A PRECONDITIONED GMRES METHOD FOR NONSYMMETRIC OR INDEFINITE PROBLEMS

JINCHAO XU AND XIAO-CHUAN CAI

ABSTRACT. A preconditioning technique is proposed for nonsymmetric or indefinite linear systems of equations. The main idea in our theory, roughly speaking, is first to use some “coarser mesh” space to correct the nonpositive portion of the eigenvalues of the underlying operator and then switch to use a symmetric positive definite preconditioner. The generality of our theory allows us to apply any known preconditioners that were originally designed for symmetric positive definite problems to nonsymmetric or indefinite problems, without losing the optimality that the original one has. Some numerical experiments based on GMRES are reported.

1. INTRODUCTION

In this paper, we are interested in developing efficient algorithms for solving nonsymmetric or indefinite linear systems of equations that are governed by a symmetric positive definite operator. A typical example is the linear system obtained by discretizing a second-order elliptic equation with first- or zero-order derivative terms by finite element or finite difference methods. If the coefficients of these lower-order terms are relatively not too large, the corresponding system is then dominated by its symmetric positive definite part. This property has been used in many of the preconditioned algorithms developed before (cf. [1, 6, 9]). The main idea there, however, is to use the symmetric positive definite part as a preconditioner and then solve the normal equation by the conjugate gradient method, in which the condition number of the system is squared. Though this is a reasonable approach in some cases, other efficient algorithms are possible. The generalized minimal residual (GMRES) method [7], among many other related algorithms, provides an alternative approach.

This work was mainly motivated by the Ph.D. thesis of the second author. In one of the algorithms in [2], a nonsymmetric linear system obtained from the finite element discretization of an elliptic partial differential equation is preconditioned by an additive domain decomposition method which involves the solution of some local subproblems and also of a global coarse-mesh problem.
An interesting phenomenon observed is that the coarse-mesh solver for the original nonsymmetric problem is crucial for the success of the algorithm. In other words, the local problems do not need to be solved exactly in order to obtain optimal convergence; however, the coarse-mesh problem has to be solved exactly. Based on our experiences with multigrid methods (cf. [10]), we tend to believe that the reason is due to the fact that the symmetric positive definite part of the system governs the equation very well on the propagation of the high-frequency modes of the error, but very poorly on the low frequencies. The importance of the coarse-mesh solver then becomes clear, since its role (as in multigrid methods) is just to liquidate those troublesome low-frequency modes of the error. Instead of using the coarse-mesh solver plus local subproblem solvers, as in [2], we show that an optimal preconditioner can be constructed by the coarse-mesh solver plus any other optimal preconditioners designed for the symmetric part of the original operator. We note that the analysis of such a more general preconditioner is quite different from that in [2].

The rest of this paper is organized as follows. In §2, we prove an abstract theorem based on a number of algebraic assumptions. In §3, we demonstrate how the theory in §2 can be applied to some algebraic iterative methods, such as GMRES. Some numerical examples involving second-order elliptic equations will be presented. Finally, some concluding remarks will be made in §4.

2. Main result

Let \( V \) be a finite-dimensional Hilbert space with an inner product \((\cdot, \cdot)\) and the corresponding norm \( \| \cdot \| = (\cdot, \cdot)^{1/2} \). For a given \( f \in V \), we are interested in solving the linear system of equations

\[
A_N u = f, 
\]

where \( A_N \) is the sum of two linear operators, \( A_N = A + N \). We assume that the equation (1) has a unique solution in \( V \). Moreover, we make the following assumptions:

A.1. The mapping \( A: V \to V \) is symmetric and positive definite with respect to the inner product \((\cdot, \cdot)\). We denote \((A\cdot, \cdot)\), which defines an inner product on \( V \), by \((\cdot, \cdot)_A\) and its induced norm by \( \| \cdot \|_A \). Furthermore, there exists a constant \( c_1 \) such that

\[
\| u \| \leq c_1 \| u \|_A \quad \forall u \in V. 
\]

A.2. The mapping \( N: V \to V \) satisfies

\[
(Nu, v) \leq c_2 \| u \| \| v \|_A \quad \forall u, v \in V, 
\]

where \( c_2 \) is a positive constant. Note that we do not assume that \((N\cdot, \cdot)\) is symmetric with respect to the inner product \((\cdot, \cdot)\).

