PCG methods applied to a system of nonlinear equations

Xiaojun Chen
Department of Mathematics, Xi'an Jiaotong University, Xi'an, China

Tetsuro Yamamoto
Department of Mathematics, Faculty of Science, Ehime University, Matsuyama 790, Japan

Received 17 August 1990
Revised 12 June 1991

Abstract
In this paper, we consider a quasi-Newton iteration for solving a nonlinear equation
\[ F(x) = Ax + g(x) = 0 \]
in \( \mathbb{R}^n \), where \( A \) is a symmetric positive definite matrix and \( g \) is a bounded continuous function. We discuss the PCG method with various preconditioners to solve the linear equation at each step of the iteration, estimate their condition numbers, and compare their computing time for a numerical example.

Keywords: Newton-like method, PCG method, nonlinear equations.

1. Introduction

In recent papers [2,3,7], we have discussed convergence of the Newton-like method
\[ B(x_k)(x_{k+1} - x_k) = -F(x_k), \quad k \geq 0, \]
for solving the equation \( F(x) = f(x) + g(x) = 0 \) in a Banach space, where \( B(x) \) is a linear operator and \( f \) is differentiable, while the differentiability of \( g \) is not assumed.

In this paper, as a model problem, we restrict our attention to a system of finite-difference equations
\[ P(x) = Ax + g(x) = 0, \quad x \in \mathbb{R}^n, \]
in \( \mathbb{R}^n \), where \( A \) is an \( n \times n \) symmetric positive definite block tridiagonal M-matrix denoted by
\[
A = \begin{pmatrix}
T_1 & A_2 & & & \\
A_2 & T_2 & A_3 & & \\
& \ddots & \ddots & \ddots & \\
& & A_{m-1} & T_{m-1} & A_m \\
& & & A_m & T_m
\end{pmatrix}
\]
\[ = (a_{ij}), \]
where \( T_i, \ i = 1, \ldots, m, \) are \( m \times m \) tridiagonal symmetric matrices and \( A_j, \ j = 2, \ldots, m, \) are \( m \times m \) diagonal. Such an equation arises from the usual discretization of the nonlinear elliptic equation

\[
- \frac{\partial}{\partial x} \left( p(x, y) \frac{\partial u}{\partial x} \right) - \frac{\partial}{\partial y} \left( q(x, y) \frac{\partial u}{\partial y} \right) = \psi \left( x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \right),
\]

in \( \Omega = (0, 1) \times (0, 1) \subset \mathbb{R}^2, \)

subject to the boundary condition

\[
u(x, y) = \psi(x, y), \quad \text{on} \quad \partial \Omega,
\]

where \( p^* \geq p(x, y) \geq p_* > 0, \ q^* \geq q(x, y) \geq q_* > 0, \ x, y \in \Omega, \) and \( \psi \) is a continuous function whose partial derivatives \( \psi_u, \psi_y, \psi_u, \) do not necessarily exist.

We use the Newton-like method (1.1) to solve (1.2). Updating matrices \( B(x_k) \) are chosen as \( B(x_k) = A + \phi(x_k) \) where \( \phi(x_k) \) are defined as follows. For \( k \geq 0, \) let \( (x_k)_i \) be the \( i \)th component of the vector \( x_k \) and \( |x_k| \) be the vector with the components \( |(x_k)_1|, \ldots, |(x_k)_n| \).

Let \( k \geq 1 \) and \( \|x_k - x_{k-1}\|_\infty = |x_k - x_{k-1}|_j \) (the \( j \)th component of the vector \( |x_k - x_{k-1}| \)).

