Equivalent operator preconditioning for linear elliptic problems

by O. Axelsson[1], J. Karátson[2]

Abstract

The numerical solution of linear elliptic partial differential equations most often involves a finite element or finite difference discretization. To preserve sparsity, the arising system is normally solved using an iterative solution method, commonly a preconditioned conjugate gradient method. Preconditioning is a crucial part of such a solution process. It is desirable that the total computational cost will be optimal, i.e. proportional to the degrees of freedom of the approximation used, which also includes mesh independent convergence of the iteration. This paper surveys the equivalent operator approach, which has proven to provide an efficient general framework to construct such preconditioners. Hereby one first approximates the given differential operator by some simpler differential operator, and then one chooses as preconditioner the discretization of this operator for the same mesh. In this survey we give a uniform presentation of this approach, including theoretical foundation and several practically important applications.

Keywords: elliptic problem, conjugate gradient method, preconditioning, equivalent operator

1 Introduction

Finite element and finite difference approximations of linear partial differential equations of elliptic type lead to algebraic systems, normally of very large size. However, since the matrices are sparse, even extremely large-sized systems can be handled in a reasonable computing time if proper solution methods are used. To save computer memory and elapsed time, such equations are normally solved by iteration, most commonly using a preconditioned conjugate gradient (PCG) method. Preconditioning techniques have emerged as an essential part of efficient solution of linear systems of equations. Preconditioning can be described as a transformation of the original system, aiming at accelerating the convergence of iterative methods. As is well-known (see, e.g., [4]), the rate of convergence depends in general on the condition number of the preconditioned operator or of its discrepancy from the identity operator.

When solving discretized partial differential equations, the rate of convergence should not slow down when the mesh is refined and hence a larger system must be solved. This is not a trivial task to achieve as the condition number of the given finite element (or finite difference) matrix increases rapidly as the mesh is refined. At the same time, the computational labour required to solve the preconditioning system should be proportional to the degrees of freedom (d.o.f.) of the approximation used. If both these requirements, i.e. a bounded number of iterations and cost proportional to the d.o.f. are fulfilled, then the total computational cost will be optimal, i.e. proportional to the d.o.f. as well.

The goal of this paper is to survey a general framework to construct such preconditioners. The main idea is as follows: instead of constructing the preconditioner directly for the given finite element (FE) or finite difference (FD) matrix, it can be more efficient to first approximate the given differential operator by some simpler differential operator, and then to use the FE or

[1] Department of Information Technology, Uppsala University, Sweden & Institute of Geonics AS CR, Ostrava, Czech Republic; owea@it.uu.se

[2] Department of Applied Analysis, ELTE University, H-1117 Budapest, Hungary; karatson@cs.elte.hu

FD matrix of this operator as preconditioner, hereby using the same discretization mesh as for the original operator. This idea can be described formally as follows. Let

$$Lu = g \tag{1.1}$$

be a given elliptic boundary value problem, let

$$L_h u_h = g_h \tag{1.2}$$

be a suitable finite dimensional discretization of (1.1), where one wishes to solve (1.2) with some preconditioned iterative method. Then one can take *another elliptic operator* S and propose its discretization S_h as preconditioner for (1.2):

$$S_h^{-1} L_h u_h = S_h^{-1} g_h\,. \tag{1.3}$$

This approach has been developed in a large number of works, of which we first mention here the early papers [42, 58] and later e.g. [37, 45, 77, 100]. Numerous papers will be cited in this survey, dealing with second order elliptic problems in the above form. A general theory for such preconditioning has been developed in [49], see also [52, 78], where the notion of *equivalent operator* has been introduced and used to give a rigorous Hilbert space background for the convergence properties. A similar preconditioning approach is used in [75, 92] where preconditioning by S is replaced by a related solution operator, for first order problems in [32] and for second order nonlinear problems in e.g. [84] and the authors' works [9, 50, 68].

To obtain favourable preconditioners in the above way, one must satisfy the previously mentioned two requirements for the preconditioning matrix. First, solving problems with S_h should be considerably simpler than those with L_h. This can be fulfilled by various choices of particular elliptic operators S, for instance, if in contrast to L:

- S is a symmetric operator
- S_h is an M-matrix or is diagonally dominant
- S_h has a favourable block structure
- S_h has a better sparsity pattern.

On the other hand, the conditioning of $S_h^{-1}L_h$ should be considerably better than the conditioning of L_h, in fact, the rate of convergence must not slow down as the mesh is refined. When preconditioning by an elliptic operator, one can indeed obtain mesh independent spectral relations between the given and the approximate FE or FD matrices, which implies that the preconditioned conjugate gradient method converges with a h-independent rate. For the study of linear convergence, a natural framework to characterize the relation between the given and the approximate operator has proved to be that of equivalent operators, developed rigorously in [49], see also [52, 77, 78]. The main idea for (1.3) is that, under proper assumptions, the condition number $\kappa(S_h^{-1}L_h)$ is bounded as $h \to 0$ (in contrast to $\kappa(L_h)$ which tends to ∞), because $\kappa(S_h^{-1}L_h)$ approaches, roughly speaking, $\kappa(S^{-1}L)$ as $h \to 0$: moreover, for FEM discretizations we usually have

$$\kappa(S_h^{-1}L_h) \le \kappa(S^{-1}L).$$

This is one of the major topics of this paper, summarized as follows: if the two operators (the original and preconditioner) are equivalent then the corresponding PCG method provides mesh independent linear convergence.

The property of equivalent operators can be refined to provide pairs of operators which are so-called compact-equivalent, i.e. for which the corresponding preconditioned operator is a compact perturbation of the identity operator. In this case the corresponding preconditioner is such that the PCG method converges with a mesh independent superlinear rate [14, 16], which normally means that much fewer iterations will be needed to achieve an increased relative accuracy, that is, loosely speaking, each additional correct digit in the approximate solution requires fewer iterations than the previous digit. To a great deal the paper will also deal with the properties and construction of compact-equivalent operators, which have their main importance when solving nonsymmetric problems. Summing up this second major topic: if the two operators (the original and preconditioner) are compact-equivalent then the corresponding PCG method provides mesh independent superlinear convergence.

In this paper we give a uniform presentation of the above topics and provide several practically important applications. The preconditioner must be such that the resulting linear algebraic systems can be solved with relatively little computational effort and demand of computer storage. We give several examples when this can occur. Both symmetric (self-adjoint) and nonsymmetric (non-self-adjoint) operators will be dealt with, therefore some properties of both the classical and the generalized conjugate gradient methods will be presented as well. Inner-outer iterations, i.e. solving the preconditioning systems themselves also by iteration, are discussed as well. Variational formulations which lead to matrices in saddle-point form will also be treated.

The paper is organized as follows. Classical and generalized conjugate gradient methods are presented in section 2. Hilbert space background is summarized in sections 3-4: equivalent operators and linear convergence properties are discussed in section 3, whereas compact-equivalent operators and superlinear convergence properties are presented in section 4. The further sections are devoted to applications to various classes of problems. Symmetric equations and systems are discussed in sections 5 and 6, respectively. The next three sections deal with nonsymmetric problems. Symmetric preconditioners for nonsymmetric equations and systems are discussed in sections 7 and 8, respectively, and some nonsymmetric preconditioners are presented in section 9. Finally some comments on inner-outer iterations are enclosed in section 10.

2 Conjugate gradient algorithms and their rate of convergence

In this section we briefly summarize well-known facts about some important conjugate gradient (CG) algorithms that we consider. For detailed discussions on CG methods, see e.g. [2, 4, 48, 56, 90, 101].

Let us consider a linear system

$$Au = b \tag{2.1}$$

with a given nonsingular matrix $A \in \mathbf{R}^{n\times n}$, $f \in \mathbf{R}^n$ and solution u. Letting $\langle .,. \rangle$ be a given inner product on $\mathbf{R}^n$ and denoting by A^* the adjoint of A w.r.t. this inner product, in what follows, we assume that

$$A + A^* > 0, \tag{2.2}$$

i.e., A is positive definite w.r.t. $\langle .,. \rangle$. We define the following quantities, to be used frequently in the study of convergence:

$$\lambda_0 := \lambda_0(A) := \inf\{\langle Ax, x\rangle : \ \|x\| = 1\} > 0, \qquad \Lambda := \Lambda(A) := \|A\|, \tag{2.3}$$

where $\|.\|$ denotes the norm induced by the inner product $\langle .,. \rangle$.

2.1 Self-adjoint matrices: the standard CG method

If A in (2.1) is self-adjoint, i.e. $\langle Au, v\rangle = \langle u, Av\rangle$ for all $u, v \in \mathbf{R}^n$, then the standard CG method reads as follows [4, 98]: let $u_0 \in \mathbf{R}^n$ be arbitrary, $d_0 := r_0$; for given u_k and d_k, with residuals $r_k := Au_k - b$, we let

$$u_{k+1} = u_k + \alpha_k d_k, \text{ where } \alpha_k = -\frac{\langle r_k, d_k\rangle}{\langle Ad_k, d_k\rangle}\,; \quad d_{k+1} = r_{k+1} + \beta_k d_k, \text{ where } \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}. \tag{2.4}$$

To save computational time, normally the residual vectors are also formed by recursion:

$$r_{k+1} = r_k + \alpha_k A d_k. \tag{2.5}$$

Further, we have $\langle r_k, d_k\rangle = -\|r_k\|^2$, hence for α_k in (2.4) we use $\alpha_k = \|r_k\|^2/\langle Ad_k, d_k\rangle$. In the study of convergence, one considers the error vector

$$e_k = u - u_k \tag{2.6}$$

and is generally interested in its energy norm

$$\|e_k\|_A = \langle Ae_k, e_k\rangle^{1/2}. \tag{2.7}$$

Below we briefly summarize the minimax property of the CG method and the derivation of two convergence estimates, based on [4].

2.1.1 Optimality and minimax properties of the CG method

Let $S_k := span\{r_0, Ar_0, \dots, A^k r_0\}$, the so-called Krylov subspace, where r_0 is the initial residual. It is seen from algorithm (2.4) that $e_k \in e_0 + S_k = \{e_0 + h : \; h \in S_k\}$. By construction, the method is optimal in the sense that the error vector is smallest in energy norm on $e_0 + S_k$, i.e. $\|e_k\|_A \le \|e\|_A$ for all $e \in e_0 + S_k$. Note that $r_k = -Ae_k$ implies $\|e_k\|_A = \|r_k\|_{A^{-1}}$, so the residuals are also minimized but in the norm $\|.\|_{A^{-1}}$.

One normally considers a preconditioned system

$$C^{-1}Au = C^{-1}b$$

or $Bu = f$, where $B = C^{-1}A$ and $f = C^{-1}b$. The preconditioner C is assumed to be symmetric and positive definite. Here C is some approximation of A which normally reduces the condition number or makes the spectrum $\sigma(C^{-1}A)$ more suitable for faster convergence of the CG algorithm, see below for details. In addition to the two inner products and three linear updates, each iteration step of the PCG method requires a solution of a linear system with the preconditioning matrix. Let the inner product $\langle x, y\rangle_C = \langle Cx, y\rangle$ be defined by C and consider algorithm (2.4) with A replaced by B and r_k by the pseudo-residuals $h_k = C^{-1}r_k = Bu_k - f$. It is readily seen that B is self-adjoint w.r.t. the inner product $\langle ., .\rangle_C$. Further,

$$f(h) := \langle B^{-1}h, h\rangle_C = \langle CA^{-1}Ch, h\rangle = \langle A^{-1}r, r\rangle = \|r\|^2_{A^{-1}}\,,$$

where $r = Ch = Au - b$. Hence it follows that the preconditioned algorithm minimizes the same functional $\|r\|^2_{A^{-1}}$ as for the unpreconditioned method, but on the Krylov subspace $\{r^0, AC^{-1}r^0, \dots, \, (AC^{-1})^k r^0\}$. By proper preconditioning, this sequence allows a sufficiently small error of the approximation of e_0 using much fewer vectors, i.e. much fewer iterations, than are required for the unpreconditioned method.

Since the Krylov basis vectors are linearly independent, the algorithm terminates with residual zero in at most n steps, where n is the order of A. However, we are interested in stopping the iterations much earlier, when a sufficient accuracy has been achieved. We need then information about the rate of convergence of the method. For this purpose we first note that the construction of the algorithm implies $e_k = P_k(A)e_0$ with some $P_k \in \pi_k^1$, where π_k^1 denotes the set of polynomials of degree k, normalized at the origin. Moreover, the optimality property shows that

$$\|e_k\|_A = \min_{P_k \in \pi_k^1} \|P_k(A)e_0\|_A \,. \tag{2.8}$$

Let $\{\lambda_i, v_i\}_1^n$ be the eigensolutions of A, where $0 < \lambda_1 \leq \ldots \leq \lambda_n$, and the eigenvectors have been taken to be orthonormal, i.e. $(v_i, v_j) = \delta_{ij}$. Let $e_0 = \sum_1^n c_i v_i$ where $c_i = (e_0, v_i)$. Then $\|P_k(A)e_0\|_A^2 = \| \sum_1^n c_i P_k(\lambda_i) v_i\|_A^2 = \sum_1^n c_i^2 P_k(\lambda_i)^2$, hence (2.8) yields the following upper bound:

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \min_{P_k \in \pi_k^1} \max_{\lambda \in \sigma(A)\}} |P_k(\lambda)| \,, \tag{2.9}$$

which is a basis for the convergence estimates of the CG method.

2.1.2 Linear convergence

If we only use the fact $\sigma(A) \subset [\lambda_1, \lambda_n]$, where $\lambda_1 = \lambda_{min}(A)$ and $\lambda_n = \lambda_{max}(A)$, then, using Chebyshev polynomials, one obtains

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \max_{\lambda_1 \leq \lambda \leq \lambda_n} \frac{T_k[(\lambda_n + \lambda_1 - 2\lambda)/(\lambda_n - \lambda_1)]}{T_k(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)} = \{T_k[(\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)]\}^{-1}$$

where T_k denotes the Chebyshev polynomial of degree k of the first kind. As is well-known,

$$T_k(x) = \frac{1}{2}[(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k].$$

Using elementary estimates, we then obtain the linear convergence estimate, see e.g. [4]:

Theorem 2.1.1 *We have*

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq 2^{1/k} \, \frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}} = 2^{1/k} \, \frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1} \qquad (k = 1, 2, ..., n), \tag{2.10}$$

where $\kappa(A) = \lambda_n / \lambda_1$ is the standard spectral condition number.

Note that for $\varepsilon > 0$, by a simple derivation from (2.10),

$$\frac{\|e_k\|_A}{\|e_0\|_A} \leq \varepsilon \quad \text{when} \;\; k = \lfloor \frac{1}{2} \sqrt{\kappa(A)} \ln \frac{2}{\varepsilon} + 1 \rfloor.$$

Therefore the upper bound of the number of iterations is proportional to both the square root of the condition number and the number of significant digits in the computed approximate solution.

2.1.3 Superlinear convergence

To show superlinear convergence rates, another useful estimate is derived if we consider the decomposition

$$A = I + E \tag{2.11}$$

or, more generally, for the preconditioned version, $B = C^{-1}A = I + E$, where $E = C^{-1}(A - C)$. Here we choose $P_k(\lambda) := \prod_{j=1}^{k}\left(1 - \frac{\lambda}{\lambda_j}\right)$ in (2.9), where $\lambda_j := \lambda_j(A)$ are ordered according to $|\lambda_1 - 1| \geq |\lambda_2 - 1| \geq ... \geq |\lambda_k - 1|$. Here $P_k(\lambda_i) = 0 \quad (i = 1, \ldots, k)$, which implies

$$\max_{\lambda\in\sigma(A)} |P_k(\lambda)| = \max_{i\geq k+1} |P_k(\lambda_i)| = \max_{i\geq k+1} \prod_{j=1}^{k} \frac{|\mu_j - \mu_i|}{\lambda_j} \leq \left(2\|A^{-1}\|\right)^k \prod_{j=1}^{k} |\mu_j| \tag{2.12}$$

where $\mu_j = \lambda_j - 1$, using that $|\mu_j - \mu_i| \leq 2|\mu_j| \quad (i \geq k+1, \quad 1 \leq j \leq k)$ and $\frac{1}{\lambda_j} \leq \|A^{-1}\|$. Since $\lambda_j(E) = \mu_j$ and using (2.9), (2.12) and the arithmetic-geometric means inequality, we finally find that

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2\|A^{-1}\|}{k} \sum_{j=1}^{k} |\lambda_j(E)| \qquad (k = 1, 2, ..., n). \tag{2.13}$$

Here by assumption $|\lambda_1(E)| \geq |\lambda_2(E)| \geq ... \geq 0$. If these eigenvalues accumulate in zero then the convergence factor is less than 1 for k sufficiently large. Moreover, the upper bound decreases, i.e. we obtain a superlinear convergence rate. For the preconditioned method, A in (2.13) is replaced by $B = C^{-1}A$ and E by $C^{-1}(A - C)$.

Remark 2.1.1 For the diagonally compensated reduction method [17] and modified preconditioning methods [60], the factor $\|B^{-1}\| = \|A^{-1}C\|$ does not take large values and is frequently even close to 1.

A magnitude $O(k^{-1/2})$ of the superlinear rate is obtained if the arithmetic mean in (2.13) is replaced by the mean of squares. Using the Frobenius norm $\|E\|_F^2 := \sum_{i=1}^{k} \lambda_i(E)^2$, one then obtains

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \frac{2\,\|E\|_F}{\lambda_0\,\sqrt{k}} \qquad (k = 1, 2, ..., n). \tag{2.14}$$

Another approach for superlinear convergence is based on the K-condition number

$$K(A) = \left(\frac{1}{k}\text{trace}(A)\right)^k / \det(A) = \left(\frac{1}{k}\sum_{i=1}^{k} \lambda_i(A)\right)^k \left(\prod_{i=1}^{k} \lambda_i(A)\right)^{-1}, \tag{2.15}$$

see e.g. [4]. It is proved in [11] that if $k \in \mathbf{N}$ is even and $k \geq 3\ln K(A)$, then

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \left(\frac{3\ln K(A)}{k}\right)^{1/2}. \tag{2.16}$$

K-condition numbers can be related to Frobenius norms via the decomposition (2.11). If, for simplicity, E is positive semidefinite, then by [65],

$$\ln K(A) \leq \frac{1}{2}\|E\|_F^2\,. \tag{2.17}$$

Then (2.16) becomes

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \le \left(\frac{3}{2k}\right)^{1/2} \|E\|_F\,. \tag{2.18}$$

Since E is positive semidefinite, we have $\lambda_0 = 1$, hence the multiplier 2 of $\|E\|_F/\sqrt{k}$ in (2.14) has been reduced to $\sqrt{3/2}$.

2.2 Nonsymmetric systems of algebraic equations

2.2.1 The generalized conjugate gradient–least square (GCG-LS) and related methods

For nonsymmetric matrices A, several CG algorithms exist (see e.g. [2, 4, 44]). First we discuss the approach that generalizes the minimization property (2.8) for nonsymmetric A but avoids the normal equation (2.34). A general form of the algorithm, which uses least-square residual minimization of $f(u) = \|Au - b\|^2$, is the *generalized conjugate gradient–least square method* (GCG-LS method) [3, 4]. Its full version uses all previous search directions. The general method involves an integer $s \in \mathbf{N}$, which limits the number of search directions used, by letting $s_k = \min\{k, s\}$ $(k \ge 0)$. Then the algorithm is as follows: let $u_0 \in \mathbf{R}^n$ be arbitrary, $d_0 := Au_0 - b$; for given u_k and d_k, with $r_k := Au_k - b$, we let

$$\left\{\begin{array}{l} u_{k+1} = u_k + \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} d_{k-j} \text{ and } d_{k+1} = r_{k+1} + \sum\limits_{j=0}^{s_k} \beta_{k-j}^{(k)} d_{k-j}, \\ \quad \text{where } \beta_{k-j}^{(k)} = -\langle Ar_{k+1}, Ad_{k-j}\rangle / \|Ad_{k-j}\|^2 \quad (j = 0, \dots, s_k) \\ \quad \text{and } \alpha_{k-j}^{(k)} \quad (j = 0, \dots, s_k) \quad \text{are the solution of the linear system} \\ \qquad \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} \langle Ad_{k-j}, Ad_{k-l}\rangle = -\langle r_k, Ad_{k-l}\rangle \qquad (0 \le l \le s_k). \end{array}\right. \tag{2.19}$$

The full version of the algorithm corresponds to formally setting $s = \infty$, and for a finite s we get a truncated version called GCG-LS(s). There are other, similar versions of such methods, such as Orthomin(s), GMRES, GCR(s), see e.g. [4, 44]. The case $s = 0$ is of particular interest as it only involves a single, namely the current search direction, and takes a similar form as the standard CG method: that is, (2.19) is replaced by

$$u_{k+1} = u_k + \alpha_k d_k, \text{ where } \alpha_k = -\frac{\langle r_k, Ad_k\rangle}{\|Ad_k\|^2}\,; \quad d_{k+1} = r_{k+1} + \beta_k d_k, \text{ where } \beta_k = -\frac{\langle Ar_{k+1}, Ad_k\rangle}{\|Ad_k\|^2}. \tag{2.20}$$

Algorithm (2.19) is developed in detail in [3]. The search directions d_j are constructed to make the vectors Ad_j orthogonal: $\langle Ad_{k+1}, Ad_{k-j}\rangle = 0$ for all $j = 0, \dots, s_k$. For $s \ge 1$, the computation of the coefficients $\alpha_{k-j}^{(k)}$ involves the solution of the linear system in the last row of (2.19), written briefly as $K^{(k)}\underline{\alpha}_k = \underline{\gamma}_k$. Here all coordinates but the first of the vector $\underline{\gamma}_k$ vanish, since the minimization property implies $\langle r_k, Ad_{k-l}\rangle = \langle Au_k - b, Ad_{k-l}\rangle = 0$ for all $l = 1, \dots, s_k$. The matrices $K^{(k)}$ are symmetric positive semidefinite. Further, under our assumption (2.2), which amounts to

$$MA + A^T M > 0 \tag{2.21}$$

when the inner product $\langle .,.\rangle$ is generated by a SPD matrix M (as in the setting of [3]), it is proved [3, 4] that $K^{(k)}$ is nonsingular if $r_k \ne 0$, hence there is no breakdown in the algorithm.

The residuals $r_{k+1} = Au_{k+1} - b$ in (2.19) can alternatively be computed by the same recursion as for u_k, i.e., $r_{k+1} = r_k + \sum_{j=0}^{s_k} \alpha_{k-j}^{(k)} Ad_{k-j}$. Unless A is very sparse, this can be computationally cheaper. Further, the matrix $K^{(k+1)}$ equals $K^{(k)}$ augmented by a row and a column (and if, due to $k > s$, a truncation takes place, then the oldest row and a column of $K^{(k)}$ are deleted). Hence no additional matrix-vector products need to be computed to form $K^{(k+1)}$, but only to form Ad_k, which is anyway needed for the update of the recursion for r_{k+1}.

Now we cite some truncation and convergence properties of the GCG-LS algorithm based on [3].

2.2.1.1. Automatic truncation. We give conditions for which the GCG-LS method truncates to short-term recurrence relations. Note first that by construction, $d_k \in V_k := span\{r_0, Ar_0, \dots, A^k r_0\}$ and $r_k \in r_0 + AV_{k-1}$, hence there is a polynomial g_{k-1} of degree $k-1$ such that $r_k = \Big(I - Ag_{k-1}(A)\Big) r_0$ $(k = 1, 2, \dots)$. We may view the matrix $g_{k-1}(A)$ as a polynomial approximation of A^{-1}. Here $g_{k-1}(A)$ depends not only on A and r_0 but also on the choice of inner product $\langle ., . \rangle$. Let the inner product be generated by some SPD matrix H. Then the adjoint of A w.r.t. this inner product (so-called H-adjoint of A) is $A^* = H^{-1}A^T H$. By [3], the following are equivalent:

(a) A is H-normal, i.e. $A^* = H^{-1}A^T H$ commutes with A.

(b) A is diagonalizable by a similarity transformation.

(c) $A^* = p(A)$ for some polynomial p.

If this holds then the smallest degree of a polynomial p for which $A^* = p(A)$ is called H-normal degree of A, and is denoted by $n(A, H)$. The following result holds, assuming exact arithmetic:

Theorem 2.2.1 [4, Th. 12.12]. *Let (2.21) hold and A be M-normal. Then the truncated GCG-LS(s) algorithm coincides with the full version if* $s \geq n(A, M) - 1$.

As a first example, we can recover the case of standard CG for symmetric matrices: let us consider the preconditioned matrix $B = C^{-1}A$ for some SPD matrices A and B. Then $C^{-1}B^T C = C^{-1}A = B$, i.e. B is self-adjoint w.r.t. the C-inner product. Hence B is C-normal with degree 1. By Theorem 2.2.1, the full GCG-LS method for the matrix B reduces to the truncated GCG-LS(0) algorithm. The latter can be rewritten to the standard PCG iteration with preconditioner C.

Second, let us consider the symmetric and antisymmetric parts $M := (A + A^T)/2$ and $N := (A - A^T)/2$, respectively, of a matrix A. Assume that M is SPD and use it as a preconditioner to A, i.e. $B = M^{-1}A = I + M^{-1}N$. Then B is M-normal w.r.t. the M-inner product with degree 1, namely, $B^* = M^{-1}B^T M = M^{-1}(I - NM^{-1})M = I - M^{-1}N = 2I - B$. Hence the full GCG-LS method for the matrix B reduces again to the truncated GCG-LS(0) algorithm. This method is closely related to the so-called CGW method [38, 100].

2.2.1.2. Convergence. The convergence estimates in the nonsymmetric case most often involve the residual

$$r_k = Ae_k = Au_k - b. \tag{2.22}$$

Then we have the following monotone convergence result:

Theorem 2.2.2 [3, 4]. *Let (2.2) hold. Then the GCG-LS(s) method satisfies*

(1) $\|r_{k+1}\| < \|r_k\|$ *unless* $r_k = 0$*, and the rate equals*

$$\|r_{k+1}\|^2 = \|r_k\|^2 - det(K^{(k)})^{-1} det(K_0^{(k)}) \langle Ar_k, r_k \rangle^2$$

where $K_0^{(k)}$ *is the first principal minor of* $K^{(k)}$*.*

(2) $\|r_{k+1}\|^2 = \|r_k\|^2 - \langle Ar_k, r_k\rangle^2 / \min\limits_{g \in W_{k-1}} \|Ar_k - g\|^2$*, where* $W_{k-1} = span\{Ad_{k-1}, \ldots, Ad_{k-s_k}\}$*.*

The second statement implies the practically useful estimate

$$\|r_{k+1}\|^2 \le \|r_k\|^2 - \frac{\langle Ar_k, r_k\rangle^2}{\|Ar_k\|^2} \ , \tag{2.23}$$

from which, using that (2.3) yields $\lambda_0 \|Ar_k\| \, \|r_k\| \le \lambda_0 \Lambda \, \|r_k\|^2 \le \Lambda \langle Ar_k, r_k\rangle$, we obtain

Corollary 2.2.1 *If (2.2)-(2.3) hold, then*

$$\|r_{k+1}\| \le \left(1 - \left(\frac{\lambda_0}{\Lambda}\right)^2\right)^{1/2} \|r_k\| \qquad (k = 1, 2, ..., n). \tag{2.24}$$

The same estimate holds for the GCR and Orthomin methods together with their truncated versions, see [44].

An important occurrence of the truncated GCG-LS(0) algorithm (2.20) arises when the decomposition

$$A = I + E \tag{2.25}$$

holds for some antisymmetric matrix E, which most often comes from symmetric part preconditioning. As discussed after Theorem 2.2.1, in this case $A^* = 2I - A$ and hence Theorem 2.2.1 is valid with $s = 0$ [3]. The convergence of this iteration is determined by E. A straightforward estimate for this is derived directly from (2.24) as follows, cf. [15, Th. 2.1]. Relation (2.25) implies $A\mathbf{c} \cdot \mathbf{c} = \|\mathbf{c}\|^2$ for all $\mathbf{c}$, hence $\lambda_0 = 1$. Since A is normal and E has imaginary eigenvalues, we have $\Lambda^2 = \|A\|^2 = |\lambda_{max}(A)|^2 = 1 + |\lambda_{max}(E)|^2 = 1 + \|E\|^2$. That is, $1 - (\lambda_0/\Lambda)^2 = \|E\|^2/(1 + \|E\|^2)$, hence (2.24) yields that the GCG-LS(0) algorithm (2.20) converges as

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{\|E\|}{\sqrt{1 + \|E\|^2}} \qquad (k = 1, 2, ..., n). \tag{2.26}$$

The optimal estimate, obtained via Chebyshev polynomials, is somewhat better, see [44, 100]:

$$\frac{\|r_k\|}{\|r_0\|} \le \varrho_k := 2 \, \frac{\|E\|^k (1 + \sqrt{1 + \|E\|^2})^k}{(1 + \sqrt{1 + \|E\|^2})^{2k} + \|E\|^{2k}} \, , \tag{2.27}$$

which asymptotically satisfies

$$\limsup \varrho_k^{1/k} \le \frac{\|E\|}{1 + \sqrt{1 + \|E\|^2}} \, . \tag{2.28}$$

Concerning the minimization property (2.8), its analogue for the GCG-LS method holds for the residuals (see [3]):

$$\|r_k\| = \min_{P_k \in \pi_k^1} \|P_k(A) r_0\| \, . \tag{2.29}$$

Let $S^{-1}AS = blockdiag(J_1, \dots, J_q)$ be the Jordan canonical form of A. Then, following [4, Ch. 13],

$$\frac{\|r_k\|}{\|r_0\|} \le \min_{P_k \in \pi_k^1} \|P_k(A)\| \le \kappa(S) \min_{P_k \in \pi_k^1} \max_i \|P_k(J_i)\| \le \kappa(S) \min_{P_k \in \pi_k^1} \max_i \sum_{j=0}^{t_i - 1} \frac{1}{j!} |P_k^{(j)}(\lambda_i)| , \quad (2.30)$$

where t_i is the maximal order of a Jordan block corresponding to the eigenvalue λ_i. To derive useful estimates, one now needs more information on the spectrum $\sigma(A)$ than the bounds (2.3). On the one hand, one can include the spectrum in suitable sets like ellipses or discs, see, e.g., [4, Section 5.4.1]. Let, for instance, $\sigma(A) \subset D(\alpha, r)$ where the latter denotes the closed disc in the complex plane with center $\alpha > 0$ and radius $r > 0$. We assume $r < \alpha$, i.e. that the disc lies in the right half-plane. Then we can take the polynomial $P_k(\lambda) := \frac{(\alpha-\lambda)^k}{\alpha^k}$ from π_k^1, for which an elementary calculation yields

$$\frac{1}{j!} |P_k^{(j)}(\lambda)| = \binom{k}{j} \frac{|\alpha - \lambda|^{k-j}}{\alpha^k} \le \frac{k^{t-1}}{r^j} \left(\frac{r}{\alpha}\right)^k ,$$

where we have used $|\alpha - \lambda| \le r$ and $1 \le j \le t_i - 1 \le t - 1$, where $t = \max_i t_i$. Then (2.30) yields

$$\frac{\|r_k\|}{\|r_0\|} \le C_1 \, k^{t-1} \left(\frac{r}{\alpha}\right)^k \qquad (k = 1, 2, ..., n), \quad (2.31)$$

where $C_1 := \kappa(S)(\sum_{j=0}^{t_i - 1} \frac{1}{r^j})$ is independent of k, i.e. the asymptotic convergence factor is

$$\limsup \left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{r}{\alpha} , \quad (2.32)$$

see [76].

