

Sharpened PCG Iteration Bound for High-Contrast Heterogeneous Scalar Elliptic PDEs

Philip Soliman^[0009-0000-1061-6659],
Filipe Cumaru^[0009-0003-9516-4226],
Alexander Heinlein^[0000-0003-1578-8104]

1 Introduction

We study the convergence of the preconditioned conjugate gradient (PCG) method for the linear system $\mathbf{A}\mathbf{u} = \mathbf{b}$ arising from a finite element discretization of a scalar elliptic partial differential equation (PDE) with a high-contrast heterogeneous coefficient function. In particular, let $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) be a bounded domain with Lipschitz boundary $\partial\Omega$, and $C \in L^\infty(\Omega)$ be scalar field defined on Ω . Let us consider the problem: find u such that

$$\begin{aligned} -\nabla \cdot (C\nabla u) &= f && \text{in } \Omega, \\ u &= u_D && \text{on } \partial\Omega. \end{aligned} \tag{1}$$

We employ a two-level overlapping additive Schwarz (2-OAS) preconditioner to the linear system $\mathbf{A}\mathbf{u} = \mathbf{b}$ resulting from discretization of Eq. 1. Let $\Omega_i, i = 1, \dots, N$ be a non-overlapping domain decomposition of Ω . Correspondingly, let $\Omega'_i \supset \Omega_i$ be subdomains with overlap δ . We denote by R_i the restriction operator such that $A_i = R_i A R_i^T$ is the submatrix of A with local Dirichlet boundary conditions on Ω'_i . Finally, let Φ be the prolongation operator whose columns span a coarse space and $A_0 = \Phi A \Phi^T$ be the Galerkin projection of A onto that coarse space. The 2-OAS preconditioner is then given by

Philip Soliman
Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands, e-mail: philipsoliman4133@gmail.com

Filipe Cumaru
Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands, e-mail: f.a.cumarasilvaalves@tudelft.nl

Alexander Heinlein
Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft, The Netherlands, e-mail: a.heinlein@tudelft.nl

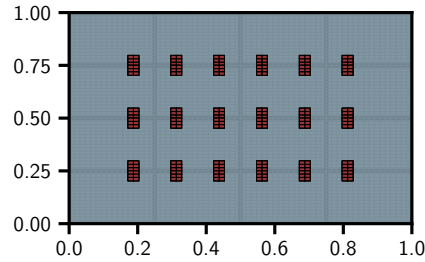
$$M_{2\text{-OAS}}^{-1} = \Phi A_0^{-1} \Phi^T + \sum_{i=1}^N R_i^T A_i^{-1} R_i, \quad (2)$$

leading to the preconditioned system

$$\tilde{A} \mathbf{u} := M_{2\text{-OAS}}^{-1} A \mathbf{u} = M_{2\text{-OAS}}^{-1} \mathbf{b} := \tilde{\mathbf{b}}. \quad (3)$$

The PCG method's convergence and scalability depends heavily on the chosen coarse space, especially for high-contrast coefficients. This work focuses on three such coarse spaces: the algebraic multiscale solver (AMS) [11], generalized Dryja-Smith-Widlund (GDSW) [5], and the reduced-dimension GDSW (RGDSW) [6].

Fig. 1: Example of a two-dimensional periodic high-contrast coefficient distribution with multiple inclusions along subdomain interfaces for a subdomain size $H = 1/4$ and local problem of size $H/h = 16$. Non-overlapping subdomain boundaries $\partial\Omega_i$ and grid elements are indicated with thick and thin lines, respectively. **Light-blue:** $C = 1$; **red:** $C = 10^8$.



We continue the work done in [4], which studied the numerical performance of the 2-OAS preconditioner with the AMS, GDSW and RGDSW coarse spaces for the high-contrast elliptic problem in Eq. 1. The key finding relevant to this work is that the traditional condition number-based bound fails to capture the observed differences in iteration counts between the coarse spaces, if C contains the regular pattern of multiple high coefficient inclusions crossing subdomain interfaces shown in Fig. 1. As shown in [4, Fig. 4], while AMS and GDSW significantly outperformed RGDSW in terms of iteration count, each of their condition numbers were comparable, indicating that the condition number alone is insufficient to predict the PCG iteration count. The authors suggest that these differences can be attributed to the distinct eigenvalue distributions of the preconditioned systems.

