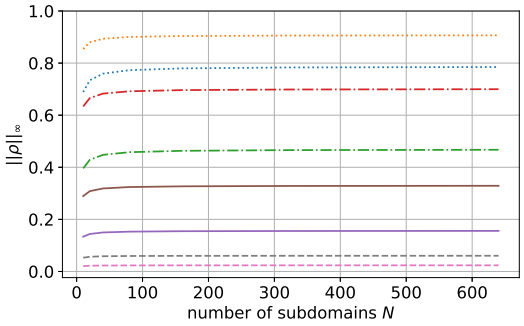

Fig. 13 Scaling of DOSM using fixed sizes of PMLs, $L = 0$, $\bar{\sigma} = 5D$ and $s \in [\sqrt{2}, \sqrt{100N^2 + 1}]$. From top to bottom each pair of lines correspond to the lower and upper bounds of $\|\rho\|_\infty$ for $D = 0.0125, 0.025, 0.05, 0.1$.

5. Chevalier, P., Nataf, F.: Symmetrized method with optimized second-order conditions for the Helmholtz equation. In: J. Mandel, C. Farhat, X.C. Cai (eds.) Domain Decomposition Methods 10, pp. 400–407. AMS (1998)
6. Chew, W.C., Jin, J.M., Michielssen, E.: Complex coordinate stretching as a generalized absorbing boundary condition. Microw. Opt. Technol. Lett. **15**, 363–369 (1997)
7. Engquist, B., Ying, L.: Sweeping preconditioner for the Helmholtz equation: Hierarchical matrix representation. Comm. Pure Appl. Math. **LXIV**, 0697–0735 (2011)
8. Engquist, B., Ying, L.: Sweeping preconditioner for the Helmholtz equation: Moving perfectly matched layers. Multiscale Model. Sim. **9**, 686–710 (2011)
9. Ernst, O., Gander, M.J.: Why it is difficult to solve Helmholtz problems with classical iterative methods. In: I. Graham, T. Hou, O. Lakkis, R. Scheichl (eds.) Numerical Analysis of Multiscale Problems, pp. 325–363. Springer-Verlag, Berlin (2012)
10. Gander, M.J.: Optimized Schwarz methods. SIAM Journal on Numerical Analysis **44**(2), 699–731 (2006)
11. Gander, M.J., Magoules, F., Nataf, F.: Optimized Schwarz methods without overlap for the Helmholtz equatio n. SIAM J. Sci. Comput. **24**, 38–60 (2002)
12. Gander, M.J., Nataf, F.: An incomplete LU preconditioner for problems in acoustics. J. Comput. Acoust. **13**, 455–476 (2005)
13. Gander, M.J., Zhang, H.: A class of iterative solvers for the Helmholtz equation: Factorizations, sweeping preconditioners, source transfer, single layer potentials, polarized traces, and optimized Schwarz methods. SIAM Review **61**(1), 3–76 (2019)
14. Stolk, C.C.: A rapidly converging domain decomposition method for the Helmholtz equation. J. Comput. Phys. **241**, 240–252 (2013)
15. Zepeda-Núñez, L., Demanet, L.: The method of polarized traces for the 2D Helmholtz equation. J. Comput. Phys. **308**, 347–388 (2016)