EÖTVÖS LORÁND UNIVERSITY

ELTE
EÖTVÖS LORÁND
TUDOMÁNYEGYETEM

DEPARTMENT OF APPLIED ANALYSIS AND COMPUTATIONAL MATHEMATICS

Superlinear convergence of iterative methods for elliptic PDEs and systems

Author:
Sebastián Josue Castillo
Advisor:
Dr. Karátson János

Budapest - November 18, 2022

1 Abstract

The conjugate gradient method (CGM) is a widespread way to find the solution of discretized elliptic partial differential equations iteratively. Furthermore, the preconditioned CGM can be competitive with multigrid methods and, under certain conditions, operator preconditioning can provide mesh-independent superlinear convergence. This project considers a self-adjoint second-order elliptic boundary value problem with variable zeroth order coefficient and its finite element discretization. We study the mesh-independent superlinear convergence of the preconditioned CGM for this type of problem see e.g [8], [3], and extend previous results of [8] to the case of unbounded reaction coefficients in some Lebesgue spaces. Our goal is to find an eigenvalue-based estimation of the rate of superlinear convergence and to show that a similar estimation can be obtained in the case of systems of PDEs.

2 General framework

Let H be a real separable Hilbert space and let us consider a linear operator equation

$$Bu = g \tag{1}$$

with some $g \in H$, under the following assumptions

(i) The operator B is decomposed as

$$B = S + Q \tag{2}$$

where S is a symmetric operator in H with dense domain D and Q is a compact self-adjoint operator defined on the domain H.

(ii) There exists $k > 0$ such that $\langle Su, u\rangle \geq k\|u\|^2$, $u \in D$.

(iii) $\langle Qu, u\rangle \geq 0$, $u \in D$.

We recall that the energy space H_S is the completion of D under the *energy inner product*

$$\langle u, v\rangle_S = \langle Su, v\rangle \tag{3}$$

, and the corresponding norm is denoted by $\|\cdot\|_S$. Assumption (ii) implies $H_S \subset H$. Then, there exists a unique operator denoted by $Q_S : H_S \mapsto H_s$ such that

$$\langle Q_S u, v\rangle_S = \langle Qu, v\rangle$$

for all $u, v \in H_S$.

We replace equation (1) by its formally preconditioned form $(I + S^{-1}Q)u = S^{-1}g$ in H_S. This is equivalent to the weak formulation

$$\langle (I + Q_S)u, v\rangle_S = \langle g, v\rangle, \quad \forall v \in H_s. \tag{4}$$

Since by assumption (iii) the bilinear form on the left is coercive on H_S, by the *Lax-Milgram theorem*, there exists a unique solution $u \in H_S$ of (4).

Now equation (4) is solved numerically using a *Galerkin discretization.*

Construction of the discretization. Let $V = \operatorname{span}\{\varphi_1, \dots, \varphi_n\} \subset H_S$ be a given finite-dimensional subspace,

$$\mathbf{S} = \{\langle \varphi_i, \varphi_j \rangle_S\}_{i,j=1}^n \quad \text{and } \mathbf{Q} = \{\langle Q\varphi_i, \varphi_j \rangle\}_{i,j=1}^n$$

the *Gram matrices* corresponding to S and Q. We look for the numerical solution $u_V \in V$ of equation (4) in V, i.e., for which

$$\langle (I + Q_S)u_v, v \rangle_S = \langle g, v \rangle, \quad \forall v \in V. \tag{5}$$

Then $u_V = \sum_{i,j=1}^n c_j \varphi_j$, where $\mathbf{c} = (c_1, \dots, c_n) \in \mathbb{R}^n$ is the solution of the system

$$(\mathbf{S} + \mathbf{Q})\mathbf{c} = \mathbf{b} \tag{6}$$

with $\mathbf{b} = \{\langle g, \varphi_j \rangle\}_{j=1}^n$ depending on V. The matrix $\boldsymbol{B} := \mathbf{S} + \mathbf{Q}$ is SPD.

By using matrix $\mathbf{S}$ as the preconditioner for the system (6), we shall work with the preconditioned system

$$(\mathbf{I} + \mathbf{S}^{-1}\mathbf{Q})\mathbf{c} = \tilde{\mathbf{b}}, \tag{7}$$

where $\tilde{\mathbf{b}} = \mathbf{S}^{-1}\mathbf{b}$ and $\mathbf{I}$ is the identity matrix in $\mathbb{R}^n$. Then we apply the CGM for the solution of this new system.

Preconditioned conjugate gradient method algorithms. The method is given by the following algorithm: Let $u_0 \in H$ arbitrary, $\rho_0 = \mathbf{B}u_0 - g$, $\mathbf{S}p_0 = \rho_0$, $r_0 = \rho_0$ and for $k \in \mathbb{N}$

$$\begin{cases} u_{k+1} = u_k + \alpha_k p_k, \\ r_{k+1} = r_k + \alpha_k \mathbf{S}^{-1}\mathbf{B}p_k, \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases}$$

with

$$\alpha_k = \frac{-\|r_k\|_{\mathbf{S}}^2}{\langle \mathbf{B}p_k, p_k \rangle}, \quad \beta_k = \frac{\|r_{k+1}\|_{\mathbf{S}}^2}{\|r_k\|_{\mathbf{S}}^2}.$$

Note that it is not necessary to compute the inverse of $\mathbf{S}$. Instead, we solve the auxiliary problem

$$\begin{cases} \mathbf{S}z_k = \mathbf{B}p_k \\ r_{k+1} = r_k + \alpha_k z_k. \end{cases}$$