In this work, we introduce a sharper iteration bound for the PCG method based on the edge eigenvalues of each cluster in the spectral distribution. Additionally, we show that, in practice, this bound can be estimated using Ritz values resulting from early PCG iterations.

2 Multi-cluster PCG iteration bound

Let \mathbf{u}_0 be an initial guess and $\mathbf{u}_1, \dots, \mathbf{u}_m$ be subsequent PCG iterates. Also, let \mathbf{u}^* be the exact solution of Eq. 3. Assuming $M_{2\text{-OAS}}^{-1}$ is symmetric positive definite (SPD),

we can define $\tilde{A}_s = M_{2\text{-OAS}}^{-1/2} A M_{2\text{-OAS}}^{-1/2}$. Then, the \tilde{A}_s -norm relative error

$$\epsilon_m = \frac{\|\mathbf{u}^* - \mathbf{u}_m\|_{\tilde{A}_s}}{\|\mathbf{u}^* - \mathbf{u}_0\|_{\tilde{A}_s}}$$

after m PCG iterations is bounded by the constrained min-max problem

$$\epsilon_m < \min_{r \in \mathcal{P}_m, r(0)=1} \max_{\lambda \in \sigma} |r(\lambda)|, \quad (4)$$

where \mathcal{P}_m is the space of polynomials of degree at most m and $\sigma := \sigma(\tilde{A}) = \{\lambda_1, \dots, \lambda_n\}$ is the spectrum of \tilde{A} , ordered from smallest to largest. Note that we used similarity of \tilde{A} and \tilde{A}_s as well as [9, Sect. 9.2 and Thm. 6.29] to obtain Eq. 4. In the classical theory of the CG method, we assume a uniformly distributed spectrum and simplify the discrete spectrum σ in Eq. 4 to a continuous interval $\sigma_1 := [\lambda_1, \lambda_n]$. Correspondingly, let $m_1 \in \mathbb{N}^+$ denote the classical PCG iteration bound. We can derive m_1 by introducing the polynomial

$$r(\lambda) \leftarrow r_1(\lambda) := \hat{C}_{m_1} := \frac{C_{m_1}(T(\lambda))}{C_{m_1}(T(0))}, \quad (5)$$

where C_{m_1} is the m_1^{th} -degree Chebyshev polynomial of the first kind and

$$T(\lambda) := \frac{2\lambda - (\lambda_1 + \lambda_n)}{\lambda_n - \lambda_1}.$$

The polynomial $r_1(\lambda)$ solves the minimization problem in Eq. 4 exactly in the interval σ_1 . Finally, let some relative error tolerance ϵ be given. The classical CG iteration bound follows by requiring $\max_{\lambda \in \sigma_1} |r_1(\lambda)| < \epsilon$ and solving this inequality for the degree m_1 , yielding

$$m_1(\kappa) := \left\lceil \frac{\sqrt{\kappa}}{2} \ln \left(\frac{2}{\epsilon} \right) + 1 \right\rceil, \quad (6)$$

where $\kappa = \lambda_n/\lambda_1$ is the condition number of \tilde{A} ; cf. [9, Eq. 6.128].

High-contrast coefficients lead to clustered eigenspectra with condition numbers on the order of the contrast in C ; see [7]. For such cases, m_1 is too pessimistic since it only depends on the relatively large condition number κ and not on the full spectral distribution. To address this overestimation of m by m_1 , we consider the simplest, non-trivial case of a spectrum with a single spectral gap; see [4, Fig. 5] for examples of such a spectrum. Let k be the *partition index* corresponding to those eigenvalues that demarcate the spectral gap as the interval $[\lambda_k, \lambda_{k+1}]$. We define the two-cluster spectrum and PCG iteration bound as $\sigma \subset [\lambda_1, \lambda_k] \cup [\lambda_{k+1}, \lambda_n] := \sigma_2$ and $m_2 \in \mathbb{N}^+$, respectively. Similar to Eq. 5, we set

$$r(\lambda) \leftarrow r_2(\lambda) := \hat{C}_p^{(1)}(\lambda) \hat{C}_{m_2-p}^{(2)}(\lambda),$$

where

$$\hat{C}_q^{(i)}(\lambda) := \frac{C_q(T_i(\lambda))}{C_q(T_i(0))} \quad q = p, m_2 - p, \tag{7}$$

is the scaled, q^{th} -degree Chebyshev polynomial for the i -th cluster with

$$T_i(\lambda) := \frac{2\lambda - (\lambda_{k_{i-1}+1} + \lambda_{k_i})}{\lambda_{k_i} - \lambda_{k_{i-1}+1}} \text{ with } k_0 = 0, k_1 = k, k_2 = n. \tag{8}$$