On the other hand, if A is normal and we have the decomposition (2.25), then the residual errors satisfy a similar estimate to (2.13) obtained in the symmetric case [4]. In fact, we now have $\kappa(S) = 1$ and $t_i = 1$ in (2.30), hence we can use the same polynomial $P_k(\lambda) := \prod_{j=1}^{k} \left(1 - \frac{\lambda}{\lambda_j}\right)$, where $\lambda_j = \lambda_j(A)$, and the same estimations as in (2.12)-(2.13) to obtain

Corollary 2.2.2 *If (2.2) and (2.25) hold, then*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{2}{k\lambda_0} \sum_{j=1}^{k} |\lambda_j(E)| \qquad (k = 1, 2, ..., n). \quad (2.33)$$

Again, this shows superlinear convergence if the eigenvalues $|\lambda_1(E)| \ge |\lambda_2(E)| \ge ...$ accumulate in zero.

If A is non-normal, then we can take $P_k(\lambda) := \prod_{j=1}^{q} \left(1 - \frac{\lambda}{\lambda_j}\right)^{t_j}$ where t_j is the order of the largest Jordan block corresponding to λ_j, and q is the number of distinct eigenvalues. This shows that the contribution of a degenerate eigenvalue λ_j to the number of necessary iterations for a given relative accuracy is at most t_j. Then the superlinear estimate remains uniform in a family of problems if t_j is bounded as $1 \le j \le n$ and $n \to \infty$.

2.2.2 The CGN method via normal equations

Another common way to solve (2.1) with nonsymmetric A is to consider the normal equation

$$A^*Au = A^*b \tag{2.34}$$

and apply the symmetric CG algorithm (2.4) for the latter [45, 62]. This approach is often called *CGN method.* In order the preserve the notation r_k for the residual $Au_k - b$, we replace r_k in (2.4) by s_k and let $r_k = A^{-*}s_k$, i.e., we have $s_k = A^*r_k$. Further, A and b are replaced by A^*A and A^*b, respectively. From this we obtain the following algorithmic form of the CGN method: let $u_0 \in \mathbf{R}^n$ be arbitrary, $r_0 := Au_0 - b$, $s_0 := d_0 := A^*r_0$; for given d_k, u_k, r_k and s_k, we let

$$\begin{cases} z_k = Ad_k, \\ \alpha_k = -\dfrac{\langle r_k, z_k\rangle}{\|z_k\|^2}\ , \quad u_{k+1} = u_k + \alpha_k d_k\ , \quad r_{k+1} = r_k + \alpha_k z_k\ ; \\ s_{k+1} = A^*r_{k+1}, \\ \beta_k = \dfrac{\|s_{k+1}\|^2}{\|s_k\|^2}\ , \quad d_{k+1} = s_{k+1} + \beta_k d_k. \end{cases} \tag{2.35}$$

The convergence estimates for this algorithm follow directly from the symmetric case. The linear convergence estimate follows from (2.10), using that $\|e_k\|_{A^*A} = \|Ae_k\| = \|r_k\|$ and $\kappa(A^*A) = \kappa(A)^2$:

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le 2^{1/k}\ \frac{\kappa(A)-1}{\kappa(A)+1} \qquad (k = 1, 2, ..., n). \tag{2.36}$$

In terms of the notation (2.3) and using

$$\|A^{-1}\| \le 1/\lambda_0, \tag{2.37}$$

we have $\kappa(A) \le \Lambda/\lambda_0$, and (2.36) implies

Corollary 2.2.3 *If (2.2)-(2.3) hold, then*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le 2^{1/k}\ \frac{\Lambda-\lambda_0}{\Lambda+\lambda_0} \qquad (k = 1, 2, ..., n). \tag{2.38}$$

On the other hand, having the decomposition (2.25), using the relation $\|(A^*A)^{-1}\| = \|A^{-1}\|^2 \le \lambda_0^{-2}$ from (2.37) and $A^*A = I + (E^* + E + E^*E)$, the analogue of the superlinear estimate (2.13) for equation $A^*Au = A^*b$ implies

Corollary 2.2.4 *If (2.2) and (2.25) hold, then*

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{2}{k\lambda_0^2}\ \sum_{i=1}^{k}\Big(\big|\lambda_i(E^* + E)\big| + \lambda_i(E^*E)\Big) \qquad (k = 1, 2, ..., n). \tag{2.39}$$

Remark 2.2.1 An alternative to the above form of the normal equations is to solve first $AA^*v = b$ and then let $u = A^*v$. Since $\|v - v_k\|^2_{AA^*} = \|u - u_k\|^2$, where $u_k = A^*v_k$, in this version we actually minimize the iteration errors instead of the residual errors at each step of the algorithm.

2.3 Variable preconditioners in inner-outer iterations

The above detailed CG algorithms for systems $Au = b$ are normally used in preconditioned versions. This usually means that the system is formally replaced by $C^{-1}Au = C^{-1}b$, where C is the preconditioner. Then in all the above-mentioned algorithms, A can be replaced by $C^{-1}A$, which in practice leads to auxiliary linear systems with matrix C. Such preconditioners are by definition fixed during the iteration. Then the iteration accuracy of the CG method can be determined from some best approximation property, such as a minimal residual, described in subsection 2.1.1.

However, to be able to fully control the accuracy of preconditioners, it turns out that variable preconditioners must be used. As here the preconditioners may vary between iteration steps, there is no Krylov set present and the accuracy of the approximations cannot be determined from the spectrum of a fixed matrix like $C^{-1}A$.

An efficient technique to control the accuracy of the preconditioner is to include some inner iterations in its construction. Inner-outer iteration methods have been considered in e.g. [20, 53, 91]. The use of inner iterations is the most common reason why the preconditioner varies. If, namely, we use different numbers of inner iterations to satisfy some variable inner iteration accuracy, then the preconditioner is not fixed. Furthermore, if we use a CG method for the inner iteration too, then the method depends on the initial vector (residual) which varies from one outer iteration to the next. In general, one can then only provide an upper bound of the convergence rate of the outer iteration method. This will be shown in this section.

To solve a linear system $Au = b$, we shall now use a GCG type method with an in general variable preconditioner. Following [23], the preconditioner will be defined by an in general nonlinear mapping $r \mapsto B[r]$ such that $AB[r] \approx r$, i.e. $B[r]$ is an approximation of $A^{-1}r$ for a given residual r. The method of forming the approximate solutions will be based on certain linear combinations of an increasing set of linearly independent search direction vectors $\{d_j\}_{j=0}^{k}$. For this purpose we shall use a generalized conjugate gradient, minimum residual (GCG-MR) method. This will be constructed as a suitable modification of (2.19), such that two inner products $\langle .,. \rangle$ and $\langle .,. \rangle_1$ are used and the residuals are defined via the variable preconditioner.

First, the new approximation is defined by the same formula from the search directions as in (2.19). Recall that this formula involves at most s search directions, with coefficients $\alpha_{k-j}^{(k)}$ determined from the linear system $K^{(k)}\underline{\alpha}_k = \underline{\gamma}_k$, where $K_{i,j}^{(k)} = \langle Ad_{k-j}, Ad_{k-l} \rangle$; further, all coordinates but the first of the right-side vector vanish. These coefficients provide that the norm of the residual, $\|r_k\|^2$, becomes minimal in the subspace spanned by the search directions.

Second, the essential step where the preconditioner is involved is in defining a new search direction. Let $d_0 := \hat{r}_0$, where $\hat{r}_0 := B[r_0]$ is the initial pseudoresidual corresponding to the initial residual r_0. At step k, the new search direction is now defined by

$$d_{k+1} = \hat{r}_{k+1} + \sum_{j=0}^{s_k} \beta_{k-j}^{(k)} \, d_{k-j}$$

where $\hat{r}_{k+1} := B[r_{k+1}]$. Here the coefficients are determined by a Gram-Schmidt orthogonalization w.r.t. the inner product $\langle .,. \rangle_1$ such that $\langle d_{k+1}, d_{k-j} \rangle_1 = 0$ ($j = 0, \ldots, s_k$), that is, $\beta_{k-j}^{(k)} = -\langle r_{k+1}, d_{k-j} \rangle_1 / \|d_{k-j}\|_1^2$. Should d_{k+1} become a zero vector, we must modify the mapping B used, i.e. by performing more inner iterations.

Altogether, we obtain the following algorithm (a modification of (2.19)). Let us fix an integer

$s \in \mathbf{N}$ and let $s_k = \min\{k, s\}$ $(k \geq 0)$. Given u_0, let $r_0 := Au_0 - b$ and $d_0 := \hat{r}_0 = B[r_0]$. Then

$$\left\{\begin{array}{l} u_{k+1} = u_k + \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} d_{k-j} \text{ and } \ d_{k+1} = \hat{r}_{k+1} + \sum\limits_{j=0}^{s_k} \beta_{k-j}^{(k)} d_{k-j}, \\ \text{ where } \hat{r}_{k+1} = B[r_{k+1}] \quad (j = 0, \dots, s_k), \qquad \beta_{k-j}^{(k)} = -\langle \hat{r}_{k+1}, d_{k-j}\rangle_1 / \|d_{k-j}\|_1^2 \\ \text{ and } \ \alpha_{k-j}^{(k)} \quad (j = 0, \dots, s_k) \quad \text{are the solution of the linear system} \\ \qquad \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} \langle Ad_{k-j}, Ad_{k-l}\rangle = \begin{cases} -\langle r_k, Ad_k\rangle & (\text{if } \ l = 0), \\ \ 0 & (\text{if } \ 1 \leq l \leq s_k). \end{cases} \end{array}\right. \tag{2.40}$$

Let us now consider two choices of the inner product $\langle ., . \rangle_1$:

(i) $\langle u, v\rangle_1 := \langle Au, Av\rangle = \langle A^T Au, v\rangle$;

(ii) $\langle u, v\rangle_1 := \langle u, v\rangle$.

(Here $\langle ., . \rangle$ can be the standard inner product.)

In the first case, the Gram-Schmidt orthogonalization implies $\langle Ad_{k+1}, Ad_{k-j}\rangle = 0$ $(j = 0, \dots, s_k)$, therefore the matrix $K^{(k)}$ (that appears in the last row of (2.40)) becomes diagonal. Hence $\alpha_{k-j}^{(k)} = 0$ for $1 \leq l \leq s_k$, and the long recursions for u_k and r_k truncate automatically to short recursions.

Then the GCG-MR method takes the following form. Let us fix an integer $s \in \mathbf{N}$ and let $s_k = \min\{k, s\}$ $(k \geq 0)$. Given u_0, let $r_0 := Au_0 - b$ and $d_0 := \hat{r}_0 = B[r_0]$. Then

$$\left\{\begin{array}{l} u_{k+1} = u_k + \alpha_k^{(k)} d_k \text{ and } \ r_{k+1} = r_k + \alpha_k^{(k)} Ad_k \\ \text{where } \quad \alpha_k^{(k)} = -\langle r_k, Ad_k\rangle / \|Ad_k\|^2\,; \\ \text{ let } \ \hat{r}_{k+1} := B[r_{k+1}] \quad \text{and} \ \text{ compute } \ A\hat{r}_{k+1}\,; \\ \text{let } \ d_{k+1} = \hat{r}_{k+1} + \sum\limits_{j=0}^{s_k} \beta_{k-j}^{(k)} d_{k-j}, \\ \text{where } \beta_{k-j}^{(k)} = -\langle A\hat{r}_{k+1}, Ad_{k-j}\rangle / \|Ad_{k-j}\|^2 \quad (j = 0, \dots, s_k). \end{array}\right. \tag{2.41}$$

It is seen that the algorithm requires two matrix-vector multiplications with A. This algorithm is similar to the generalized conjugate residual (GCR) method in [44]. In this algorithm one can replace the computation of $A\hat{r}_{k+1}$ by a long recursion, in this way avoiding the second matrix-vector multiplication with A. This can, however, not be applied in our case, when the preconditioner is variable.

Algorithm (2.41) achieves the same result as the well-known GMRES method [89], but has a simpler form since the so-called upper Hessenberg matrix need not be used. The GMRES method is further based on an Arnoldi process to form the orthogonal search vectors. Here also two matrix-vector multiplications with A are needed.

Rounding errors may cause some loss of orthogonality of the search vectors used, hence the vectors used in the recursions for u_k and r_k may not be exactly orthogonal, implying that the recursion for such inexact vectors may actually not truncate. To some extent we may correct for this by computing the right-hand side components $-\langle r_k, Ad_{k-l}\rangle$ $(1 \leq l \leq s_k)$ in the last row of (2.40), which would equal zero in exact arithmetic, and solve the arising system. To save computational labour, this correction should be done only in, say, every s step.

To avoid the influence of loss of orthogonality and to save one matrix-vector multiplication with A, one can replace algorithm (2.41) by the following algorithm, which is similar to (2.19) but the inner product $\langle u,v\rangle_1 := \langle u,v\rangle$ is used to compute the coefficients $\beta_{k-j}^{(k)}$. Let us fix an integer $s \in \mathbf{N}$ and let $s_k = \min\{k,s\}$ $(k \geq 0)$. Given u_0, let $r_0 := Au_0 - b$ and $d_0 := \hat{r}_0 = B[r_0]$. Then

$$\left\{\begin{array}{l} u_{k+1} = u_k + \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} d_{k-j} \text{ and } r_{k+1} = r_k + \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} A d_{k-j} \\ \text{where } \alpha_{k-j}^{(k)} \quad (j = 0,\dots,s_k) \quad \text{are the solution of the linear system} \\ \quad \sum\limits_{j=0}^{s_k} \alpha_{k-j}^{(k)} \langle Ad_{k-j}, Ad_{k-l}\rangle = -\langle r_k, Ad_{k-l}\rangle \qquad (0 \leq l \leq s_k); \\ \text{compute } \hat{r}_{k+1} := B[r_{k+1}]; \\ \text{let } d_{k+1} = \hat{r}_{k+1} + \sum\limits_{j=0}^{s_k} \beta_{k-j}^{(k)} d_{k-j}, \\ \text{where } \beta_{k-j}^{(k)} = -\langle \hat{r}_{k+1}, d_{k-j}\rangle / \|d_{k-j}\|^2 \quad (j = 0,\dots,s_k) \end{array}\right. \tag{2.42}$$

The GCG algorithms (2.41)-(2.42) require one action of $B[\,.\,]$ at each iteration step. In addition, algorithm (2.41) requires two matrix-vector multiplications with A, while algorithm (2.42) requires only one such multiplication. The latter is the minimum number required for any method. Frequently, as in our applications, the action of $B[\,.\,]$ is the most expensive part of the iteration method and, unless s is large, may dominate the labour involved in the long recursions for the updates of vectors.

As shown in [23], an upper bound of the rate of convergence can be determined by two parameters, those of coercivity and boundedness. That is, we assume that the preconditioning mapping $B[\,.\,]$ satisfies

$$\langle AB[v], v\rangle \geq \lambda_0 \|v\|^2, \qquad \|AB[v]\| \leq \Lambda \|v\| \qquad (\forall v \in \mathbf{R}^n) \tag{2.43}$$

for some constants $\Lambda \geq \lambda_0 > 0$. In practice on frequently estimates $\|AB - I\| \leq \delta < 1$, in which case $\lambda_0 \geq 1 - \delta$ and $\Lambda \leq 1 + \delta$. The bounds (2.43) are the analogues of (2.3), and in the same way as in Corollary 2.2.1, we obtain

Proposition 2.3.1 *The variable-step GCG method converges monotonically with the following convergence rate:*

$$\|r_{k+1}\| \leq \left(1 - \left(\frac{\lambda_0}{\Lambda}\right)^2\right)^{1/2} \|r_k\| \qquad (k = 1, 2, ..., n). \tag{2.44}$$

Remark 2.3.1 The above upper bound holds also for the fully truncated versions of the methods, including the steepest descent method which corresponds to $s = 0$. As shown in [72], under certain (though special) conditions these upper bounds may actually be sharp, showing that the method for variable preconditioners in these cases cannot converge faster than for the corresponding steepest descent method. On the other hand, if a sufficient number of inner iterations are performed, then the method approaches the corresponding method with a fixed preconditioner.

Remark 2.3.2 There is a simple way to automatically determine if the preconditioner is sufficiently accurate, namely, to check the sign of $-\langle r_k, Ad_k\rangle$, which equals $\langle AB[r_k], r_k\rangle$ and hence must be positive. Should it be negative, then the last row of (2.40) and Cramer's rule imply

$$\alpha_k^{(k)} = -\langle r_k, Ad_k\rangle \; det(K^{(k)})/det(K^{(k+1)}) < 0$$

(since $K^{(k)}$ and $K^{(k+1)}$ are positive definite and hence have positive determinants). Then no convergence can occur, and one must repeat the last step with a more accurate preconditioner, e.g. by performing more inner iterations.

3 Equivalent operators and linear convergence

3.1 On the general theory of equivalent operators

The basic idea of our paper is that the discretization A_h of a linear operator A can be preconditioned by the discretization B_h (under the same process as for A) of another linear operator B. Under the requirement that systems with B_h are easier to solve than systems with A_h, the main goal of this approach is to ensure that the condition numbers of the preconditioned operators are bounded in h as $h \to 0$. The proper general framework, suitable to treat this independence property, has proved to be the concept of equivalence of operators. It has been introduced in [49] where a rigorous study of this framework has been given, also followed by [52, 64, 77, 78]. We briefly outline some notions and related results from this work.

Let $B : W \to V$ and $A : W \to V$ be linear operators between the Hilbert spaces W and V. For our purposes it suffices to consider the case when B and A are one-to-one and $D = D(A) \cap D(B)$ is dense. The operator A is said to be equivalent in V-norm to B on D if there exist constants $K \geq k > 0$ such that

$$k \leq \frac{\|Au\|_V}{\|Bu\|_V} \leq K \qquad (u \in D \setminus \{0\}). \tag{3.1}$$

If (3.1) holds, then under suitable density assumptions on D, the condition number of AB^{-1} in V is bounded by K/k. The W-norm equivalence of B^{-1} and A^{-1} implies this bound similarly for $B^{-1}A$.

The analogous property for the discretized problems is uniform norm equivalence defined as follows. Let us consider families of operators A_h and B_h (indexed by $h > 0$) such that there exist the pointwise limit operators A and B as $h \to 0$. The families A_h and B_h are said to be V-norm uniformly equivalent if there exist constants $\tilde{K} \geq \tilde{k} > 0$, independent of h, such that

$$\tilde{k} \leq \frac{\|A_h u\|_V}{\|B_h u\|_V} \leq \tilde{K} \qquad (u \in D \setminus \{0\},\, h > 0). \tag{3.2}$$

Analogously to the above, this implies that the condition numbers of the family $A_h B_h^{-1}$ are bounded uniformly in h, and the similar uniform equivalence of B_h^{-1} and A_h^{-1} implies that the condition numbers of the family $B_h^{-1} A_h$ are bounded uniformly in h.

Using the above notions, the following general results hold. First, the V-norm equivalence of A and B is necessary for the V-norm uniform equivalence of the families A_h and B_h. Second, sufficiency of the above statement is also true if the families A_h and B_h are obtained via orthogonal projections from A and B and, further, if A and B are equivalent to the families A_h and B_h. For details and various special and related cases see [49, Chap. 2].

The above setting is mostly intended to handle L_2-norm equivalence for elliptic operators. However, it is often more convenient to use H^1-norm equivalence [49, 78] based on a weak formulation, since this helps to avoid regularity requirements. Namely, in the usual definition of an elliptic operator A in strong form in $L^2(\Omega)$, its domain $D(A)$ is a subspace of $H^2(\Omega)$ and, moreover, the above equivalence theory is based on the H^2-regularity assumption for the involved operators.

The notion of H^1-norm equivalence is based on the weak form of elliptic operators as follows, see [78] for details. In a standard way, using Green's formula, one can define the bilinear form $a(.,.)$ corresponding to an elliptic operator A on a subspace $H^1_D(\Omega)$ of $H^1(\Omega)$ (associated with the boundary conditions), and this form satisfies $a(u,v) = \langle Au, v\rangle_{L^2}$ for $u, v \in D(A)$. The bounded bilinear form a gives rise to an operator A_w from $H^1_D(\Omega)$ into its dual satisfying $A_w u(v) = a(u,v)$. We note that the dual of $H^1_D(\Omega)$ can be identified with $H^1_D(\Omega)$ itself by the Riesz theorem, which will be convenient for our purposes as we can consider A_w as mapping into $H^1_D(\Omega)$ and satisfying

$$\langle A_w u, v\rangle_{H^1_D} = \langle Au, v\rangle_{L^2} \qquad (u, v \in D(A)). \tag{3.3}$$

The fundamental result on H^1-norm equivalence in [78] reads as follows: if A and B are invertible uniformly elliptic operators, then A_w^{-1} and B_w^{-1} are H^1-norm equivalent if and only if A and B have homogeneous Dirichlet boundary conditions on the same portion of the boundary.

In the present paper, in what follows, we build on the above result and develop a simpler Hilbert space setting of equivalent operators that is a priori suited for invertible elliptic operators with identical Dirichlet boundary. This setting (of so-called S-bounded and S-coercive operators) is developed in the next subsection based on [16, section 3.1], and it will rely on a coercivity assumption and a relation analogous to (3.3). Here the Hilbert space model is also followed by the description of the considered corresponding class of elliptic operators. Then in the last section, mesh independent linear convergence results are presented in a Hilbert space setting for S-bounded and S-coercive equivalent operators. These results will be used in the later parts of the paper to derive mesh independent linear convergence results for FEM discretizations using various equivalent preconditioners.

Concerning the desired uniform equivalence results for discretized operators A_h and B_h, it turns out from [49, 78] that one can mostly derive general results for FEM type discretizations only, whereas FDM discretizations require a case by case study. Accordingly, in our setting of S-bounded and S-coercive operators, we can give general results for Galerkin discretizations in Hilbert space and apply them later for general elliptic problems, whereas for FDM discretizations we only cite certain case by case investigations.

3.2 A coercive setting in weak form

In this section we first discuss S-bounded and S-coercive operators in Hilbert space, based on [16, section 3.1]. The main ingredient of this setting is a coercivity assumption. Then, using this background, we describe the class of elliptic operators that will be mostly considered throughout this paper. Their main property is coercivity in the corresponding Sobolev space, which ensures well-posedness of the related boundary value problems. The weak formulation will allow us to treat the equivalence of operators in an easy, that is, not very technical form.

3.2.1 S-bounded and S-coercive operators in Hilbert space

In what follows, let H be a real Hilbert space. We are interested in solving an operator equation

$$Lu = g \tag{3.4}$$

for an unbounded linear operator L in H, where $g \in H$. In order to define a weak form, we recast the required properties of L to the energy space of a suitable symmetric and coercive operator S.

Therefore, let S be an (also unbounded) linear symmetric operator in H which is coercive, i.e., there exists $p > 0$ such that $\langle Su, u\rangle \geq p\|u\|^2$ $(u \in D(S))$. We recall that the energy space H_S is the completion of $D(S)$ under the inner product $\langle u, v\rangle_S = \langle Su, v\rangle$, and the coercivity of S implies $H_S \subset H$. The corresponding S-norm is denoted by $\|u\|_S$, and the space of bounded linear operators on H_S by $B(H_S)$.

Definition 3.2.1 Let S be a linear symmetric coercive operator in H. A linear operator L in H is said to be *S-bounded and S-coercive*, and we write $L \in BC_S(H)$, if the following properties hold:

(i) $D(L) \subset H_S$ and $D(L)$ is dense in H_S in the S-norm;

(ii) there exists $M > 0$ such that

$$|\langle Lu, v\rangle| \leq M\|u\|_S\|v\|_S \qquad (u, v \in D(L));$$

(iii) there exists $m > 0$ such that

$$\langle Lu, u\rangle \geq m\|u\|_S^2 \qquad (u \in D(L)).$$

Definition 3.2.2 For any $L \in BC_S(H)$, let $L_S \in B(H_S)$ be defined by

$$\langle L_S u, v\rangle_S = \langle Lu, v\rangle \qquad (u, v \in D(L)).$$

Remark 3.2.1 (a) The above definition makes sense since L_S is the bounded linear operator on H_S that represents the unique extension to H_S of the densely defined S-bounded bilinear form $u, v \mapsto \langle Lu, v\rangle$.

(b) The density of $D(L)$ implies

$$|\langle L_S u, v\rangle_S| \leq M\|u\|_S\|v\|_S, \qquad \langle L_S u, u\rangle_S \geq m\|u\|_S^2 \qquad (u, v \in H_S). \tag{3.5}$$

(c) If $R(L) \subset R(S)$ (where $R(.)$ denotes the range), then the restriction of the operator L_S to $D(L)$ satisfies

$$L_S\big|_{D(L)} = S^{-1}L. \tag{3.6}$$

(d) Definition 3.2.2 uses the idea of weak form of operators from [78] as explained in section 3.1, see equation (3.3).

We verify now that our setting leads to a special case of equivalent operators.

Proposition 3.2.1 *Let N and L be S-bounded and S-coercive operators for the same S. Then*
(a) N_S and L_S are H_S-norm equivalent,
(b) N_S^{-1} and L_S^{-1} are H_S-norm equivalent.

Proof. (a) By (3.1), we must find $K \geq k > 0$ such that

$$k\|N_S u\|_S \leq \|L_S u\|_S \leq K\|N_S u\|_S \qquad (u \in H_S). \tag{3.7}$$

Since $L \in BC_S(H)$, there exists constants $M_L \geq m_L > 0$ such that for all $u \in H_S$,

$$m_L \|u\|_S \leq \frac{\langle L_S u, u\rangle_S}{\|u\|_S} \leq \|L_S u\|_S = \sup_{v \in H_S \setminus \mathbf{0}} \frac{\langle L_S u, v\rangle_S}{\|v\|_S} \leq M_L \|u\|_S \tag{3.8}$$

and the analogous estimate holds for N with some $M_N \geq m_N > 0$. The two estimates yield (3.7) with $K = \frac{M_L}{m_N}$ and $k = \frac{m_L}{M_N}$.

(b) Properties (3.5) imply that L_S is invertible in $B(H_S)$, hence for all $v \in H_S$ we can set $u = L_S^{-1} v$ in (3.8) to obtain

$$m_L \|L_S^{-1} v\|_S \leq \|v\|_S \leq M_L \|L_S^{-1} v\|_S \qquad (v \in H_S).$$

This and its analogue for N yield the required estimate similarly as in (a), now with $K = \frac{M_N}{m_L}$ and $k = \frac{m_N}{M_L}$. ■

Let us now return to the operator equation (3.4) for $L \in BC_S(H)$.

Definition 3.2.3 For given $L \in BC_S(H)$, we call $u \in H_S$ the *weak solution* of equation (3.4) if

$$\langle L_S u, v\rangle_S = \langle g, v\rangle \qquad (v \in H_S). \tag{3.9}$$

Remark 3.2.2 For all $g \in H$ the weak solution of (3.4) exists and is unique. This follows in the usual way from the Lax-Milgram theorem since, by the coercivity of S, the mapping $v \mapsto \langle g, v\rangle$ is a bounded linear functional on H_S. In particular, if $u \in D(L)$, then u satisfies (3.4) (i.e. it is a strong solution) if and only if it is a weak solution.

3.2.2 Coercive elliptic operators

Let us define the elliptic operator

$$Lu \equiv -\mathrm{div}\,(A\,\nabla u) + \mathbf{b} \cdot \nabla u + cu \qquad \text{for} \;\; u_{|\Gamma_D} = 0, \; \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \tag{3.10}$$

where $\frac{\partial u}{\partial \nu_A} = A\,\nu \cdot \nabla u$ denotes the weighted form of the normal derivative. (The formal domain of L to be used in Definition 3.2.2 consists of those $u \in H^2(\Omega)$ that satisfy the above boundary conditions, however, this is nowhere used elsewhere in this paper.) The following properties are assumed to hold:

Assumptions 3.2.2.1

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subsets of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$;

(ii) $A \in C^1(\overline{\Omega}, \mathbf{R}^{d \times d})$ and for all $x \in \overline{\Omega}$ the matrix $A(x)$ is symmetric; $\mathbf{b} \in C^1(\overline{\Omega})^d$, $c \in L^\infty(\Omega)$, $\alpha \in L^\infty(\Gamma_N)$;

(iii) we have the following properties which will imply coercivity: there exists $p > 0$ such that $A(x)\xi \cdot \xi \geq p\,|\xi|^2$ for all $x \in \overline{\Omega}$ and $\xi \in \mathbf{R}^d$; $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b} \geq 0$ in Ω and $\hat{\alpha} := \alpha + \frac{1}{2}\,(\mathbf{b} \cdot \nu) \geq 0$ on Γ_N;

(iv) either $\Gamma_D \neq \emptyset$, or $\hat{c}$ or $\hat{\alpha}$ has a positive lower bound.