By setting $w_k = z_k - p_k$, the previous system is equivalent to

$$\begin{cases} \mathbf{S}w_k = \mathbf{Q}p_k, \\ r_{k+1} = r_k + \alpha_k z_k. \end{cases}$$

The next step is to find superlinear convergence rates for the CGM. Let $\mathbf{A} = (\mathbf{I} + \mathbf{S}^{-1}\mathbf{Q})$ and $\mathbf{E} = \mathbf{S}^{-1}\mathbf{Q}$. Assume that $\lambda_j = \lambda_j(\mathbf{A})$ are ordered according to $|\lambda_1 - 1| \geq |\lambda_2 - 1| \geq \cdots \geq |\lambda_n - 1|$. Then $\lambda_j(\mathbf{E}) = \lambda_j - 1$ and the *error vectors* $e_k = c_k - c$ satisfy [1]

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A} \right)^{1/k} \leq \frac{2\|\mathbf{A}^{-1}\|}{k} \sum_{j=1}^k |\lambda_j(\mathbf{S}^{-1}\mathbf{Q})|, \quad k = 1, 2, \dots, n. \tag{8}$$

The following result allows us to give a convergence rate for the upper bound of (8) through the eigenvalues of the operator Q_S. This is a modification of Theorem 1 in [8] where the square of eigenvalues was considered.

Theorem 1. *For any* $k = 1, 2, \dots, n$

$$\sum_{j=1}^{k} |\lambda_j(\mathbf{S}^{-1}\mathbf{Q})| \le \sum_{j=1}^{k} \lambda_j(Q_S), \tag{9}$$

Proof. Let $\lambda_m = \lambda_m(\mathbf{S}^{-1}\mathbf{Q})$. Let $\mathbf{c}^m = (c_1^m, \dots, c_n^m) \in \mathbb{R}^n$ be the corresponding eigenvectors. Then

$$\mathbf{Q}\mathbf{c}^m = \lambda_m \mathbf{S}\mathbf{c}^m \tag{10}$$

for all m. Since $\mathbf{S}^{-1}\mathbf{Q}$ is self-adjoint with respect to the $\mathbf{S}-$inner product, therefore all eigenvalues $\lambda_1, \dots, \lambda_n$ are real, counting with multiplicity. Furthermore, the corresponding eigenvectors are orthogonal in $\mathbb{R}^n$ with respect to the $\mathbf{S}-$inner product. Let us choose them such that they are also orthonormal:

$$\mathbf{S}c^m \cdot c^l = \delta_{ml}, \quad m, l = 1, \dots, n,$$

where δ_{ml} is the Kronecker delta.

Let $u_m = \sum_{i=1}^{n} c_i^m \varphi_i \in V$, $m = 1, \dots, n$. Then for all $m, l = 1, \dots, n$ we have that

$$\langle u_m, u_l \rangle_S = \sum_{i,j=1}^{n} \langle \varphi_i, \varphi_j \rangle_S c_i^m c_j^l = \mathbf{S}c^m \cdot c^l, \tag{11}$$

hence (10) implies that $u_1, \dots, u_n$ form an orthonormal basis in $V \subset H_S$ with respect to the H_S-inner product. Then (10),(11) yield

$$\mathbf{Q}c^m \cdot c^l = \lambda_m \delta_{ml}, \quad m, l = 1, \dots, n.$$

Hence, we obtain

$$\langle Q_S u_m, u_l \rangle_S = \lambda_m \delta_{ml}, \quad m, l = 1, \dots, n. \tag{12}$$

Using Corollary 3.3 of [7] and since Q_S is a positive compact self-adjoint operator on the Hilbert space H_S, we have that

$$\sum_{m=1}^{n} |\langle Q_S u_m, u_m \rangle_S| \le \sum_{m=1}^{n} s_j(Q_S) = \sum_{m=1}^{n} \lambda_j(Q_S), \tag{13}$$

where $s_j(Q_S)$ are the singular values of Q_S. Then, by (12) and (13) we arrive at

$$\sum_{m=1}^{n} |\lambda_m| = \sum_{m=1}^{n} |\langle Q_S u_m, u_m \rangle_S| \le \sum_{m=1}^{n} \lambda_j(Q_S).$$

□

An immediate consequence of this theorem is the following mesh-independent bound.

Corollary 1. *For any $k = 1, 2, \ldots, n$*

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} \lambda_j(Q_S), \quad k = 1, 2, \ldots, n. \tag{14}$$

Proof. By [2, Prop. 4.1] we are able to estimate $\|\mathbf{A}\|$ to obtain

$$\|(\mathbf{I} + \mathbf{S}^{-1}\mathbf{Q})^{-1}\| \leq \|(I + Q_S)^{-1}\|.$$

This, together with the previous result and (8) completes the proof. □

Since $|\lambda_1(Q_S)| \geq |\lambda_2(Q_S)| \geq \cdots \geq 0$ and the eigenvalues tend to 0, the convergence factor is less than 1 for k sufficiently large. Hence the upper bound decreases as $k \longrightarrow \infty$ and we obtain superlinear convergence rate.

3 The main results

Let $d \geq 2$, $p > 2$ and $\Omega \subset \mathbb{R}^d$ be a bounded domain. We consider the elliptic problem

$$\begin{cases} -\mathrm{div}(G\nabla u) + \eta u = g, \\ u|_{\partial\Omega} = 0, \end{cases} \tag{15}$$

under the standard assumptions listed below. We shall focus on the case when the principal part has constant or separable coefficients, i.e.,

$$G(x) \equiv G \in \mathbb{R}^d \times \mathbb{R}^d \quad \text{or } G(x) \equiv \mathrm{diag}\{G_i(x_i)\}_{i=1}^N$$

whereas $\eta = \eta(x)$ is a general variable (i.e. nonconstant) coefficient. Let problem (15) satisfy the following assumptions:

(i) The symmetric matrix-valued function $G \in \mathrm{L}^\infty(\overline{\Omega}, \mathbb{R}^d \times \mathbb{R}^d)$ satisfies

$$G(x)\xi \cdot \xi \geq m|\xi|^2$$

for all $\xi \in \mathbb{R}^d$, $m > 0$ independent of ξ.