The notations \( a^+ \) and \( a^- \) are defined by

\[
a^+ = \begin{cases} \frac{1}{a}, & a \neq 0, \\ 0, & a = 0, \end{cases} \quad a^- = \begin{cases} 0, & a \neq 0, \\ 1, & a = 0. \end{cases}
\]

Then we put \( \phi(x_0) = 0 \) and for \( k \geq 1, \)

\[
\phi_1(x_k) = \text{diag}((x_k - x_{k-1})_i^+),
\]

\[
\phi_2(x_k) = (x_k - x_{k-1})_j^1 \sum_{i=1}^n (e_i e'_i + e_i e'_j)(x_k - x_{k-1})_i^{-}
\]

and

\[
\phi(x_k) = (\phi_1(x_k) + \phi_2(x_k)) \text{ diag}((g(x_k) - g(x_{k-1})))_i,
\]

where \( e_i \) stands for the \( i \)th column of the \( n \times n \) identity. Then \( B(x_k) = A + \phi(x_k) \) satisfy the quasi-Newton equations

\[
B(x_k)(x_k - x_{k-1}) = F(x_k) - F(x_{k-1}), \quad k \geq 1,
\]

so that \( \{x_k\} \) converges to a solution of (1.2), if \( g(x) \) satisfies a Lipschitz condition (see [2]).

Here, we are interested in the preconditioned conjugate gradient (PCG) method for solving the linear system

\[
B(x_k)y = (A + \phi(x_k))y = -F(x_k), \quad k = 0, 1, 2, \ldots,
\]

at each step of the quasi-Newton iteration. We shall choose a preconditioner \( M \) based on the structure of \( A \) and fix it for all \( k \geq 0. \) Let \( D = \text{diag}(a_{ii}), \ T = \text{diag}(T_i) \) (block diagonal) and \( L \) and \( L_c \) be lower triangular matrices such that

\[
L + L^T = A - D \quad \text{and} \quad L_c + L^T_c = A - T.
\]

Then the following matrices \( M \) are considered:

(1) \( M = D, \) \quad Jacobi, \quad (1.4)

(2) \( M = T, \) \quad Block Jacobi, \quad (1.5)
We first estimate the spectral condition number \( \kappa(M^{-1}B(x_k)) = \lambda_n / \lambda_1 \) (\( \geq 1 \)) with different \( M \), where \( \lambda_1 \) and \( \lambda_n \) are the smallest and largest eigenvalues of \( M^{-1}B(x_k) \), respectively, under the condition that \( B(x_k) \) are positive definite. As is well known, the PCG method converges rapidly if \( \lambda_n / \lambda_1 \) is small. However, the total computing time throughout the Newton-like iteration may increase, since solving linear equations with coefficient matrix \( M \) may be necessary, which needs considerable amount of work if \( n \) is large. Hence, the total number of operations will be counted, and we shall show that efficiency of PCG methods applied to nonlinear equations depends not only on the preconditioning matrix \( M \) but also on the dimension \( n \) and a stopping constant \( \epsilon \). Finally, in Section 4, the results are illustrated with a numerical example.

2. Construction of preconditioners

For the sake of simplicity, we denote \( \phi(x_k) \), \( B(x_k) \) and \( -F(x_k) \) by \( \phi \), \( B \) and \( b \), respectively, and consider the PCG methods with the preconditioners \( M \) applied to the linear system \( By = b \), which are defined as follows [1]. Choose \( y_0 = x_k \), calculate \( r_0 = By_0 - b \) and \( q_0 = M^{-1}r_0 \) and put \( p_0 = -q_0 \). For \( l \geq 0 \):

\[
\alpha_l = \frac{(r_l, q_l)}{(p_l, Bp_l)}, \quad y_{l+1} = y_l + \alpha_l p_l, \quad r_{l+1} = r_l + \alpha_l Bp_l, \\
q_{l+1} = M^{-1}r_{l+1}, \quad \beta_l = \frac{(r_{l+1}, q_{l+1})}{(r_l, q_l)}, \quad p_{l+1} = -q_{l+1} + \beta_l p_l.
\]

The following iterative methods for solving linear equations \( Ax = b \) are well known:

1. Jacobi:
\[
y_{l+1} = (I - D^{-1}A)y_l + D^{-1}b,
\]

2. Block Jacobi:
\[
y_{l+1} = (I - T^{-1}A)y_l + T^{-1}b,
\]

3. SSOR:
\[
y_{l+1/2} = \omega D^{-1}\left\{ -Ly_{l+1/2} - L'y_l + b \right\} + (1 - \omega)y_l, \\
y_{l+1} = \omega D^{-1}\left\{ -Ly_{l+1/2} - L'y_{l+1} + b \right\} + (1 - \omega)y_{l+1/2},
\]