Let us also define a symmetric elliptic operator on the same domain Ω with otherwise analogous properties:

$$Su \equiv -\mathrm{div}\,(G\,\nabla u) + \sigma u \qquad \text{for} \;\; u_{|\Gamma_D} = 0, \; \tfrac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{3.11}$$

which satisfies

Assumptions 3.2.2.2

(i) Substituting G for A, Ω, Γ_D, Γ_N and G satisfy Assumptions 3.2.2.1;

(ii) $\sigma \in L^\infty(\Omega)$ and $\sigma \geq 0$; $\beta \in L^\infty(\Gamma_N)$ and $\beta \geq 0$; further, if $\Gamma_D = \emptyset$ then σ or β has a positive lower bound.

For the study of these operators we define the space

$$H^1_D(\Omega) := \{u \in H^1(\Omega) : \, u_{|\Gamma_D} = 0\} \quad \text{with} \; \langle u, v\rangle_S := \int_\Omega (G\,\nabla u \cdot \nabla v + \sigma uv) + \int_{\Gamma_N} \beta uv \, d\sigma \tag{3.12}$$

which is the energy space H_S of the operator S.

Proposition 3.2.2 *If Assumptions 3.2.2.1-2 hold, then the operator L is S-bounded and S-coercive in $L^2(\Omega)$, i.e., $L \in BC_S(L^2(\Omega))$.*

Proof. We must verify the properties in Definition 3.2.1 from the above assumptions. The domain of definition of L is $D(L) := \{u \in H^2(\Omega) : \, u_{|\Gamma_D} = 0, \, \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0\}$ in the Hilbert space $L^2(\Omega)$, so $D(L) \subset H_S = H^1_D(\Omega)$ and $D(L)$ is dense in $H^1_D(\Omega)$ in the S-inner product (3.12). Further, for $u, v \in D(L)$, by Green's formula, we have

$$\langle Lu, v\rangle_{L^2(\Omega)} = \int_\Omega \Big(A\,\nabla u \cdot \nabla v + (\mathbf{b} \cdot \nabla u)v + cuv \Big) \; + \int_{\Gamma_N} \alpha uv \, d\sigma \,. \tag{3.13}$$

Using this and (3.12), one can check properties (ii)-(iii) of Definition 3.2.1 with a standard calculation as follows. First, Assumptions 3.2.2.2 imply that the S-norm related to (3.12) is equivalent to the usual H^1-norm, and accordingly, there exist constants $C_{\Omega,S} > 0$ and $C_{\Gamma_N,S} > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq C_{\Omega,S}\|u\|_S \quad \text{and} \quad \|u\|_{L^2(\Gamma_N)} \leq C_{\Gamma_N,S}\|u\|_S \qquad (u \in H^1_D(\Omega)), \tag{3.14}$$

see, e.g., [97]. Further, the uniform spectral bounds of A and G also imply the existence of constants $p_1 \geq p_0 > 0$ such that

$$p_0 \; (G(x)\xi \cdot \xi) \leq A(x)\xi \cdot \xi \leq p_1 \; (G(x)\xi \cdot \xi) \qquad (x \in \overline{\Omega}, \; \xi \in \mathbf{R}^d), \tag{3.15}$$

and there exists $q > 0$ such that

$$q \, \|\nabla u\|^2_{L^2(\Omega)} \leq \int_\Omega G\,\nabla u \cdot \nabla u \leq \|u\|^2_S \qquad (u \in H^1_D(\Omega)). \tag{3.16}$$

Then from (3.13) we obtain

$$\begin{aligned} \langle Lu, v\rangle \;\leq\; & p_1\|u\|_S\|v\|_S + \|\mathbf{b}\|_{L^\infty(\Omega)^d}\|\nabla u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\ & + \|c\|_{L^\infty(\Omega)}\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} + \|\alpha\|_{L^\infty(\Gamma_N)}\|u\|_{L^2(\Gamma_N)}\|v\|_{L^2(\Gamma_N)} \end{aligned}$$

$$\leq \Big(p_1 + C_{\Omega,S}\, q^{-1/2} \|\mathbf{b}\|_{L^\infty(\Omega)^d} + C_{\Omega,S}^2 \|c\|_{L^\infty(\Omega)} + C_{\Gamma_N,S}^2 \|\alpha\|_{L^\infty(\Gamma_N)}\Big)\ \|u\|_S \|v\|_S\,. \tag{3.17}$$

On the other hand, for any $u \in H^1_D(\Omega)$, using the definition of $\hat{c}$ and $\hat{\alpha}$ from Assumptions 3.2.2.1 (iii), a standard calculation with Green's formula yields (see, e.g., [66]) that

$$\langle Lu, u\rangle_{L^2(\Omega)} = \int_\Omega (A\ \nabla u \cdot \nabla u + \hat{c} u^2) + \int_{\Gamma_N} \hat{\alpha} u^2\, d\sigma\ =: \|u\|_L^2\,. \tag{3.18}$$

Assumptions 3.2.2.1 imply that the L-norm, defined above on the right, is equivalent to the usual H^1-norm, hence there exist constants $C_{\Omega,L} > 0$ and $C_{\Gamma_N,L} > 0$ such that the analogue of (3.14) holds for the L-norm instead of the S-norm. Therefore

$$\begin{aligned}\|u\|_S^2 &= \int_\Omega (G\ \nabla u \cdot \nabla u + \sigma u^2) + \int_{\Gamma_N} \beta u^2\, d\sigma \\ &\leq p_0^{-1} \int_\Omega\ A\ \nabla u \cdot \nabla u + \|\sigma\|_{L^\infty(\Omega)} \int_\Omega u^2 +\ \|\beta\|_{L^\infty(\Gamma_N)} \int_{\Gamma_N} u^2\, d\sigma \\ &\leq \Big(p_0^{-1} + C_{\Omega,L}^2 \|\sigma\|_{L^\infty(\Omega)} + C_{\Gamma_N,L}^2 \|\beta\|_{L^\infty(\Gamma_N)}\Big)\ \langle Lu, u\rangle_{L^2(\Omega)} \qquad (u \in H^1_D(\Omega)).\end{aligned} \tag{3.19}$$

Summing up, estimates (3.17) and (3.19) yield that properties (ii)-(iii) of Definition 3.2.1 are valid with

$$\begin{aligned} M &:= p_1 + C_{\Omega,S}\, q^{-1/2} \|\mathbf{b}\|_{L^\infty(\Omega)^d} + C_{\Omega,S}^2 \|c\|_{L^\infty(\Omega)} + C_{\Gamma_N,S}^2 \|\alpha\|_{L^\infty(\Gamma_N)}\ , \\ m &:= \Big(p_0^{-1} + C_{\Omega,L}^2 \|\sigma\|_{L^\infty(\Omega)} + C_{\Gamma_N,L}^2 \|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1}\,. \qquad \blacksquare \end{aligned} \tag{3.20}$$

Remark 3.2.3 It is clear from Definition 3.2.2 and the standard application of Green's formula that equation (3.9) for operators (3.10)-(3.11) and some $g \in L^2(\Omega)$ coincides with the usual weak formulation

$$\int_\Omega \Big(A\, \nabla u \cdot \nabla v + (\mathbf{b} \cdot \nabla u) v + cuv\Big)\ + \int_{\Gamma_N} \alpha uv\, d\sigma = \int_\Omega gv \qquad (v \in H^1_D(\Omega)), \tag{3.21}$$

hence by Remark 3.2.2, the weak solution exists and is unique.

Remark 3.2.4 The constants $C_{\Omega,S}$ and $C_{\Gamma_N,S}$ in (3.20) can be calculated as follows. (The same holds for $C_{\Omega,L}$ and $C_{\Gamma_N,L}$.)

In order to find $C_{\Omega,S}$, first let $\Gamma_D \neq \emptyset$. Then it suffices to determine $C_\Omega > 0$ such that

$$\|u\|_{L^2(\Omega)} \leq C_\Omega \|\nabla u\|_{L^2(\Omega)} \qquad (u \in H^1_D(\Omega)), \tag{3.22}$$

in which case $C_{\Omega,S} = q^{-1/2} C_\Omega$ from (3.16). Here such a C_Ω exists because for $\Gamma_D \neq \emptyset$, the usual H^1-norm is equivalent to $\|\nabla u\|^2_{L^2(\Omega)}$. Its sharp value satisfies $C_\Omega = \lambda_1^{-1/2}$, where λ_1 is the smallest eigenvalue of $-\Delta$ under boundary conditions $u_{|\Gamma_D} = 0$, $\frac{\partial u}{\partial \nu}_{|\Gamma_N} = 0$. For Dirichlet boundary conditions, one can use the estimate

$$C_\Omega \leq \Big(\sum_{i=1}^d \big(\frac{\pi}{a_i}\big)^2\Big)^{-1/2}$$

if Ω is embedded in a brick with edges $a_1, \ldots, a_d$, see, e.g., [81]. If $\Gamma_D = \emptyset$ then similarly as above, $C_{\Omega,S} \leq p_0^{-1/2} \hat{C}_\Omega$, where $\hat{C}_\Omega$ is the smallest eigenvalue of the operator $-\Delta u + (\sigma_0/p_0)u$ under boundary conditions $u_{|\Gamma_D} = 0$, $\frac{\partial u}{\partial \nu} + (\beta_0/p_0)_{|\Gamma_N} = 0$, in which $\sigma_0 := \inf \sigma$ and $\beta_0 := \inf \beta$. Here it is advisable to choose σ to satisfy $\sigma_0 > 0$, in which case $\|u\|^2_{L^2(\Omega)} \le \sigma_0^{-1} \int_\Omega \sigma u^2 \le \sigma_0^{-1} \|u\|^2_S$, i.e. $C_{\Omega,S} \le \sigma_0^{-1/2}$.

For $C_{\Gamma_N,S}$, one should first find $C_{\Gamma_N} > 0$ such that

$$\|u\|_{L^2(\Gamma_N)} \le C_{\Gamma_N} \|u\|_{H^1(\Omega)} \qquad (u \in H^1_D(\Omega)),$$

in which case $C_{\Gamma_N,S} = \left(1 + C_\Omega^2\right)^{1/2} q^{-1/2} C_{\Gamma_N}$ from (3.16) and (3.22). For polygonal domains in 2D, explicit estimates for C_{Γ_N} are given in [86].

3.3 Mesh independence properties of linear convergence in Hilbert space

In this subsection we give some mesh independent bounds for condition numbers and corresponding linear convergence factors, when general Galerkin discretizations are considered in Hilbert space. These results will be used in the later parts of the paper to derive mesh independent linear convergence results for FEM discretizations. As noted in subsection 3.1, FDM type discretizations do not have such a general abstract background, therefore they are not yet considered here.

Let us then return to a Hilbert space H and consider the operator equation

$$Lu = g \tag{3.23}$$

where L is S-bounded and S-coercive in the sense of Definition 3.2.1, and $g \in H$. Equation (3.23) can be solved numerically using a Galerkin discretization: let

$$V_h = span\{\varphi_1, \dots, \varphi_n\} \subset H_S\,,$$

where φ_i are linearly independent, be a given finite-dimensional subspace and

$$\mathbf{L}_h := \left\{ \langle L_S \varphi_i, \varphi_j \rangle_S \right\}_{i,j=1}^n .$$

Finding the discrete solution $u_h \in V_h$ requires solving the $n \times n$ system

$$\mathbf{L}_h\, \mathbf{c} = \mathbf{b}_h \tag{3.24}$$

with $\mathbf{b}_h = \{\langle g, \varphi_j \rangle\}_{j=1}^n$. Since $L \in BC_S(H)$, the symmetric part of $\mathbf{L}_h$ is positive definite, hence system (3.24) has a unique solution. Moreover, if a sequence of such subspaces V_h satisfies $\inf_{v \in V_h} \|u - v\|_S \to 0$ for all $u \in H_S$, then the coercivity of L_S implies in the standard way [34] that u_h converges to the exact weak solution in H_S-norm.

In what follows, we outline some abstract preconditioning concepts based on the equivalent operator idea, and give corresponding mesh independent bounds for the convergence of proper CG algorithms.

3.3.1 Symmetric problems

Let us first consider the case when L itself is a symmetric operator, in which case its S-coercivity and S-boundedness simply turns into the spectral equivalence relation

$$m\|u\|^2_S \le \langle L_S u, u \rangle_S \le M \|u\|^2_S \qquad (u \in H_S). \tag{3.25}$$

Then $\mathbf{L}_h$ is symmetric too.

Let S be the symmetric coercive operator from Definition 3.2.1, and introduce the stiffness matrix of S,

$$\mathbf{S}_h = \left\{ \langle \varphi_i, \varphi_j \rangle_S \right\}_{i,j=1}^{n} ,$$

as preconditioner for system (3.24). This yields the preconditioned system

$$\mathbf{S}_h^{-1} \mathbf{L}_h \, \mathbf{c} = \tilde{\mathbf{b}_h} \tag{3.26}$$

(with $\tilde{\mathbf{b}_h} = \mathbf{S}_h^{-1} \mathbf{b}_h$). Now $\mathbf{S}_h^{-1} \mathbf{L}_h$ is self-adjoint w.r.t. the inner product $\langle \mathbf{c}, \mathbf{d} \rangle_{\mathbf{S}_h} := \mathbf{S}_h \, \mathbf{c} \cdot \mathbf{d}$. Hence we can apply algorithm (2.4) for system (3.26), in which we endow $\mathbf{R}^n$ with the $\mathbf{S}_h$-inner product $\langle .,. \rangle_{\mathbf{S}_h}$. Then estimate (2.10) holds with $\kappa(A) = \kappa(\mathbf{S}_h^{-1} \mathbf{L}_h)$.

The main result here is in general terms

$$\kappa(\mathbf{S}_h^{-1} \mathbf{L}_h) \le \kappa(S^{-1} L),$$

but to be precise, in order to have an operator on the whole space H_S on the r.h.s., we must replace $\kappa(S^{-1}L)$ by $\kappa(L_S)$ (cf. (3.6)); further, with the available bounds from (3.25), we shall use the estimate M/m for the latter:

Proposition 3.3.1 *If the symmetric operator L satisfies (3.25), then for any subspace $V_h \subset H_S$*

$$\kappa(\mathbf{S}_h^{-1} \mathbf{L}_h) \le \frac{M}{m} \tag{3.27}$$

independently of V_h.

Proof. This property follows trivially since for arbitrary $\mathbf{c} \in \mathbf{R}^n$, setting $u = \sum\limits_{j=1}^{n} c_j \varphi_j \in V_h$ in (3.25) yields

$$m \ (\mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c}) \le \mathbf{L}_h \, \mathbf{c} \cdot \mathbf{c} \le M \ (\mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c}) . \qquad \blacksquare$$

Consequently, the CG algorithm (2.4) for system (3.26) converges as

$$\left(\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}} \right)^{1/k} \le 2^{1/k} \, \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \qquad (k = 1, 2, ..., n) \tag{3.28}$$

independently of V_h. This idea of preconditioning (yet in the context of simple iterations) goes back to [39, 43].

Remark 3.3.1 (i) An important application of this result arises for inner iterations used in the inexact Newton solution of nonlinear variational problems. Let a nonlinear operator $F : H_S \to H_S$ satisfy the uniform ellipticity property

$$m \|v\|_S^2 \le \langle F'(u)v, v \rangle_S \le M \|v\|_S^2 \qquad (u, v \in H_S) \tag{3.29}$$

with $M, m > 0$, which ensures well-posedness of equation $F(u) = 0$. If u_n is the nth outer Newton iterate and $L_S := F'(u_n)$, then Proposition 3.3.1 provides the mesh independent convergence rate $\frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}$ for the inner CG iteration.

(ii) The following class of operators forms the most common special case to satisfy (3.29), see [50, Remark 6.1]. Let H_S be a given Sobolev space over some bounded domain $\Omega \subset \mathbf{R}^d$, such that its inner product is expressed as

$$\langle h, v \rangle_S = \int_\Omega B(h,v)$$

for some given bilinear mapping $B : H_S \times H_S \to L^1(\Omega)$. Let the operator $F : H_S \to H_S$ have the form

$$\langle F(u), v \rangle_S = \int_\Omega \Big(a(B(u,u))\, B(u,v) - fv \Big) \qquad (u, v \in H_S), \tag{3.30}$$

where $a : \mathbf{R}^+ \to \mathbf{R}^+$ is a scalar C^1 function and there exist constants $M \geq m > 0$ such that

$$0 < m \leq a(r) \leq M, \qquad 0 < m \leq \tfrac{d}{dr}\Big(a(r^2) r \Big) \leq M \qquad (r \geq 0), \tag{3.31}$$

further, $f \in L^2(\Omega)$. In order to verify (3.29), let

$$p(r^2) = \min\Big\{ a(r^2),\, \tfrac{d}{dr}\Big(a(r^2) r \Big) \Big\}, \quad q(r^2) = \max\Big\{ a(r^2),\, \tfrac{d}{dr}\Big(a(r^2) r \Big) \Big\} \qquad (r \geq 0), \tag{3.32}$$

where by (3.31): ,

$$0 < m \leq p(r) \leq q(r) \leq M \qquad (r \geq 0). \tag{3.33}$$

It is easy to see that for all $u, h, v \in H_S$

$$\langle F'(u)h, v \rangle_S = \int_\Omega \Big(a(B(u,u))\, B(h,v) + 2a'(B(u,u))\, B(u,h)\, B(u,v) \Big) \qquad (u, h, v \in H_S) \tag{3.34}$$

and hence

$$m \int_\Omega B(v,v) \leq \int_\Omega p(B(u,u))\, B(v,v) \leq \langle F'(u)v, v \rangle_S \leq \int_\Omega q(B(u,u))\, B(v,v) \leq M \int_\Omega B(v,v)\,, \tag{3.35}$$

which coincides with (3.29).

For a corresponding boundary value problem, the FEM solution u_h in some subspace $V_h \subset H$ must satisfy

$$\langle F(u_h), v \rangle_S = 0 \qquad (v \in V_h) \tag{3.36}$$

or $F(u_h) = 0$. If u_n is the nth Newton iterate, then the correction term $p_n \in V_h$ is found by solving the linearized problem

$$\langle F'(u_n)p_n, v \rangle_S = -\langle F(u_n), v \rangle_S \qquad (v \in V_h), \tag{3.37}$$

which now reads as follows: for all $v \in V_h$

$$\int_\Omega \Big(a(B(u_n,u_n))\, B(p_n,v) + 2a'(B(u_n,u_n))\, B(u_n,p_n)\, B(u_n,v) \Big) = -\int_\Omega \Big(a(B(u_n,u_n))\, B(u_n,v) - fv \Big) \tag{3.38}$$

As stated above after (3.29), Proposition 3.3.1 provides mesh independent convergence for the inner CG iteration for problem (3.38).

3.3.2 Nonsymmetric problems: symmetric equivalent preconditioners

In general, when L is nonsymmetric, we can take again the symmetric coercive operator S from Definition 3.2.1 and introduce the stiffness matrix of S,

$$\mathbf{S}_h = \Big\{ \langle \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n} , \tag{3.39}$$

as preconditioner for system (3.24). To solve the preconditioned system $\mathbf{S}_h^{-1} \mathbf{L}_h \, \mathbf{c} = \tilde{\mathbf{b}_h}$, one can propose the CG methods in section 2.2, using the $\mathbf{S}_h$-inner product $\langle ., . \rangle_{\mathbf{S}_h}$. As follows from (2.24) and (2.38), the convergence estimates depend on the bounds

$$\lambda_0 = \lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) := \inf\{\mathbf{L}_h \, \mathbf{c} \cdot \mathbf{c} : \ \mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c} = 1\}, \qquad \Lambda = \Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h) := \|\mathbf{S}_h^{-1}\mathbf{L}_h\|_{\mathbf{S}_h} ,$$

defined as in (2.3). Moreover, the convergence factor is determined by the ratio Λ/λ_0.

The mesh independence result here is an extension of (3.27), now using the S-coercivity and S-boundedness relations

$$m\|u\|_S^2 \leq \langle L_S u, u \rangle_S, \quad |\langle L_S u, v \rangle_S| \leq M \|u\|_S \|v\|_S \qquad (u, v \in H_S). \tag{3.40}$$

coming from (3.5).

Proposition 3.3.2 *If the operator L satisfies (3.40), then for any subspace $V_h \subset H_S$*

$$\frac{\Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h)}{\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h)} \leq \frac{M}{m} \tag{3.41}$$

independently of V_h.

Proof. We can in fact verify

$$\Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h) \leq M \ \text{ and } \ \lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) \geq m \, , \tag{3.42}$$

which follows similarly to Proposition 3.3.1. Namely, for arbitrary $\mathbf{c}, \mathbf{d} \in \mathbf{R}^n$, setting $u = \sum\limits_{j=1}^{n} c_j \varphi_j \ \in V_h$ and $z = \sum\limits_{j=1}^{n} d_j \varphi_j \ \in V_h$ in (3.40) yields

$$m \ (\mathbf{S}_h \, \mathbf{c} \cdot \mathbf{c}) \leq \mathbf{L}_h \, \mathbf{c} \cdot \mathbf{c}, \qquad |\mathbf{L}_h \, \mathbf{c} \cdot \mathbf{d}| \leq M \|\mathbf{c}\|_{\mathbf{S}_h} \|\mathbf{d}\|_{\mathbf{S}_h} \tag{3.43}$$

which implies (3.42). ∎

Estimate (2.37) for $\mathbf{S}_h^{-1}\mathbf{L}_h$ then immediately implies

Proposition 3.3.3 *If the operator L satisfies (3.40), then for any subspace $V_h \subset H_S$*

$$\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \leq \frac{M}{m}$$

independently of V_h.

Then (2.24) implies

Corollary 3.3.1 *If (3.40) holds, then the GCG-LS algorithm (2.19) for system (3.26) satisfies*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \left(1 - \left(\frac{m}{M}\right)^2\right)^{1/2} \qquad (k = 1, 2, ..., n) \tag{3.44}$$

independently of V_h.

This holds as well for the GCR and Orthomin methods together with their truncated versions. Further, by (2.36) and (2.38), we have

Corollary 3.3.2 *If (3.40) holds, then the CGN algorithm (2.35) for system (3.26) satisfies*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le 2^{1/k}\, \frac{M - m}{M + m} \qquad (k = 1, 2, ..., n) \tag{3.45}$$

independently of V_h.

Remark 3.3.2 The boundedness of $\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h)$ in h is established in [49] in a more general setting than our S-coercive operators: essentially, the second inequality in (3.5) can be replaced by the two weaker statements

$$\sup_{v \in H_S} \frac{\langle L_S u, v\rangle_S}{\|v\|_S} \ge m\|u\|_S \quad (u \in H_S), \qquad \sup_{u \in H_S} \langle L_S u, v\rangle_S > 0 \quad (v \in H_S). \tag{3.46}$$

However, in contrast to (3.5), these inequalities do not automatically hold in the subspaces V_h with the same constants: the latter has to be assumed to prove the boundedness of $\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h)$.

Concerning the GCG algorithm, of particular interest is the case when $\mathbf{S}_h$ is the symmetric part of $\mathbf{L}_h$:

$$\mathbf{S}_h := \frac{1}{2}(\mathbf{L}_h + \mathbf{L}_h^T), \tag{3.47}$$

which arises from the choice $S := \frac{1}{2}(L + L^*)$ when $D(L) = D(L^*)$. (When $D(L) \neq D(L^*)$, the definition of S requires a more general weak approach, see [66]). Then, see [3], the $\mathbf{S}_h$-adjoint of $\mathbf{S}_h^{-1}\mathbf{L}_h$ is $2\mathbf{I}_h - \mathbf{S}_h^{-1}\mathbf{L}_h$, hence, as discussed in subsection 2.2.1.1, in this case Theorem 2.2.1 is valid with $s = 0$. Therefore the full GCG algorithm reduces to the simple truncated version (2.20), which is one of the main reasons for choosing this preconditioning approach. Here we can simplify the derivation of the convergence factor as follows. Let us decompose $\mathbf{L}_h$ into its symmetric and antisymmetric parts:

$$\mathbf{L}_h = \mathbf{S}_h + \mathbf{Q}_h$$

where $\mathbf{Q}_h := \frac{1}{2}(\mathbf{L}_h - \mathbf{L}_h^T)$, and accordingly, let

$$L_S = I + Q_S \tag{3.48}$$

where I is the identity operator on H_S and Q_S is antisymmetric in H_S. Then $\mathbf{S}_h^{-1}\mathbf{L}_h = \mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h$ where $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is antisymmetric w.r.t. the inner product $\langle .,.\rangle_{\mathbf{S}_h}$, hence (2.26) holds for $C = \mathbf{S}_h^{-1}\mathbf{Q}_h$. Further, it is readily seen that $\mathbf{Q}_h$ is the stiffness matrix of Q_S in V_h, hence, similarly to (3.42), we can obtain the estimate

$$\|\mathbf{S}_h^{-1}\mathbf{Q}_h\|_{\mathbf{S}_h} \le \|Q_S\|.$$

Using (2.26)-(2.28), the GCG-LS(0) algorithm (2.20) for system (3.26) then satisfies

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \frac{\|\mathbf{S}_h^{-1}\mathbf{Q}_h\|_{\mathbf{S}_h}}{\sqrt{1+\|\mathbf{S}_h^{-1}\mathbf{Q}_h\|^2_{\mathbf{S}_h}}} \leq \frac{\|Q_S\|}{\sqrt{1+\|Q_S\|^2}} \qquad (k=1,2,...,n) \tag{3.49}$$

and for the best possible estimate we have asymptotically

$$\limsup \left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \frac{\|Q_S\|}{1+\sqrt{1+\|Q_S\|^2}}\,, \tag{3.50}$$

where both ratios are independent of V_h.

3.3.3 Nonsymmetric problems: nonsymmetric equivalent preconditioners

Let us consider a nonsymmetric preconditioning operator N for equation (3.23). We assume that N is S-bounded and S-coercive, i.e. $N \in BC_S(H)$ in the sense of Definition 3.2.1, for the same symmetric operator S as is L. Then we introduce the stiffness matrix of N_S,

$$\mathbf{N}_h = \Big\{\langle N_S\varphi_i, \varphi_j\rangle_S\Big\}_{i,j=1}^{n}, \tag{3.51}$$

as preconditioner for the discretized system (3.24). To solve the preconditioned system

$$\mathbf{N}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h} \tag{3.52}$$

(with $\tilde{\mathbf{b}_h} = \mathbf{N}_h^{-1}\mathbf{b}_h$), we apply the CGN method as described in subsection 2.2.2, using the $\mathbf{S}_h$-inner product $\langle .,. \rangle_{\mathbf{S}_h}$. By (2.36), this algorithm converges as

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq 2^{1/k}\,\frac{\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h)-1}{\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h)+1} \qquad (k=1,2,...,n). \tag{3.53}$$

In the convergence analysis of nonsymmetric preconditioners, we must distinguish between the bounds of L and N, i.e., (3.40) is replaced by

$$\begin{aligned} m_L\|u\|_S^2 &\leq \langle L_S u, u\rangle_S, \quad |\langle L_S u, v\rangle_S| \leq M_L\|u\|_S\|v\|_S, \\ m_N\|u\|_S^2 &\leq \langle N_S u, u\rangle_S, \quad |\langle N_S u, v\rangle_S| \leq M_N\|u\|_S\|v\|_S \end{aligned} \tag{3.54}$$

for all $u, v \in H_S$.

Proposition 3.3.4 *If the operators L and N satisfy (3.54), then for any subspace $V_h \subset H_S$*

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \leq \frac{M_L M_N}{m_L m_N} \tag{3.55}$$

independently of V_h.

Proof. Let $\mathbf{c} \in \mathbf{R}^n$ be arbitrary, $\mathbf{d} := \mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}$, i.e. $\mathbf{N}_h\mathbf{d} = \mathbf{L}_h\mathbf{c}$, further, let $u = \sum_{j=1}^{n} c_j\varphi_j \in V_h$ and $z = \sum_{j=1}^{n} d_j\varphi_j \in V_h$. Then $m_L\|u\|_S^2 \leq \langle L_S u, u\rangle_S = \mathbf{L}_h\,\mathbf{c}\cdot\mathbf{c} = \mathbf{N}_h\,\mathbf{d}\cdot\mathbf{c} = \langle N_S z, u\rangle_S \leq \|N_S z\|_S\|u\|_S$, hence $m_L\|u\|_S \leq \|N_S z\|_S \leq M_N\|z\|_S$, and by exchanging L and N resp. u and z, we similarly obtain $m_N\|z\|_S \leq \|L_S u\|_S \leq M_L\|u\|_S$. Hence, altogether,

$$\frac{m_L}{M_N} \leq \frac{\|\mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}\|_{\mathbf{S}_h}}{\|\mathbf{c}\|_{\mathbf{S}_h}} = \frac{(\mathbf{S}_h\,\mathbf{d}\cdot\mathbf{d})^{1/2}}{(\mathbf{S}_h\,\mathbf{c}\cdot\mathbf{c})^{1/2}} = \frac{\|z\|_S}{\|u\|_S} \leq \frac{M_L}{m_N}\,. \qquad \blacksquare \tag{3.56}$$

The above result is a direct extension of Proposition 3.3.3: the latter is recovered by the case $N = S$, for which $M_N = m_N = 1$. However, if both N and L have a large ratio M/m, then the upper bound in (3.55) becomes large even if N is an accurate approximation of L. In this case it is more useful to involve the difference of N and L in the bound:

Proposition 3.3.5 *If the operators L and N satisfy (3.54), then for any subspace $V_h \subset H_S$*

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \leq \Big(1 + \frac{m_L + m_N}{2m_L m_N} \, \|L_S - N_S\|\Big)^2$$

independently of V_h.