(ii) $\eta \in \mathrm{L}^{p/(p-2)}(\Omega)$ and $\eta \geq 0$.

(iii) $\partial\Omega$ is a Lipschitz boundary.

(iv) $g \in \mathrm{L}^2(\Omega)$.

Then problem (15) has a unique weak solution in $\mathrm{H}_0^1(\Omega)$.

Let $V_h \subset \mathrm{H}_0^1(\Omega)$ be a given FEM subspace. We look for the numerical solution u_h of (15) in V_h:

$$\int_\Omega (G\nabla u_h \cdot \nabla v + \eta u_h v) = \int_\Omega g v, \quad v \in V_h. \tag{16}$$

The corresponding linear algebraic system has the form

$$(\mathbf{G}_h + \mathbf{D}_h)\mathbf{c} = \mathbf{g}_h,$$

where $\mathbf{G}_h$ and $\mathbf{D}_h$ are the corresponding stiffness and mass matrices, respectively. We apply the matrix $\mathbf{G}_h$ as preconditioner, thus the preconditioned form of (16) is given by

$$(\mathbf{I}_h + \mathbf{G}_h^{-1}\mathbf{D}_h)\mathbf{c} = \tilde{\mathbf{g}}_h \tag{17}$$

with $\tilde{\mathbf{g}}_h = \mathbf{G}_h^{-1}\mathbf{g}_h$. Then we apply the CGM to (17) and the auxiliary systems with $\mathbf{G}_h$ can be solved efficiently with fast solvers.

Theorem 2. *Let* $2 < p < \frac{2d}{d-2}$*, and* m *the lower spectral bound of* G *given by assumption* (i)*. Then there exists* $C > 0$ *such that for all* $k \in \mathbb{N}$

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{\frac{1}{k}} \leq Ck^{-\alpha}, \tag{18}$$

where $\alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}$.

Proof. Let us consider the Hilbert space $\mathrm{L}^2(\Omega)$ endowed with the usual inner product. Let $D = \{u \in \mathrm{H}_0^1(\Omega) \cap \mathrm{H}^2(\Omega) \ / \ G\nabla u \in \mathrm{H}^1(\Omega)^N\}$. We define the operators

$$Su \equiv -\mathrm{div}(G\nabla u), \quad u \in D \quad \text{and} \quad Qu \equiv \eta u, \quad u \in \mathrm{H}_0^1(\Omega)$$

and since $p < 2^* = \frac{2N}{N-2}$, the embedding $\mathcal{I} : \mathrm{H}_0^1(\Omega) \to \mathrm{L}^p(\Omega)$ is compact, in particular, there exists $\hat{c} > 0$ such that for all $u \in \mathrm{H}_0^1(\Omega)$

$$\|u\|_{\mathrm{L}^p(\Omega)} \leq \hat{c}\|u\|_{\mathrm{H}_0^1(\Omega)}. \tag{19}$$

Then

$$\langle Su, u\rangle \geq m\int_\Omega |\nabla u|^2 \geq m\nu\int_\Omega u^2, \qquad u \in D,$$

where ν is the Sobolev constant. Hence, the energy space H_S is a well-defined Hilbert space with $\langle u, v\rangle_S = \int_\Omega G\nabla u\cdot\nabla v$. It is easy to see that $H_S = \mathrm{H}_0^1(\Omega)$ and that the following inequality

$$\sqrt{m}\|u\|_{\mathrm{H}_0^1(\Omega)} \leq \|u\|_{H_S} \tag{20}$$

holds for all $u \in H_S$. Furthermore,

$$\begin{aligned}
\|Q_S v\|_{H_S} = \sup_{\|u\|_{H_S}=1} |\langle Q_S v, u\rangle_S| &= \sup_{\|u\|_{H_S}=1} \langle Q v, u\rangle \\
&= \sup_{\|u\|_{H_S}=1} \int_\Omega \eta v u \\
&\le \sup_{\|u\|_{H_S}=1} \left(\int_\Omega |\eta|^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}} \left(\int_\Omega |v|^p\right)^{\frac{1}{p}} \left(\int_\Omega |u|^p\right)^{\frac{1}{p}} \\
&\le \hat{c} \sup_{\|u\|_{H_S}=1} \|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)} \|v\|_{\mathrm{L}^p(\Omega)} \|u\|_{\mathrm{H}_0^1(\Omega)} \\
&\le \frac{\hat{c}}{\sqrt{m}} \sup_{\|u\|_{H_S}=1} \|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)} \|v\|_{\mathrm{L}^p(\Omega)} \|u\|_{H_S} \\
&= \frac{\hat{c}M}{\sqrt{m}} \|v\|_{\mathrm{L}^p(\Omega)},
\end{aligned} \tag{21}$$

where $M = \|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)}$. Here we applied the extension of Hölder's inequality ([4, Th. 4.6]) with

$$1 = \frac{1}{p} + \frac{1}{p} + \left(\frac{p-2}{p}\right).$$

Hence Q_S is compact and self-adjoint in H_S.