4. Block SSOR:
\[
y_{l+1/2} = \omega T^{-1}\left\{ -Lcy_{l+1/2} - L'cy_l + b \right\} + (1 - \omega)y_l, \\
y_{l+1} = \omega T^{-1}\left\{ -Lcy_{l+1/2} - L'cy_{l+1} + b \right\} + (1 - \omega)y_{l+1/2}.
\]
They can be rewritten in the form $M(y_i - y_{i+1}) = r_i$, where $r_i = Ay_i - b$ and $M$ is a symmetric positive definite matrix defined in (1.4)-(1.7).

We are now interested in constructing $H$, an incomplete block Cholesky factorization of $A$. Being motivated by the fact

$$A = (\Sigma + L_c)\Sigma^{-1}(\Sigma + L_c^t),$$

where $\Sigma$ is the symmetric block diagonal matrix with $m \times m$ blocks $\Sigma_i$ satisfying

$$\Sigma_1 = T_1, \quad \Sigma_i = T_i - A_i\Sigma_{i-1}\Sigma_{i-1}^t, \quad i = 2, \ldots, m,$$

we construct the matrix $H$ as follows. Put

$$\Delta_1 = T_1, \quad \Delta_i = T_i - A_i\Delta_{i-1}A_i, \quad i = 2, \ldots, m,$$

where $A_{i-1}$ is a tridiagonal matrix (denoted by $\text{trid}(\Delta_{i-1}^t)$) whose tridiagonal elements are those of $\Delta_{i-1}^t$.

Decompose the matrices $\Delta_i$ and $\Lambda_i$:

$$\Delta_i = P_iP_i^t, \quad \Lambda_i = Q_iQ_i^t, \quad i = 1, \ldots, m,$$

where $P_i$ and $Q_i$ are lower bidiagonal. Put $W_i = A_iQ_i^t$, $i = 2, \ldots, m$.

Decompose the matrices $W_i$ and $A_i$:

$$W_i = P_i, \quad A_i = Q_iQ_i^t,$$

where $P_i$ and $Q_i$ are lower bidiagonal. Put $K = A_iQ_i^t$, $i = 2, \ldots, m$, $U = U^tU$.

We can prove that all the $\Delta_i$ are positive definite M-matrices so that $P_i$ are nonsingular. Hence, $H = U^tU$ is a symmetric positive definite matrix. Similarly, let

$$Z = \begin{pmatrix} Q_1P_1^t \\ \vdots \\ Q_mP_m^t \end{pmatrix} = (z_{ij}).$$

Then $Z$ is a nonsingular tridiagonal matrix and $H$ can be written as $H = T + L_cZ + Z^tL_c$.

Here $\Delta_i$ are computed by the following method (see [8]). Let

$$\Delta_i = T_0 = \begin{pmatrix} b_1 & a_2 \\ a_2 & \ddots & \ddots \\ \vdots & \ddots & \ddots \\ a_{m-1} & a_m \\ a_m & b_m \end{pmatrix}, \quad a_2, \ldots, a_m \neq 0.$$

Define two sequences $\{u_i\}, \{v_i\}$ as follows:

$$u_0 = 0, \quad u_1 = h_1, \quad u_i = -\frac{1}{a_i}(a_{i-1}u_{i-2} + b_{i-1}u_{i-1}), \quad 2 \leq i \leq m, \quad (2.1)$$

$$v_{m+1} = 0, \quad v_m = h_2, \quad v_i = -\frac{1}{a_{i+1}}(b_{i+1}v_{i+1} + a_{i+2}v_{i+2}), \quad 0 \leq i \leq m - 1. \quad (2.2)$$
where \( h_1, h_2, a_1 \) and \( a_{m+1} \) may be chosen arbitrarily, but \( h_1, h_2 \) and \( a_1 \) may not be zero. Then \( \Lambda_i = \text{trid}(\Delta_i^{-1}) = \langle \tau_{ij} \rangle \) is given by

\[
\langle \tau_{ij} \rangle = \frac{-1}{a_1 h_1 v_0} \begin{pmatrix}
  u_1 v_1 & u_1 v_2 \\
  u_1 v_2 & \cdots & \cdots \\
  \cdots & \cdots & \cdots \\
  u_{m-1} v_{m} & u_m v_m
\end{pmatrix}.
\]