Proof. We follow the proof of Proposition 3.3.4. Let $\mathbf{c}, \mathbf{d} \in \mathbf{R}^n$ and $u, z \in V_h$ be as therein, $\mathbf{k} := \mathbf{d} - \mathbf{c}$ and $h := \sum\limits_{j=1}^{n} k_j \varphi_j = z - u$. Then $m_N \|h\|_S^2 \leq \langle N_S h, h\rangle_S = \mathbf{N}_h \, \mathbf{k} \cdot \mathbf{k} = \mathbf{N}_h \, \mathbf{d} \cdot \mathbf{k} - \mathbf{N}_h \, \mathbf{c} \cdot \mathbf{k} = (\mathbf{L}_h - \mathbf{N}_h) \, \mathbf{c} \cdot \mathbf{k} = \langle (L_S - N_S)u, h\rangle_S \leq \|L_S - N_S\| \, \|u\|_S \|h\|_S$. Hence $\|z\|_S \leq \|u\|_S + \|h\|_S \leq \|u\|_S \Big(1 + \frac{1}{m_N}\|L_S - N_S\| \Big)$. Exchanging L and N resp. u and z, we obtain $\|u\|_S \leq \|z\|_S \Big(1 + \frac{1}{m_L}\|L_S - N_S\| \Big)$. In view of (3.56), the obtained bounds on the ratio $\|z\|_S / \|u\|_S$ imply

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \leq \Big(1 + \frac{1}{m_N}\|L_S - N_S\| \Big)\Big(1 + \frac{1}{m_L}\|L_S - N_S\| \Big) \leq \Big(1 + \frac{m_L + m_N}{2m_L m_N} \, \|L_S - N_S\|\Big)^2$$

where the second estimate uses the arithmetic-geometric mean inequality. ■

Consequently, by (3.53), the CGN algorithm (2.35) for system (3.52) converges with a ratio bounded independently of V_h.

4 Compact-equivalent operators and superlinear convergence

As shown in Chapter 3, equivalence of operators provides mesh independent linear convergence of CG iterations. Our further goal is to complete the above results by showing that for a proper class of problems, superlinear convergence is also mesh independent. This means that a bound on the rate of superlinear convergence is given in the form of a sequence which is mesh independent and is determined only by the underlying operators. To describe the suitable class of problems, we introduce the notion of compact-equivalent operators which expresses that preconditioning one of them with the other yields a compact perturbation of the identity. This notion and the convergence results provide a refinement of the case of equivalent operators: roughly speaking, if the two operators (the original and preconditioner) are equivalent then the corresponding PCG method provides mesh independent linear convergence, whereas if the two operators are compact-equivalent then the PCG method provides mesh independent superlinear convergence.

Results for mesh independent superlinear convergence are essentially available for FEM type discretizations only, even more so than has been pointed out for linear convergence in the previous chapter. The general results below concern Galerkin discretizations in Hilbert space, based on [14, 16], and will be applied to FEM discretizations of elliptic problems in later parts of the paper.

4.1 Compact-equivalent operators in Hilbert space

In this chapter we shall consider compact operators in Hilbert space, i.e., linear operators C such that the image (Cv_i) of any bounded sequence (v_i) contains a convergent subsequence. Recall that the eigenvalues of a compact self-adjoint operator cluster at the origin.

Definition 4.1.1 (i) We call $\lambda_i(F)$ $(i = 1, 2, \dots)$ the *ordered eigenvalues* of a compact self-adjoint linear operator F in H if each of them is repeated as many times as its multiplicity and $|\lambda_1(F)| \geq |\lambda_2(F)| \geq ...$

(ii) The *singular values* of a compact operator C in H are

$$s_i(C) := \lambda_i(C^*C)^{1/2}, \qquad (i = 1, 2, \dots)$$

where $\lambda_i(C^*C)$ are the ordered eigenvalues of C^*C. In particular, if C is self-adjoint then $s_i(C) = |\lambda_i(C)|$.

It follows that the singular values of a compact operator cluster at the origin.

Superlinear convergence results will require a certain refinement of the notion of equivalence of operators, which we introduce in the class defined in Definition 3.2.1:

Definition 4.1.2 Let L and N be S-bounded and S-coercive operators in H. We call L and N *compact-equivalent in H_S* if

$$L_S = \mu N_S + Q_S \tag{4.1}$$

for some constant $\mu > 0$ and compact operator $Q_S \in B(H_S)$.

Remark 4.1.1 (i) It follows in a straightforward way that the property of compact-equivalence is an equivalence relation.

(ii) In the special case $R(L) \subset R(N)$, compact-equivalence of L and N means that $N^{-1}L$ is a compact perturbation of constant times the identity in the space H_S. Indeed, it is easy to see that here $N^{-1}L = N_S^{-1}L_S\big|_{D(L)}$, and, by definition, the latter is the perturbation of μI with the operator $N_S^{-1}Q_S\big|_{D(L)}$, which is compact since N_S^{-1} is bounded. (In the general case the 'weakly preconditioned' form $N_S^{-1}L_S$ is also a compact perturbation.)

4.2 A characterization of compact-equivalent elliptic operators

We shall now characterize compact-equivalence for elliptic operators. For this, let us consider the class of coercive elliptic operators defined in subsection 3.2.2. That is, let us take two operators as in (3.10):

$$\begin{aligned} L_1 u &\equiv -\mathrm{div}\,(A_1\,\nabla u) + \mathbf{b}_1 \cdot \nabla u + c_1 u \qquad \text{for } \ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_{A_1}} + \alpha_1 u_{|\Gamma_N} = 0, \\ L_2 u &\equiv -\mathrm{div}\,(A_2\,\nabla u) + \mathbf{b}_2 \cdot \nabla u + c_2 u \qquad \text{for } \ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_{A_2}} + \alpha_2 u_{|\Gamma_N} = 0 \end{aligned}$$

where we assume that L_1 and L_2 satisfy Assumptions 3.2.2.1. Then by Proposition 3.2.2, the operators L_1 and L_2 are S-bounded and S-coercive in $L^2(\Omega)$, where S is the symmetric operator from (3.11). The corresponding energy space $H_S = H_D^1(\Omega)$ with S-inner product has been given in (3.12). Then it makes sense to study the compact-equivalence of L_1 and L_2 in $H_D^1(\Omega)$, and the following result is available:

Proposition 4.2.1 *Let the elliptic operators L_1 and L_2 satisfy Assumptions 3.2.2.1. Then L_1 and L_2 are compact-equivalent in $H^1_D(\Omega)$ if and only if their principal parts coincide up to some constant $\mu > 0$, i.e. $A_1 = \mu A_2$.*

The sufficiency of the principal part condition is partly verified in [52], and the whole proposition is proved in [16].

4.3 Mesh independent superlinear convergence in Hilbert space

Now we present mesh independent superlinear convergence estimates in the case of compact-equivalent preconditioning. For simplicity, in what follows, we will consider compact-equivalence with $\mu = 1$ in (4.1). This is clearly no restriction, since if a preconditioner N_S satisfies $L_S = \mu N_S + Q_S$ then we can consider the preconditioner μN_S instead.

4.3.1 Symmetric compact-equivalent preconditioners

Let us consider operators L and S as in Definition 3.2.1, and assume in addition that L and S are compact-equivalent with $\mu = 1$. In this special case (4.1) holds with $N_S = I$:

$$L_S = I + Q_S \tag{4.2}$$

with a compact operator Q_S. We apply the stiffness matrix $\mathbf{S}_h$ of S, see (3.39), as preconditioner for system (3.24). By (4.2), letting

$$\mathbf{Q}_h = \left\{ \langle Q_S \varphi_i, \varphi_j \rangle_S \right\}_{i,j=1}^{n}, \tag{4.3}$$

the preconditioned system (3.26) takes the form

$$(\mathbf{I}_h + \mathbf{S}_h^{-1} \mathbf{Q}_h)\, \mathbf{c} = \tilde{\mathbf{b}_h} \tag{4.4}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix.

We have the estimates (2.33) and (2.39) for the GCG-LS algorithm (2.19) and the CGN algorithm (2.35), respectively. In order to have mesh independent bounds for these estimates in the case $A = \mathbf{S}_h^{-1}\mathbf{L}_h$, we first observe that (3.42) provides the bound $\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h) \geq m$, hence the task left is to find bounds for the sums of eigenvalues in these expressions in the case $C = \mathbf{S}_h^{-1}\mathbf{Q}_h$. Using Definition 4.1.1, the following two results give the required bounds in terms of the operator Q_S.

Proposition 4.3.1 [14]. *Let H be a complex Hilbert space. If Q_S is a normal compact operator in H_S and the matrix $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-normal, then*

$$\sum_{i=1}^{k} \left| \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h) \right| \leq \sum_{i=1}^{k} \left| \lambda_i(Q_S) \right| \qquad (k = 1, 2, \ldots, n).$$

If H is a real Hilbert space (as it is throughout this paper) then H and H_S can be extended to a complex Hilbert space, as well as Q_S can be extended to be defined on the complex space, in an obvious way by linearity.

Then (2.33) implies

Corollary 4.3.1 *Under the conditions of Proposition 4.3.1, the GCG-LS algorithm (2.19) for system (4.4) yields*

$$\left(\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}}\right)^{1/k} \leq \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{4.5}$$

where

$$\varepsilon_k := \frac{2}{km} \sum_{j=1}^{k} \left|\lambda_j(Q_S)\right| \to 0 \quad \textit{as} \quad k \to \infty \tag{4.6}$$

and ε_k is a sequence independent of V_h.

The most important special case here is symmetric part preconditioning, when both normality assumptions are readily satisfied. In fact, Q_S is antisymmetric in H_S and similarly, $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-antisymmetric, see (3.48) and afterwards. Then the GCG-LS algorithm reduces to the truncated GCG-LS(0) version (2.20). Moreover, the $\mathbf{L}_h$-norm equals the $\mathbf{S}_h$-norm in (4.5) and $m = 1$ in (4.6).

In the general case without normality, we have the following bounds for (2.39):

Proposition 4.3.2 [16]. *Any compact operator Q_S in H_S satisfies the following relations:*

$$(a) \qquad \sum_{i=1}^{k} \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\, \mathbf{S}_h^{-1}\mathbf{Q}_h) \leq \sum_{i=1}^{k} s_i(Q_S)^2 \qquad (k = 1, 2, \ldots, n),$$

$$(b) \qquad \sum_{i=1}^{k} \left|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T + \mathbf{S}_h^{-1}\mathbf{Q}_h)\right| \leq \sum_{i=1}^{k} \left|\lambda_i(Q_S^* + Q_S)\right| \qquad (k = 1, 2, \ldots, n),$$

Then (2.39) implies

Corollary 4.3.2 *If L and S are compact-equivalent with $\mu = 1$, then the CGN algorithm (2.35) for system (4.4) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{4.7}$$

where

$$\varepsilon_k := \frac{2}{km^2} \sum_{i=1}^{k} \Big(\left|\lambda_i(Q_S^* + Q_S)\right| + \lambda_i(Q_S^* Q_S) \Big) \to 0 \quad \textit{as} \quad k \to \infty \tag{4.8}$$

and ε_k is a sequence independent of V_h.

Remark 4.3.1 A self-adjoint compact operator Q_S is called a Hilbert-Schmidt operator if $|||Q_S|||^2 \equiv \sum\limits_{i=1}^{\infty} \lambda_i(Q_S)^2 < \infty$. (Here $|||S^{-1}Q|||$ is called the Hilbert-Schmidt norm of $S^{-1}Q$.) In this case it is proved in [65] that

$$\sum_{i=1}^{k} \left|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h)\right|^2 \leq |||Q_S|||^2 \qquad (k = 1, 2, \ldots, n).$$

Then (2.18) yields

$$\left(\frac{\|e_k\|_A}{\|e_0\|_A}\right)^{1/k} \leq \left(\frac{3}{2k}\right)^{1/2} |||Q_S|||\,, \tag{4.9}$$

which shows a superlinear rate of convergence of magnitude $O(k^{-1/2})$.

4.3.2 Nonsymmetric compact-equivalent preconditioners

Now let N be a nonsymmetric S-bounded and S-coercive operator which is compact-equivalent to L with $\mu = 1$, i.e., (4.1) becomes

$$L_S = N_S + Q_S\,. \tag{4.10}$$

We apply the stiffness matrix $\mathbf{N}_h$ of N_S, see (3.51), as preconditioner for the discretized system (3.24). Since N is nonsymmetric, in order to define an inner product on $\mathbf{R}^n$ we preserve the stiffness matrix of S on V_h, i.e. using (3.39) we endow $\mathbf{R}^n$ with the $\mathbf{S}_h$-inner product $\langle \mathbf{c}, \mathbf{d}\rangle_{\mathbf{S}_h} := \mathbf{S}_h\,\mathbf{c}\cdot\mathbf{d}$ as earlier. Then the $\mathbf{S}_h$-adjoint of $\mathbf{N}_h^{-1}\mathbf{L}_h$ is $\mathbf{S}_h^{-1}\mathbf{L}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h$, hence we apply the CGN algorithm (2.35) with $A = \mathbf{N}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T\mathbf{N}_h^{-T}\mathbf{S}_h$.

Using (4.3) and (4.10), the preconditioned system (3.52) takes the form

$$(\mathbf{I}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \hat{\mathbf{b}_h} \tag{4.11}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix. The CGN algorithm (2.35) then provides

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{2\mu_h}{k}\sum_{i=1}^{k}\Big(\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h) + \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h)\Big) \tag{4.12}$$

($k = 1, 2, ..., n$), where (2.39) is used but the value $\mu_h := \|(\mathbf{N}_h^{-1}\mathbf{L}_h)^{-1}\|^2_{\mathbf{S}_h}$ is now preserved instead of the estimate (2.37). Again, our goal is to give a bound on (4.12) that is independent of V_h. Since (3.56) implies $\mu_h \le (M_N/m_L)^2$ with the bounds in (3.54), our task is again to find bounds for the sum of eigenvalues in (4.12).

Proposition 4.3.3 [16]. *Let L and N be S-bounded and S-coercive operators, and let (4.10) hold with a compact operator Q_S on H_S. Let $s_i(Q_S)$ $(i = 1, 2, \dots)$ denote the singular values of Q_S. Then the following relations hold:*

(*a*)
$$\sum_{i=1}^{k}\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h) \le \frac{1}{m_N^2}\sum_{i=1}^{k}s_i(Q_S)^2 \qquad (k = 1, 2, \dots, n),$$

(*b*)
$$\sum_{i=1}^{k}\big|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)\big| \le \frac{2}{m_N}\sum_{i=1}^{k}s_i(Q_S) \qquad (k = 1, 2, \dots, n),$$

Corollary 4.3.3 *For compact-equivalent operators L and N, the CGN algorithm (2.35) for system (4.11) yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \varepsilon_k \qquad (k = 1, 2, ..., n) \tag{4.13}$$

$$\textit{where} \quad \varepsilon_k = \frac{2M_N^2}{km_L^2}\sum_{i=1}^{k}\Big(\frac{2}{m_N}\,s_i(Q_S) + \frac{1}{m_N^2}\,s_i(Q_S)^2\Big) \to 0 \qquad (\textit{as } k \to \infty) \tag{4.14}$$

and ε_k is a sequence independent of V_h.

5 Spectrally equivalent preconditioners for symmetric elliptic equations

In this chapter we consider some methods based on the concept of spectral equivalence to construct preconditioners for elliptic problems with variable coefficients. Both second and fourth order scalar equations are discussed, whereas second order systems will be considered in Chapter 6. We also refer to some early use of equivalent operators in iterative solution methods.

5.1 Second order equations

Let us consider the special case when the operator (3.10) is symmetric:

$$Lu \equiv \ -\mathrm{div}\,(A\,\nabla u) + cu \qquad \text{for} \ \ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \tag{5.1}$$

and define another symmetric elliptic operator like in (3.11) as preconditioning operator:

$$Su \equiv \ -\mathrm{div}\,(G\,\nabla u) + \sigma u \qquad \text{for} \ \ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{5.2}$$

where both operators satisfy Assumptions 3.2.2.2.

Being symmetric special cases of (3.10) and (3.11), Proposition 3.3.1 holds for the operators (5.1)-(5.2). Hence by (3.28), for FEM discretizations, the preconditioned CG algorithm (2.4) for system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h}$ (where $\tilde{\mathbf{b}_h} = \mathbf{S}_h^{-1}\mathbf{b}_h$) converges with a ratio independent of V_h. Similar results on FDM discretizations are available on rectangular domains for certain special cases of (5.1)-(5.2), cited later. In what follows, we are interested in efficient applications of such preconditioning, i.e. when systems with $\mathbf{S}_h$ are easier to solve than systems with $\mathbf{L}_h$, for both FEM and FDM discretizations.

For ease of presentation, we will mainly deal with problems with only the principal parts of (5.1) and with Dirichlet boundary conditions:

$$\begin{cases} Lu := -\mathrm{div}\,\big(A\nabla u\big) = g \\ \ u_{|\partial\Omega} = 0. \end{cases} \tag{5.3}$$

We note that most of the practical applications of spectrally equivalent preconditioning has been developed for problems like (5.3). This operator is a special case of (5.1), hence we assume that $A \in C^1(\overline{\Omega}, \mathbf{R}^{d\times d})$ is a symmetric and uniformly positive definite matrix function. The latter means that there exist constants $p_1 \ge p_0 > 0$ such that

$$p_0\ |\xi|^2 \le A(x)\xi\cdot\xi \le p_1\ |\xi|^2 \qquad (x \in \overline{\Omega},\ \xi \in \mathbf{R}^d). \tag{5.4}$$

An important special case of (5.3) arises for the linearization of a nonlinear problem of the form

$$\begin{cases} -\mathrm{div}\,\Big(a(|\nabla u|^2\,\nabla u\Big) = f \\ \ u_{|\partial\Omega} = 0, \end{cases} \tag{5.5}$$

where $a : \mathbf{R}^+ \to \mathbf{R}^+$ is a scalar C^1 function satisfying

$$0 < p_0 \le a(r) \le p_1, \qquad 0 < p_0 \le \tfrac{d}{dr}\Big(a(r^2)r\Big) \le p_1 \qquad (r \ge 0) \tag{5.6}$$

for some constants $p_1 \ge p_0 > 0$. Then the weak form of (5.5) can be written as $F(u) = 0$, where the operator F is of the type (3.30) on $H_S = H_0^1(\Omega)$ with inner product $[u, v] = \int_\Omega \nabla u\cdot\nabla v$, mentioned in Remark 3.3.1. The Newton linearization of (5.5) leads to linear elliptic problems of the type (5.3), and this will be discussed in the next subsection.

In general, suitable preconditioning operators for problem (5.3) are proposed with diagonal coefficients:

$$Su := -\mathrm{div}\,\big(a(x)\nabla u\big) \tag{5.7}$$

for u also satisfying $u_{|\partial\Omega} = 0$, where $a \in C^1(\overline{\Omega}, \mathbf{R})$ is some scalar function also bounded between two positive constants. We will refer to possible extensions to more general problems. Requiring a diagonal coefficient in (5.7) implies that for a wide class of discretizations, the matrices of the auxiliary linear systems have favourable properties such as the M-matrix property [36, 74].

5.1.1 The Laplacian and related preconditioning operators

The simplest preconditioning operator in this context is

$$Su := -\Delta u, \tag{5.8}$$

which is historically important as being the first application of the equivalent operator idea for discretized elliptic problems. In the papers [42] and [58], the centered finite difference discretization of (5.3) for scalar A was studied on a rectangle, and the same discretization of the minus Laplacian was proposed as preconditioner. This preconditioning was constructed in the context of simple (or Richardson) iterations, and was later termed as D'yakonov-Gunn iteration. We note that the Laplacian preconditioner in an infinite-dimensional setting has first been applied in [39], whereas various modifications of the D'yakonov-Gunn iteration have been given in [33, 37, 59, 87, 99]).

The implementation of the Laplacian preconditioner is based on the variety of available efficient solvers. The above-mentioned earlier applications were based on fast direct solvers, see e.g. [94] for a summary of classical fast methods and [88, 96] for later applications, including the extension of fast solvers from rectangular domains to general ones via the fictitious domain method, and see e.g. [51] for spectral methods. More recent implementations involve multigrid or multilevel methods [21, 22, 61] for solving the auxiliary problems.

Concerning the rate of convergence, mesh independence of the spectral condition number (i.e., spectral equivalence) has been proved in [40, 42, 58, 85] for FDM discretizations and in [18] for FEM discretizations. For the latter, the mesh independence result is a special case of subsection 3.3, i.e., in our setting, (5.4) yields

$$p_0 \int_\Omega |\nabla u|^2 \le \int_\Omega A\ \nabla u \cdot \nabla u \le p_1 \int_\Omega |\nabla u|^2 \qquad (u \in H_0^1(\Omega)), \tag{5.9}$$

which is nothing but (3.25) since the operator $S = -\Delta$ satisfies $H_S = H_0^1(\Omega)$. Hence Proposition 3.3.1 is valid with $M = p_1$ and $m = p_0$, and by (3.28), the CG algorithm (2.4) for system (3.26) converges as

$$\left(\frac{\|e_k\|_{\mathbf{L}_h}}{\|e_0\|_{\mathbf{L}_h}}\right)^{1/k} \le 2^{1/k}\ \frac{\sqrt{p_1} - \sqrt{p_0}}{\sqrt{p_1} + \sqrt{p_0}} \qquad (k = 1, 2, ..., n) \tag{5.10}$$

independently of V_h.

As an important application of the above, let us consider the FEM discretization of the nonlinear problem (5.5), solved by an outer Newton iteration. Based on Remark 3.3.1, if u_n is the nth Newton iterate then the correction term $p_n \in V_h$ is found by solving the linearized problem (3.38), in which we now set $[h, v] = \int_\Omega \nabla h \cdot \nabla v$. This problem is the FEM discretization of a linear elliptic problem (5.3), in which the coefficient matrix is the Jacobian of the nonlinearity at u_n, that is,

$$A = \ a(|\nabla u|^2) \cdot I + 2a'(|\nabla u|^2)\, \nabla u_n^T \cdot \nabla u_n\,. \tag{5.11}$$

Let us apply the CG algorithm as an inner iteration for this linearized problem. Here assumption (5.6) means that (3.31) holds with $M = p_1$ and $m = p_0$, hence by (3.35), the matrix A satisfies (5.9). Therefore, as stated above, Proposition 3.3.1 and the mesh independent convergence estimate (3.28) are valid with $M = p_1$ and $m = p_0$ for the inner CG iteration. For details, see [18] and [87] where this approach is used. In fact, one can consider mixed boundary conditions $u_{|\Gamma_D} = 0$, $\frac{\partial u}{\partial \nu}_{|\Gamma_N} = 0$ in exactly the same framework [18], and pure Neumann boundary conditions in the subspace $\{u \in H^1(\Omega) : \int_\Omega u = 0\}$ as in [87].

Let us now turn to more general preconditioning operators. Some direct extensions of (5.8) followed the D'yakonov-Gunn-idea along with the development of fast solvers: Helmholtz preconditioners [37], scaled Laplacians [55] and in particular, operators with separable coefficients [25, 57]:

$$Su = -\sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left(a_i(x_i) \frac{\partial u}{\partial x_i} \right)$$

where the functions $a_i(x_i)$ are separable approximations of the original coefficients. The implementation is based on the fact that the scope of the above-mentioned Laplacian solvers includes general separable problems as well. In particular, fast direct solvers are available such as [88], and have been developed on general domains as well using the fictitious domain approach [29, 87]. We note that the development of multilevel and multigrid methods have in general decreased the need for the use of preconditioned conjugate gradient methods when solving (5.3). Still for complicated domains or coefficients, it can be more efficient to use a PCG iteration with the above preconditioning operators, and to apply multigrid only to the inner auxiliary Poisson or Helmholtz problems, than to apply multigrid directly to the original problem (5.3) with variable coefficients.

When applying Helmholtz preconditioners for problems with Laplacian principal part and variable zeroth-order term as in [37], one can even verify mesh independent superlinear convergence. Namely, it is proved in [65] that such preconditioning leads to the estimate (4.9).

Further generalizations of the Laplacian preconditioning operator are operators with a diagonal coefficient like in (5.7). This may be motivated by ill-conditioned problems where the bounds p_1 and p_0 in (5.4) are much separated. We consider a particular preconditioning operator of this type in the next subsection.

5.1.2 Piecewise constant coefficient preconditioning operators

When the spectral bounds of A are much separated, i.e. p_1/p_0 is large, the convergence factor in (3.28) deteriorates. For such problems a suitable generalization of the discrete Laplacian to 'local Laplacians' is a preconditioning operator with piecewise constant coefficient [9]. Formally we write

$$Su := -\text{div}\,\big(w(x)\nabla u\big) \tag{5.12}$$

where w is a piecewise constant function, that is, the domain Ω is decomposed in subdomains $\Omega_i \quad (i = 1, \ldots, s)$, and for all i,

$$w_{n\,|\Omega_i} \equiv c_i \ > 0. \tag{5.13}$$

In fact we only have to use the corresponding inner product

$$\langle u,v\rangle_S = \int_\Omega w\,\nabla u\cdot\nabla v \qquad (u,v\in H_0^1(\Omega)).$$

The estimate of the condition number follows from the proper choice of the constants c_i. Let us introduce the spectral bounds m_i and M_i of A relative to Ω_i, i.e. such that $\sigma(A(x))\subset[m_i,M_i]$ for all $x\in\Omega_i$. Then one should choose c_i between m_i and M_i. The definition of m_i and M_i implies that

$$\min_i \frac{m_i}{c_i}\int_\Omega w|\nabla h|^2 \le \int_\Omega A\;\nabla h\cdot\nabla h \le \max_i \frac{M_i}{c_i}\int_\Omega w|\nabla h|^2 \qquad (h\in H_0^1(\Omega)),$$

i.e., introducing $m:=\min\limits_i m_i/c_i$ and $M:=\max\limits_i M_i/c_i$, estimate (3.25) holds. In particular, if c_i is some (arithmetic, geometric or harmonic) mean of m_i and M_i, then $M/m=\max\limits_i M_i/m_i$. Altogether, using these values of M and m, Proposition 3.3.1 and the mesh independent convergence estimate (3.28) are valid

An important application arises for the nonlinear problem (5.5) when the ratio of the bounds p_1 and p_0 in (5.6) is large. Similarly to the previous subsection, we consider the FEM discretization of (5.5), solved by an outer Newton iteration, further, we use the results given in Remark 3.3.1. If u_n is the nth Newton iterate then the correction term $p_n\in V_h$ is found by solving the linearized problem (3.38), in which we now set $[h,v]=\int_\Omega\nabla h\cdot\nabla v$. This is the FEM solution of an elliptic problem of the type (5.3) with coefficient matrix (5.11). Estimate (3.35) then yields

$$\int_\Omega p(|\nabla u_n|^2)\,|\nabla h|^2 \le \int_\Omega A\;\nabla h\cdot\nabla h \le \int_\Omega q(|\nabla u_n|^2)\,|\nabla h|^2 \qquad (h\in V_h),$$

hence for relative spectral bounds m_i and M_i on Ω_i, we can define

$$m_i := \inf_{\Omega_i} p(|\nabla u_n|^2), \qquad M_i := \sup_{\Omega_i} q(|\nabla u_n|^2).$$

It is also shown in [9] that prescribed condition numbers M/m can be achieved via a suitable recursive definition of the subdomains in a form

$$\Omega_i := \{x\in\Omega:\ r_{i-1}\le|\nabla u_n(x)|<r_i\} \qquad (i=1,...,s)$$

with prescribed ratios r_i/r_{i-1}, which reduces the conditioning analysis to the scalar functions p and q. In practice favourable condition numbers have been achieved with 6 or 9 subdomains even for almost singular problems, see this and further details in [9, 68].

The corresponding stiffness matrix $\mathbf{S}_h$ arises by modifying the discrete Laplacian such that the corresponding blocks are multiplied by the constants c_i. In the case of few subdomains, this structure property only slightly increases the complexity of a Laplacian solver. The solution of the auxiliary linear systems can then rely on various well-developed methods, including ones designed especially for piecewise constant coefficient problems [21, 41, 54]. We stress here that algebraic multilevel preconditioners [21] for the auxiliary problems can produce an optimal condition number independent of the variation of w.

5.2 Fourth order equations

An analogue of the Laplacian preconditioner occurs for fourth order problems when the biharmonic operator is used as preconditioning operator for fourth order variable coefficient equations.

Problems of the latter kind occur in the Newton linearization of nonlinear fourth order equations, which model the elasto-plastic bending of plates (see e.g. [80]). In this section we follow the setting of [50].

A general form of such problems is

$$\left\{ \begin{array}{l} \operatorname{div}^2 \left(g(E(D^2u))\ \tilde{D}^2 u \right) \ = \ \alpha(x) \\ u_{|\partial\Omega} = \frac{\partial u}{\partial \nu}|_{\partial\Omega} = 0 \end{array} \right. \tag{5.14}$$

(using clamped boundary conditions for simplicity), where $\Omega \subset \mathbf{R}^2$ and the following notations are used: D^2u is the Hessian, $\tilde{D}^2u := \frac{1}{2}(D^2u + \Delta u \cdot I)$ where I is the 2×2 identity matrix, $E(D^2u) = \frac{1}{2}(|D^2u|^2 + (\Delta u)^2)$ where $|A|^2 = A : A$ and the latter means elementwise matrix product, and the scalar material function $g \in C^1(\mathbf{R}^+)$ satisfies

$$0 < \mu_1 \le g(r) \le \mu_2\,, \quad 0 < \mu_1 \le \tfrac{d}{dr}\Big(g(r^2)r\Big) \le \mu_2 \tag{5.15}$$

with suitable constants $\mu_1, \mu_2 > 0$.