Let $\lambda_n = \lambda_n(Q_S)$. Since Q_S is a compact self-adjoint operator in H_S, by [7, Ch.6, Th.1.5] we have the following characterization of the eigenvalues of Q_S:

$$\forall n \in \mathbb{N}: \quad \lambda_n(Q_S) = \min\{\|Q_S - L_{n-1}\| \;/\; L_{n-1} \in \mathcal{L}(H_S), \operatorname{rank}(L_{n-1}) \le n-1\}. \tag{22}$$

By taking the minimum over a smaller subset of finite rank operators, we obtain

$$\lambda_n(Q_S) \le \min\{\|Q_S - Q_S L_{n-1}\| \;/\; L_{n-1} \in \mathcal{L}(H_S), \operatorname{rank}(L_{n-1}) \le n-1\}. \tag{23}$$

Now, by (21) and (20) we get

$$\begin{aligned}
\|Q_S - Q_S L_{n-1}\| &= \sup_{u \in H_S} \frac{\|(Q_S - Q_S L_{n-1})u\|_{H_S}}{\|u\|_{H_S}} \\
&= \sup_{u \in H_S} \frac{\|Q_S(u - L_{n-1}u)\|_{H_S}}{\|u\|_{H_S}} \\
&\le \frac{\hat{c}M}{\sqrt{m}} \sup_{u \in H_S} \frac{\|u - L_{n-1}u\|_{\mathrm{L}^p(\Omega)}}{\|u\|_{H_S}} \\
&\le \frac{\hat{c}M}{\sqrt{m}\sqrt{m}} \sup_{u \in \mathrm{H}_0^1(\Omega)} \frac{\|u - L_{n-1}u\|_{\mathrm{L}^p(\Omega)}}{\|u\|_{\mathrm{H}_0^1(\Omega)}}.
\end{aligned}$$

This, together with (23) yields

$$\begin{aligned}
\lambda_n(Q_S) &\le \frac{\hat{c}M}{m} \min\{\|\mathcal{I} - L_{n-1}\| \;/\; L_{n-1} \in \mathcal{L}(\mathrm{H}_0^1(\Omega), \mathrm{L}^p(\Omega)), \operatorname{rank}(L_{n-1}) \le n-1\} \\
&:= \frac{\hat{c}M}{m} a_n(\mathcal{I}),
\end{aligned} \tag{24}$$

where $a_n(\mathcal{I})$ denotes the approximation numbers of the compact embedding $\mathcal{I}\colon \mathrm{H}_0^1(\Omega) \mapsto \mathrm{L}^p(\Omega)$, [11]. Furthermore, we have the estimation [6]

$$a_n(\mathcal{I}) \leq \hat{C} n^{-\alpha}, \quad \alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p},$$

for some constant $\hat{C} > 0$. Therefore, we arrive at the inequality

$$\lambda_n(Q_S) \leq \frac{\hat{C}\hat{c}M}{m} n^{-\alpha}.$$

Now, taking the arithmetic mean on both sides and estimating the sum from above by an integral we obtain

$$\frac{1}{k}\sum_{n=1}^{k} \lambda_n(Q_S) \leq \frac{\hat{C}\hat{c}M}{m}\frac{1}{k}\left(1 + \int_1^k \frac{1}{x^\alpha}\right) \leq \frac{\hat{C}\hat{c}M}{m(1-\alpha)}\frac{1}{k^\alpha}. \tag{25}$$

Then, by (14), we conclude. □

Remark 1. *The auxiliary problem* $\mathbf{S}w_k = \mathbf{Q}p_k$ *for the PCGM can be solved easily with fast solvers due to the special structure of* S*, [9], [5].*

3.1 Elliptic systems

In this section, we prove that the previous results can be extended to systems of the form

$$\begin{cases} -\Delta u_i + \eta_{i1}u_1 + \ldots \eta_{is}u_s = g_i, \\ u_i|_{\partial\Omega} = 0, \quad (i = 1, \ldots, s), \end{cases} \tag{26}$$

where $\boldsymbol{H} = \{\eta_{ij}\}_{i,j=1}^s$ is a symmetric positive semidefinite variable coefficient matrix such that

$$\forall i, j \in \{1, \ldots, s\}: \quad \eta_{ij} \in \mathrm{L}^{p/(p-2)}(\Omega).$$

We work with the space $\mathrm{L}^p(\Omega)^s$ with the norm

$$\|u\|_{\mathrm{L}^p(\Omega)^s} = \left(\sum_{j=1}^{s} \|u_j\|_{\mathrm{L}^p(\Omega)}^2\right)^{1/2}, \quad u = (u_1, \ldots, u_s) \in \mathrm{L}^p(\Omega)^s.$$

Let $H = \mathrm{L}^2(\Omega)^s$. Let $u = (u_1 \ldots u_s) \in \boldsymbol{D} = (\mathrm{H}_0^1(\Omega) \cap \mathrm{H}^2(\Omega))^s$, we define the operators

$$Su = \begin{pmatrix} -\Delta u_1 \\ \cdot \\ \cdot \\ \cdot \\ -\Delta u_s \end{pmatrix}, \qquad Qu = \boldsymbol{H}u, \qquad u \in \mathrm{H}_0^1(\Omega)^s. \tag{27}$$

Clearly, S is a uniformly positive symmetric operator in H. In fact, by Poincare's inequality

$$\langle Su, u\rangle \geq \frac{1}{\nu^2}\sum_{i=1}^{s}\|u_i\|^2_{\mathrm{L}^2(\Omega)} = \frac{1}{\nu^2}\|u\|^2_H, \tag{28}$$

where ν is the Sobolev constant. Then, the energy space H_S is well defined with

$$\langle u, v\rangle_S = \sum_{i=1}^{s}\int_\Omega \nabla u_i \nabla v_i, \quad \|u\|^2_{H_S} = \sum_{i=1}^{s}\int_\Omega |\nabla u_i|^2$$

and so $H_S = \mathrm{H}^1_0(\Omega)^s$. Furthermore, by (19) we have that

$$\|u\|^2_{H_S} \geq \frac{1}{\hat{c}^2}\sum_{i=1}^{s}\|u_i\|^2_{\mathrm{L}^p(\Omega)} = \frac{1}{\hat{c}^2}\|u\|^2_{\mathrm{L}^p(\Omega)^s}. \tag{29}$$

