Eötvös Loránd University

Directed studies 1

Superlinear convergence of the conjugate gradient method for elliptic partial differential equations with unbounded reaction coefficient

Author:

Sebastián Josue Castillo

Adviser:

Dr. Karátson János

Budapest - May 16, 2022

1 Summary

We consider a self-adjoint second order elliptic boundary value problem with variable zeroth order coefficient and its finite element discretization. In this project, we study the mesh-independent superlinear convergence of the preconditioned conjugate gradient method (CGM) for this type of problem. Our goal is to find an eigenvalue-based estimation of the rate of the superlinear convergence when the reaction coefficient of the elliptic boundary value problem belongs to a general Sobolev space. This work extends the results done in [1] where the coefficient was assumed to be continuous.

2 General framework

Let H be a 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 $B = S + Q$ where S is a self-adjoint 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* $\langle u, v\rangle_S = \langle Su, v\rangle$, and the corresponding norm is denoted by $\|\cdot\|_S$. Assumptions (ii) implies $H_S \subset H$. Moreover, assumptions $(i) - (ii)$ on S imply that $\mathrm{R}(S) = H$, hence $S^{-1}Q$ makes sense.

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

$$\langle u, v\rangle_S + \langle Qu, v\rangle = \langle g, v\rangle, \quad \forall v \in H_s. \tag{2}$$

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 (2).

Now equation (2) is solved numerically using a *Galerkin discretization.*

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

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

the *Gram matrices* corresponding to S and Q. We look for the numerical solution $u_V \in V$ of equation (2) in V, i.e., for which

$$\langle u_V, v\rangle_S + \langle Qu, v\rangle = \langle g, v\rangle, \quad \forall v \in V. \tag{3}$$

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

$$(\mathbf{S} + \mathbf{Q})\mathbf{c} = \mathbf{b} \tag{4}$$

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

By using matrix $\mathbf{S}$ as the preconditioner for the system (4), we shall work with the preconditioned system

$$(\mathbf{I} + \mathbf{S}^{-1}\mathbf{Q})\mathbf{c} = \tilde{\mathbf{b}}, \tag{5}$$

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

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 $|\lambda_1 - 1| \geq |\lambda_2 - 1| \geq \cdots \geq |\lambda_k - 1|$. Then $\lambda_j(\mathbf{E}) = \lambda_j - 1$ and the *error vectors* $e_k = c_k - c$ satisfy [2]

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A} \right)^{1/k} \leq \frac{2\|\mathbf{A}^{-1}\|}{k} \sum_{j=1}^{k} |\lambda_j(\mathbf{S}^{-1}\mathbf{Q})|, \quad k = 1, 2, \dots, n. \tag{6}$$

The next result allows us to give a convergence rate for the upper bound of (6) through the eigenvalues of the operator $Q_S = S^{-1}Q$.

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

$$\sum_{j=1}^{k} |\lambda_j(\mathbf{S}^{-1}\mathbf{Q})| \leq \sum_{j=1}^{k} \lambda_j(S^{-1}Q), \tag{7}$$

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

$$\mathbf{Q}\mathbf{c}^m = \lambda_m \mathbf{S}\mathbf{c} \tag{8}$$

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

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

where δ_{ml} is the Kronecker delta.

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

$$\langle u_m, u_l \rangle_S = \sum_{i,j=1}^{k} \langle \varphi_i, \varphi_j \rangle_S c_i^m c_j^l = \mathbf{S}c^m \cdot c^l, \tag{9}$$

hence (8) implies that $u_1, \dots, u_k$ form an orthonormal basis in $V \subset H_S$ with respect to the H_S-inner product. Then (8),(9) yield

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

Hence, we obtain

$$\langle Q_S u_m, u_l \rangle_S = \lambda_m \delta_{ml}, \quad m, l = 1, \ldots, k. \tag{10}$$

Using Corollary 3.3 of [3] and since $Q_S = S^{-1}Q$ is a compact self-adjoint operator on the Hilbert space H_S, we have that

$$\sum_{m=1}^{k} |\langle Q_S u_m, u_m \rangle_S| \le \sum_{m=1}^{k} s_j(Q_S) = \sum_{m=1}^{k} \lambda_j(Q_S), \tag{11}$$

where $s_j(S^{-1}Q)$ are the singular values of $S^{-1}Q$. Then, by (10) and (11) we arrive at