The weak form of (5.14) can be written as $F(u) = 0$, where the operator F falls into the type of (3.30), discussed in Remark 3.3.1. Here $H_S = H_0^2(\Omega)$ with inner product satisfying

$$\langle u, v\rangle_S := \int_\Omega D^2u \,:\, D^2v = \int_\Omega \tilde{D}^2u \,:\, D^2v \qquad (u, v \in H_0^2(\Omega)), \tag{5.16}$$

further, (3.31) holds with constants $m = \mu_1$ and $M = \mu_2$. The FEM discretization and Newton linearization of such equations lead to problems of the form (3.38), where u_n is the nth Newton iterate and the correction term $p_n \in V_h$ is looked for. Now problem (3.38) is the FEM discretization of the 4th order linear elliptic BVP

$$L\,p_n \equiv \operatorname{div}^2 \Big(\mathcal{K}\ \tilde{D}^2 p_n \Big) \ = \ r_n \tag{5.17}$$

with the array

$$\mathcal{K} \equiv \ g(E(D^2u_n)) \cdot I + 2g'(E(D^2u_n))\,(\tilde{D}^2u_n \times \tilde{D}^2u_n),$$

where r_n is the residual at u_n. Here, for simplicity, we have fixed n and did not indicate the dependence of $\mathcal{K}$ (and L) on n.

Let us introduce the biharmonic preconditioning operator

$$S = \Delta^2$$

for problem (5.17) with the same boundary conditions as in (5.14). Its energy space is $H_S = H_0^2(\Omega)$, whose standard inner product is (5.16). Then estimate (3.35) is valid with $m = \mu_1$ and $M = \mu_2$, which yields

$$\mu_1 \int_\Omega |D^2h|^2 \le \int_\Omega \mathcal{K}\ \tilde{D}^2h \,:\, D^2h \le \mu_2 \int_\Omega |D^2h|^2 \qquad (h \in H_0^2(\Omega)),$$

i.e., estimate (3.25) holds in $H_0^2(\Omega)$. Hence Proposition 3.3.1 and the mesh independent convergence estimate (3.28) are valid with $m = \mu_1$ and $M = \mu_2$.

The auxiliary biharmonic problems can be solved by various fast solvers (e.g. [47, 79]) or alternatively, using a mixed variable variational formulation, e.g. [31].

6 Separate displacement preconditioning for symmetric elliptic systems

In this chapter we consider symmetric elliptic systems instead of a single symmetric equation. An important advantage of the equivalent operator idea is that one can define independent operators for the preconditioner, and thereby reduce the size of auxiliary systems to that of one elliptic equation.

The most important mathematical model that leads to such a system is the elasticity problem in three dimensions, and therefore we use this model to present the preconditioning with independent operators (called separate displacements in this context). The description of an elastic body in structural mechanics leads to an elliptic system of three equations. There are several models that describe these equations, formulated in terms of the displacement or deformation [28, 35, 82]. We consider a model that involves a slight (i.e. material but not geometric) nonlinearity, see the monograph [82] for the foundation of this model and [28] and the references therein for further discussion. The Newton linearization of this model leads to a coupled 3D linear system which contains the standard linear elasticity system as a special case. We give a brief derivation of this system and then present a preconditioning method based on separate displacements. The detailed development of this method is given in [5, 27] for linear elasticity and in [13] in the nonlinear case. We rely on [13] and refer the reader there whenever more details are required.

6.1 Problem formulation and reduction to linear systems

The basic system of equations is

$$\left\{ \begin{aligned} -\operatorname{div} \sigma_i &= \varphi_i(x) \quad \text{in } \Omega \\ \sigma_i \cdot \nu &= \gamma_i(x) \quad \text{on } \Gamma_N \\ u_i &= 0 \qquad\;\; \text{on } \Gamma_D \end{aligned} \right\} \qquad (i = 1, 2, 3) \tag{6.1}$$

where $\Omega \subset \mathbf{R}^3$ is a bounded domain, $x \in \Omega$ is the space variable, $\sigma_i = (\sigma_{i1}, \sigma_{i2}, \sigma_{i3})$ $(i = 1, 2, 3)$ is the ith row of the stress tensor $\sigma : \Omega \to \mathbf{R}^{3\times3}$, the functions $\varphi : \Omega \to \mathbf{R}^3$ and $\gamma : \Gamma_N \to \mathbf{R}^3$ describe the body and boundary force vectors, respectively, further, $\partial\Omega = \Gamma_N \cup \Gamma_D$ is a disjoint measurable subdivision and $\Gamma_D \neq \emptyset$. The body is clamped on Γ_D.

Problem (6.1) can be formulated as a second order system in terms of the displacement vector $\mathbf{u} : \Omega \to \mathbf{R}^3$ and the corresponding strain tensor $\varepsilon = \varepsilon(\mathbf{u})$,

$$\mathbf{u} = (u_1, u_2, u_3), \qquad \varepsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^t\right)$$

respectively, where $\nabla\mathbf{u}^t(x)$ denotes the transpose of the matrix $\nabla\mathbf{u}(x) \in \mathbf{R}^{3\times3}$ for $x \in \Omega$. Namely, let us introduce the following notations: for any $A, B \in \mathbf{R}^{3\times3}$ let

$$\operatorname{vol} A = \frac{1}{3} trA \cdot I, \qquad \operatorname{dev} A = A - \operatorname{vol} A, \qquad A : B = \sum_{i,k=1}^{3} A_{ik}B_{ik}, \qquad |A|^2 = A : A \tag{6.2}$$

where $trA = \sum_{i=1}^{3} A_{ii}$ is the trace of A and I is the identity matrix. Using these notations, in the involved model the connection of strain and stress is given by the matrix-valued nonlinear expression

$$\sigma(x) = T(x, \varepsilon(\mathbf{u}(x))) \tag{6.3}$$

with $T : \Omega \times \mathbf{R}^{3\times 3} \to \mathbf{R}^{3\times 3}$ given by

$$T(x,A) = 3k(x, |\mathrm{vol}\, A|^2)\ \mathrm{vol}\, A + 2\mu(x, |\mathrm{dev}\, A|^2)\ \mathrm{dev}\, A \qquad (x \in \Omega,\ A \in \mathbf{R}^{3\times 3}), \tag{6.4}$$

where $k(x,s)$ is the bulk modulus of the material and $\mu(x,s)$ is Lamé's coefficient. Here the functions $k, \mu : \Omega \times \mathbf{R}^+ \to \mathbf{R}$ are measurable and bounded w.r.t. x and C^1 w.r.t. the variable s and, moreover, $\partial k/\partial s$ and $\partial\mu/\partial s$ are assumed to be Lipschitz continuous. Further, they satisfy

$$\begin{aligned} &0 < k_0 \le 2\mu(x,s) \le 3k(x,s) \le K_0\,, \\ &0 < k_0 \le \tfrac{\partial}{\partial s}\big(3\,k(x,s^2)s\big) \le K_0\,, \qquad 0 < k_0 \le \tfrac{\partial}{\partial s}\big(2\,\mu(x,s^2)s\big) \le K_0\,, \end{aligned} \tag{6.5}$$

with suitable constants $K_0 \ge k_0 > 0$ independent of (x,s). Altogether, substituting (6.3) into (6.1) and letting T_i denote the ith column of T, we obtain the quasilinear elliptic system

$$\left\{ \begin{aligned} -\operatorname{div} T_i(x, \varepsilon(\mathbf{u})) &= \varphi_i(x) \quad \text{in } \Omega \\ T_i(x, \varepsilon(\mathbf{u})) \cdot \nu &= \gamma_i(x) \quad \text{on } \Gamma_N \\ u_i &= 0 \qquad\ \text{on } \Gamma_D \end{aligned} \right\} \qquad (i = 1,2,3). \tag{6.6}$$

For numerical treatment we rather need the weak formulation, which is done in the real Sobolev space

$$H^1_D(\Omega) := \{u \in H^1(\Omega) : u_{|\Gamma_D} = 0\}$$

that corresponds to the Dirichlet boundary conditions. Then the weak formulation reads as follows: find $\mathbf{u} = (u_1, u_2, u_3) \in H^1_D(\Omega)^3$ such that

$$\int_\Omega T(x, \varepsilon(\mathbf{u})) : \varepsilon(\mathbf{v}) - \int_\Omega \varphi \cdot \mathbf{v} - \int_{\Gamma_N} \gamma \cdot \mathbf{v}\, d\sigma = 0 \qquad (\mathbf{v} \in H^1_D(\Omega)^3) \tag{6.7}$$

where, with the representation (6.4) for $T(x, \varepsilon(\mathbf{u}))$, one can derive

$$T(x,\varepsilon(\mathbf{u})) : \varepsilon(\mathbf{v}) = 3k(x, |\mathrm{vol}\,\varepsilon(\mathbf{u})|^2)\ \mathrm{vol}\,\varepsilon(\mathbf{u}) : \mathrm{vol}\,\varepsilon(\mathbf{v}) + 2\mu(x, |\mathrm{dev}\,\varepsilon(\mathbf{u})|^2)\ \mathrm{dev}\,\varepsilon(\mathbf{u}) : \mathrm{dev}\,\varepsilon(\mathbf{v}) \tag{6.8}$$

Well-posedness for (6.7) can be shown using monotone potential operators [28, 82].

The above model includes linear elasticity as a special case when the functions k and μ only depend on x but do not depend on s, or, in particular, when they are constant. In this case the correspondence $\lambda := (3k - 2\mu)/3$ leads to the more standard formulation of linear elasticity with two Lamé coefficients:

$$\left\{ \begin{aligned} &-\mu\,\Delta\mathbf{u} - (\lambda + \mu)\,\nabla \operatorname{div} \mathbf{u} = \varphi \\ &\mathbf{u}_{|\partial\Omega} = 0\,. \end{aligned} \right.$$

The numerical solution of the nonlinear problem (6.7) is approached in a standard way. First a finite element discretization is done in a fixed FEM subspace $V_h \subset H^1_D(\Omega)$. Looking for the FE solution $\mathbf{u}_h \in V_h^3$ satisfying problem (6.7) only for all test functions $\mathbf{v} \in V_h^3$, the corresponding finite dimensional problem is written as

$$\langle F_h(\mathbf{u}_h), \mathbf{v} \rangle = 0 \qquad (\mathbf{v} \in V_h^3). \tag{6.9}$$

Then problem (6.9) is solved by a damped inexact Newton (DIN) iteration

$$\mathbf{u}_{n+1} = \mathbf{u}_n + \tau_n \mathbf{p}_n \tag{6.10}$$

in V_h^3, where $\tau_n \in (0,1]$ is a suitable damping parameter and $\mathbf{p}_n$ is the numerical solution of the linear problem

$$\langle F_h'(\mathbf{u}_n)\mathbf{p}_n, \mathbf{v}\rangle = -\langle F_h(\mathbf{u}_n), \mathbf{v}\rangle \qquad (\mathbf{v} \in V_h^3). \tag{6.11}$$

Altogether, the original system (6.6) is reduced to the solution of discretized linear elliptic systems (6.11), hence we are now concerned with the iterative solution of (6.11) using a suitable preconditioning.

6.2 Separate displacement preconditioning

The discrete problem (6.11) is the discretization of the linear elliptic system

$$L_i^n\,\mathbf{p}_n \equiv -\mathrm{div}\ (\mathcal{K}_n\ \varepsilon(\mathbf{p}_n))_i \ = \ -r_i^n \qquad (i=1,2,3) \tag{6.12}$$

with the array

$$\mathcal{K}_n \equiv\ T'(x,\varepsilon(\mathbf{u}_n)) \tag{6.13}$$

where, for brevity, the arrays $T'(x, .)$ denote the derivative of T w.r.t. its second (tensor) variable, and $r_i^n \equiv -\,\mathrm{div}\ T_i(x,\varepsilon(\mathbf{u}_n)) - \varphi_i$ are the residuals at $\mathbf{u}_n$, further, the boundary conditions

$$p_i^n = 0 \quad \text{on } \Gamma_D \qquad \text{and} \qquad (\mathcal{K}_n\,\varepsilon(\mathbf{p}_n))_i \cdot \nu = -T_i(x,\varepsilon(\mathbf{u}_n)) \cdot \nu + \gamma_i \qquad \text{on } \Gamma_N \tag{6.14}$$

are satisfied. In the sequel we fix n and study one fixed problem (6.12), hence for simplicity we omit the index n and the coordinates i, and simply write (6.12) as

$$L\,\mathbf{p} \ = \ r\,. \tag{6.15}$$

Accordingly, its discretization (6.11) is written as

$$L_h\,p_h = \ r_h\,. \tag{6.16}$$

Corresponding to the three coordinates, the operator L can be represented in a tensor form $L = \{L^{ij}\}_{i,j=1}^3$, and the stiffness matrix L_h has the analogous block partitioned form

$$L_h = \begin{pmatrix} L_h^{11} & L_h^{12} & L_h^{13} \\ L_h^{21} & L_h^{22} & L_h^{23} \\ L_h^{31} & L_h^{32} & L_h^{33} \end{pmatrix}\ . \tag{6.17}$$

Separate displacement preconditioning is based on the above partitioning. As preconditioning operator we propose the triplet of independent negative Laplacians:

$$S\,\mathbf{p} := \big(-\Delta p_i\big)_{i=1,2,3}$$

in $H_D^1(\Omega)^3$, whence as preconditioning matrix we have the corresponding block diagonal stiffness matrix

$$S_h = \begin{pmatrix} -\Delta_h & 0 & 0 \\ 0 & -\Delta_h & 0 \\ 0 & 0 & -\Delta_h \end{pmatrix}\ , \tag{6.18}$$

where $-\Delta_h$ is the discretization of $-\Delta$ in V_h:

$$\{-\Delta_h\}_{i,j} = \int_\Omega \nabla\psi_i \cdot \nabla\psi_j. \tag{6.19}$$

(Concerning boundary conditions, note that $p \in H^1_D(\Omega)^3$ satisfies $p_i = 0$ on Γ_D for $i = 1, 2, 3$, further, in the PCG iteration for (6.12) we have $\frac{\partial p_i}{\partial \nu} = (\mathcal{K}_n\, \varepsilon(\mathbf{d}_k^{(n)}))_i \cdot \nu$ on Γ_N where $\mathbf{d}_k^{(n)}$, $k = 1, 2, ...$, is the sequence of CG search directions in the nth outer DIN step.)

The auxiliary problems with this preconditioner are therefore decoupled to three independent Poisson equations. For each of these equations one can apply efficient Poisson solvers, mentioned in section 5.1.1, such as multigrid or multilevel methods or fast direct solvers.

Remark 6.2.1 An alternative separate displacement preconditioner instead of (6.18) is

$$S_h = \begin{pmatrix} L_h^{11} & 0 & 0 \\ 0 & L_h^{22} & 0 \\ 0 & 0 & L_h^{33} \end{pmatrix} ,$$

which can be directly composed from the elements of the original stiffness matrix L_h and, accordingly, can produce better conditioning properties than (6.18) when there is anisotropy and/or large variation of the coefficients [27]. However, the above-mentioned fast Poisson solvers cannot be used here, and incomplete factorization is also easier to construct for the preconditioner (6.18), see also [27].

To derive the achieved mesh independent condition number with the preconditioner (6.18), first note that Definition 3.2.2 and (6.12) imply

$$\langle L_S \mathbf{p}, \mathbf{v} \rangle_S = \int_\Omega \mathcal{K}_n\, \varepsilon(\mathbf{p}) : \varepsilon(\mathbf{v}) \qquad (\mathbf{p}, \mathbf{v} \in H^1_D(\Omega)^3). \tag{6.20}$$

Here $H_S = H^1_D(\Omega)^3$ with inner product $\langle \mathbf{p}, \mathbf{v} \rangle_S = \sum_{i=1}^{3} \int_\Omega \nabla p_i \cdot \nabla v_i$. From (6.5), (6.8) and (6.13) a suitable calculation yields

$$k_0 \int_\Omega |\varepsilon(\mathbf{v})|^2 \le \langle L_S \mathbf{v}, \mathbf{v} \rangle_S \le K_0 \int_\Omega |\varepsilon(\mathbf{v})|^2 \qquad (\mathbf{v} \in H^1_D(\Omega)^3),$$

which is an analogue of (3.35) for the sum of two nonlinearities. The obvious relation $\int_\Omega |\varepsilon(\mathbf{v})|^2 \le \|\mathbf{v}\|_S^2$ and Korn's inequality $\|\mathbf{v}\|_S^2 \le \kappa \int_\Omega |\varepsilon(\mathbf{v})|^2$, see [63], yield the analogue of spectral equivalence relation (3.25):

$$\frac{k_0}{\kappa} \|\mathbf{v}\|_S^2 \le \langle L_S \mathbf{v}, \mathbf{v} \rangle_S \le K_0 \|\mathbf{v}\|_S^2 \qquad (\mathbf{v} \in H^1_D(\Omega)^3).$$

Hence Proposition 3.3.1 implies

Proposition 6.2.1 *The preconditioner (6.18) yields* $cond(S_h^{-1} L_h) \le \kappa \dfrac{K_0}{k_0}$ *independently of the subspace* $V_h \subset H^1_D(\Omega)$.

Accordingly, the CG algorithm (2.4) for system $S_h^{-1} L_h p_h = S_h^{-1} r_h$ converges with ratio $\frac{\sqrt{\kappa K_0} - \sqrt{k_0}}{\sqrt{\kappa K_0} + \sqrt{k_0}}$ independently of V_h.

We finally note that the ratio $\frac{K_0}{k_0}$ may deteriorate when $k >> \mu$, which corresponds to the locking phenomenon. A possible remedy is a suitable mixed formulation, for which an outer-inner iteration scheme can still preserve mesh independent condition numbers, see [13] for details.

7 Symmetric equivalent preconditioners for nonsymmetric equations

In this chapter we consider nonsymmetric elliptic problems

$$\begin{cases} Lu := -\text{div}\,(A\,\nabla u) + \,\mathbf{b}\cdot\nabla u + cu = g \\ u_{|\Gamma_D} = 0,\; \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \end{cases} \tag{7.1}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$, where $\frac{\partial u}{\partial \nu_A} = A\,\nu\cdot\nabla u$ denotes the weighted normal derivative. We assume that the operator L satisfies Assumptions 3.2.2.1, that is, L is of the type (3.10), further, that $g \in L^2(\Omega)$. Defining the corresponding Sobolev space $H^1_D(\Omega) = \{u \in H^1(\Omega) : u_{|\Gamma_D} = 0\}$, these assumptions then ensure that problem (7.1) has a unique weak solution $u \in H^1_D(\Omega)$ (see also Remark 3.2.3).

We note that the homogeneity of the boundary conditions only serve convenience, the nonhomogeneous case can be reduced to it in a standard way, see Remark 7.1.2. For problem (7.1) one can rely directly on the results of chapters 3-4.

Problem (7.1) is most often solved numerically with a finite difference or finite element method. Our concern is to define preconditioners for the arising linear systems as the discretizations of suitable equivalent operators, and to derive mesh independence results for the convergence of proper PCG iterations. Concerning the latter, as noted in subsection 3.1, it turns out from [49, 78] that one can mostly derive general results for FEM type discretizations only, whereas FDM discretizations require a case by case study. Accordingly, we give general results below for FEM discretizations, based on chapters 3-4, whereas for FDM discretizations we cite certain case by case investigations.

When FEM is used, we define in general a subspace $V_h = span\{\varphi_1, \dots, \varphi_n\} \subset H^1_D(\Omega)$ and seek the FEM solution $u_h \in V_h$, which requires solving the $n \times n$ system

$$\mathbf{L}_h\,\mathbf{c} = \mathbf{g}_h \tag{7.2}$$

where

$$\big(\mathbf{L}_h\big)_{i,j} = \int_\Omega \Big(A\,\nabla\varphi_i\cdot\nabla\varphi_j + (\mathbf{b}\cdot\nabla\varphi_i)\varphi_j + c\varphi_i\varphi_j\Big) \;+ \int_{\Gamma_N} \alpha\varphi_i\varphi_j\,d\sigma \tag{7.3}$$

and $(\mathbf{g}_h)_j = \int_\Omega g\varphi_j$ $(j = 1, 2, ..., n)$. Since L is coercive, the symmetric part of $\mathbf{L}_h$ is positive definite, hence system (7.2) has a unique solution. Moreover, if a sequence of such subspaces V_h satisfies $\inf_{v\in V_h} \|u - v\|_{H^1} \to 0$ for all $u \in H^1_D(\Omega)$, then it follows in a standard way [34] that u_h converges to the exact weak solution in H^1-norm. The equivalent operator idea proposes to define a preconditioner for (7.2) as the discretization in V_h of another suitable elliptic operator, equivalent to L.

In this chapter we consider symmetric preconditioning operators for the discretizations of (7.1). It is a natural idea to involve symmetric operators when easier equivalent problems are looked for, since the solution of symmetric discrete elliptic problems is in general considerably easier than that of nonsymmetric ones. As pointed out e.g. in [30], the matrices of discretized elliptic operators often have properties that allow the use of sparse matrix packages, further, fast direct solvers are often available on various domains. In addition, symmetric part preconditioning yields an automatic truncation of the GCG-LS algorithm to short term recursion. These properties have made symmetric equivalent preconditioning an attractive strategy, see, e.g., [25, 30, 33, 37, 38, 45, 100].

General convergence results are presented in section 7.1, then important particular symmetric preconditioning operators are discussed in sections 7.2-section 7.3.

7.1 General symmetric preconditioners

We give general convergence results in our S-bounded and S-coercive setting: namely, we verify linear and superlinear mesh independent convergence for two corresponding classes of symmetric preconditioning operators.

7.1.1 Linear convergence with symmetric preconditioners

Let S be the symmetric elliptic operator introduced in (3.11):

$$Su \equiv -\text{div}\,(G\,\nabla u) + \sigma u \qquad \text{for } \; u_{|\Gamma_D} = 0,\; \tfrac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{7.4}$$

assumed to satisfy Assumptions 3.2.2.2. The corresponding inner product on $H^1_D(\Omega)$ is

$$\langle u, v\rangle_S := \int_\Omega (G\,\nabla u \cdot \nabla v + \sigma uv) + \int_{\Gamma_N} \beta uv\, d\sigma\,. \tag{7.5}$$

Let us first consider the FEM discretization (7.2) of problem (7.1). Then we introduce the stiffness matrix of S

$$\mathbf{S}_h = \Big\{ \langle \varphi_i, \varphi_j\rangle_S \Big\}_{i,j=1}^{n} \tag{7.6}$$

that is,

$$\big(\mathbf{S}_h\big)_{i,j} = \int_\Omega (G\,\nabla\varphi_i \cdot \nabla\varphi_j + \sigma\varphi_i\varphi_j) + \int_{\Gamma_N} \beta\varphi_i\varphi_j\, d\sigma\,,$$

as preconditioner for system (7.2), and then solve the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h} \tag{7.7}$$

(with $\tilde{\mathbf{g}_h} = \mathbf{S}_h^{-1}\mathbf{g}_h$) with a CG algorithm taken from section 2.2. The basic conditioning estimate is as follows:

Proposition 7.1.1 *For any subspace $V_h \subset H^1_D(\Omega)$, the matrices (7.3) and (7.6) satisfy*

$$\frac{\Lambda(\mathbf{S}_h^{-1}\mathbf{L}_h)}{\lambda_0(\mathbf{S}_h^{-1}\mathbf{L}_h)} \le \frac{M}{m} \tag{7.8}$$

independently of V_h, where

$$\begin{aligned} M &:= p_1 + C_{\Omega,S}\, q^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)^d} + C^2_{\Omega,S}\|c\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,S}\|\alpha\|_{L^\infty(\Gamma_N)}\;, \\ m &:= \Big(p_0^{-1} + C^2_{\Omega,L}\|\sigma\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,L}\|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1}\,. \end{aligned} \tag{7.9}$$

(The meaning of the constants in the above formulas is given in Proposition 3.2.2 and Remark 3.2.4.)

Proof. Since Assumptions 3.2.2.1-2 hold, Proposition 3.2.2 yields that the operator L is S-bounded and S-coercive in $L^2(\Omega)$. In particular, (3.20) implies that the S-bounds of L are those in (7.9). Therefore Proposition 3.3.2 can be used in $H_S = H^1_D(\Omega)$ to obtain the required result. ■

Similarly, Proposition 3.3.3 yields

$$\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h) \le \frac{M}{m}$$

independently of V_h. Now let us apply the CG algorithms from section 2.2, such that we endow $\mathbf{R}^n$ with the $\mathbf{S}_h$-inner product $\langle .,.\rangle_{\mathbf{S}_h}$. The above conditioning estimates and (3.44)-(3.45) yield

Proposition 7.1.2 *For system (7.7), the GCG-LS algorithm (2.19) satisfies*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq \left(1-\left(\frac{m}{M}\right)^2\right)^{1/2} \qquad (k=1,2,...,n), \tag{7.10}$$

which holds as well for the GCR and Orthomin methods together with their truncated versions; further, the CGN algorithm (2.35), where $A=\mathbf{S}_h^{-1}\mathbf{L}_h$ and $A^*=\mathbf{S}_h^{-1}\mathbf{L}_h$, satisfies

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq 2^{1/k}\,\frac{M-m}{M+m} \qquad (k=1,2,...,n), \tag{7.11}$$

where both ratios are independent of V_h.

We note that the boundedness of $\kappa(\mathbf{S}_h^{-1}\mathbf{L}_h)$ in h for FEM discretizations is established in [49] in the setting mentioned in Remark 3.3.2.

Remark 7.1.1 The above preconditioned CG algorithms lead to one (for GCG-LS) or two (for CGN) auxiliary problems with matrix $\mathbf{S}_h$. For instance, in the case of CGN, finding the correction terms in algorithm (2.35) with the present choice $A=\mathbf{S}_h^{-1}\mathbf{L}_h$ and $A^*=\mathbf{S}_h^{-1}\mathbf{L}_h^T$ are equivalent to the auxiliary problems

$$\begin{aligned} &\text{find } z_k \in V_h: \qquad \langle z_k, v\rangle_S = \langle L_S d_k, v\rangle_S \qquad (v\in V_h),\\ &\text{find } s_{k+1} \in V_h: \qquad \langle s_{k+1}, v\rangle_S = \langle L_S^* r_{k+1}, v\rangle_S \qquad (v\in V_h), \end{aligned}$$

i.e., z_k and s_{k+1} are the FEM solutions in V_h of the symmetric elliptic problems of the form $Sz_k = Ld_k$ and $Ss_{k+1} = L^* r_{k+1}$ with corresponding boundary conditions. For z_k, using notation $Pd_k := \frac{\partial d_k}{\partial \nu_A} + \alpha d_k$ on Γ_N, this amounts to the FEM solution in V_h of the problem

$$\left\{ \begin{array}{l} -\mathrm{div}\,(G\,\nabla z_k) + \sigma z_k = Ld_k \\ z_{k\,|\Gamma_D} = 0, \;\; \frac{\partial z_k}{\partial \nu_G} + \beta z_{k\,|\Gamma_N} = Pd_k\,, \end{array}\right. \tag{7.12}$$

and a similar problem is solved for s_{k+1}. Such symmetric problems can be solved by a variety of efficient solvers such as multigrid or multilevel methods, sparse matrix packages or fast direct solvers.

Remark 7.1.2 In the above, we have considered homogeneous boundary conditions in the original problem (7.1). The results can be extended to the non-homogeneous case in a standard way as follows.

Let us first consider non-homogeneity only for the Neumann boundary condition on Γ_N of (7.1), i.e.

$$u_{|\Gamma_D} = 0, \quad Pu \equiv \tfrac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = \gamma$$

for some $\gamma \in L^2(\Gamma_N)$. In FEM discretization, the algebraic system (7.2) then becomes modified only on the right-hand side: $(\mathbf{g}_h)_j = \int_\Omega g\varphi_j + \int_{\Gamma_N} \gamma\varphi_j$, but $\mathbf{L}_h$ remains unchanged. Therefore Propositions 7.1.1-7.1.2 remain valid. The CG algorithm is modified in the auxiliary problems such that the Neumann right-hand side Pd_k in (7.12) is replaced by $Pd_k - \gamma$.

If we also have non-homogeneity for the Dirichlet boundary condition on Γ_D of (7.1): $u_{|\Gamma_D} = \varphi$, then we first choose an $u_0 \in H^1(\Omega)$ to satisfy only $u_{0\,|\Gamma_D} = \varphi$. Then we solve the weak form of problem $Lv = g - Lu_0$ with $v_{|\Gamma_D} = 0$ and $Pv_{|\Gamma_N} = \gamma - Pu_0$, and finally let $u = u_0 + v$. In finite element applications one normally chooses u_0 as a sum of basis functions for the nodepoints on Γ_D.

For FDM discretizations of problem (7.1), boundedness of $\kappa(S_h^{-1}L_h)$ is proved for Dirichlet problems in [45] and for mixed problems in [49] on the unit square. In particular, similar mesh independence results for the Orthomin and CGN methods as in Proposition 7.1.2 are presented in [45], wherein the analogue of (7.10)-(7.11) is formulated with other suitable constants in the bounds. The constants in the above results are, however, not all a priori defined in contrast to m and M above, which is due to the lack of orthogonality compared to the FEM.