Let

\[
\alpha = \max_{1 \leq i \leq m} \frac{|b_{i-1}|}{|a_i|}, \quad \beta = \max_{1 \leq i \leq m} \frac{|a_{i-1}|}{|a_i|}, \quad \bar{\alpha} = \max_{1 \leq i \leq m-1} \frac{|b_i|}{|a_i|}, \quad \bar{\beta} = \max_{1 \leq i \leq m-1} \frac{|a_{i+1}|}{|a_i|}.
\]

Then we have the following theorem which improves the estimates for bounds of \( |u_i| \) and \( |v_i| \) in [4].

**Theorem 1.** Let \( T_0 \) be diagonally dominant and \( |b_1| > |a_2|, \ |b_m| > |a_m| \). Then (i) \( T_0^{-1} \) exists and the sequences \( \{ u_i \} \) and \( \{ v_i \} \) satisfy

\[
|u_1| < |u_2| < \cdots < |u_m|, \quad |v_0| > |v_1| > \cdots > |v_m|;
\]

(ii) there exist positive constants \( s, \sigma, \tilde{s}, \tilde{\sigma} \) for which

\[
|u_i| \leq s \bar{t}_1^{i-1} + \sigma \bar{t}_2^{i-1}, \quad i = 1, 2, \ldots, m,
\]

\[
|v_i| \leq \tilde{s} \tilde{t}_1^{m-i} + \tilde{\sigma} \tilde{t}_2^{m-i}, \quad i = 0, 1, 2, \ldots, m,
\]

where \( \bar{t}_1 \) and \( \tilde{t}_2 \) are the roots of \( t^2 - \sigma t - \beta = 0 \), and \( \tilde{t}_1 \) and \( \tilde{t}_2 \) are the roots of \( t^2 - \tilde{\sigma} t - \tilde{\beta} = 0 \), which satisfy

\[-1 < \bar{t}_2 < 0 < 1 < \tilde{t}_1, \quad -1 < \tilde{t}_2 < 0 < 1 < \tilde{t}_1;\]

(iii) the inequalities (2.3) and (2.4) hold with equality if \( b = b_1 = \cdots = b_m \) and \( a = a_2 = \cdots = a_m \). Furthermore, \( |u_i| = |v_{m-i+1}| \) if \( |h_1| = |h_2| \).

**Proof.** From (2.1), we have

\[
|u_2| = \frac{|b_1|}{|a_2|} |u_1| > |u_1|,
\]

and for \( i \geq 3 \),

\[
|u_i| \geq \frac{1}{|a_i|} (|b_{i-1}| |u_{i-1}| - |a_{i-1}| |u_{i-2}|) > \frac{1}{|a_i|} (|b_{i-1}| - |a_{i-1}|) |u_{i-1}| \geq |u_{i-1}|,
\]

if \( |u_{i-1}| > |u_{i-2}| \). Hence the sequence \( \{ u_i \} \) is strictly increasing. Now we prove that (2.3) is true. From (2.1) we have

\[
|u_i| \leq \frac{|b_{i-1}|}{|a_i|} |u_{i-1}| + \frac{|a_{i-1}|}{|a_i|} |u_{i-2}| \leq \alpha |u_{i-1}| + \beta |u_{i-2}|.
\]

The linear difference equation \( z_{i+1} = \alpha z_i + \beta z_{i-1} \) has the general solution \( z_i = st_1^i + \sigma t_2^i \), where \( s \) and \( \sigma \) are constants, and it is a simple matter to show that