A.3. There exists a subspace \( V_0 \subset V \) such that for any given \( u \in V \), there exists a unique \( u_0 \in V_0 \) such that

\[
(A_N u_0, v) = (A_N u, v) \quad \forall v \in V_0. 
\]

The above equation defines an operator \( P_0: V \to V_0 \) so that \( P_0 u = u_0 \).
Corresponding to the subspace $V_0$ introduced in Assumption A.3, we define a linear operator $A_0 : V_0 \to V_0$ and a projection $Q_0 : V \to V_0$ by

$$
(A_0 v_0, v_0) = (A_N v_0, v_0) \quad \forall v_0, v_0 \in V_0
$$

and

$$
(Q_0 v, v_0) = (u, v_0) \quad \forall u \in V, v_0 \in V_0,
$$

respectively. By Assumption A.3, $A_0$ is invertible.

Following [10], by the definitions of $P_0$, $A_0$, and $Q_0$, we can easily see that

$$
A_0 P_0 = Q_0 A_N.
$$

Based on the operators $A_0$ and $Q_0$, we propose a preconditioner for $A_N$ as follows:

$$
B_N = A_0^{-1} Q_0 + \beta B,
$$

where $B$ is a given symmetric positive definite preconditioner for $A$, and $\beta$ is a given positive constant depending on $B$. A proper choice of $\beta$ is necessary, since both terms on the right-hand side of (6) should somehow be “balanced”.

Combining (5) with (6) gives that

$$
B_N A_N = P_0 + \beta B A_N.
$$

To study the property of the operator $B_N A_N$, we introduce the following parameter:

$$
\delta_0 = \sup_{v \in V} \frac{\| (I - P_0) v \|_A}{\| v \|_A}.
$$

We begin our analysis with a simple lemma.

**Lemma 1.** There holds

$$
\| P_0 u \|_A \leq (1 + c_2 \delta_0) \| u \|_A \quad \forall u \in V,
$$

and

$$
\| P_0 u \|_A^2 \leq 2 (P_0 u, u)_A + c_2^2 \delta_0^2 \| u \|_A^2 \quad \forall u \in V.
$$

**Proof.** By definitions, and A.2,

$$
\| P_0 u \|_A^2 = A(P_0 u, P_0 u) = (A_N u, P_0 u) - (N P_0 u, P_0 u)
= (A u, P_0 u) + (N(I - P_0) u, P_0 u)
\leq \| u \|_A \| P_0 u \|_A + c_2 \delta_0 \| u \|_A \| P_0 u \|_A.
$$

The first inequality then follows.

To prove the second inequality, we have

$$
\| P_0 u \|_A^2 = (P_0 u, u)_A + (P_0 u, P_0 u - u)_A
= (P_0 u, u)_A + (N(I - P_0) u, P_0 u)
\leq (P_0 u, u)_A + c_2 \delta_0 \| u \|_A \| P_0 u \|_A
\leq (P_0 u, u)_A + \frac{1}{2} c_2^2 \delta_0^2 \| u \|_A^2 + \frac{1}{2} \| P_0 u \|_A^2.
$$

The desired result then follows. \qed

We now present the main result of the paper.
**Theorem 1.** There exist positive constants \( \varepsilon, \alpha, \beta, \) and \( \mu, \) depending on \( \lambda_0 \) and \( \lambda_1 \) (the minimal and maximal eigenvalues of \( BA \), respectively), such that, for the \( B_N \) given by (6), if \( \delta_0 \leq \varepsilon, \)

\[
\|B_N A_N u\|_A < \mu \|u\|_A \quad \forall u \in V,
\]

and

\[
(B_N A_N u, u)_A \geq \alpha (u, u)_A \quad \forall u \in V.
\]

**Proof.** We first show the estimate (9). By A.2, we have

\[
\|u\|_A^2 = (Au, u) \leq \lambda_0^{-1} (BAu, u)_A \leq \lambda_0^{-1} (BANu, u)_A - \lambda_0^{-1} (Nu, BAu)_A \leq \lambda_0^{-1} (BANu, u)_A + c_2 \lambda_0^{-1} \|u\| \|BAu\|_A \leq \lambda_0^{-1} (BANu, u)_A + c_2 \lambda_1 \lambda_0^{-1} \|u\| \|u\|_A.
\]

Therefore,

\[
\|u\|_A^2 \leq 2 \lambda_0^{-1} (BANu, u)_A + c_2^2 \left( \frac{\lambda_1}{\lambda_0} \right)^2 \|u\|^2.
\]