Then there exists a unique operator $Q_S\colon \mathrm{H}^1_0(\Omega)^s \to \mathrm{L}^2(\Omega)^s$ such that

$$\langle Q_S u, v\rangle_S = \int_\Omega \sum_{i,j=1}^{s} \eta_{ij} u_j v_i. \tag{30}$$

It is easy to see that Q_S is self-adjoint in H_S. Analogous to (21), by (29), (28) and Hölder's inequality we get

$$\begin{aligned}
\|Q_S v\|_{H_S} &= \sup_{\|u\|_S=1} |\langle Q_S v, u\rangle_S| \\
&\leq \sup_{\|u\|_{H_S}=1} \sum_{i,j=1}^{s}\int_\Omega |\eta_{ij}||v_j||u_i| \\
&\leq \sup_{\|u\|_{H_S}=1} \sum_{i,j=1}^{s} \|\eta_{ij}\|_{\mathrm{L}^{p/(p-2)}(\Omega)}\|v_j\|_{\mathrm{L}^p(\Omega)}\|u_i\|_{\mathrm{L}^p(\Omega)} \\
&\leq M \sup_{\|u\|_{H_S}=1} \sum_{j=1}^{s}\|v_j\|_{\mathrm{L}^p(\Omega)} \sum_{i=1}^{s}\|u_i\|_{\mathrm{L}^p(\Omega)} \\
&\leq M \sup_{\|u\|_{H_S}=1} \sqrt{s}\left(\sum_{j=1}^{s}\|v_j\|^2_{\mathrm{L}^p(\Omega)}\right)^{1/2} \sqrt{s}\left(\sum_{i=1}^{s}\|u_i\|^2_{\mathrm{L}^p(\Omega)}\right)^{1/2} \\
&= Ms \sup_{\|u\|_{H_S}=1} \|v\|_{\mathrm{L}^p(\Omega)^s}\|u\|_{\mathrm{L}^p(\Omega)^s} \\
&\leq Ms\hat{c}\|v\|_{\mathrm{L}^p(\Omega)^s},
\end{aligned} \tag{31}$$

where $M = \max_{i,j}\|\eta_{ij}\|_{\mathrm{L}^{p/(p-2)}(\Omega)}$. Hence, we have proved that Q_S is a compact self-adjoint operator in H_S. Then, the characterization (22) of the eigenvalues of Q_S holds. The rest of the proof follows by modifying the scalar case. In this case, we take the minimum over a smaller subset of finite rank operators to obtain

$$\lambda_n(Q_S) \leq \min\{\|Q_S - Q_S L_{n-1}\| \;/\; L_{n-1} \in \mathcal{L}_{\mathrm{diag}}(H_S), \mathrm{rank}(L_{n-1}) \leq n-1\},$$

with $L_{n-1} \in \mathcal{L}_{\mathrm{diag}}(H_S)$ if and only if

$$L_{n-1}u = \begin{pmatrix} L^s_{n-1}u_1 \\ \cdot \\ \cdot \\ \cdot \\ L^s_{n-1}u_s \end{pmatrix}, \text{ such that } L^s_{n-1} \in \mathcal{L}(\mathrm{H}^1_0(\Omega)) \text{ and } \mathrm{rank}(L^s_{n-1}) \le \left[\frac{n-1}{s}\right].$$

Furthermore, we shall use the approximation numbers

$$a_{\left[\frac{n-1}{s}\right]} = \min\left\{\|I - T_{n-1}\| \ / \ T_{n-1} \in \mathcal{L}(\mathrm{H}^1_0(\Omega), \mathrm{L}^p(\Omega)), \mathrm{rank}(T_{n-1}) \le \left[\frac{n-1}{s}\right]\right\}.$$

Note that if $n \le s$, then we can use $\lambda_n(Q_S) \le \|Q_S\|$, and for $n \ge s+1$ the above numbers are estimated by

$$a_{\left[\frac{n-1}{s}\right]} \le \hat{C}\left[\frac{n-1}{s}\right]^{-\alpha}, \tag{32}$$