\[-1 < t_2 < 0 < 1 < t_1.\]
For example, we have
\[ t_2 = \frac{1}{2}(\alpha - \sqrt{\alpha^2 + 4\beta}) < 0 \]
and
\[ |t_2| = \frac{1}{2}(\sqrt{\alpha^2 + 4\beta} - \alpha) \leq \frac{1}{2}\left(\sqrt{\alpha^2 + 4\alpha} - \alpha\right) \leq \frac{1}{2}\left(\sqrt{(\alpha + 2)^2} - \alpha\right) = 1, \text{ etc.} \]
The constants \( s \) and \( \sigma \) can uniquely be determined by the relations
\[ z_0 = s + \sigma = |u_1|, \quad z_1 = st_1 + \sigma t_2 = |u_2|. \]
Furthermore, it can easily be shown that \( s, \sigma > 0 \). We now obtain by induction that \( |u_i| \leq z_{i-1}, i = 1, 2, \ldots, m \). In fact, if this is true up to some \( i \geq 2 \), then
\[ |u_{i+1}| \leq \alpha|u_i| + \beta|u_{i-1}| \leq \alpha z_{i-1} + \beta z_{i-2} = z_i. \]
By the same way, we can obtain \( |v_0| > \cdots > |v_m| \) and (2.4). \( \Box \)

The following corollary justifies our procedure which approximates \( \Delta_j^{-1} \) by the tridiagonal matrix \( \Lambda_j \).

**Corollary 2.** Suppose that the conditions of Theorem 1 hold. Then we have
\[ |\tau_{ij}| \geq |\tau_{ij+1}|, \quad \text{for } i \leq j, \quad (2.5) \]
\[ |\tau_{ij}| \geq |\tau_{ij-1}|, \quad \text{for } j \leq i, \quad (2.6) \]
and
\[ |\tau_{ij}| \leq \frac{|b_i| |R|}{|a_i| |a_2|} R^{-\left(\frac{R}{r}\right)} \frac{1}{R^{i-j}}, \quad (2.7) \]
where
\[ r = \min_{2 \leq i \leq m-1} \left\{ \frac{|b_i| - |a_i+1|}{|a_i|}, \frac{|b_i| - |a_i|}{|a_i+1|} \right\} \geq 1 \]
and
\[ R = \max_{2 \leq i \leq m-1} \left\{ \frac{|b_i| + |a_i+1|}{|a_i|}, \frac{|b_i| + |a_i|}{|a_i+1|} \right\} \geq 1. \]

**Proof.** The inequalities (2.5) and (2.6) are direct consequences from the definition of \( (\tau_{ij}) \) and the assertion (i) of Theorem 1. To prove (2.7), take \( h_1 \) arbitrarily. Then, for \( u_1 = h_1 \), we have
\[ |u_2| = \left| \frac{1}{a_2} (a_1 u_0 + b_1 u_1) \right| = \left| \frac{b_1}{a_2} \right| |u_1| \neq |u_1|. \]
If \( |u_{i-1}| \geq |u_{i-2}|, i \geq 3 \), then
\[ |u_i| = \left| \frac{1}{a_1} (a_{i-1} u_{i-2} + b_{i-1} u_{i-1}) \right| \geq \left| \frac{1}{a_1} \right| |\frac{|b_{i-1}|}{|a_{i-1}|} |u_{i-2}| \geq r^2 \left| \frac{r-1}{2 b_1 a_2} \right| |u_1| \]

and

\[ |u_i| \leq \left( \frac{|b_{i-1}| + |a_{i-1}|}{|a_i|} \right) |u_{i-1}| \leq R |u_{i-1}| \leq R^{i-2} \frac{b_1}{a_2} |u_1|. \]

By the same way, we obtain

\[ |v_j| \geq r^{m-j-1} \frac{b_m}{a_m} |v_m| \]

and

\[ |v_j| \leq R^{m-j-1} \frac{b_m}{a_m} |v_m|, \quad i = 0, 1, \ldots, m. \]

Hence, if \( i \leq j \), then we have

\[
|\tau_{ij}| = \left| \frac{u_i v_j}{a_i h_1 v_0} \right| \leq \frac{1}{a_i h_1} \frac{a_m}{b_{m-1} b_m v_m} R^{m-j-1} \frac{b_m}{a_m} v_m R^{i-2} \frac{b_1}{a_2} u_1.
\]

\[ = \left| \frac{b_1}{a_1 a_2} \right| \left( \frac{R}{r} \right)^{m-1} R^{i-j-2}. \]