On the other hand, by A.1, Lemma 1, and the definition of \( \delta_0, \) we have

\[
\|u\|_A^2 \leq 2 \|P_0 u\|_A^2 + 2 \|u - P_0 u\|_A^2 \leq 2 c_1^2 (2 (P_0 u, u)_A + c_2^2 \delta_0^2 \|u\|_A^2) + 2 \beta^2 \|u\|_A^2 = 4 c_1^2 (P_0 u, u)_A + 2 (c_1^2 c_2^2 + 1) \delta_0^2 \|u\|_A^2.
\]

Consequently,

\[
\|u\|_A^2 \leq 2 \lambda_0^{-1} (BANu, u)_A + 4 c_1^2 c_2^2 \left( \frac{\lambda_1}{\lambda_0} \right)^2 (P_0 u, u)_A + 2 c_2^2 \left( \frac{\lambda_1}{\lambda_0} \right)^2 (c_1^2 c_2^2 + 1) \delta_0^2 \|u\|_A^2.
\]

Let \( \varepsilon > 0 \) be such that

\[
2 c_2^2 \left( \frac{\lambda_1}{\lambda_0} \right)^2 (c_1^2 c_2^2 + 1) \varepsilon^2 = \frac{1}{2};
\]

then, if \( \delta_0 \leq \varepsilon, \)

\[
\|u\|_A^2 \leq 4 \lambda_0^{-1} (BANu, u)_A + 8 c_1^2 c_2^2 \left( \frac{\lambda_1}{\lambda_0} \right)^2 (P_0 u, u)_A.
\]

Estimate (9) then follows if we take

\[
\alpha = \frac{\lambda_0^2}{8 c_1^2 c_2^2 \lambda_1^2}, \quad \beta = \frac{\lambda_0}{2 c_1^2 c_2^2 \lambda_1^2}.
\]

We are now in a position to prove (8). It is easy to see that

\[
\|BANu\|_A \leq \|BAu\|_A + \|BNU\|_A \leq \lambda_1 \|u\|_A + \|BNU\|_A.
\]

By using Assumption A.2, we have

\[
\|BNU\|_A^2 = (ABNU, BNU) = (NU, BABNU) \leq c_2 \|u\| \|BABNU\|_A \leq c_2 \lambda_1 \|u\| \|BNU\|_A.
\]
Therefore,
\[ \| B N u \|_A \leq c_2 \lambda_1 \| u \| \leq c_1 c_2 \lambda_1 \| u \|_A. \]
Combining this with (7) and Lemma 1, we then get
\[ \| B N A_N u \|_A \leq \| P_0 u \|_A + \beta \| B A_N \|_A \leq (1 + c_2 \delta_0 + \beta c_1 c_2 \lambda_1) \| u \|_A. \]
Estimate (8) now follows with \( \mu = 1 + c_2 \epsilon + \beta c_1 c_2 \lambda_1 \).

3. Applications and numerical examples

The theory developed in this paper can be applied together with, among many others (see Remark 3.1 below), the GMRES method. The GMRES method, introduced in [7], is mathematically equivalent to the generalized conjugate residual (GCR) method [8] and can be used to solve the linear system of algebraic equations

\[ Gu = g, \]
where \( G \) is a linear operator which may be nonsymmetric or indefinite, defined on a finite-dimensional vector space \( V \), and \( g \) is a given vector in \( V \).

The method begins with an initial approximate solution \( u_0 \in V \) and an initial residual \( r_0 = g - Gu_0 \). At the \( m \)th iteration, a correction vector \( z_m \) is computed in the Krylov subspace
\[ \mathcal{X}_m(r_0) = \text{span}\{r_0, Gr_0, \ldots, G^{m-1}r_0\} \]
which minimizes the residual, \( \min_{z \in \mathcal{X}_m(r_0)} \| g - G(u_0 + z) \| \) for some appropriate norm \( \| \cdot \| \). The \( m \)th iteration is then \( u_m = u_0 + z_m \). According to the theory of [8], the rate of convergence of the GMRES method can be estimated by the ratio of the minimal eigenvalue of the symmetric part of the operator to the norm of the operator. These two quantities are defined by
\[ c_G = \inf_{u \neq 0} \frac{[u, Gu]}{[u, u]} \quad \text{and} \quad C_G = \sup_{u \neq 0} \frac{\| Gu \|}{\| u \|}, \]
where \([\cdot, \cdot]\) is an inner product on \( V \) that induces the norm \( \| \cdot \| \). We have the following theorem for the rate of convergence.