Although $r_2(\lambda)$ does not necessarily solve the minimization problem in Eq. 4 in the union of intervals, σ_2 , one can still derive an expression for m_2 that is sharper than m_1 for clustered spectra. Namely,

$$m_2(\kappa, \kappa_1, \kappa_2) := \left[1 + \frac{\sqrt{\kappa_2}}{2} \ln\left(\frac{4\kappa}{\kappa_1}\right) + \frac{1}{2} \ln\left(\frac{2}{\epsilon}\right) \left(\sqrt{\kappa_1} + \sqrt{\kappa_2} + \frac{\sqrt{\kappa_1 \kappa_2}}{2} \ln\left(\frac{4\kappa}{\kappa_1}\right) \right) \right], \tag{9}$$

where $\kappa = \lambda_n/\lambda_1$, $\kappa_1 = \lambda_k/\lambda_1$ and $\kappa_2 = \lambda_n/\lambda_{k+1}$; cf. [1, Eq. 4.4].

We generalize the approach to obtain m_2 in [1] to multi-cluster spectra with s clusters, as shown in Fig. 2. We introduce the partition indices k_i , $i = 0, \dots, s$, and

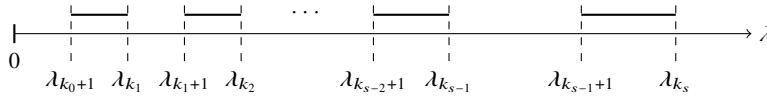


Fig. 2: Example of an eigenvalue spectrum consisting of multiple clusters.

define the i^{th} cluster as the interval $I_i = [\lambda_{k_{i-1}+1}, \lambda_{k_i}]$. Then, we have $\sigma \subset \cup_{i=1}^s I_i := \sigma_s$, and approximate the solution of Eq. 4 by the polynomial

$$r_s(\lambda) := \prod_{i=1}^s \hat{C}_{p_i}^{(i)}(\lambda). \tag{10}$$

Next, we bound Eq. 10 by ϵ using the following Thm. 1, which ensures that the i^{th} cluster’s polynomial $\hat{C}_{p_i}^{(i)}(\lambda)$ does not amplify the value of $|r_s(\lambda)|$ on I_j with $j < i$.

Theorem 1 *Let k_i , $i = 0, \dots, s$ be the partition indices for a spectrum with s clusters σ_s , $\hat{C}_{p_i}^{(i)}(\lambda)$ be as in Eq. 7 and let the i^{th} cluster be denoted as $I_i = [\lambda_{k_{i-1}+1}, \lambda_{k_i}]$. Then, for any $i > 1$ and $1 \leq j < i$, it holds that*

$$\max_{\lambda \in I_j} \left| \hat{C}_{p_i}^{(i)}(\lambda) \right| < 1.$$

Proof. The local extrema of $C_q(x)$ and the roots of the first derivative of $C_q(x)$ for $x \in \mathbb{R}$ are given by $x_l = \cos(l\pi/q)$ with $C_q(x_l) = (-1)^l$ for $l = 1, \dots, q - 1$; c.f. [8, Ch. 2]. By the fundamental theorem of algebra, the first derivative of $C_q(x)$ has no roots other than $x_l \in (-1, 1)$ for $l = 1, \dots, q - 1$. Since $T_i(\lambda) \in [-1, 1]$

for $\lambda \in I_i$, all roots of the first derivative of $\hat{C}_{p_i}^{(i)}(\lambda)$ are in I_i . Therefore, $\hat{C}_{p_i}^{(i)}(\lambda)$ is strictly monotone outside I_i . Since $|\hat{C}_{p_i}^{(i)}(0)| = 1$ and $|\hat{C}_{p_i}^{(i)}(\lambda_{k_{i-1}+1})| < 1$, $|\hat{C}_{p_i}^{(i)}(\lambda)|$ is monotonically decreasing in $[0, \lambda_{k_{i-1}+1}]$. In conclusion, $|\hat{C}_{p_i}^{(i)}(\lambda)| < 1$ on I_j with $I_j \subset (0, \lambda_{k_{i-1}+1}]$, which holds for i, j as stated in the theorem. \square