with $\alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}$. Then

$$\begin{aligned}
\|Q_S - Q_S L_{n-1}\| &= \sup_{u \in H_S} \frac{\|(Q_S - Q_S L_{n-1})u\|_{H_S}}{\|u\|_{H_S}} \\
&= \sup_{u \in H_S} \frac{\|Q_S(u - L_{n-1}u)\|_{H_S}}{\|u\|_{H_S}} \\
&\le Ms\hat{c} \sup_{u \in H_S} \frac{\|u - L_{n-1}u\|_{\mathrm{L}^p(\Omega)^s}}{\|u\|_{H_S}} \\
&= Ms\hat{c} \sup_{u \in H_S} \frac{\left(\sum_{j=1}^s \|u_i - L^s_{n-1}u_i\|^2_{\mathrm{L}^p(\Omega)}\right)^{1/2}}{\left(\sum_{j=1}^s \|u_i\|^2_{\mathrm{H}^1_0(\Omega)}\right)^{1/2}} \\
&\le Ms\hat{c} \sup_{u \in H_S} \frac{\left(\|I - L^s_{n-1}\|^2_{\mathcal{L}(\mathrm{H}^1_0(\Omega),\mathrm{L}^p(\Omega))} \sum_{j=1}^s \|u_i\|^2_{\mathrm{H}^1_0(\Omega)}\right)^{1/2}}{\left(\sum_{j=1}^s \|u_i\|^2_{\mathrm{H}^1_0(\Omega)}\right)^{1/2}} \\
&= Ms\hat{c}\|I - L^s_{n-1}\|_{\mathcal{L}\left(\mathrm{H}^1_0(\Omega),\mathrm{L}^p(\Omega)\right)}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\lambda_n(Q_S) &\le Ms\hat{c}\min\left\{\|I - L^s_{n-1}\|_{\mathcal{L}\left(\mathrm{H}^1_0(\Omega),\mathrm{L}^p(\Omega)\right)} \ / \ L^s_{n-1} \in \mathcal{L}(\mathrm{H}^1_0(\Omega), \mathrm{L}^p(\Omega)), \mathrm{rank}(L^s_{n-1}) \le \left[\frac{n-1}{s}\right]\right\} \\
&= Ms\hat{c}a_{\left[\frac{n-1}{s}\right]}.
\end{aligned}$$

Hence, by (32) we obtain the estimation

$$\lambda_n(Q_S) \le Ms\hat{c}\hat{C}\left[\frac{n-1}{s}\right]^{-\alpha}, \quad n \ge s+1. \tag{33}$$

$$\lambda_n(Q_S) \leq \|Q_S\| \leq Ms\hat{c} \quad n \leq s. \tag{34}$$

Note that there exists $k_0, k_1 > 0$ such that

$$k_0 \leq \frac{[x]}{x} \leq k_1, \quad \forall x > 1.$$

Thus, for $n \geq s+1$

$$\begin{aligned} \left[\frac{n-1}{s}\right]^{-\alpha} &\leq \frac{1}{k_0^\alpha}\frac{s^\alpha}{(n-1)^\alpha} \\ &= \left(\frac{s}{k_0}\right)^\alpha \left(\frac{n^\alpha}{(n-1)^\alpha}\right)\frac{1}{n^\alpha} \\ &\leq \left(\frac{(s+1)}{k_0}\right)^\alpha \frac{1}{n^\alpha}. \end{aligned}$$

Hence, (33) becomes

$$\lambda_n(Q_S) \leq Ms\hat{c}\hat{C}\left(\frac{(s+1)}{k_0}\right)^\alpha \frac{1}{n^\alpha} := C_1\frac{1}{n^\alpha}.$$

and by taking arithmetic meaning on both sides and splitting the sum we get

$$\begin{aligned} \frac{1}{k}\sum_{n=1}^{k}\lambda_n(Q_S) &\leq \frac{1}{k}\left(s\|Q_S\| + \sum_{n=s+1}^{k}\lambda_n(Q_S)\right) \\ &\leq \frac{1}{k}\left(s\|Q_S\| + C_1\sum_{n=s+1}^{k}\frac{1}{n^\alpha}\right) \\ &\leq \frac{1}{k}\left(s\|Q_S\| + C_1\int_s^k \frac{1}{x^\alpha}\right) \\ &\leq \frac{s}{k}\|Q_S\| + \frac{C_1}{1-\alpha}\frac{1}{k^\alpha} \\ &\leq C_2\frac{1}{k^\alpha}, \end{aligned}$$

where $C_2 = \max\{s\|Q_S\|, C_1(1-\alpha)^{-1}\}$. Finally, by Corollary 1, we have proved there exists $\boldsymbol{C} > 0$ such that for all $k \in \mathbb{N}$

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{\frac{1}{k}} \leq \boldsymbol{C}k^{-\frac{1}{\alpha}}. \tag{35}$$

3.2 Extension to non-symmetric systems

Let us now study (26) for $\boldsymbol{H} = \{\eta_{i,j}\}_{i,j=1}^{s}$ non-symmetric. We apply the *generalized minimal residual (GMRES) method* to the corresponding discretized system. This method is an extension of the CG method to non-symmetric systems, [10].

First, we note that in the proof of Theorem 1 we show that (9) also holds if we exchange the eigenvalues of Q_S with its singular values. Furthermore, by **[Robust Super. conv. paper]** we have an analogue of Corollary 1 when A is non-hermitian. In this case, the

GMRES method is applied to the system and we obtain superlinear converge estimates for the residuals $r_k = Au_k - g$:

$$\left(\frac{\|r_k\|_A}{\|r_0\|_A}\right)^{1/k} \leq \frac{\|A^{-1}\|}{k} \sum_{j=1}^{k} s_j(Q_S), \quad \forall k = 1, 2, \ldots, n. \tag{36}$$

To show that Theorem 2 still holds in this case, we follow the same steps as we did previously. We define the operators S, Q, Q_S as before, (27), (30). Here, Q_S is no longer self-adjoint and its eigenvalues do not coincide with its singular values. Nonetheless, by [7, Ch.6, Th.1.5] we have the following characterization of the singular values of Q_S:

$$\forall n \in \mathbb{N}: \quad s_n(Q_S) = \min\{\|Q_S - L_{n-1}\| \;/\; L_{n-1} \in \mathcal{L}(H_S), \operatorname{rank}(L_{n-1}) \leq n-1\}. \tag{37}$$

Then, similarly to the proof for symmetric systems, we can show that there exists $C_1 > 0$ such that

$$\frac{1}{k} \sum_{n=1}^{k} s_n(Q_S) \leq C_1 \frac{1}{k^\alpha}, \quad \alpha = \frac{1}{d} - \frac{1}{2} + \frac{1}{p}. \tag{38}$$

Therefore, by (36), we obtain that there exists $C_2 > 0$ such that

$$\left(\frac{\|r_k\|_A}{\|r_0\|_A}\right)^{1/k} \leq C_2 \frac{1}{k^\alpha}. \tag{39}$$

Finally, note that $r_k = Ae_k$. Then $\|e_k\|_A \leq \|A^{-1}\| \|r_k\|_A$ and $\|r_0\| \leq \|A\| \|e_0\|_A$. Hence

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq C_2 \frac{1}{k^\alpha} \operatorname{cond}(A)^{1/k} \leq C_2 \frac{1}{k^\alpha}.$$

where $\operatorname{cond}(A) = \|A\| \|A^{-1}\| < 1$ denotes the conditioning number of A.