7.1.2 Superlinear convergence with symmetric preconditioners

Based on Proposition 4.2.1, we now define S to have the same principal part as L, i.e.,

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \sigma u \qquad \text{for } \; u_{|\Gamma_D} = 0,\; \tfrac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{7.13}$$

assumed to satisfy Assumptions 3.2.2.2. The corresponding inner product on $H^1_D(\Omega)$ is the same as (7.5) with $G := A$.

General superlinear results are only available for FEM discretizations. (The FDM discretization of a particular problem will be mentioned in this respect at the end of section 7.3.) Therefore, our goal now is to verify that the CG algorithm for system (7.7) provides mesh independent superlinear convergence. This can be easily derived from subsection 4.3.1. Using the decomposition $\mathbf{L}_h = \mathbf{S}_h + \mathbf{Q}_h$, system (7.7) can be rewritten as in (4.4):

$$(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \tilde{\mathbf{g}_h} \tag{7.14}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix. Here

$$\mathbf{Q}_h = \Big\{ \langle Q_S \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n} \tag{7.15}$$

for the operator Q_S on $H^1_D(\Omega)$ defined via

$$\langle Q_S u, v\rangle_S = \int_\Omega \Big((\mathbf{b}\cdot\nabla u)v + (c-\sigma)uv \Big) \; + \int_{\Gamma_N} (\alpha - \beta) uv\, d\sigma \qquad (u, v \in H^1_D(\Omega)), \tag{7.16}$$

which satisfies (4.2) in $H^1_D(\Omega)$. Then, by Proposition 4.2.1, the operator Q_S is compact, hence Propositions 4.3.1-4.3.2 are valid as well as estimates (4.5)-(4.8). We note here that the normality assumptions in Proposition 4.3.1 are too strong in general, it practically occurs in the case of symmetric part preconditioning, which will be discussed in section 7.2. Hence we only consider the second case that uses the CGN algorithm.

Summing up, let us consider system (7.7) or (7.14), where $\mathbf{L}_h$ is from (7.3) and $\mathbf{S}_h$ is from (7.6) for the operator (7.13), i.e. for $G = A$. Then (4.7)-(4.8) yield

Proposition 7.1.3 [16]. *Let us apply the CGN algorithm (2.35), where $A = \mathbf{S}_h^{-1}\mathbf{L}_h$, $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h$ and $\mathbf{R}^n$ is endowed with the $\mathbf{S}_h$-inner product $\langle .,. \rangle_{\mathbf{S}_h}$. Then*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \le \; \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{7.17}$$

where

$$\varepsilon_k := \frac{2}{km^2} \sum_{i=1}^{k} \Big(\big|\lambda_i(Q_S^* + Q_S)\big| + \lambda_i(Q_S^* Q_S) \Big) \; \to 0 \quad \textit{as} \quad k \to \infty \tag{7.18}$$

(with m from (7.9) and Q_S from (7.16)), and ε_k is a sequence independent of V_h.

The above sequence ε_k is not a priori computable in practice, but the magnitude in which $\varepsilon_k \to 0$ can be determined in some cases, namely, when the asymptotics for symmetric eigenvalue problems

$$Su = \mu u, \quad u_{|\Gamma_D} = 0, \quad r\left(\tfrac{\partial u}{\partial \nu_A} + \beta u\right)_{|\Gamma_N} = \mu u \tag{7.19}$$

are known, as is the case for Dirichlet problems where $\mu_i = O(i^{2/d})$.

Proposition 7.1.4 [16]. *The sequence ε_k in (7.18) satisfies $\varepsilon_k \leq (4s/k) \sum_{i=1}^{k} (1/\mu_i)$ for some constants $s, r > 0$, where μ_i $(i \in \mathbf{N}^+)$ are the solutions of (7.19). When the asymptotics $\mu_i = O(i^{2/d})$ holds, in particular, for Dirichlet boundary conditions,*

$$\varepsilon_k \leq O\Big(\frac{\log k}{k}\Big) \quad \textit{if} \ \ d = 2 \quad \textit{and} \quad \varepsilon_k \leq O\Big(\frac{1}{k^{2/d}}\Big) \quad \textit{if} \ \ d \geq 3. \tag{7.20}$$

7.2 Symmetric part preconditioning

A famous preconditioning strategy for solving (7.2) is symmetric part preconditioning, introduced in [38, 100], see also [12, 14]. We outline this approach for the FEM discretization (7.2) of problem (7.1). FDM discretization on the unit square is considered in the mentioned papers [38, 100]: mesh independent linear convergence is derived in a special case and confirmed by numerical tests in [100].

Let us define

$$\mathbf{S}_h := \frac{1}{2}(\mathbf{L}_h + \mathbf{L}_h^T), \quad \mathbf{Q}_h := \frac{1}{2}(\mathbf{L}_h - \mathbf{L}_h^T), \tag{7.21}$$

that is, the symmetric and antisymmetric parts of $\mathbf{L}_h$, respectively. Then $\mathbf{L}_h = \mathbf{S}_h + \mathbf{Q}_h$. If this $\mathbf{S}_h$ is chosen as preconditioner for (7.2), then the preconditioned system $\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h}$ becomes

$$(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h)\,\mathbf{c} = \tilde{\mathbf{g}_h}\,, \tag{7.22}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix, such that $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is antisymmetric w.r.t. the $\mathbf{S}_h$-inner product $\langle .,. \rangle_{\mathbf{S}_h}$. This fact has important advantages, described in subsection 2.2.1: in fact, the decomposition (7.22) is of the type (2.25). As pointed out there, for such antisymmetric perturbations of the identity, one can avoid normal equations to construct a simple CG iteration: by subsection 2.2.1.1, the full GCG-LS algorithm reduces to the truncated version GCG-LS(0) by automatic truncation, i.e. one can use the one-step recurrence (2.20). This is the main advantage of symmetric part preconditioning.

Concerning the convergence of the GCG-LS(0) iteration under symmetric part preconditioning, one can specify the results of section 7.1 by using the estimates of chapters 3-4 developed for abstract symmetric part preconditioning. To this end, we must define an appropriate elliptic operator S such that $\mathbf{S}_h$ is the stiffness matrix of S. For Dirichlet boundary conditions, where $D(L) = D(L^*)$, one can simply set $S = (L+L^*)/2$, but for mixed problems where $D(L) \neq D(L^*)$, the definition of S requires a more general weak approach, see [66]. Based on this, one constructs the symmetric part of L as

$$Su \equiv -\mathrm{div}\,(A\,\nabla u) + \hat{c}u \qquad \text{for} \ \ u_{|\Gamma_D} = 0, \ \tfrac{\partial u}{\partial \nu_G} + \hat{\alpha}u_{|\Gamma_N} = 0, \tag{7.23}$$

where $\hat{c} := c - \frac{1}{2}\,\mathrm{div}\,\mathbf{b}$ and $\hat{\alpha} := \alpha + \frac{1}{2}\,(\mathbf{b} \cdot \nu)$. Since L satisfies Assumptions 3.2.2.1, we just obtain (by setting $\sigma := \hat{c}$ and $\beta := \hat{\alpha}$) that S satisfies Assumptions 3.2.2.2. The corresponding inner product on $H_D^1(\Omega)$ is

$$\langle u, v \rangle_S := \int_\Omega (A\ \nabla u \cdot \nabla v + \hat{c}uv) + \int_{\Gamma_N} \hat{\alpha}uv\,d\sigma\,. \tag{7.24}$$

This inner product satisfies

$$\langle u,v\rangle_S = \frac{1}{2}\Big(\langle L_S u,v\rangle_S + \langle u, L_S v\rangle_S\Big) \qquad (u,v\in H^1_D(\Omega)). \tag{7.25}$$

(This is seen by Green's formula, and in fact, the construction in [66] is based on this equality.) Setting $u=\varphi_i$ and $v=\varphi_j$ in (7.25), it is readily seen that $\mathbf{S}_h$ is the symmetric part of $\mathbf{L}_h$, i.e. the required equality in (7.21) is satisfied.

We then have a decomposition (3.48):

$$L_S = I + Q_S \tag{7.26}$$

where I is the identity operator and Q_S is an antisymmetric operator on H_S. The desired mesh independent convergence estimates are determined by the operator Q_S. Here, analogously to (7.25),

$$\langle Q_S u,v\rangle_S = \frac{1}{2}\Big(\langle L_S u,v\rangle_S - \langle u, L_S v\rangle_S\Big) = \frac{1}{2}\int_\Omega \Big((\mathbf{b}\cdot\nabla u)\,\overline{v} - u\,(\mathbf{b}\cdot\nabla\overline{v})\Big) \qquad (u,v\in H^1_D(\Omega)). \tag{7.27}$$

First, the corresponding linear convergence estimate is given in (3.49)-(3.50) in terms of $\|Q_S\|$. It follows easily from (7.27) that $\|Q_S\| \le (p\,\nu_0)^{-1/2}\,\|\mathbf{b}\|_\infty$ where $\|\mathbf{b}\|_\infty = \max_{\overline{\Omega}}|\mathbf{b}|$, p is the spectral lower bound of A from Assumptions 3.2.2.1 (iii), and $\nu_0>0$ is the smallest eigenvalue of S (cf. e.g. [12]). Hence we obtain that for any FEM subspace $V_h\subset H^1_D(\Omega)$, the GCG-LS(0) algorithm (2.20) for system (7.22) satisfies

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{\|\mathbf{b}\|_\infty}{\sqrt{p\,\nu_0 + \|\mathbf{b}\|^2_\infty}} \qquad (k=1,2,...,n) \tag{7.28}$$

and for the best possible estimate we have asymptotically

$$\limsup \left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le \frac{\|\mathbf{b}\|_\infty}{\sqrt{p\ \nu_0} + \sqrt{p\,\nu_0 + \|\mathbf{b}\|^2_\infty}}\,, \tag{7.29}$$

where both ratios are independent of V_h.

For superlinear convergence, one can use the simple estimates (4.5)-(4.6) and the comments afterwards, since Q_S is antisymmetric in H_S and similarly, $\mathbf{S}_h^{-1}\mathbf{Q}_h$ is $\mathbf{S}_h$-antisymmetric. Then the GCG-LS(0) algorithm (2.20) for system (7.22) yields

$$\left(\frac{\|e_k\|_{\mathbf{S}_h}}{\|e_0\|_{\mathbf{S}_h}}\right)^{1/k} \le\ \varepsilon_k \qquad (k=1,2,...,n) \tag{7.30}$$

for the error vector e_k, where

$$\varepsilon_k := \frac{2}{k}\,\sum_{j=1}^{k}\big|\lambda_j(Q_S)\big| \ \to 0 \quad \text{as} \quad k\to\infty \tag{7.31}$$

and ε_k is a sequence independent of V_h, see [14]. We note that for 2D problems the asymptotic magnitude

$$\varepsilon_k = O\big(\frac{1}{\sqrt{k}}\big) \tag{7.32}$$

holds for (7.31), which is derived in [14] for the unit square and extended in [67] to general bounded domains.

We finally point out a significant limitation, namely, symmetric part preconditioning is only favourable for diffusion-dominated problems, i.e. when $|\mathbf{b}| = O(1)$. For convection-dominated problems, i.e. when $|\mathbf{b}| >> 1$, the estimates (7.28)-(7.29) deteriorate, as well as (7.30)-(7.31) since the magnitude of Q_S grows with $\mathbf{b}$. In this case the proper choice of symmetric preconditioning operator requires an additional large zeroth-order term, as discussed in the next section.

7.3 Helmholtz preconditioners

A widespread general approach for defining preconditioners for the original variable coefficient operator L in (7.1) is to introduce a constant coefficient operator S, see, e.g., [25, 33, 37, 45]. Due to the constant coefficients, efficient solution methods are available for the auxiliary problems with S, such as multigrid or multilevel methods, or (if Ω is rectangular, or the boundary conditions allow the problem to be easily embedded into a rectangular domain) fast direct solvers for separable equations, see e.g. [88, 93]. The scope of these methods usually includes separable problems as well, hence S can be more generally a separable approximation of L; for simplicity we restrict our presentation for constant coefficients, where in particular some explicit results can be cited.

Preconditioning with constant coefficient operators can also be motivated by the previously mentioned shortcoming of the symmetric part preconditioner. That is, for convection-dominated problems, i.e. when $|\mathbf{b}| >> 1$ in (7.1), the symmetric part preconditioning operator (7.23) produces deteriorating convergence factors (7.28), i.e. the latter is close to 1. A proper choice of preconditioning operator for such problems has been proposed in [77] with exact derivations for constant coefficient problems: namely, large values of the zeroth order term in S can compensate for the large values of $\mathbf{b}$ in L, as detailed below. It is useful now to define constant diffusion terms in the preconditioning operator, since if S is no more the exact symmetric part of L (i.e. the simple one-step recurrence is not applicable) then there is no need to preserve the original principal part.

Therefore, we now propose the Helmholtz preconditioning operator

$$Su \equiv -k\,\Delta u + \sigma u \qquad \text{for } \; u_{|\Gamma_D} = 0,\; \frac{\partial u}{\partial \nu_G} + \beta u_{|\Gamma_N} = 0, \tag{7.33}$$

where $k > 0$, $\sigma, \beta \geq 0$ are constants such that $\sigma > 0$ or $\beta > 0$ if $\Gamma_D = \emptyset$. For FEM discretizations, the mesh independent convergence estimates of section 7.1 are valid for (7.33) as a special case: for general L as in (7.1), linear convergence is established by Proposition 7.1.2, whereas if L itself has a Laplacian principal part (which often occurs in convection-diffusion problems) then Propositions 7.1.3-7.1.4 provide the corresponding superlinear convergence result. Mesh independent linear convergence for FDM discretizations on the unit square follows from the results [45, 49], mentioned at the end of subsection 7.1.1.

Concerning the case of convection-dominated problems, i.e. when $|\mathbf{b}| >> 1$ in (7.1), the optimal choice of σ is revealed by another bound, the asymptotic convergence factor (2.32). Based on this, an exact derivation of the optimal Helmholtz preconditioner has been derived in [77] for FDM discretizations of the following constant coefficient Dirichlet problem.

Let in (7.1) the domain Ω be the unit square in 2D,

$$Lu \equiv -\Delta u + \mathbf{b}\cdot\nabla u + cu \quad \text{and} \quad Su \equiv -\Delta u + \sigma u \qquad \text{for } \; u_{|\partial\Omega} = 0, \tag{7.34}$$

where $\mathbf{b} \in \mathbf{R}^2$ and $c, \sigma \geq 0$ are constants. It has been proved in [77] that the explicit optimal value of σ equals $\sigma_{opt} = |\mathbf{b}|^2(1 + \sqrt{1 + 8(\pi/|\mathbf{b}|)^2})/2$, that is,

$$\sigma_{opt} = O(|\mathbf{b}|^2)$$

as $|\mathbf{b}| \to \infty$; furthermore, in this case the spectrum of $S^{-1}L$ is contained in a disc with radius $1/4$ and center $3/4$ for any $\mathbf{b} \in \mathbf{R}^2$. The same bound has been proved for upwind FDM discretizations of the operators S and L via explicit calculation of the eigenvalues of $S_h^{-1}L_h$. Hence, cf. (2.32), one obtains the asymptotic convergence factor

$$\limsup \left(\frac{\|r_k\|}{\|r_0\|} \right)^{1/k} \leq \frac{1}{3} \tag{7.35}$$

independently of both $\mathbf{b}$ and h. (On the other hand, as suggested by (2.31) and shown by the experiments of [77], the above result is only asymptotic and the ratios $(\|r_k\|/\|r_0\|)^{1/k}$ do not behave in a mesh independent way.)

We finally note that using the above-mentioned explicit eigenvalues of $S_h^{-1}L_h$ from [77], mesh independence for the superlinear convergence of the PCG method has also been derived in [70] for the same problem, i.e. for FDM discretizations for the operators L and S in (7.34).

Remark 7.3.1 One can achieve mesh independent superlinear convergence even if the original operator has a variable but scalar diffusion coefficient, i.e., we have

$$Lu \equiv -\mathrm{div}\,(a\,\nabla u) + \mathbf{b} \cdot \nabla u + cu = g$$

in (7.1) for some scalar function $a \in C^1(\overline{\Omega})$, $a(x) \geq p > 0$. To this end, one has to apply the method of scaling, which was originally introduced for symmetric operators [37, 55]. Namely, let us rewrite our equation as

$$a^{-1/2}Lu = a^{-1/2}g =: \hat{g} \tag{7.36}$$

and introduce the new unknown function $v := a^{1/2}u$. Then, by a direct calculation [37],

$$a^{-1/2}\mathrm{div}\,(a\,\nabla u) + qu \;=\; \Delta v$$

where $q = \Delta(a^{1/2})$, which implies that

$$a^{-1/2}Lu = -\Delta v + \text{ lower order terms},$$

that is, (7.36) becomes

$$Nv \equiv -\Delta v + \hat{\mathbf{b}} \cdot \nabla v + \hat{c}v = \hat{g}\,. \tag{7.37}$$

Here $\hat{\mathbf{b}} = a^{-1}\mathbf{b}$ and $\hat{c} = a^{-1}c - (1/2a^2)\mathbf{b} \cdot \nabla a + a^{-1/2}\Delta(a^{1/2})$.

The relation $Nv \equiv a^{-1/2}Lu$ shows that

$$\langle Nv, v\rangle_{L^2} = \langle a^{-1/2}Lu,\, a^{1/2}u\rangle_{L^2} = \langle Lu, u\rangle_{L^2}$$

for all $u \in D(L)$ and $v := a^{1/2}u$. Further, using the uniform positivity of a, it is easy to see that the norms $\|u\|_{H^1}$ and $\|v\|_{H^1}$ are equivalent. Therefore N inherits the H^1-coercivity of L, i.e. the relation $\langle Lu, u\rangle_{L^2} \geq m\|u\|_{H^1}^2$ is replaced by $\langle Nv, v\rangle_{L^2} \geq \hat{m}\|v\|_{H^1}^2$ for some other proper constant $\hat{m} > 0$. This implies that the scaled problem is of the type (7.1), and hence the theory of subsection 7.1.2 can be applied. That is, the preconditioning operator S from (7.34) provides mesh independent superlinear convergence of the CGN method by Propositions 7.1.3-7.1.4 .

8 Decoupled symmetric preconditioners for nonsymmetric systems

In this chapter we consider certain nonsymmetric elliptic systems. Here, using the idea of chapter 6, an important advantage of the equivalent operator idea is that one can define decoupled (that is, independent) operators for the preconditioner, thereby reducing the size of auxiliary systems to that of one elliptic equation. This is a considerable advantage when the elliptic system consists of many equations, which case arises in important models involving reactions between several components. For such problems the decoupled preconditioners allow efficient parallelization for the solution of the auxiliary systems. The choice of symmetric preconditioning operators provides an additional simplification, similarly to the previous chapter.

Preconditioning for convection-diffusion-reaction systems with several components is discussed in section 8.1. Then some applications of the equivalent operator approach for saddle-point problems are mentioned in section 8.2.

8.1 Convection-diffusion-reaction systems

Let us consider an elliptic system

$$\left.\begin{array}{l} L_i u \equiv -\mathrm{div}\,(A_i\,\nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + \sum\limits_{j=1}^{l} V_{ij} u_j = g_i \\ u_{i\,|\Gamma_D} = 0, \qquad \frac{\partial u_i}{\partial \nu_A} + \alpha_i u_{i\,|\Gamma_N} = 0 \end{array}\right\} \qquad (i = 1, \dots, l) \tag{8.1}$$

where Ω, A_i and α_i are as in Assumptions 3.2.2.1, $\mathbf{b}_i \in C^1(\overline{\Omega})^N$, $g_i \in L^2(\Omega)$, $V_{ij} \in L^\infty(\Omega)$. We assume that $\mathbf{b}_i$ and the matrix $V = \{V_{ij}\}_{i,j=1}^{l}$ satisfy the coercivity property

$$\lambda_{min}(V + V^T) - \max_i \mathrm{div}\,\mathbf{b}_i \geq 0 \tag{8.2}$$

pointwise on Ω, where λ_{min} denotes the smallest eigenvalue. These conditions imply that the operator

$$L = (L_1, \dots, L_l)$$

is coercive in $H^1_D(\Omega)^l$, hence system (8.1) has a unique weak solution $u \in H^1_D(\Omega)^l$. Such systems arise e.g. from suitable time discretization and Newton linearization of transport systems, which often consist of a huge number of equations [102]. Condition (8.2) is then satisfied by choosing sufficiently small time-steps in the time discretization.

In this section we briefly consider a decoupled equivalent preconditioning for (8.1). Such systems have been studied by the authors in [16, 69, 71]: the efficiency of decoupled equivalent preconditioners is confirmed by numerical experiments in [69] and extended to parallel computers in [71], further, embedded in the above-mentioned Newton linearization of nonlinear transport systems in [1]. Although the experiments in [69, 71] use the GCG algorithm, we now involve the CGN algorithm which is developed for a much wider scope in [16] than the GCG method.

Similarly as in the previous chapter, our main interest is the FEM solution of (8.1). Choosing a FEM subspace $V_h \subset H^1_D(\Omega)$, the discretization of (8.1) in V_h^l leads to the corresponding algebraic system

$$\mathbf{L}_h\,\mathbf{c} = \mathbf{g}_h\,. \tag{8.3}$$

Let us define the preconditioning operator

$$S = (S_1, \dots, S_l)$$

as the l-tuple of independent operators

$$S_i u_i := -\text{div}\,(A_i\,\nabla u) + h_i u \qquad \text{for } \ u_{i\,|\Gamma_D} = 0,\ \frac{\partial u_i}{\partial \nu_A} + \beta_i u_{i\,|\Gamma_N} = 0 \qquad (i = 1, \dots, l) \tag{8.4}$$

such that each S_i satisfies Assumptions 3.2.2.2. The preconditioner for the discrete system (8.3) is defined as the stiffness matrix $\mathbf{S}_h$ of S in $H^1_D(\Omega)^l$, and we apply the CGN algorithm (2.35) for the preconditioned system

$$\mathbf{S}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{g}_h} \tag{8.5}$$

(with $\tilde{\mathbf{g}_h} = \mathbf{S}_h^{-1}\mathbf{g}_h$).

The convergence of this iteration is analogous to the case of a single equation in subsection 7.1.2. Note that S_i and L_i have the same principal part and they satisfy Assumptions 3.2.2.1-2, respectively. Therefore, as an analogue of Proposition 4.2.1, it is easy to verify that L and S are compact-equivalent, moreover, with $\mu = 1$. Therefore the Hilbert space results of subsection 4.3.1 and, in particular, estimates (4.7)-(4.8), are valid for the convergence of the CGN method. This implies that the superlinear convergence of the CGN algorithm (2.35) for (8.5) is mesh independent in the sense of Proposition 7.1.3, i.e., estimates of the form (7.17)–(7.18) hold.

The realization of the iteration with this preconditioning benefits by the fact that S_i are decoupled. In fact, finding the correction terms in algorithm (2.35) with the present choice $A = \mathbf{S}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T$ are equivalent to the following two auxiliary problems: find $z_k \in V_h^l$ and $s_{k+1} \in V_h^l$ such that

$$\langle z_k, v\rangle_S = \langle L_S d_k, v\rangle_S \quad \text{and} \quad \langle s_{k+1}, v\rangle_S = \langle L_S^* d_k, v\rangle_S \qquad (\forall v \in V_h^l),$$

i.e., z_k and s_{k+1} are the FEM solutions in V_h^l of the symmetric elliptic systems of the form $Sz_k = Ld_k$ and $Ss_{k+1} = L^* r_{k+1}$ (cf. also Remark 7.1.1). Since S consists of the independent operators S_i, its stiffness matrix is block diagonal:

$$\mathbf{S}_h = \begin{pmatrix} \mathbf{S}_h^1 & 0 & \dots & \dots & 0 \\ 0 & \mathbf{S}_h^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & 0 & \mathbf{S}_h^l \end{pmatrix}, \tag{8.6}$$

that is, the auxiliary systems consist of l independent discrete symmetric elliptic equations in V_h (with the boundary conditions of S_i, respectively). Therefore the auxiliary problems have smaller size than the original one, and their solution admits parallelization. For instance, in transport systems with pollutants there may even be hundreds of equations, see [102], hence for such problems the decoupled preconditioners lead to a considerable simplification.

8.2 An excursion to saddle-point systems

Saddle-point problems form an important mathematical model in various applications, and have been studied extensively in the literature, see, e.g. [26, 46] and the references therein. It is beyond the scope of this paper to give any summary of this wide area. Instead, our goal is only to indicate that the operator level for such problems provides a natural and useful background for their numerical solution, further, this approach is closely related to the equivalent operator idea. Furthermore, in some special cases one can apply directly some methods presented previously in this paper.

Our brief discussion mostly involves the Stokes problem, which is one of the most important models in saddle-point form. Its classical formulation reads as

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{f} \\ \operatorname{div} \mathbf{u} = 0 \\ \mathbf{u}_{|\partial\Omega} = 0 \end{cases} \tag{8.7}$$

in a bounded domain $\Omega \subset \mathbf{R}^d$ ($d = 2$ or 3) with $\mathbf{f} \in L^2(\Omega)^d$. Here $\mathbf{u}$ is the displacement and p is the pressure. Since p in (8.7) is determined up to an additive constant only, for uniqueness one introduces the space

$$L_0^2(\Omega) := \{p \in L^2(\Omega) : \int_\Omega p = 0\}. \tag{8.8}$$

The standard weak formulation then reads as follows: find $(\mathbf{u}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ satisfying

$$\begin{cases} \displaystyle\int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} - \int_\Omega p \,(\operatorname{div} \mathbf{v}) = \int_\Omega \mathbf{f} \cdot \mathbf{v} & (\forall \mathbf{v} \in H_0^1(\Omega)^d) \\ \displaystyle\int_\Omega q \,(\operatorname{div} \mathbf{u}) = 0 & (\forall q \in L_0^2(\Omega)), \end{cases} \tag{8.9}$$

where the notation $\nabla \mathbf{u} \cdot \nabla \mathbf{v} := \sum_{i=1}^d \nabla u_i \cdot \nabla v_i$ is used. Then problem (8.9) has a unique weak solution, see, e.g., [95].

Applying the Schur complement to treat the Stokes equations reduces them to a symmetric and positive definite problem, this approach will be discussed in subsections 8.2.1-8.2.2. The nonsymmetric formulation and symmetric part preconditioning will be outlined in subsection 8.2.3. In the last subsection we will also mention Navier's system of elasticity equations, which is closely related to (8.7), in fact the momentum equation is modified only by adding a constant factor of the pressure. Symmetric part preconditioning can be applied here in a similar way as for the regularized Stokes problem.

8.2.1 Stokes problem: the Schur complement approach on continuous level

A common way to treat saddle-point problems involves the Schur complement, both on continuous and discrete level. In this subsection we consider the continuous Stokes problem (8.7) and outline the Schur complement approach for theoretical purposes, concerning both solvability and iterations in function space.

Problem (8.7) can be recast in the pressure variable p by inverting the d-tuple of Laplacians in the first equation and eliminating $\mathbf{u}$ in the second one. One thus obtains formally the equation

$$S^0 p \equiv -\operatorname{div} (-\Delta)^{-1} \nabla p = -\operatorname{div} (-\Delta)^{-1} \mathbf{f} \equiv \hat{\mathbf{f}}\,, \tag{8.10}$$

where $(-\Delta)^{-1}$ is understood as the d-tuple of inverse Laplacians mapping from $L^2(\Omega)^d$ to $H_0^1(\Omega)^d$ (i.e. with the Dirichlet boundary conditions). A more precise definition of the Schur operator $S^0 : L^2(\Omega) \to L^2(\Omega)$ uses the weak formulation: for any $p \in L^2(\Omega)$, let

$$S^0 p := \operatorname{div} \mathbf{w} \tag{8.11}$$

where $\mathbf{w} \in H_0^1(\Omega)^d$ is the unique solution of

$$\int_\Omega \nabla \mathbf{w} \cdot \nabla \mathbf{v} = \int_\Omega p \,(\operatorname{div} \mathbf{v}) \qquad (\forall \mathbf{v} \in H_0^1(\Omega)^d). \tag{8.12}$$

(Here $\mathbf{w} = \Delta^{-1}\nabla p$ in terms of (8.10).) Then system (8.9) is equivalent to the equation

$$S^0 p = -\operatorname{div} \mathbf{h}, \tag{8.13}$$

where $\mathbf{h} \in H_0^1(\Omega)^d$ is uniquely determined by the equation

$$\int_\Omega \nabla \mathbf{h} \cdot \nabla \mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad (\forall \mathbf{v} \in H_0^1(\Omega)^d), \tag{8.14}$$

that is, $\mathbf{h} = (-\Delta)^{-1}\mathbf{f}$. Indeed, $p \in L^2(\Omega)$ satisfies (8.13) if and only if the corresponding $\mathbf{w}$ from (8.11) fulfils $\operatorname{div}(\mathbf{w} + \mathbf{h}) = 0$, in other words, $\mathbf{w} + \mathbf{h} = \mathbf{u}$ for some $\mathbf{u} \in H_0^1(\Omega)^d$ with

$$\operatorname{div} \mathbf{u} = 0 \tag{8.15}$$

where, adding (8.12) to (8.14), $\mathbf{u}$ must also satisfy

$$\int_\Omega \nabla \mathbf{u} \cdot \nabla \mathbf{v} = \int_\Omega (\nabla \mathbf{w} + \nabla \mathbf{h}) \cdot \nabla \mathbf{v} = \int_\Omega p\,(\operatorname{div} \mathbf{v}) + \int_\Omega \mathbf{f} \cdot \mathbf{v} \qquad (\forall \mathbf{v} \in H_0^1(\Omega)^d). \tag{8.16}$$

That is, as shown by (8.15) and (8.16), $\mathbf{u}$ is just the solution of the Stokes problem (8.9).