A consequence of the proof of Thm. 1 is that, for all i, j such that $1 \leq j \leq i \leq s$, $|\hat{C}_{p_j}^{(j)}(\lambda)|$ is monotonically increasing in the interval $[\lambda_{k_j}, \infty)$. Therefore, the maximum value of the polynomial $r_i(\lambda)$ on an entire i -cluster spectrum σ_i is attained at the largest edge of the i -th cluster, i.e., at λ_{k_i} . Hence, we can iteratively bound $|r_s(\lambda)|$ by requiring

$$\left| \prod_{j=1}^i \hat{C}_{p_j}^{(j)}(\lambda_{k_i}) \right| < \epsilon \tag{11}$$

for all $i = 1, \dots, s$ and in that order. The iterative bounding process via Eq. 11 yields expressions for the degrees

$$p_i = \left\lceil \log_{\gamma_i} \left(\frac{\epsilon}{2} \right) + \sum_{j=1}^{i-1} p_j \log_{\gamma_i} \left(\frac{T_j(0) + \sqrt{T_j(0)^2 - 1}}{T_j(\lambda_{k_i}) - \sqrt{T_j(\lambda_{k_i})^2 - 1}} \right) \right\rceil, \tag{12}$$

where $\gamma_i = \frac{\sqrt{\kappa_i} - 1}{\sqrt{\kappa_i} + 1}$ and $\kappa_i = \frac{\lambda_{k_i}}{\lambda_{k_{i-1}+1}}$; c.f. [10]. Finally, the multi-cluster PCG iteration bound is given as the total degree of $r_s(\lambda)$ from Eq. 10

$$m_s(p_1, \dots, p_s) := \sum_{i=1}^s p_i, \tag{13}$$

with p_i as in Eq. 12.

Calculating m_s from Eq. 13 requires partition indices k_i that split the eigenvalues into disjoint clusters. To obtain k_i , we define a candidate two-cluster split using the largest relative gap

$$k^*(\sigma) := \arg \max_i \frac{\lambda_{i+1}}{\lambda_i}, \tag{14}$$

Next, we require the two-cluster bound from Eq. 9 resulting from this candidate split to be smaller than the classical bound from Eq. 6. In doing so, we derive a condition number threshold via the Lambert W function, the inverse of $y \mapsto ye^y$, which arises during solving for κ in $m_2 < m_1$. For a real argument $x \in [-1/e, 0)$ the equation $ye^y = x$ has two real solutions: the principal branch $y = W_0(x) > -1$ and the negative branch $y = W_{-1}(x) \leq -1$. When solving the inequality $m_2 < m_1$, we employ the principal negative branch W_{-1} and its asymptotic expansion for $x \rightarrow 0^-$; see [3, Eq. 4.19]. Letting $\kappa_1 = \lambda_{k^*}/\lambda_1$, $\kappa_2 = \lambda_n/\lambda_{k^*+1}$, $x = -\left(2\sqrt{\kappa_2} \exp\left(\frac{1}{\sqrt{\kappa_2}}\right)\right)^{-1}$, $L = \ln(-x)$ and $l = \ln(-L)$ gives the threshold explicitly as

$$\kappa > 4\kappa_1\kappa_2 \left(L - l + \frac{l}{L} \right)^2 + \mathcal{O} \left(\frac{\kappa_1\kappa_2 l^4}{L^4} \right), \quad (15)$$

If the condition in Eq. 15 holds, we accept the candidate split and apply a greedy recursion on each subcluster until no more splits are accepted or clusters consist of singletons; see [10, Sect. 4.2] for the full derivation of Eq. 15, [10, Alg. 8] for the partitioning algorithm and [10, Alg 9] for the complete computation of m_s . The Lambert- W threshold in Eq. 15 ensures that partitioning the spectrum is beneficial and avoids unnecessary refinement when clusters are not sufficiently separated.