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

□

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

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

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A} \right)^{1/k} \le \frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} \lambda_j(S^{-1}Q), \quad k = 1, 2, \ldots, n. \tag{12}$$

Proof. By [4, Prop. 4.1] we are able to estimate $\|\mathbf{A}\|$ to obtain

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

This, together with the previous result and (6) completes the proof. □

Since $|\lambda_1(S^{-1}Q)| \ge |\lambda_2(S^{-1}Q)| \ge \cdots \ge 0$ and the eigenvalues tend to 0, the convergence factor is less than 1 for k sufficiently large. Hence the upper bound decreases as $k \to \infty$ and we obtain superlinear convergence rate.

3 Main result

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

$$\begin{cases} -\mathrm{div}(G\nabla u) + \eta u = g, \\ u_{\partial\Omega} = 0, \end{cases} \tag{13}$$

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

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

whereas $\eta = \eta(x)$ is a general variable (i.e. nonconstant) coefficient. Let problem (13) satisfy the following assumptions:

(i) The symmetric matrix-valued function $G \in \mathrm{C}^1(\overline{\Omega}, \mathbb{R}^N \times \mathbb{R}^N)$ satisfies

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

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

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

(iii) $\partial\Omega$ is piecewise C^2 and Ω is locally convex at the corners.

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

Then problem (13) has a unique weak solution in $\mathrm{H}^1_0(\Omega)$.

Let $V_h \subset \mathrm{H}^1_0(\Omega)$ be a given FEM subspace. We look for the numerical solution u_h of (13) in V_h:

$$\int_\Omega (G\nabla u_h \cdot \nabla v + du_h v) = \int_\Omega gv, \quad v \in V_h. \tag{14}$$

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 (14) is given by

$$(\mathbf{I}_h + \mathbf{G}_h^{-1}\mathbf{D}_h)\mathbf{c} = \tilde{\mathbf{g}}_h \tag{15}$$

with $\tilde{\mathbf{g}}_h = \mathbf{G}_h^{-1}\mathbf{g}_h$. Now, we apply the CGM for the system (15).

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

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{\frac{1}{k}} \leq \boldsymbol{C} k^{-\frac{1}{s}}, \tag{16}$$

where $\alpha = \frac{1}{N} - \frac{1}{2} + \frac{1}{p}$ and $s > \frac{1}{\alpha}$.

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

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

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

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

Then

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

where v is the Sobolev constant. By assumption (*iii*) the symmetric operator S maps onto $\mathrm{L}^2(\Omega)$. Furthermore,

$$\begin{aligned}\|Q_S v\|_{H_S} = \sup_{\|u\|_S=1} |\langle Q_S v, u\rangle_S| &= \sup_{\|u\|_S=1} \langle Qv, u\rangle \\ &= \sup_{\|u\|_S=1} \int_\Omega \eta v u \\ &\le \sup_{\|u\|_S=1} \left(\int_\Omega |\eta|^{\frac{p}{p-2}}\right)^{\frac{p-2}{p}} \left(\int_\Omega |v|^p\right)^{\frac{1}{p}} \left(\int_\Omega |u|^p\right)^{\frac{1}{p}} \\ &\le c \sup_{\|u\|_S=1} \|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)} \|v\|_{\mathrm{L}^p(\Omega)} \|u\|_S \\ &= C\|v\|_{\mathrm{L}^p(\Omega)},\end{aligned} \tag{17}$$

where $C = c\|\eta\|_{\mathrm{L}^{p/(p-2)}(\Omega)}$. Here we apply the extension of Hölder's inequality ([5, Th. 4.6]) with

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

Hence $Q_S = S^{-1}Q$ is compact and self-adjoint in $H_S = \mathrm{H}_0^1(\Omega)$ with $\langle u, v\rangle_S = \int_\Omega G\nabla u \cdot \nabla v$.