The solvability of (8.13) is due to the inf-sup property, stating that there exists a constant $\gamma > 0$ such that for all $p \in L^2(\Omega)$,

$$\sup_{\substack{\mathbf{v} \in H_0^1(\Omega)^d \\ \mathbf{v} \neq \mathbf{0}}} \frac{\int_\Omega p\,(\operatorname{div} \mathbf{v})}{\|\mathbf{v}\|_{H_0^1(\Omega)^d}} \;\geq\; \gamma\, \|p\|_{L^2(\Omega)}, \tag{8.17}$$

where $\|\mathbf{v}\|^2_{H_0^1(\Omega)^d} := \|\nabla \mathbf{v}\|^2_{L^2(\Omega)^d} = \int_\Omega \sum_{i=1}^d |\nabla v_i|^2$. In fact, the corresponding $\mathbf{w} \in H_0^1(\Omega)^d$ from (8.12) then satisfies

$$\gamma\, \|p\|_{L^2(\Omega)} \leq \sup_{\substack{\mathbf{v} \in H_0^1(\Omega)^d \\ \mathbf{v} \neq \mathbf{0}}} \frac{\int_\Omega \nabla \mathbf{w} \cdot \nabla \mathbf{v}}{\|\nabla \mathbf{v}\|_{L^2(\Omega)^d}} \;=\; \|\nabla \mathbf{w}\|_{L^2(\Omega)^d}$$

which, by setting $\mathbf{v} := \mathbf{w}$ in (8.12), yields that

$$\gamma^2\, \|p\|^2_{L^2(\Omega)} \leq\; \|\nabla \mathbf{w}\|^2_{L^2(\Omega)^d} \;= \int_\Omega p\,(\operatorname{div} \mathbf{w}) = \langle S^0 p, p \rangle_{L^2(\Omega)} \qquad (p \in L^2(\Omega)), \tag{8.18}$$

that is, S^0 is coercive in $L^2(\Omega)$. It is easily seen that S^0 is a symmetric and bounded operator in $L^2(\Omega)$. In particular, from (8.18),

$$\|\nabla \mathbf{w}\|^2_{L^2(\Omega)^d} \;= \int_\Omega p\,(\operatorname{div} \mathbf{w}) \leq \|p\|_{L^2(\Omega)} \|\operatorname{div} \mathbf{w}\|_{L^2(\Omega)} \leq \|p\|_{L^2(\Omega)} \|\nabla \mathbf{w}\|_{L^2(\Omega)^d}$$

using that the boundary condition $\mathbf{w}_{|\partial\Omega} = 0$ implies $\|\operatorname{div} \mathbf{w}\|_{L^2(\Omega)} \leq \|\nabla \mathbf{w}\|_{L^2(\Omega)^d}$ [5, 15]. Hence, also by (8.18),

$$\langle S^0 p, p \rangle_{L^2(\Omega)} = \|\nabla \mathbf{w}\|^2_{L^2(\Omega)^d} \;\leq \|p\|^2_{L^2(\Omega)} \qquad (p \in L^2(\Omega)),$$

that is, the spectral bounds of S^0 are γ^2 and 1. Altogether, the properties of S^0 imply that problem (8.13) has a unique solution $p \in L^2(\Omega)$. If this is found then $\mathbf{u}$ is obtained by solving (8.16). Further, as best expressed by (8.10), the operator S^0 includes a factor $(-\Delta)^{-1}$ which can be considered as a kind of inner preconditioning operator.

In theory one can define iterations for (8.7) on the continuous level by applying standard iterations for equation (8.13). Such iterations are useful as they exploit the structure of the Stokes system and can be directly adapted to FEM discretizations. In the construction one can use the fact that for given $p \in L^2(\Omega)$, the residual $r := S^0\, p + \operatorname{div} \mathbf{h}$ for equation (8.13) satisfies

$$r = \operatorname{div} \mathbf{u}, \quad \text{where} \quad -\Delta \mathbf{u} + \nabla p = \mathbf{f}.$$

First, simple (or Richardson) iterations for (8.13) of the form

$$p_{k+1} = p_k + \alpha r_k \qquad (k \in \mathbf{N})$$

(where $p_0 \in L^2(\Omega)$ is arbitrary, $\alpha \in \mathbf{R}$ is constant and $r_k = S^0\, p_k + \operatorname{div} \mathbf{h}$) can be rewritten as follows: if p_k is found then

$$\begin{cases} -\Delta \mathbf{u}_k + \nabla p_k = \mathbf{f}, & \mathbf{u}_{k\,|\partial\Omega} = 0 \\ p_{k+1} = p_k + \alpha \operatorname{div} \mathbf{u}_k = 0\,, \end{cases} \tag{8.19}$$

which is nothing but the well-known Uzawa iteration. Since S^0 has spectral bounds γ^2 and 1, letting $\alpha = -2/(1+\gamma^2)$ the Uzawa iteration converges with ratio $(1-\gamma^2)/(1+\gamma^2)$ in L^2-norm.

On the other hand, we can use the CG algorithm (2.4)-(2.5) for (8.13) in $L^2(\Omega)$: for arbitrary $p_0 \in L^2(\Omega)$, let $d_0 = r_0$, and if r_k and d_k are found then

$$r_{k+1} = r_k + \alpha_k S^0\, d_k, \text{ where } \alpha_k = \frac{\|r_k\|^2_{L^2}}{\langle S^0\, d_k, d_k\rangle_{L^2}}\,; \quad d_{k+1} = r_{k+1} + \beta_k d_k, \text{ where } \beta_k = \frac{\|r_{k+1}\|^2_{L^2}}{\|r_k\|^2_{L^2}}. \tag{8.20}$$

This can be rewritten similarly as above: given $p_0 \in L^2(\Omega)$, we first let $d_0 = r_0 = \operatorname{div} \mathbf{u}_0$ where $\mathbf{u}_0$ solves $-\Delta \mathbf{u}_0 + \nabla p_0 = \mathbf{f}$, $\mathbf{u}_{0\,|\partial\Omega} = 0$; further, if r_k and d_k are found then the next iterates are determined by

$$\begin{cases} -\Delta \mathbf{z}_k + \nabla d_k = 0, & \mathbf{z}_{k\,|\partial\Omega} = 0 \\ r_{k+1} = r_k + \alpha_k \operatorname{div} \mathbf{z}_k, \\ d_{k+1} = r_{k+1} + \beta_k d_k\,. \end{cases} \tag{8.21}$$

The constants α_k and β_k are as in (8.20), except that $S^0\, d_k$ in α_k is replaced by $\operatorname{div} \mathbf{z}_k$. Using the spectral bounds γ^2 and 1 of S^0, the iteration (8.21) converges with ratio $(1-\gamma)/(1+\gamma)$ in L^2-norm.

In fact one uses the weak formulation of the auxiliary problems in the above iterations. For instance, determining $\mathbf{z}_k$ in the CG algorithm (8.21) requires the solution of the variational problem

$$\int_\Omega \nabla \mathbf{z}_k \cdot \nabla \mathbf{v} = \int_\Omega d_k\, (\operatorname{div} \mathbf{v}) \qquad (\forall \mathbf{v} \in H^1_0(\Omega)^d). \tag{8.22}$$

8.2.2 Stokes problem: the Schur complement approach for FEM discretizations

The ideas of the previous subsection have their exact analogues when finite elements are used. In order to apply FEM to (8.9), one chooses suitable FE subspaces $V_h \subset H^1_0(\Omega)^d$ and $P_h \subset L^2_0(\Omega)$ and replaces $H^1_0(\Omega)^d$ and $L^2_0(\Omega)$ in (8.9) by V_h and P_h, respectively. That is, we wish to find $(\mathbf{u}_h, p_h) \in V_h \times P_h$ satisfying

$$\begin{cases} \displaystyle\int_\Omega \nabla \mathbf{u}_h \cdot \nabla \mathbf{v}_h - \int_\Omega p_h\, (\operatorname{div} \mathbf{v}_h) = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h & (\forall \mathbf{v}_h \in V_h) \\ \displaystyle\int_\Omega q_h\, (\operatorname{div} \mathbf{u}_h) = 0 & (\forall q_h \in P_h). \end{cases} \tag{8.23}$$

Then the corresponding algebraic system has the block form

$$\begin{pmatrix} \mathbf{A}_h & \mathbf{B}_h^T \\ \mathbf{B}_h & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{u}_h \\ \mathbf{p}_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ \mathbf{0} \end{pmatrix} , \tag{8.24}$$

where $\mathbf{A}_h$ and $\mathbf{B}_h$ are the Gram matrices of the operators $-\Delta$ and $-\mathrm{div}$, respectively. Analogously to (8.10), system (8.24) can be rewritten as

$$\mathbf{B}_h \mathbf{A}_h^{-1} (\mathbf{B}_h^T \mathbf{p}_h - \mathbf{f}_h) = \mathbf{0}$$

in which

$$\mathbf{S}_h^0 = \mathbf{B}_h \mathbf{A}_h^{-1} \mathbf{B}_h^T$$

is the Schur complement.

Now one can benefit by what has been developed on the continuous level. Since problem (8.23) only differs from (8.9) in that the function spaces are replaced by corresponding FEM subspaces, one can similarly execute this replacement in the formulas of the previous subsection. (In particular, the resulting discrete operator S_h^0 on the subspace P_h corresponds to the Schur complement matrix $\mathbf{S}_h^0$.) However, when function spaces have been replaced by corresponding FEM subspaces in subsection 8.2.1, care must be taken: in contrast to the other formulas, due to taking supremum on a smaller set, the inequality (8.17) is not automatically satisfied. On the contrary, it is a crucial point in the FEM discretization to define the subspaces V_h and P_h such that (8.17) is preserved with (possibly another) constant $\hat{\gamma}$, required to be independent of h as $h \to 0$. This property is the famous LBB-condition: that is, there must exist a constant $\hat{\gamma} > 0$ such that for the given family of subspaces V_h and P_h, for all $h > 0$ and $p \in P_h$

$$\sup_{\substack{\mathbf{v} \in V_h \\ \mathbf{v} \neq \mathbf{0}}} \frac{\int_\Omega p \, (\mathrm{div}\, \mathbf{v})}{\|\mathbf{v}\|_{H_0^1(\Omega)^d}} \;\geq\; \hat{\gamma} \, \|p\|_{L^2(\Omega)}. \tag{8.25}$$

The construction of such stable pairs of subspaces V_h and P_h is well-known, see, e.g., [18, 46]. Examples are piecewise linear (and continuous) polynomial basis functions for V_h and piecewise constants for P_h, or piecewise quadratic (and continuous) polynomials for V_h and piecewise linear for P_h (the latter containing discontinuous elements).

As a conclusion of what has been said above, we obtain the following results. Consider the FEM discretization (8.23) of the Stokes problem (8.7), such that the subspaces V_h and P_h satisfy the LBB-condition (8.25). Then one can define the analogue of the Uzawa and CG iterations (8.19) and (8.21), respectively, for the corresponding algebraic system. These algorithms have the same form as (8.19) and (8.21), except that in the weak formulation of the auxiliary problems the function spaces are replaced by the corresponding FEM subspaces. For instance, in the case of the CG algorithm, if $r_k \in P_h$ and $d_k \in P_h$ are found then the next iterates are determined as follows:

$$\begin{cases} \text{find } \; \mathbf{z}_k \in V_h \; \text{ such that } \int_\Omega \nabla \mathbf{z}_k \cdot \nabla \mathbf{v} = \int_\Omega d_k \, (\mathrm{div}\, \mathbf{v}) \qquad (\forall \mathbf{v} \in V_h), \\ r_{k+1} = r_k + \alpha_k \, \mathrm{div}\, \mathbf{z}_k, \\ d_{k+1} = r_{k+1} + \beta_k d_k \, . \end{cases}$$

Furthermore, if $\hat{\gamma}$ is the inf-sup constant from the LBB-condition (8.25), then the Uzawa iteration with parameter $\alpha = -2/(1+\hat{\gamma}^2)$ converges with ratio $(1-\hat{\gamma}^2)/(1+\hat{\gamma}^2)$, and the CG algorithm converges with ratio $(1-\hat{\gamma})/(1+\hat{\gamma})$, both independently of V_h.

8.2.3 Stokes problem: regularization and symmetric part preconditioning

When one defines the discretized Stokes problem (8.23), a crucial issue with the choice of V_h and P_h is to satisfy the LBB-condition (8.25), which is not always straightforward (see [6] for a discussion). Therefore an important effort has been done to circumvent this problem and define regularized versions of (8.23), which are consistent with the original problem but allow equal-order approximation (i.e. both the velocity and pressure are looked for in H^1).

The following regularized version of (8.23) is taken from [6], see also the references therein. (It is now modified such that the real Sobolev spaces are replaced by complex ones, which is required to apply the theory of normal operators from [14] to our problem). Let the FE subspaces

$$V_h \subset H_0^1(\Omega)^d, \qquad P_h \subset \dot{H}^1(\Omega) := H^1(\Omega) \cap L_0^2(\Omega)$$

consist of piecewise linear functions. We fix a parameter $\sigma > 0$, and want to find $(\mathbf{u}_h, p_h) \in V_h \times P_h$ satisfying

$$\left\{ \begin{array}{lll} \displaystyle\int_\Omega \nabla \mathbf{u}_h \cdot \nabla \overline{\mathbf{v}}_h \; - \; \int_\Omega p_h \,(\operatorname{div} \overline{\mathbf{v}}_h) = \int_\Omega \mathbf{f} \cdot \overline{\mathbf{v}}_h & & (\forall \mathbf{v}_h \in V_h) \\ \displaystyle\int_\Omega (\operatorname{div} \mathbf{u}_h)\, \overline{q}_h + \sigma \int_\Omega \nabla p_h \cdot \nabla \overline{q}_h \; = \sigma \int_\Omega \mathbf{f} \cdot \nabla \overline{q}_h & & (\forall q_h \in P_h). \end{array} \right. \tag{8.26}$$

The nonsymmetric problem (8.26) is the special case of (3.5)-(3.6) in [6] with the choice (3.8) there. Formally, in terms of the strong form (8.7) and assuming sufficient regularity, the regularization in the second row comes from adding the relation $-\sigma \Delta p = -\sigma \operatorname{div} \mathbf{f}$, which follows from taking the divergence of $-\Delta \mathbf{u} + \nabla p = \mathbf{f}$ and using $\operatorname{div} \mathbf{u} = 0$.

One can rewrite problem (8.26) by letting $s_h := \sigma^{1/2} p_h$ in order to balance the parameter σ in the diagonal. To formulate the corresponding algebraic system, the following obvious notations will be used for the discrete operators: let ∇_h, div_h, Δ_h^0 and Δ_h^ν denote the Gram matrices of the operators ∇, div, Δ with Dirichlet boundary conditions and Δ with Neumann boundary conditions, respectively, in the considered subspaces. Further, let $diag_d(-\Delta_h^0)$ denote the block diagonal matrix with $-\Delta_h^0$ blocks repeated d times. Then the algebraic system corresponding to the rewritten discrete problem takes the following, nonsymmetric form:

$$\mathbf{L}_h \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} \equiv \begin{pmatrix} diag_d(-\Delta_h^0) & \sigma^{-1/2} \nabla_h \\ \sigma^{-1/2} \operatorname{div}_h & -\Delta_h^\nu \end{pmatrix} \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ \sigma^{-1/2} \operatorname{div} \mathbf{f}_h \end{pmatrix} \tag{8.27}$$

where ξ_h and η_h are the coefficient vectors of $\mathbf{u}_h$ and p_h in the given basis of V_h and P_h, respectively. Here $\mathbf{L}_h \in \mathbf{R}^{n \times n}$ where $n = dim(V_h) + dim(P_h)$, and its symmetric and antisymmetric parts are

$$\mathbf{M}_h = \begin{pmatrix} diag_d(-\Delta_h^0) & \mathbf{0} \\ \mathbf{0} & -\Delta_h^\nu \end{pmatrix} \qquad \text{and} \qquad \mathbf{N}_h = \begin{pmatrix} \mathbf{0} & \sigma^{-1/2} \nabla_h \\ \sigma^{-1/2} \operatorname{div}_h & \mathbf{0} \end{pmatrix}. \tag{8.28}$$

Then by [15], one can propose symmetric part preconditioning and obtain superlinear convergence. Namely, let us consider the complex separable Hilbert space and corresponding inner product

$$H_M := H_0^1(\Omega)^d \times \dot{H}^1(\Omega), \qquad \left\langle \begin{pmatrix} \mathbf{u} \\ s \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right\rangle_M := \int_\Omega \nabla \mathbf{u}_h \cdot \nabla \overline{\mathbf{v}}_h \; + \int_\Omega \nabla s_h \cdot \nabla \overline{q}_h \, . \tag{8.29}$$

One can verify that the following relation defines a compact linear operator $C : H_M \to H_M$ which is antisymmetric:

$$\left\langle C \begin{pmatrix} \mathbf{u} \\ s \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \right\rangle_M = - \int_\Omega s\,(\operatorname{div} \overline{\mathbf{v}}) + \int_\Omega (\operatorname{div} \mathbf{u})\, \overline{q} \qquad \left(\forall \begin{pmatrix} \mathbf{u} \\ s \end{pmatrix}, \begin{pmatrix} \mathbf{v} \\ q \end{pmatrix} \in H_M \right). \tag{8.30}$$

It is readily seen that system (8.27) corresponds to the FEM discretization of the operator equation

$$(I + \sigma^{-1/2} C) \begin{pmatrix} \mathbf{u} \\ s \end{pmatrix} = \begin{pmatrix} \mathbf{g} \\ r \end{pmatrix} \tag{8.31}$$

(where I is the identity operator on H_M) in the subspace $V_h \times P_h$, that is, $\mathbf{M}_h$ and $\mathbf{N}_h$ are the stiffness matrices corresponding to the inner product (8.29) and $\sigma^{-1/2}$ times the operator (8.30), respectively. Consequently, by (4.5)-(4.6) and the comments afterwards on symmetric part preconditioning, setting $\mathbf{S}_h = \mathbf{M}_h$ and $\mathbf{Q}_h = \mathbf{N}_h$, we obtain the following result [15]: the preconditioned GCG-LS(0) algorithm (2.20), using $\mathbf{M}_h$ as preconditioner for system (8.27), yields

$$\left(\frac{\|e_k\|_{\mathbf{M}_h}}{\|e_0\|_{\mathbf{M}_h}} \right)^{1/k} \leq \varepsilon_k \qquad (k = 1, 2, ..., n), \tag{8.32}$$

where

$$\varepsilon_k := \frac{2}{\sigma^{1/2} k} \sum_{j=1}^{k} \big| \lambda_j(C) \big| \to 0 \quad \text{as} \quad k \to \infty\,.$$

8.2.4 Navier's elasticity system

Another famous saddle-point system, closely related to the Stokes problem (8.7), is Navier's system for elasticity equations. We consider the mixed formulation based on [19]:

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \dfrac{1}{\mu}\, \mathbf{f} \\ \operatorname{div} \mathbf{u} + (1 - 2\nu) p = 0 \\ \mathbf{u}_{|\partial\Omega} = 0 \end{cases} \tag{8.33}$$

where ν is the Poisson ratio satisfying $0 < \nu < 1/2$. Similarly to subsection 8.2.2, for the finite element solution of (8.33) one chooses suitable FE subspaces $V_h \subset H_0^1(\Omega)^d$, $P_h \subset L_0^2(\Omega)$. We will use symmetric part preconditioning as in subsection 8.2.3, hence the Sobolev spaces are taken to be complex (again to apply the theory of normal operators). One then looks for $(\mathbf{u}_h, p_h) \in V_h \times P_h$ satisfying

$$\begin{cases} \displaystyle\int_\Omega \nabla \mathbf{u}_h \cdot \nabla \overline{\mathbf{v}}_h - \int_\Omega p_h\, (\operatorname{div} \overline{\mathbf{v}}_h) = \int_\Omega \frac{1}{\mu}\, \mathbf{f} \cdot \overline{\mathbf{v}}_h & (\forall \mathbf{v}_h \in V_h) \\ \displaystyle\int_\Omega (\operatorname{div} \mathbf{u}_h)\, \overline{q}_h + (1 - 2\nu) \int_\Omega p_h\, \overline{q}_h = 0 & (\forall q_h \in P_h). \end{cases} \tag{8.34}$$

We quote two results from [15] using symmetric part preconditioning. First, the presence of the term $(1 - 2\nu)p$ in (8.33) (compared to (8.7) enables us to avoid regularization. Balancing system (8.34) by letting $s_h := (1 - 2\nu)^{1/2} p_h$, we obtain a corresponding algebraic system similarly to (8.27), using the additional notation I_h for the mass matrix corresponding to the subspace P_h:

$$\mathbf{L}_h \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} \equiv \begin{pmatrix} diag_N(-\Delta_h^0) & (1 - 2\nu)^{-1/2}\, \nabla_h \\ (1 - 2\nu)^{-1/2} \operatorname{div}_h & I_h \end{pmatrix} \begin{pmatrix} \xi_h \\ \eta_h \end{pmatrix} = \begin{pmatrix} \frac{1}{\mu}\, \mathbf{f}_h \\ 0 \end{pmatrix}. \tag{8.35}$$

By a suitable application of estimate (3.49), one can derive the following linear convergence result for the GCG-LS(0) algorithm with symmetric part preconditioning :

$$\left(\frac{\|e_k\|_{\mathbf{M}_h}}{\|e_0\|_{\mathbf{M}_h}}\right)^{1/k} \leq \frac{1}{\sqrt{2(1-\nu)}} \qquad (k=1,\dots,n). \tag{8.36}$$

Second, when ν is close to 1/2, it is worthwile to rather use a regularization in the same way as in subsection 8.2.3. Let us now consider (8.34) with FE subspaces $V_h \subset H_0^1(\Omega)^d$, $P_h \subset \dot{H}^1(\Omega) := H^1(\Omega) \cap L_0^2(\Omega)$ consisting of piecewise linear functions. Fixing a parameter $\sigma > 0$, the regularized version reads as follows: find $(\mathbf{u}_h, p_h) \in V_h \times P_h$ satisfying

$$\begin{cases} \displaystyle\int_\Omega \nabla \mathbf{u}_h \cdot \nabla \overline{\mathbf{v}}_h - \int_\Omega p_h \,(\operatorname{div} \overline{\mathbf{v}}_h) = \int_\Omega \frac{1}{\mu}\, \mathbf{f} \cdot \overline{\mathbf{v}}_h \qquad (\forall \mathbf{v}_h \in V_h) \\ \displaystyle\int_\Omega (\operatorname{div} \mathbf{u}_h)\, \overline{q}_h + 2\sigma(1-\nu) \int_\Omega \nabla p_h \cdot \nabla \overline{q}_h + (1-2\nu) \int_\Omega p_h\, \overline{q}_h = \frac{\sigma}{\mu} \int_\Omega \mathbf{f} \cdot \nabla \overline{q}_h \qquad (\forall q_h \in P_h). \end{cases} \tag{8.37}$$

Forming the corresponding algebraic system, one can then prove an analogoue of (8.32): there exists a sequence $\tilde{\varepsilon}_k \to 0$ (as $k \to \infty$) such that the PCG algorithm with symmetric part preconditioning yields

$$\left(\frac{\|e_k\|_{\mathbf{M}_h}}{\|e_0\|_{\mathbf{M}_h}}\right)^{1/k} \leq \tilde{\varepsilon}_k \qquad (k=1,\dots,n),$$

and here $\tilde{\varepsilon}_k \to 0$ depends on σ and Ω but is independent of ν.

9 Nonsymmetric equivalent preconditioners

Chapters 7-8 have been based on the idea that symmetric problems are considerably simpler than nonsymmetric ones regarding standard efficient solvers, therefore the former may serve as proper preconditioners for the latter. However, if the original problem has large nonsymmetric (first-order) terms, then this approach may not work satisfactorily and it may still be advisable to include nonsymmetric terms in the preconditioning operator. In this chapter we briefly discuss some nonsymmetric preconditioners: first, general convergence results are derived from subsection 3.3.3, then some efficient realizations are outlined.

9.1 General nonsymmetric preconditioners

Let us consider again the nonsymmetric elliptic equation (7.1):

$$\begin{cases} Lu := -\operatorname{div}(A\,\nabla u) + \mathbf{b} \cdot \nabla u + cu = g \\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0 \end{cases} \tag{9.1}$$

on a bounded domain $\Omega \subset \mathbf{R}^d$, where L satisfies Assumptions 3.2.2.1 and $g \in L^2(\Omega)$. As before, we are mainly interested in FEM discretization: given a FEM subspace $V_h \subset H_D^1(\Omega)$, we then seek the solution of the corresponding algebraic system $\mathbf{L}_h\,\mathbf{c} = \mathbf{g}_h$ as in (7.2). Some results on FDM will be also mentioned regarding linear convergence.

9.1.1 Linear convergence with nonsymmetric preconditioners

We consider the following general form of nonsymmetric preconditioning operator:

$$Nu := -\mathrm{div}\,(K\,\nabla u) + \mathbf{w}\cdot\nabla u + zu \qquad \text{for } \; u\in H^2(\Omega):\; u_{|\Gamma_D}=0,\; \tfrac{\partial u}{\partial\nu_K}+\eta u_{|\Gamma_N}=0 \tag{9.2}$$

for some properly chosen functions $\mathbf{w}, z, \eta$, such that N satisfies Assumptions 3.2.2.1 in the obvious sense.

Let us first consider the FEM discretization $\mathbf{L}_h\,\mathbf{c}=\mathbf{g}_h$ of problem (9.1) as in (7.2). Then we introduce the nonsymmetric stiffness matrix of N as preconditioner:

$$\big(\mathbf{N}_h\big)_{i,j} = \int_\Omega\Big(K\,\nabla\varphi_i\cdot\nabla\varphi_j + (\mathbf{w}\cdot\nabla\varphi_i)\,\varphi_j + z\varphi_i\varphi_j\Big) \;+\; \int_{\Gamma_N}\eta\varphi_i\varphi_j\,d\sigma\,. \tag{9.3}$$

We use the same energy space H_S as in the symmetric case, i.e. $H_S=H^1_D(\Omega)$ with inner product (7.5). We then solve the preconditioned system

$$\mathbf{N}_h^{-1}\mathbf{L}_h\,\mathbf{c} = \tilde{\mathbf{b}_h} \tag{9.4}$$

(with $\tilde{\mathbf{b}_h}=\mathbf{N}_h^{-1}\mathbf{b}_h$) using the CGN algorithm (2.35) with the $\mathbf{S}_h$-inner product and by setting $A=\mathbf{N}_h^{-1}\mathbf{L}_h$ and $A^*=\mathbf{S}_h^{-1}\mathbf{L}_h^T\mathbf{N}_h^{-T}\mathbf{S}_h$.

In the convergence analysis we need the bounds of L and N as in (3.54): now there holds

$$\begin{aligned} m_L\|u\|_S^2 &\le \langle L_S u,u\rangle_S, \quad |\langle L_S u,v\rangle_S| \le M_L\|u\|_S\|v\|_S,\\ m_N\|u\|_S^2 &\le \langle N_S u,u\rangle_S, \quad |\langle N_S u,v\rangle_S| \le M_N\|u\|_S\|v\|_S \end{aligned} \tag{9.5}$$

for all $u,v\in H^1_D(\Omega)$, where

$$\begin{aligned} M_L &:= p_1 + C_{\Omega,S}\,q^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)^d} + C^2_{\Omega,S}\|c\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,S}\|\alpha\|_{L^\infty(\Gamma_N)}\;,\\ m_L &:= \Big(p_0^{-1} + C^2_{\Omega,L}\|\sigma\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,L}\|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1}\;,\\ M_N &:= \hat{p}_1 + C_{\Omega,S}\,q^{-1/2}\|\mathbf{w}\|_{L^\infty(\Omega)^d} + C^2_{\Omega,S}\|z\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,S}\|\eta\|_{L^\infty(\Gamma_N)}\;,\\ m_N &:= \Big(\hat{p}_0^{-1} + C^2_{\Omega,N}\|\sigma\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,N}\|\beta\|_{L^\infty(\Gamma_N)}\Big)^{-1}\,. \end{aligned} \tag{9.6}$$

The meaning of the constants in the above formulas is given in Proposition 3.2.2 and Remark 3.2.4, where these bound are derived for a general operator. In our present case with two operators, p_0 and p_1 are the uniform spectral bounds of A w.r.t. G from (3.15), and $\hat{p}_0$ and $\hat{p}_1$ are the uniform spectral bounds of K w.r.t. G in the analogous sense.

In order to apply the results of subsection 3.3.3, we also need a bound for $\|L_S-N_S\|$. Since $M_L=\|L_S\|$ and $M_N=\|N_S\|$ above, we obtain in a similar way that

$$\|L_S-N_S\| \le \tilde{p}_1 + C_{\Omega,S}\,q^{-1/2}\|\mathbf{b}-\mathbf{w}\|_{L^\infty(\Omega)^d} + C^2_{\Omega,S}\|c-z\|_{L^\infty(\Omega)} + C^2_{\Gamma_N,S}\|\alpha-\eta\|_{L^\infty(\Gamma_N)} \tag{9.7}$$

where $\tilde{p}_1>0$ is a uniform upper spectral bound of $A-K$ w.r.t. G analogously to the right side of (3.15), that is, it satisfies

$$\big(A(x)-K(x)\big)\,\xi\cdot\xi \le \tilde{p}_1\;(G(x)\xi\cdot\xi) \qquad (x\in\overline{\Omega},\;\xi\in\mathbf{R}^d).$$

Then Propositions 3.3.4-3.3.5 yield

Proposition 9.1.1 *For any subspace $V_h \subset H^1_D(\Omega)$*

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \leq \frac{M_L M_N}{m_L m_N}$$

and

$$\kappa(\mathbf{N}_h^{-1}\mathbf{L}_h) \leq \Big(1 + \frac{m_L + m_N}{2m_L m_N}\, \|L_S - N_S\|\Big)^2 ,$$

with values of the bounds from (9.6)-(9.7) independently of V_h.