**Theorem 2** ([8]). If \( c_G > 0 \), then the GMRES method converges, and at the \( m \)th iteration the residual is bounded as
\[ \| \| r_m \| \| \leq \left( 1 - \frac{c_2^2 G}{\lambda_1} \right)^{m/2} \| r_0 \|, \]
where \( r_m = g - Gu_m \).

We shall now demonstrate how the results in the preceding section can be used with the GMRES method described above. We are interested in solving the equation (1). To this end, we use the preconditioner \( B_N \) defined by (6) and consider the equivalent preconditioned system

\[ B_N A_N u = B_N f. \]
Therefore, we come to the equation (10) with \( G = B_N A_N \) and \( g = B_N f \). The next thing is to make a proper choice of the inner product \([\cdot, \cdot]\), which will...
be \((A, \cdot, \cdot)\). It follows from Theorems 1 and 2 that the corresponding GMRES method has the following convergence rate estimate:

\[
||r_m||_A \leq \left( 1 - \frac{\alpha^2}{\mu^2} \right)^{m/2} ||r_0||_A.
\]

Consequently, if \(B\) is a good preconditioner for \(A\) (e.g., the condition number of \(BA\) is uniformly bounded), \(\alpha^2/\mu^2\) will be bounded from above by a positive number less than 1 (uniformly). As a result, the GMRES method for (11) will have a good convergence rate.

**Remark 3.1.** There are many other conjugate gradient type methods obeying the estimate given in Theorem 2, such as the so-called Orthomin\((k)\), GCR\((k)\), and MR methods; we refer to [8] for details. Our preconditioning technique can thus be applied to these algorithms as well.

We consider the Dirichlet boundary value problem

\[
\begin{cases}
LU = F & \text{in } \Omega, \\
U = 0 & \text{on } \partial \Omega,
\end{cases}
\]

where \(\Omega\) is a bounded domain in \(\mathbb{R}^2\) with boundary \(\partial \Omega\), and \(L\) is a second-order elliptic operator:

\[
LU(x) = -\sum_{i,j=1}^{2} \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial U(x)}{\partial x_j} \right) + \sum_{i=1}^{2} b_i(x) \frac{\partial U(x)}{\partial x_i} + c(x)U(x).
\]

We assume that all the coefficients are sufficiently smooth and the matrix \(\{a_{ij}(x)\}\) is symmetric and uniformly positive definite for any \(x \in \Omega\). We also assume that (12) is uniquely solvable.

If a finite element or finite difference method is used to solve equation (12), the discretized system will take the form of (1), with \(A\) corresponding to the symmetric positive definite part (namely the second-order terms) of the operator \(L\), and \(N\) to the remaining (lower-order terms) parts. It is routine to show that the assumptions described in the preceding section are all satisfied (cf. [2, 4] for details). In particular, when using a finite element method, the subspace \(V_0\) corresponds to the space defined on a coarser triangulation of \(\Omega\) and the assumption that \(\delta_0\) is sufficiently small is equivalent to the assumption that the coarse triangulation is sufficiently fine. Examples of \(B\) can be obtained by using domain decomposition methods, multigrid methods, multilevel preconditioners such as hierarchical basis, and multilevel nodal basis preconditioners (cf. [11] for a discussion of these examples).

In the following, we present some numerical experiments. The domain \(\Omega\) is the unit square in \(\mathbb{R}^2\), and the subdivision is shown in Figure 1. We use piecewise linear finite element approximation in all the examples. The symmetric positive definite part will be preconditioned by the additive Schwarz method (including a coarse space solver) described in [5] with two grid sizes of overlap in both \(x\)- and \(y\)-directions. It is known, from [5], that the additive Schwarz preconditioner is optimal for symmetric positive definite operators, i.e., the constants \(\lambda_0\) and \(\lambda_1\) that appeared in Theorem 1 are independent of the mesh parameters used in the finite element subdivision. It follows from our theory that by adding a nonsymmetric (or indefinite) coarse-mesh operator to the properly scaled optimal symmetric positive definite preconditioner, we can construct
A PRECONDITIONED GMRES METHOD

Figure 1. A sample two-level triangulation of $\Omega$. The left figure shows the fine mesh and also the decomposition of the domain into nine nonoverlapping subregions, and each is extended to a larger subregion as the shadowed region. The right figure shows the coarse mesh.

an optimal preconditioner for the original nonsymmetric or indefinite operator. In our experiments, a full GMRES method without restarting is used with an initial guess of zero. The stopping criterion is set to be $\|r_i\|_A/\|r_0\|_A < 10^{-5}$, where $r_i$ is the $i$th step residual.