Remark 2. *For elliptic symmetric systems, the auxiliary problem* $\mathbf{S}w_k = \mathbf{Q}p_k$ *for the PCGM becomes*

$$\begin{cases} -\Delta(w_k)_1 & = \sum_{j=1}^{s} \eta_{1j}(p_k)_j, \\ -\Delta(w_k)_2 & = \sum_{j=1}^{s} \eta_{2j}(p_k)_j, \\ & . \\ & . \\ & . \\ -\Delta(w_k)_s & = \sum_{j=1}^{s} \eta_{sj}(p_k)_j, \\ (w_i)|_{\partial\Omega} & = 0, \quad \forall i = 1, \ldots, s. \end{cases}$$

Note that these equations are independent of one another. Hence, they can be solved in parallel. Furthermore, in practice, these types of systems can be large, e.g in [12], long-range transport of air pollution models are described by a system of PDEs with $s = 30$*. That is,* $\mathbf{S}$ *is considerably simpler than* $\mathbf{B}$*.*

4 A numerical example

Let us solve the following PDEs numerically

$$\begin{cases} -\Delta u + \eta u = f_i, & \text{in } \Omega = [0,1]^2, \\ u|_{\partial\Omega} = 0 \end{cases} \tag{E_i}$$

with $i = 1, 2$. Here $\eta \in \mathrm{L}^{\frac{p}{p-2}}(\Omega)$ is defined as

$$\eta(x, y) = (x^2 + y^2)^{-\beta}, \qquad 0 < \beta < \frac{p-2}{p}$$

and

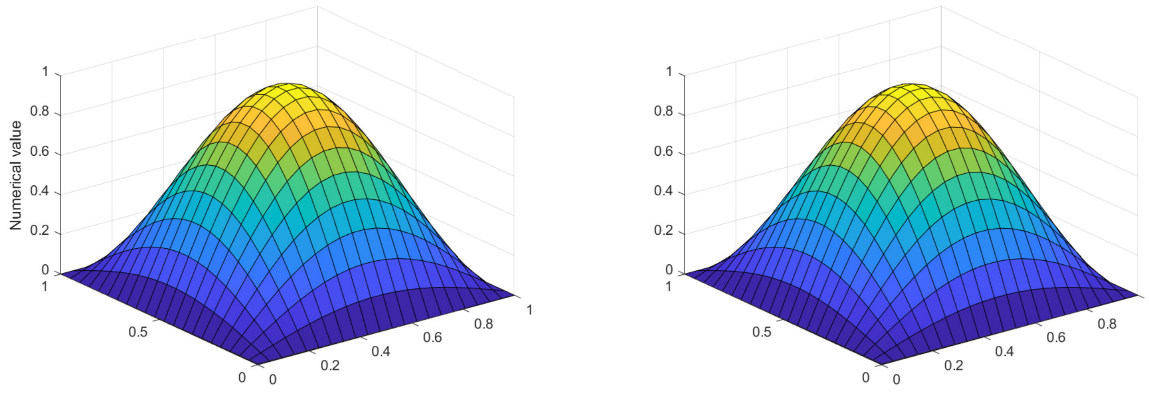
$$f_1(x, y) = 2\pi^2 \sin(\pi x)\sin(\pi y) + \eta(x, y)\sin(\pi x)\sin(\pi y),$$
$$f_2(x, y) = 1.$$

The exact solution of (E_i) with $i = 1$ is $u(x, y) = \sin(\pi x)\sin(\pi y)$.

Applying the finite element method to (E_i) with stepsize $h = 1/(N + 1)$ we obtain the algebraic system

$$(\mathbf{G}_h + \mathbf{D}_h)\mathbf{c}_i = \mathbf{g}_h^i, \quad i = 1, 2. \tag{E_i'}$$

Then, we apply $\mathbf{G}_h$ as a preconditioner and we solve the preconditioned system using the CGM.



(a) Numerical solution with $N = 20$.

(b) Exact solution

Figure 1: Graphics of the numerical and exact solution of (E_i) with $i = 1$ and $\beta = 1/4$.

To measure the error of the PCGM, we use the energy norm

$$\|e\|_{A_h} = \langle (G_h + D_h)e, e\rangle^{\frac{1}{2}} \quad e \in \mathbb{R}^d.$$

Table 1 shows the error and the residual obtained at each iteration k of the method applied to (E_i') for $i = 1, 2$ respectively. We see that it takes 7 steps to reach a $\mathcal{O}(10^{-14})$ error.