The computational overhead of the spectral partitioning and bound evaluation is comparable to the calculation of Ritz values. Indeed, at PCG iteration k , computing the k Ritz values amounts to solving the eigenvalue problem of the $k \times k$ symmetric, tridiagonal Lanczos matrix and costs between $\mathcal{O}(k \log k)$ and $\mathcal{O}(k^2)$, depending on the algorithm; cf. [2]. The subsequent recursive splitting identifies clusters by iteratively locating the largest relative gap; its cost depends on the spectral structure, being $\mathcal{O}(k \log k)$ for balanced splits and $\mathcal{O}(k^2)$ for maximally unbalanced splits, with practical behavior typically between these extremes. Additionally, evaluation of the condition in Eq. 15 and calculation of m_s are at most $\mathcal{O}(k)$ and $\mathcal{O}(k^2)$, respectively. Overall, the complexity of partitioning and bound evaluation remains comparable to Ritz-value computation, making it feasible for $k \ll n$.

The quality of the resulting bound m_s depends crucially on the accuracy of the Ritz values used for partitioning and bound evaluation. For certain preconditioner configurations, Ritz values stabilize relatively quickly during PCG iterations, yielding reliable early estimates of the spectral structure and thus accurate bounds. For others, slower Ritz-value convergence can lead to inaccurate cluster identification in early iterations, potentially resulting in an underestimation of the true iteration count. In the following section, we investigate these effects numerically for a high-contrast elliptic PDE introduced in Sect. 1.

3 Numerical experiments

We solve the system from Eq. 3 for the coefficient distribution shown in Fig. 1 using the PCG method with the $M_{2\text{-OAS}}$ preconditioner from Eq. 2 with the AMS, GDSW, and RGDSW coarse spaces and for subdomain sizes $H = 1/4, 1/8, 1/16, 1/32, 1/64$, grid size $h = H/16$, overlap $\delta = 2h$ and relative error tolerance $\epsilon = 10^{-8}$. All numerical experiments presented here are for two-dimensional model problems.

First, we compute our multi-cluster bound m_s using Ritz values obtained at convergence. Fig. 3 shows that the classical bound m_1 severely overestimates the iteration count and fails to differentiate between the coarse spaces. In contrast, m_s is consistently accurate, staying within a factor of 1-10 of the true iteration count m and correctly discerning the coarse spaces' performance.

In Table 1, we evaluate m_s as an early estimator by computing it during the initial PCG iterations, i.e., within t_{\max} iterations or a fraction r of the total iteration count

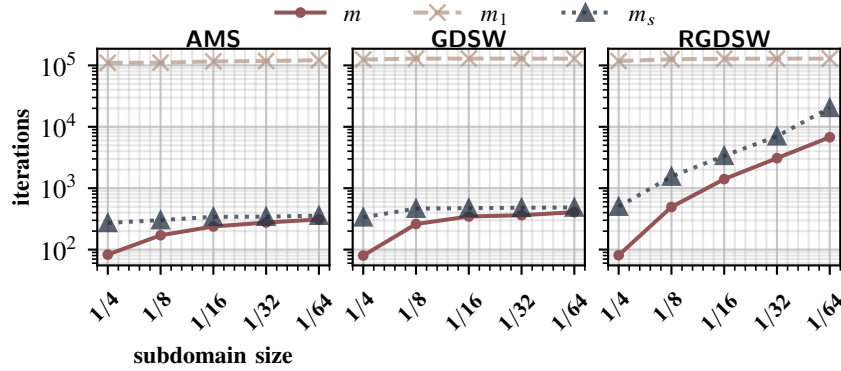


Fig. 3: Actual PCG iterations m versus the classical (m_1) and multi-cluster (m_s) bounds computed at convergence for varying subdomain sizes H .

m . We calculate m_s only if the edge Ritz values in each cluster have stabilized. To lessen computational costs, this check is performed every $\eta = 5$ iterations against a tolerance $\tau = 0.1$

$$\frac{\lambda_{k_j^{(i)}}}{\lambda_{k_j^{(i-\eta)}}} < 1 + \tau, \quad j = 0, \dots, s. \tag{16}$$

The results in Table 1 show that m_s , computed using early Ritz values, provides a good estimate of the final iteration count m for AMS and GDSW for $i < \min\{100, 0.5m\}$. Although the performance of m_s as an upper bound deteriorates slightly for smaller subdomain sizes $H \leq 1/16$ for GDSW and even more so for RGDSW, it is still able to reflect the differences in convergence speed between the three coarse spaces. In contrast, the classical bound m_1 is not able to do so, and again, largely overestimates m in all cases. The underestimation of the number of iterations for RGDSW by m_s is due to the fact that the early Ritz values do not approximate the full spectrum of the preconditioned system matrix well in these cases, as shown in [10, Sec. 6.2].