Our first example is a convection-diffusion equation,

$$
\begin{cases}
-\Delta U + \sigma \cdot \nabla U = F & \text{in } \Omega, \\
U = 0 & \text{on } \partial \Omega.
\end{cases}
$$

(13)

$F$ is so chosen that the solution is $e^{xy} \sin(\pi x) \sin(\pi y)$. We use a $60 \times 60$ grid, i.e., the fine-mesh size $h = 1/60$, and the results are summarized in Table 1.

We first set $\sigma^T = (10, 10)$. The iteration counts are shown in Table 1 with two different coarse-mesh sizes $H$, but the same fine-mesh size $h$. The parameter $1/\beta$ (cf. (6)) ranges from 1.0 to 10.0.

Note that a smaller $H$ means a smaller $\delta_0$. We observe that the algorithm speeds up if the size of $H$ is decreased, which agrees with our theoretical estimates. An important fact we can see is that the algorithm is robust for the choice of $\beta$, and a balance between the symmetric positive definite preconditioner and the nonsymmetric coarse-mesh preconditioner can be achieved easily, at least for this example. This is perhaps due to the fact that the convection terms are relatively small compared with the diffusion terms.

<table>
<thead>
<tr>
<th>$1/\beta$</th>
<th>1.0</th>
<th>2.0</th>
<th>3.0</th>
<th>4.0</th>
<th>5.0</th>
<th>6.0</th>
<th>7.0</th>
<th>8.0</th>
<th>9.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H = 1/5$, $h = 1/60$</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>16</td>
<td>16</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>$H = 1/10$, $h = 1/60$</td>
<td>13</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>12</td>
<td>13</td>
<td>13</td>
</tr>
</tbody>
</table>

License or copyright restrictions may apply to redistribution; see https://www.ams.org/journal-terms-of-use
In the next set of experiments, we set \( \sigma^T = (50, 50) \), and the results are shown in Table 2.

Here the appropriate \( \beta \)-values are obtained by testing a series of \( 1/\beta \) ranging from 0.1 to 20.0. It can be seen from the above results that if the coarse-mesh size \( H \) is small enough, then a well-chosen \( \beta \) can keep the iteration count small. The interesting thing is that the algorithm behaves almost the same for those \( \beta \) in the intervals shown in Table 2. This means that the best possible constant \( \beta \) is not necessarily unique, but contained in a small interval. This phenomenon has not been well understood theoretically. Comparing this with the first example, we can see that as the convection terms are getting larger, the balancing parameter \( \beta \) becomes more and more important in order to keep a low iteration count of the GMRES. Nevertheless, our numerical results have also shown that the algorithm is always convergent for \( \beta = 1 \), although a few more iterations are needed sometimes.

As the last example, we consider

\[
\begin{align*}
-\Delta U + \sigma \cdot \nabla U + \delta U &= F \quad \text{in } \Omega, \\
U &= 0 \quad \text{on } \partial \Omega.
\end{align*}
\]

The numerical results shown in Table 3 correspond to \( \beta = 1 \), \( h = 1/128 \), the size of overlap = \( H \times 25\% \), and \( \sigma^T = (10, 10) \).

A detailed comparison of this algorithm with other domain decomposition algorithms and the ILU preconditioners is given in [3].

### 4. Concluding remarks

The preconditioning technique proposed in this paper appears quite promising for solving nonsymmetric or indefinite elliptic partial differential equations. If the coefficients of the nonsymmetric part of the partial differential operator are relatively not too big, we can always obtain a preconditioner by using any good preconditioner for the corresponding symmetric positive definite part together with a coarse-mesh solver. Our approach allows one to adapt any existing codes developed for solving symmetric positive definite problems in constructing a preconditioner for nonsymmetric positive definite problems. Nevertheless, the robustness of the algorithm depends on a proper choice of \( \beta \). We shall further investigate how \( \beta \) can be selected for specific applications in future work.
Finally, we would like to remark that a multiplicative version of the preconditioner proposed in this paper can give rise to a new class of iterative algorithms. For details, we refer to Xu [12].

ACKNOWLEDGMENT

The first author would like to express his thanks to Professor James Bramble for his suggestions for simplifying the proof of the main theorem of the paper.

REFERENCES