To test Theorem 2, note that $d = 2$ and so $\alpha = \frac{1}{p}$. Furthermore, recall that

$$\eta \in \mathrm{L}^{\frac{p}{p-2}}(\Omega) \quad \text{if } \beta < \frac{p-2}{p} = 1 - 2\alpha.$$

That is, if $p > \frac{2}{1-\beta}$, we get that the theorem holds when $\alpha < \frac{1-\beta}{2}$. Table 2 shows the values of

$$\delta_k = \left(\frac{\|e_k\|_{A_h}}{\|e_0\|_{A_h}}\right)^{\frac{1}{k}} k^{\alpha}, \quad \hat{\delta}_k = \left(\frac{\|r_k\|_{A_h}}{\|r_0\|_{A_h}}\right)^{\frac{1}{k}} k^{\alpha}$$

for different choices of β (and hence of α) with $i = 1, 2$ respectively, while fixing a mesh size. The value of δ_k corresponds to the system (E_i') with $i = 1$ and the value of $\hat{\delta}_k$ corresponds to (E_i') with $i = 2$. This demonstrates that (18) holds in these cases since the values of δ_k and $\hat{\delta}_k$ are bounded by a constant.

Finally, Table 3 shows the values of δ_k and $\hat{\delta}_k$ for different mesh sizes while fixing the values of β. Here we verify that the results of Theorem 2 are not sensitive to the size of the mesh.

Table 1: Error and residual obtained with PCGM applied to the system (E_i') with $N = 40$, $\beta = 1/2$.

	$\|u_k - c\|_{A_h}$	$\|r_k\|_{A_h}$
k	f_1	f_2
1	0.029824319963556	0.193919601149356
2	0.000187444098497	0.003450112947903
3	0.000000867801395	0.000042352080426
4	0.000000003253692	0.000000414294358
5	0.000000000011404	0.000000003016228
6	0.000000000000049	0.000000000016330
7	0.000000000000018	0.000000000000058

Table 2: Values of δ_k and $\hat{\delta}_k$ for different α's and β's for a fixed mesh size $N = 40$.

	$\beta = 1/4, \alpha = 0.374$		$\beta = 1/3, \alpha = 0.31$		$\beta = 1/2, \alpha = 0.24$		$\beta = 2/3, \alpha = 0.15$		$\beta = 3/4, \alpha = 0.12$	
k	f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2	f_1	f_2
1	0.0050	0.1383	0.0072	0.1414	0.0129	0.1575	0.0209	0.1768	0.0261	0.1904
2	0.0036	0.0622	0.0054	0.0660	0.0107	0.0784	0.0182	0.0917	0.0235	0.1019
3	0.0042	0.0475	0.0056	0.0478	0.0094	0.0533	0.0157	0.0616	0.0208	0.0697
4	0.0052	0.0342	0.0059	0.0347	0.0086	0.0404	0.0139	0.0473	0.0184	0.0538
5	0.0061	0.0291	0.0064	0.0292	0.0081	0.0322	0.0124	0.0370	0.0163	0.0430
6	0.0095	0.0255	0.0083	0.0248	0.0081	0.0259	0.0112	0.0311	0.0145	0.0360
7	0.0217	0.0230	0.0186	0.0221	0.0154	0.0224	0.0135	0.0263	0.0135	0.0299

Table 3: Values of δ_k for different mesh sizes with $\beta = 3/4, \alpha = 0.12$.

	$N = 20$		$N = 40$		$N = 80$	
k	f_1	f_2	f_1	f_2	f_1	f_2
1	0.0259	0.1892	0.0261	0.1904	0.0262	0.1907
2	0.0229	0.1005	0.0235	0.1019	0.0237	0.1023
3	0.0197	0.0678	0.0208	0.0697	0.0211	0.0702
4	0.0168	0.0515	0.0184	0.0538	0.0189	0.0546
5	0.0145	0.0401	0.0163	0.0430	0.0170	0.0440
6	0.0127	0.0331	0.0145	0.0360	0.0155	0.0372
7	0.0124	0.0272	0.0135	0.0299	0.0157	0.0315

5 Bibliography

References

[1] Owe Axelsson. *Iterative Solution Methods.* Cambridge University Press, 1994.

[2] Owe Axelsson and János Karátson. Mesh independent superlinear PCG rates via compact-equivalent operators. *SIAM Journal on Numerical Analysis*, 45(4):1495–1516, 2007.

[3] Owe Axelsson and János Karátson. Equivalent operator preconditioning for elliptic problems. *Numerical Algorithms*, 50(3):297–380, 2009.

[4] Haim Brézis. *Functional analysis, Sobolev spaces and partial differential equations*, volume 2. Springer, 2011.

[5] Gustavo Chávez, George Turkiyyah, Stefano Zampini, Hatem Ltaief, and David Keyes. Accelerated cyclic reduction: A distributed-memory fast solver for structured linear systems. *Parallel Computing*, 74:65–83, 2018.

[6] David E Edmunds and Hans Triebel. Entropy numbers and approximation numbers in function spacess. *Proceedings of the London Mathematical Society*, 3(1):137–152, 1989.

[7] Israel Gohberg, Seymour Goldberg, and Marinus A Kaashoek. Operator theory: Advances and applications. *Classes of Linear Operators*, 49, 1992.

[8] Janos Karatson. Mesh independent superlinear convergence estimates of the conjugate gradient method for some equivalent self-adjoint operators. *Applications of Mathematics*, 50(3):277–290, 2005.

[9] Tuomo Rossi and Jari Toivanen. A parallel fast direct solver for the discrete solution of separable elliptic equations. In *PPSC*. Citeseer, 1997.

[10] Youcef Saad and Martin H Schultz. Gmres: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7(3):856–869, 1986.

[11] Jan Vybíral. Widths of embeddings in function spaces. *Journal of Complexity*, 24(4):545–570, 2008.

[12] Zahari Zlatev. Numerical treatment of large air pollution models. In *Computer Treatment of Large Air Pollution Models*, pages 69–109. Springer, 1995.