Consequently, by (3.53), the CGN algorithm (2.35) for system (9.4) converges with a ratio bounded independently of V_h.

For FDM discretizations of problem (7.1), mesh independent convergence of CG iterations under nonsymmetric preconditioning is presented by numerical tests in [45], and the boundedness of $\kappa(N_h^{-1}L_h)$ is proved in [49], both on the unit square. Similarly to the symmetric case, the bounds are not a priori defined in contrast to the FEM case like Proposition 9.1.1.

9.1.2 Superlinear convergence with nonsymmetric preconditioners

Based on Proposition 4.2.1, similarly to subsection 7.1.2, we now define N in (9.2) to have the same principal part as L, i.e.,

$$Nu := -\mathrm{div}\,(A\,\nabla u) + \mathbf{w}\cdot\nabla u + zu \qquad \text{for}\ \ u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_K} + \eta u_{|\Gamma_N} = 0 \qquad (9.8)$$

and assumed to satisfy Assumptions 3.2.2.1. Superlinear results are here only available for FEM discretizations. Accordingly, the preconditioner is the stiffness matrix $\mathbf{N}_h$ such that we set $K := A$ in (9.3). Similíarly to subsection 9.1.1, we apply the CGN algorithm (2.35) with the $\mathbf{S}_h$-inner product and by setting $A = \mathbf{N}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T\mathbf{N}_h^{-T}\mathbf{S}_h$.

In order to apply the results of subsection 4.3.2, we use the bounds in (9.5) and define the operator Q_S in $H^1_D(\Omega)$ via

$$\langle Q_S u, v\rangle_S = \int_\Omega \Big(((\mathbf{b}-\mathbf{w})\cdot\nabla u)v + (c-z)uv\Big) \ + \int_{\Gamma_N} (\alpha-\eta)uv\, d\sigma \qquad (u,v \in H^1_D(\Omega)), \ (9.9)$$

which obviously satisfies (4.10) in $H^1_D(\Omega)$. Then (4.13)-(4.14) yield

Proposition 9.1.2 [16]. *The superlinear convergence of the preconditioned CGN method is mesh independent, i.e.,*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq\ \varepsilon_k \qquad (k = 1,2,...,n) \qquad (9.10)$$

$$where \quad \varepsilon_k = \frac{2M_N^2}{km_L^2}\ \sum_{i=1}^{k}\Big(\frac{2}{m_N}\, s_i(Q_S) + \frac{1}{m_N^2}\, s_i(Q_S)^2\Big) \ \to 0 \qquad (as\ k \to \infty) \qquad (9.11)$$

and ε_k is a sequence independent of V_h.

9.2 Separable or constant coefficient preconditioners

In this section we give some nonsymmetric preconditioning operators of the type discussed in the previous section 9.1, based on [16]. Since the convergence results of the previous section are valid, here we are interested in the complexity of the preconditioning operators.

In general, the operator L in problem (7.1) has variable coefficients:

$$Lu = -\mathrm{div}\,(A(x)\,\nabla u) + \mathbf{b}(x)\cdot\nabla u + c(x)u \qquad \text{for} \quad u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_A} + \alpha(x)u_{|\Gamma_N} = 0, \tag{9.12}$$

where for clearness, the dependence of the coefficients on x has now been indicated. For convection-dominated problems (i.e. when $|\mathbf{b}|$ is large), the inclusion of nonsymmetric terms in N may turn it into a much better approximation of L than a symmetric preconditioner like (7.4). Although the preconditioner N thus becomes nonsymmetric as L itself, the solution of the auxiliary problems can still remain considerably simpler than the original one. For this, one can propose a preconditioning operator with constant coefficients:

$$Nu = \ -k\,\Delta u + \ \mathbf{w}\cdot\nabla u + zu \qquad \text{for} \quad u\in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0, \tag{9.13}$$

where $k > 0$, $\mathbf{w}\in\mathbf{R}^d$, $z, \eta \geq 0$ are constants such that $z > 0$ or $\eta > 0$ if $\Gamma_D = \emptyset$. Owing to the fact that N has constant coefficients, one can rely on efficient solution methods for the auxiliary problems. Here one can use either multigrid or multilevel methods, or (if Ω is rectangular, or the boundary conditions allow the problem to be easily embedded into a rectangular domain) fast direct solvers for separable equations are available, see e.g. [93].

The preconditioning operator (9.13) can be further simplified if one convection coefficient is dominating [14]. Assume that, say, $b_1(x)$ has considerably larger values than $b_j(x)$ $(j \geq 2)$. Then one can include only one nonsymmetric coefficient, i.e. propose the preconditioning operator

$$Nu = \ -k\,\Delta u + \ w_1\,\tfrac{\partial u}{\partial x_1} + zu \qquad \text{for} \quad u\in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0, \tag{9.14}$$

where $k, w_1, z, \eta \in \mathbf{R}$ are constants with the same properties as required for (9.13). In this case (above all, if $b_1(x)$ are large), the presence of the term $w_1\,\frac{\partial u}{\partial x_1}$ itself may turn N into a much better approximation of L. Nevertheless, since this term is one-dimensional, the solution of the auxiliary problems remains considerably simpler than the original one, e.g. via local 1D Green's functions [10].

The mesh independent linear convergence follows from Proposition 9.1.1. In particular, it is clear from (9.7) that the constants in N are best to be chosen as proper mean values of the corresponding coefficients of L, understood coordinatewise for $\mathbf{b}$ and as $k = (1/2)(\max_{x,\lambda}(\lambda(A(x)) + \min_{x,\lambda}(\lambda(A(x)))$ for A. Further, L itself has often a Laplacian principal part in convection-diffusion problems, in which case Proposition 9.1.2 provides mesh independent superlinear convergence.

We finally note that most of what has been said above holds for separable preconditioners as well, i.e. when $k, \mathbf{w}, z, \eta$ in N are not constants but suitable separable approximations of the corresponding coefficients of L. Mesh independence results for FDM discretizations under separable nonsymmetric preconditioning are presented by numerical tests in the previously mentioned paper [45].

9.3 Nonsymmetric preconditioners for systems

Analogously to the symmetric case in section 8.1, the above results can be extended to systems of convection-diffusion equations in a natural way. Namely, let us consider system (8.1). If there are large convection terms $\mathbf{b}_i$, then one can propose a nonsymmetric preconditioning operator N as an l-tuple of decoupled operators N_i, where each N_i is of the type (9.2). Then the linear convergence of the preconditioned CGN method is mesh independent in the sense of Proposition 9.1.1, further, if L_i have constant times Laplacian principal parts themselves (as well as chosen for N_i) then the superlinear convergence is also mesh independent in the sense of Proposition 9.1.2.

Since N_i are decoupled, the resulting algorithm is parallelizable. This turns it into an efficient method if, in particular, each N_i is like (9.13) or (9.14), or the problem itself is in 1D which may occur e.g. after using some method of splitting in meteorological models with several components, see [102].

10 Inner-outer iterations and element by element preconditioners

To realize the solution of a discretized elliptic problem using an equivalent operator preconditioner, it is in practice most efficient to use inner-outer iterations. That is, if L is the original and S is the preconditioning operator, then at each iteration step we can solve the arising systems with the discretization S_h of the operator S using some preconditioned iterative method, which will then be the inner iteration. The assumption is that it is easier to construct an efficient preconditioner for S_h than for L_h. For instance, L may be an elliptic operator with variable coefficients while S may be chosen as one with (piecewise) constant coefficients.

As shown in section 2.3, inner iterations can be treated in the framework of variable preconditioning, in which we use different numbers of inner iterations to satisfy some variable inner iteration accuracy. Following [23], such preconditioners can be defined by an in general nonlinear mapping $r \mapsto B[r]$ such that $L_h B[r] \approx r$, i.e. $B[r]$ is an approximation of $L_h^{-1} r$ for a given residual r. Let us assume that the preconditioner S_h satisfies

$$\langle L_h S_h^{-1} v, v \rangle \geq m \|v\|^2, \qquad \|L_h S_h^{-1} v\| \leq M \|v\| \qquad (\forall v \in \mathbf{R}^n) \tag{10.1}$$

for some constants $M \geq m > 0$. If we solve the arising systems with S_h sufficiently accurately, i.e. using a sufficient number of inner iterations, then we may assume that the mapping $B[\,.\,]$ that corresponds to the inner iterations satisfies

$$\langle L_h B[v], v \rangle \geq (m - \varepsilon) \|v\|^2, \qquad \|L_h B[v]\| \leq (M + \varepsilon) \|v\| \qquad (\forall v \in \mathbf{R}^n),$$

where $0 \leq \varepsilon < m$. By Proposition 2.3.1, it follows that the outer iteration converges with a rate

$$\frac{\|r_{k+1}\|}{\|r_k\|} \leq \left(1 - \left(\frac{m - \varepsilon}{M + \varepsilon} \right)^2 \right)^{1/2} = \frac{\sqrt{(M+m)(M-m+2\varepsilon)}}{M + \varepsilon} \qquad (k = 1, 2, ..., n). \tag{10.2}$$

An important particular application of inner-outer iterations arises when we use elementwise constructed preconditioners. For various reasons, this can be a very efficient technique [24]. The original idea goes back to [20, 73], later developments can be found in [7, 8]. Such methods can, however, never give optimal order, i.e. h-independent convergence rates, as they are similar to block diagonal preconditioning methods. More recently, a new type of elementwise preconditioners has been developed, based on a partitioning of the node set in two subsets [24]. For simplicity, we consider an elliptic problem on a bounded, planar domain which has been divided into non-overlapping triangles. Each triangle which forms a macroelement may itself be subdivided into smaller triangular elements (microelements). Then we let the vertices of the coarse triangulation define the node set S_0, and the remaining (edge and interior) points form the other node set S_1. The finite element matrix can be ordered accordingly and split in a two by two block form

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where A_{22} corresponds to the coarse node set S_0 and accordingly, A_{11} corresponds to the node points arising due to the refinement of the coarse mesh. The matrix A can be factored as

$$A = \begin{bmatrix} A_{11} & 0 \\ A_{21} & I_2 \end{bmatrix} \begin{bmatrix} I_1 & A_{11}^{-1}A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

where $A_{22} - A_{21}A_{11}^{-1}A_{12} =: S_2$ is the Schur complement matrix.

When solving the arising systems with A, we may apply some inner iteration method for the arising system with A_{11} (which appears twice, once for each block factor). Similarly, systems with S_2 may be solved by inner iterations (possibly involving inner iterations for A_{11} when computing the matrix-vector actions for S_2). For reasons of efficiency, we must apply accurate but not too costly preconditioners when solving these systems. The preconditioners we advocate here are based on local element approximations for the arising matrices, and can be understood in the differential operator framework. Namely, let us first consider the matrix A_{11}. The continuous counterpart of A_{11} is the differential operator L defined on the whole domain but where we have imposed homogeneous Dirichlet boundary conditions on the vertex node set S_0. To construct a preconditioner B_{11} to A_{11}, the domain is partitioned in macroelements as defined above where on each macroelement edge we impose homogeneous Neumann boundary conditions. In this approach there is no need to assemble the corresponding global finite element matrices, as all matrix-vector multiplications with A_{11} and solutions of the arising block diagonal systems with B_{11} can be done elementwise.

In an alternative approach, one first assembles A_{11} before the local element matrices in its approximation B_{11} are constructed, as this can make the preconditioning more robust w.r.t. jumps in coefficients and anisotropy.

Acknowledgements. The second author thanks for the support by the Hungarian National Office for Research and Technology NKTH, by the Hungarian Research Grant OTKA No.K 67819 and by HAS under the Bolyai János Scholarship.

References

[1] Antal I., Karátson J., A mesh independent superlinear algorithm for some nonlinear nonsymmetric elliptic systems, to appear in *Comput. Math. Appl.*, 2007

[2] Ashby, S. F., Manteuffel, T. A., Saylor, P. E., A taxonomy for conjugate gradient methods, *SIAM J. Numer. Anal.* 27 (1990), no. 6, 1542–1568.

[3] Axelsson, O., A generalized conjugate gradient least square method, *Numer. Math.* 51 (1987), 209-227.

[4] Axelsson, O., *Iterative Solution Methods,* Cambridge University Press, 1994.

[5] Axelsson, O., On iterative solvers in structural mechanics; separate displacement orderings and mixed variable methods, *Math. Comput. Simulation* 50 (1999), no. 1-4, 11–30.

[6] Axelsson, O., Barker, V. A., Neytcheva, M., Polman, B., Solving the Stokes problem on a massively parallel computer, *Math. Model. Anal.* 6 (2001), no. 1, 7–27.

[7] Axelsson, O., Blaheta, R. and Neytcheva, M., Preconditioning of boundary value problems using elementwise Schur complements. Technical Report 2006-048, Department of Information Technology, Uppsala University, November 2006.

[8] Axelsson, O., Blaheta, R. and Neytcheva, M., A black–box generalized conjugate gradient minimum residual method based on variable preconditioners and local element approximations, submitted to *ETNA*.

[9] Axelsson, O., Faragó I., Karátson J., Sobolev space preconditioning for Newton's method using domain decomposition, *Numer. Lin. Alg. Appl.*, **9** (2002), 585-598.

[10] Axelsson, O., Gololobov, S. V., A combined method of local Green's functions and central difference method for singularly perturbed convection-diffusion problems, *J. Comput. Appl. Math.* 161 (2003), no. 2, 245–257.

[11] Axelsson, O., Kaporin, I., On the sublinear and superlinear rate of convergence of conjugate gradient methods. Mathematical journey through analysis, matrix theory and scientific computation (Kent, OH, 1999), *Numer. Algorithms* 25 (2000), no. 1-4, 1–22.

[12] Axelsson, O., Karátson J., Symmetric part preconditioning for the conjugate gradient method in Hilbert space, *Numer. Funct. Anal.* **24** (2003), No. 5-6, 455-474.

[13] Axelsson, O., Karátson J., Conditioning analysis of separate displacement preconditioners for some nonlinear elasticity systems, *Math. Comput. Simul.* **64** (2004), No.6, pp. 649-668.

[14] Axelsson, O., Karátson J., Superlinearly convergent CG methods via equivalent preconditioning for nonsymmetric elliptic operators, *Numer. Math.* 99 (2004), No. 2, 197-223.

[15] Axelsson, O., Karátson J., Symmetric part preconditioning of the CGM for Stokes type saddle-point systems, *Numer. Funct. Anal.* 28 (2007), 9-10, pp. 1027-1049

[16] Axelsson, O., Karátson J., Mesh independent superlinear PCG rates via compact-equivalent operators, *SIAM J. Numer. Anal.*, 45 (2007), No.4, pp. 1495-1516 (electronic)

[17] Axelsson, O., Kolotilina, L., Diagonally compensated reduction and related preconditioning methods, *Numer. Linear Algebra Appl.* 1 (1994), no. 2, 155–177.

[18] Axelsson, O., Maubach, J., On the updating and assembly of the Hessian matrix in finite element methods, *Comp. Meth. Appl. Mech. Engrg.*, 71 (1988), pp. 41-67.

[19] Axelsson, O., Neytcheva, M., Scalable algorithms for the solution of Navier's equations of elasticity, *J. Comput. Appl. Math.* 63 (1995), no. 1-3, 149–178.

[20] Axelsson, O., Neytcheva, M., An iterative solution method for Schur complement systems with inexact inner solver. In O. Iliev, M. Kaschiev, S. Margenov, Bl. Sendov, P.S. Vassilevski eds., *Recent Advances in Numerical Methods and Applications II*, World Scientific, 1999, 795–803.

[21] Axelsson, O., Vassilevski, P.S., Algebraic multilevel preconditioning methods I., *Numer. Math.* 56 (1989), pp. 157-177.

[22] Axelsson, O., Vassilevski, P.S., Algebraic multilevel preconditioning methods II., *SIAM J. Numer. Anal.* 27 (1990), pp. 1569-1590.

[23] Axelsson, O., Vassilevski, P.S., Variable-step multilevel preconditioning methods. I. Selfadjoint and positive definite elliptic problems, *Numer. Linear Algebra Appl.* 1 (1994), no. 1, 75–101.

[24] Bängtsson, E., Neytcheva, M., Finite Element Block-Factorized Preconditioners, Technical reports from the Department of Information Technology, Uppsala University, No. 2007-008, March 2007

[25] Bank, R.E., Marching algorithms for elliptic boundary value problems. II. The variable coefficient case, *SIAM J. Numer. Anal.* 14 (1977), no. 5, 950–970.

[26] Benzi, M., Golub, G. H., Liesen, J., Numerical solution of saddle point problems, *Acta Numer.* 14 (2005), 1–137.

[27] Blaheta, R., Displacement decomposition—incomplete factorization preconditioning techniques for linear elasticity problems, *Numer. Linear Algebra Appl.* 1 (1994), no. 2, 107–128.

[28] Blaheta, R., Multilevel Newton methods for nonlinear problems with applications to elasticity, Copernicus 940820, Technical report. Ostrava, 1997.

[29] Börgers, C., Widlund, O. B., On finite element domain imbedding methods, *SIAM J. Numer. Anal.* 27 (1990), no. 4, 963–978.

[30] Bramble, J. H., Pasciak, J. E., Preconditioned iterative methods for nonselfadjoint or indefinite elliptic boundary value problems, in: *Unification of finite element methods*, 167–184, North-Holland Math. Stud., 94, North-Holland, Amsterdam, 1984.

[31] Brezzi, F., Raviart, P.-A., Mixed finite element methods for 4th order elliptic equations, in: *Topics in numerical analysis III*, Proc. Roy. Irish Acad. Conf., Trinity Coll., Dublin (1976), pp. 33–56.

[32] W. Cao, R.D. Haynes, M. R. Trummer, Preconditioning for a class of spectral differentiation matrices, *J. Sci. Comput.* 24 (2005), No. 3, 343-371.

[33] Carey, G.F., Jiang, B.-N., Nonlinear preconditioned conjugate gradient and least-squares finite elements, *Comp. Meth. Appl. Mech. Engrg.*, 62 (1987), pp. 145-154.

[34] Ciarlet, P. G., *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978

[35] Ciarlet, Ph., *Mathematical elasticity.* Vol. I. Three-dimensional elasticity. Studies in Mathematics and its Applications, 20. North-Holland, 1988.

[36] Ciarlet, P. G., Raviart, P.-A., Maximum principle and uniform convergence for the finite element method, *Comput. Methods Appl. Mech. Engrg.* 2 (1973), 17–31.

[37] Concus, P., Golub, G.H., Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations, *SIAM J. Numer. Anal.* 10 (1973), 1103–1120.

[38] Concus, P., Golub, G.H., A generalized conjugate method for non-symmetric systems of linear equations, in: *Lect. Notes Math. Syst.* 134 (eds. Glowinski, R., Lions, J.-L.), pp. 56-65, Springer, 1976.

[39] Czách, L., *The steepest descent method for elliptic differential equations* (in Russian), C.Sc. thesis, 1955.

[40] Dryja, M. A priori estimates in W_2^2 in a convex domain for systems of elliptic difference equations (Russian), *Ž. Vyčisl. Mat. i Mat. Fiz.* 12 (1972), 1595–1601, 1632.

[41] Dryja, M., An iterative substructuring method for elliptic mortar finite element problems with discontinuous coefficients, in *Domain decomposition methods* 10 (Boulder, CO, 1997), Contemp. Math., 218, AMS-Providence, RI, 1998; 94–103.

[42] D'yakonov, E. G., On an iterative method for the solution of finite difference equations (in Russian), *Dokl. Akad. Nauk SSSR* 138 (1961), 522–525.

[43] D'yakonov, E. G., The construction of iterative methods based on the use of spectrally equivalent operators, *USSR Comput. Math. and Math. Phys.*, 6 (1965), pp. 14-46.

[44] Eisenstat, S.C., Elman, H.C., Schultz. M.H., Variational iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal.* 20 (1983), no. 2, 345–357.

[45] Elman, H.C., Schultz. M.H., Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations, *SIAM J. Numer. Anal.,* 23 (1986), 44-57.

[46] Elman, H. C., Silvester, D. J., Wathen, A. J., *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2005.

[47] Ewing, R. E., Margenov, S. D., Vassilevski, P. S., Preconditioning the biharmonic equation by multilevel iterations, *Math. Balkanica* (N.S.) 10 (1996), no. 1, 121–132.

[48] Faber, V., Manteuffel, T., Parter, S.V., Necessary and sufficient conditions for the existence of a conjugate gradient method, *SIAM J. Numer. Anal.* 21 (1984), no. 2, 352–362.

[49] Faber, V., Manteuffel, T., Parter, S.V., On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations, *Adv. in Appl. Math.,* 11 (1990), 109-163.

[50] Faragó, I., Karátson, J., *Numerical solution of nonlinear elliptic problems via preconditioning operators. Theory and applications.* Advances in Computation, Volume 11, NOVA Science Publishers, New York, 2002.

[51] Funaro, D., *Polynomial approximation of differential equations,* Lecture Notes in Physics, New Series, Monographs 8, Springer, 1992.

[52] Goldstein, C. I., Manteuffel, T. A., Parter, S. V., Preconditioning and boundary conditions without H_2 estimates: L_2 condition numbers and the distribution of the singular values, *SIAM J. Numer. Anal.* 30 (1993), no. 2, 343–376.

[53] Golub G.H., Ye Q., Inexact preconditioned conjugate gradient method with inner–outer iteration. *SIAM J. Sci. Comput.*, 21(4): 1305–1320, 1999/00.

[54] Graham, I. G., Hagger, M. J. Unstructured additive Schwarz-conjugate gradient method for elliptic problems with highly discontinuous coefficients, *SIAM J. Sci. Comput.* 20 (1999), 2041–2066 (electronic).

[55] Greenbaum, A., Diagonal scalings of the Laplacian as preconditioners for other elliptic differential operators, *SIAM J. Matrix Anal. Appl.,* 13 (1992), 826-846.

[56] Greenbaum, A., *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, PA (1997).

[57] Guillard, H., Désidéri, J.-A., Iterative methods with spectral preconditioning for elliptic equations, *Comput. Methods Appl. Mech. Engrg.* 80 (1990), no. 1-3, 305–312.

[58] Gunn, J. E., The numerical solution of $\nabla \cdot a \nabla u = f$ by a semi-explicit alternating direction iterative method, *Numer. Math.* 6 (1964), 181-184.

[59] Gunn, J. E., The solution of elliptic difference equations by semi-explicit iterative techniques, *SIAM J. Numer. Anal. Ser. B* 2 (1965), 24–45.

[60] Gustafsson, I., A class of first order factorization methods, *BIT* 18 (1978), no. 2, 142–156.

[61] Hackbusch, W., *Multigrid Methods and Applications*, Springer Series in Computational Mathematics 4, Springer, Berlin, 1985.

[62] Hestenes, M.R., Stiefel, E., Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards*, Sect. B, 49 (1952) No.6., 409-436.

[63] Horgan, C. O., Korn's inequalities and their applications in continuum mechanics. *SIAM Rev.* 37 (1995), no. 4, 491–511.

[64] Joubert, W., Manteuffel, T. A., Parter, S., Wong, S.-P., Preconditioning second-order elliptic operators: experiment and theory, *SIAM J. Sci. Statist. Comput.* 13 (1992), no. 1, 259–288.

[65] Karátson J., Mesh independent superlinear convergence estimates of the conjugate gradient method for some equivalent self-adjoint operators *Appl. Math.* (Prague) 50 (2005), No. 3, 277-290.

[66] Karátson J., Superlinear PCG algorithms: symmetric part preconditioning and boundary conditions, Preprint, ELTE Dept. Appl. Anal. Comp. Math., http://www.cs.elte.hu/applanal/preprints; 2006-10

[67] Karátson J., On the superlinear convergence rate of the preconditioned CGM for some nonsymmetric elliptic problems, *Numer. Funct. Anal.* 28 (2007), 9-10, pp. 1153-1164

[68] Karátson J., Faragó I., Variable preconditioning via quasi-Newton methods for nonlinear problems in Hilbert space, *SIAM J. Numer. Anal.* **41** (2003), No. 4, 1242-1262.

[69] Karátson J., Kurics T., Superlinearly convergent PCG algorithms for some nonsymmetric elliptic systems, *J. Comp. Appl. Math.* (2007), available online at http://dx.doi.org/doi:10.1016/j.cam.2006.12.004

[70] Karátson J., Kurics T., Superlinear PCG Methods for FDM Discretizations of Convection-Diffusion Equations, Preprint, ELTE Dept. Appl. Anal. Comp. Math., http://www.cs.elte.hu/applanal/preprints; 2006-13

[71] Karátson J., Kurics T., Lirkov, I., A Parallel Algorithm for Systems of Convection-Diffusion Equations, in: *NMA 2006*, eds. T. Boyanov et al., *Lecture Notes Comp. Sci.* 4310, pp. 65-73, Springer, 2007.

[72] Knyazev, A, Lashuk, I., Steepest descent and conjugate gradient methods with variable preconditioning. Electronic. Math. NA/0605767, arXiv. org., http://arxiv.org/abs/math/0605767, 2006-2007.

[73] J. Kraus, Algebraic multilevel preconditioning of finite element matrices using local Schur complements, *Numer. Lin. Algebra Appl.*, 13 (2006),49–70.

[74] Křížek, M., Lin Qun, On diagonal dominance of stiffness matrices in 3D, *East-West J. Numer. Math.* 3 (1995), 59–69.

[75] Loghin, D., Green's Functions for Preconditioning, DPhil Thesis, Oxford, 1999.

[76] Manteuffel, T., The Tchebychev iteration for nonsymmetric linear systems, *Numer. Math.* 28 (1977), no. 3, 307–327.

[77] Manteuffel, T., Otto, J., Optimal equivalent preconditioners, *SIAM J. Numer. Anal.,* 30 (1993), 790-812.

[78] Manteuffel, T., Parter, S. V., Preconditioning and boundary conditions, *SIAM J. Numer. Anal.* 27 (1990), no. 3, 656–694.

[79] Mayo, A., Greenbaum, A., Fast parallel iterative solution of Poisson's and the biharmonic equations on irregular regions, *SIAM J. Sci. Statist. Comput.* 13 (1992), no. 1, 101–118.

[80] Mikhlin, S.G., *The Numerical Performance of Variational Methods,* Walters-Noordhoff, 1971

[81] Mikhlin, S.G., *Constants in some inequalities of analysis* (translated from the Russian by R. Lehmann), John Wiley and Sons, Ltd., Chichester, 1986.

[82] Nečas, J., Hlaváček, I., *Mathematical Theory of Elastic and Elasto-plastic Bodies: an Introduction,* Studies in Applied Mechanics 3, Elsevier Scientific Publishing Co., Amsterdam-New York, 1980.

[83] Nevanlinna, O., *Convergence of Iterations for Linear Equations,* Birkhäuser, Basel, 1993.

[84] Neuberger, J. W., *Sobolev Gradients and Differential Equations*, Lecture Notes in Math., No. 1670, Springer, 1997.

[85] Nitsche, J., Nitsche, J. C. C., Error estimates for the numerical solution of elliptic differential equations, *Arch. Rational Mech. Anal.* 5 (1960), 293–306.

[86] Repin, S., Sauter, S., Smolianski, A., A posteriori error estimation for the Poisson equation with mixed Dirichlet/Neumann boundary conditions, *J. Comput. Appl. Math.* 164-165 (2004), 601–612.

[87] Rossi, T., Toivanen, J., Parallel fictitious domain method for a non-linear elliptic Neumann boundary value problem, Czech-US Workshop in Iterative Methods and Parallel Computing, Part I (Milovy, 1997), *Numer. Linear Algebra Appl.* 6 (1999), no. 1, 51–60.

[88] Rossi, T., Toivanen, J., A parallel fast direct solver for block tridiagonal systems with separable matrices of arbitrary dimension, *SIAM J. Sci. Comput.* 20 (1999), no. 5, 1778–1796.

[89] Saad, Y., Schultz, M.H., GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Statist. Comput.* 7 (1986), 856–869.

[90] Y. Saad, *Iterative methods for sparse linear systems*, PWS Publ Co, Boston 1996

[91] Simoncini V., Szyld D.B., Flexible inner-outer Krylov subspace methods, *SIAM J. Numer. Anal.*, 40 (2003), 2219–2239

[92] Sundqvist, P., Numerical Computations with Fundamental Solutions, PhD thesis, Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology, Uppsala University, 2005.

[93] Swarztrauber, P. N., A direct method for the discrete solution of separable elliptic equations, *SIAM J. Numer. Anal.* 11 (1974), 1136–1150.

[94] Swarztrauber, P. N., The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle, *SIAM Rev.* 19 (1977), no. 3, 490–501.

[95] Temam, R., *Navier-Stokes equations. Theory and numerical analysis*, Studies in Mathematics and its Applications, Vol. 2. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1977.

[96] Vassilevski, P. S. Fast algorithm for solving a linear algebraic problem with separable variables, *C. R. Acad. Bulgare Sci.* 37 (1984), no. 3, 305–308.

[97] Vladimirov, V. S., *Equations of mathematical physics* (translated from the Russian by E. Yankovsky), Mir, Moscow, 1984.

[98] Winter, R., Some superlinear convergence results for the conjugate gradient method, *SIAM J. Numer. Anal.*, 17 (1980), 14-17.

[99] Widlund, O., On the use of fast methods for separable finite difference equations for the solution of general elliptic problems, in *Sparse Matrices and Applications*, D.J. Rose and R.A. Willoughby (eds.), Plenum Press, N.Y. 1972, pp. 121–134.

[100] Widlund, O., A Lanczos method for a class of non-symmetric systems of linear equations, *SIAM J. Numer. Anal.,* 15 (1978), 801-812.

[101] Young, D. M., *Iterative Solution of Large Linear Systems,* Academic Press, New York-London, 1971.

[102] Zlatev, Z., *Computer Treatment of Large Air Pollution Models*, Kluwer Academic Publishers, Dordrecht-Boston-London, 1995.