Table 1: The number of PCG iterations required to achieve convergence m , corresponding iteration bounds m_1 , m_s and the iteration at which the bounds are computed i . Cell colors indicate if bounds are larger (blue) or smaller (red) than m , with shading proportional to absolute difference. Bounds are calculated with $\eta = 5$, $\tau = 0.1$, $i_{\max}=100$ (AMS), 100 (GDSW), 300 (RGDSW) and fraction of total run $r = 0.5$.

	H = 1/4			H = 1/16			H = 1/64		
	AMS	GDSW	RGDSW	AMS	GDSW	RGDSW	AMS	GDSW	RGDSW
m	83	80	81	238	346	1,406	310	407	6,766
m_1	98,421	124,727	117,699	110,969	123,872	122,195	114,629	121,215	69,645
m_s	191	88	133	310	306	2,185	324	312	2,612
i	41	36	36	81	76	241	91	81	291

Allowing a modest increase in PCG iterations before evaluating m_s improves the early estimate for the GDSW coarse space. Indeed, in [10, Tab. 6.3] i_{\max} was raised to 300, and produced $m_s = 485$ for $H = 1/64$ at iteration $i = 186$, a valid upper bound for $m = 407$. Similar behavior is suggested for RGDSW, although Ritz values require substantially more iterations to stabilize and yield reliable estimates.

The proposed multi-cluster bound m_s and the associated algorithm are independent of the spatial dimension. However, their practical performance may differ in three dimensions due to changes in the spectral properties of the problem, with and without Schwarz preconditioning, which can affect clustering behavior and the stabilization of Ritz values.

In summary, the results suggest that the PCG iterations are able to handle unresolved bad eigenmodes as long as they are grouped into clusters. Less computationally complex coarse spaces may be chosen in those cases. Future work could focus on refining the spectral partitioning strategy, applying the bound to more complex PDE systems, and, if possible, developing *a priori* cluster edge eigenvalue estimates to remove the dependency on Ritz values.

References

1. Axelsson, O.: A class of iterative methods for finite element equations. *Comput. Methods Appl. Mech. Eng.* **9**(2), 123–137 (1976)
2. Coakley, E.S., Rokhlin, V.: A fast divide-and-conquer algorithm for computing the spectra of real symmetric tridiagonal matrices. *Appl. Comput. Harmon. Anal.* **34**(3), 379–414 (2013)
3. Corless, R., Gonnet, G., Hare, D., Jeffrey, D., Knuth, D.: On the Lambert W function. *Adv. Comput. Math.* **5**, 329–359 (1996)
4. Cumaru, F., Heinlein, A., Hajibeygi, H.: A computational study of algebraic coarse spaces for two-level overlapping additive schwarz preconditioners. In: *Proceedings of the 28th International Conference on Domain Decomposition Methods at KAUST, Saudi Arabia (2024)*. Accepted for publication.
5. Dohrmann, C.R., Klawonn, A., Widlund, O.B.: A family of energy minimizing coarse spaces for overlapping Schwarz preconditioners. In: *Domain Decomposition Methods in Science and Engineering XVII, Lecture Notes in Computational Science and Engineering*, vol. 60, pp. 247–254 (2008)
6. Dohrmann, C.R., Widlund, O.B.: On the design of small coarse spaces for domain decomposition algorithms. *SIAM J. Sci. Comput.* **39**(4), A1466–A1488 (2017)
7. Graham, I.G., Lechner, P.O., Scheichl, R.: Domain decomposition for multiscale PDEs. *Numer. Math.* **106**, 589–626 (2007)
8. Mason, J.C., Handscomb, D.C.: *Chebyshev Polynomials*, 1st edn. Chapman and Hall/CRC (2002)
9. Saad, Y.: *Iterative Methods for Sparse Linear Systems*, second edn. Society for Industrial and Applied Mathematics (2003)
10. Soliman, P.: *Sharpened CG iteration bound for high-contrast heterogeneous scalar elliptic PDEs: going beyond condition number*. Master thesis, TU Delft (2025)
11. Wang, Y., Hajibeygi, H., Tchelepi, H.: Algebraic multiscale solver for flow in heterogeneous porous media. *J. Comput. Phys.* **259**, 284–303 (2014)