Let $\lambda_n = \lambda_n(S^{-1}Q)$. Since $S^{-1}Q$ is a compact self-adjoint operator in H_S, by [3, Ch.6, Th.1.5] we have the following characterization of the eigenvalues of Q_S:

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

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

$$\lambda_n(Q_S) \le \min\{\|Q_S - Q_S L_{n-1}\| \ / \ L_{n-1} \in \mathcal{L}(H_S), \mathrm{rank}(L_{n-1}) \le n-1\}. \tag{18}$$

Now, by (17) we get

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

where in the last step we use the inequality $\sqrt{m}\|u\|_{\mathrm{H}_0^1(\Omega)} \le \|u\|_{H_S}$. This, together with (18) yields

$$\lambda_n(Q_S) \le \frac{C}{\sqrt{m}} \min\{\|\mathcal{I} - L_{n-1}\| / \ L_{n-1} \in \mathcal{L}(\mathrm{H}_0^1(\Omega), \mathrm{L}^p(\Omega)), \mathrm{rank}(L_{n-1}) \le n-1\} := a_n(\mathcal{I}), \tag{19}$$

where $a_n(\mathcal{I})$ denotes the approximation numbers of the compact embedding $\mathcal{I}\colon \mathrm{H}_0^1(\boldsymbol{\Omega}) \mapsto \mathrm{L}^p(\boldsymbol{\Omega})$, [6]. Furthermore, we have the estimation [7]

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

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

$$s_n(Q_S) \le \frac{C\hat{C}}{\sqrt{m}} n^{-\alpha}.$$

Now, taking the arithmetic mean on both sides and by Hölder's inequality, we obtain

$$\frac{1}{k}\sum_{n=1}^{k} s_n(Q_S) \le \frac{C\hat{C}}{\sqrt{m}} \frac{1}{k} \left(\sum_{n=1}^{k} \frac{1}{n^{\alpha s}}\right)^{\frac{1}{s}} k^{\frac{1}{t}} = \frac{C\hat{C}}{\sqrt{m}} \left(\sum_{n=1}^{k} \frac{1}{n^{\alpha s}}\right)^{\frac{1}{s}} \frac{1}{k^{\frac{1}{s}}}, \tag{20}$$

where $\frac{1}{t} + \frac{1}{s} = 1$. Let $s\alpha > 1$, then we obtain

$$\frac{1}{k}\sum_{n=1}^{k} s_n(Q_S) \le \frac{C\hat{C}}{\sqrt{m}} \left(\sum_{n=1}^{\infty} \frac{1}{n^{s\alpha}}\right)^{\frac{1}{s}} \frac{1}{k^{\frac{1}{s}}} = \frac{\boldsymbol{C}}{k^{\frac{1}{s}}}.$$

Then, by (12), we conclude. $\square$

4 Bibliography

References

[1] J. Karatson, “Mesh independent superlinear convergence estimates of the conjugate gradient method for some equivalent self-adjoint operators,” *Applications of Mathematics*, vol. 50, no. 3, pp. 277–290, 2005.

[2] O. Axelsson and J. Karátson, “Equivalent operator preconditioning for elliptic problems,” *Numerical Algorithms*, vol. 50, no. 3, pp. 297–380, 2009.

[3] I. Gohberg, S. Goldberg, and M. A. Kaashoek, “Operator theory: Advances and applications,” *Classes of Linear Operators*, vol. 49, 1992.

[4] O. Axelsson and J. Karátson, “Mesh independent superlinear PCG rates via compact-equivalent operators,” *SIAM Journal on Numerical Analysis*, vol. 45, no. 4, pp. 1495–1516, 2007.

[5] H. Brezis and H. Brézis, *Functional analysis, Sobolev spaces and partial differential equations*. Springer, 2011, vol. 2, no. 3.

[6] J. Vybíral, “Widths of embeddings in function spaces,” *Journal of Complexity*, vol. 24, no. 4, pp. 545–570, 2008.

[7] D. E. Edmunds and H. Triebel, “Entropy numbers and approximation numbers in function spacess,” *Proceedings of the London Mathematical Society*, vol. 3, no. 1, pp. 137–152, 1989.