Furthermore, if \( j \leq i \), then

\[ |\tau_{ij}| = -\left| \frac{u_i v_i}{a_i h_1 v_0} \right| \leq \left| \frac{b_1}{a_1 a_2} \right| \left( \frac{R}{r} \right)^{m-1} R^{i-j-2}, \]

so that (2.7) holds. \( \square \)

3. Estimates of spectral condition number and number of operations

Let \( P \) be an \( n \times n \) matrix, \( \lambda_1(P) \) and \( \lambda_n(P) \) the smallest and largest eigenvalues of \( P \), respectively. We discretize the nonlinear equation in Section 1 by the usual finite-difference method with \( h = 1/(m + 1) \) and put \( n = m \times m. \)

In this section, we estimate the spectral condition number \( \kappa(M^{-1}B) = \lambda_n(M^{-1}B)/\lambda_1(M^{-1}B) \) with different preconditioners \( M \).

We first consider the two cases \( M = I \) and \( M = A. \)

**Theorem 3.** If there exists a positive constant \( \alpha \) such that \( \| \phi \|_\infty \leq \alpha h^2 < 4(p_* + q_*)\sin^2 \frac{1}{2} \pi h, \) then as \( h \to 0, \) we have

\[ \kappa(B) \geq \frac{4(p_* + q_*) \sin^2 \frac{1}{2} \pi(1 - h) - \alpha h^2}{4(p_* + q_*) \sin^2 \frac{1}{2} \pi h + \alpha h^2} \to \infty \quad (3.1) \]

and

\[ \kappa(A^{-1}B) \leq \frac{4(p_* + q_*) \sin^2 \frac{1}{2} \pi h + \alpha h^2}{4(p_* + q_*) \sin^2 \frac{1}{2} \pi h - \alpha h^2} \to \frac{(p_* + q_*) \pi^2 + \alpha}{(p_* + q_*) \pi^2 - \alpha}. \quad (3.2) \]
Proof. We first consider two splittings of $A$ such that

$$A = \Gamma_1 - \nabla_1 \quad \text{and} \quad A = \Gamma_2 + \nabla_2,$$

where

$$\Gamma_1 = \begin{pmatrix}
2(p^* + q^*)I_m - p^*C & -q^*I_m \\
-q^*I_m & 2(p^* + q^*)I_m - p^*C \\
-4I_m & 2(p^* + q^*)I_m - p^*C \\
\end{pmatrix},$$

$$\Gamma_2 = \begin{pmatrix}
2(p^* + q^*)I_m - p^*C & -q^*I_m \\
-q^*I_m & 2(p^* + q^*)I_m - p^*C \\
-4I_m & 2(p^* + q^*)I_m - p^*C \\
\end{pmatrix},$$

$I_m$ is the $m \times m$ identity and

$$C = \begin{pmatrix}
0 & 1 & 0 & \cdots & 0 \\
1 & 0 & 1 & \cdots & 0 \\
& \ddots & \ddots & \ddots & \ddots \\
& & 1 & 0 & 1 \\
& & & 1 & 0 \\
\end{pmatrix} \quad (m \times m).$$

We have

$$\lambda_1(\Gamma_1) = 2(p^* + q^*)(1 - \cos \pi h) = 4(p^* + q^*) \sin^2 \frac{\pi}{2} h,$$

$$\lambda_1(\Gamma_2) = 2(p^* + q^*)(1 - \cos \pi h) = 4(p^* + q^*) \sin^2 \frac{\pi}{2} h,$$

and

$$\lambda_\sigma(\Gamma_2) = 2(p^* + q^*)(1 - \cos m \pi h) = 4(p^* + q^*) \sin^2 \frac{\pi}{2} (1 - h)$$

(see [5]). It can easily be shown that the matrices $\nabla_1 = ((\nabla_1)_{ij})$ and $\nabla_2 = ((\nabla_2)_{ij})$ are symmetric positive definite, since

$$(\nabla_1)_{ii} = (\Gamma_1 - A)_{ii} = 2(p^* + q^*) - a_{ii} \geq \sum_{i \neq j} |(\nabla_1)_{ij}| > 0$$

and

$$(\nabla_2)_{ii} = (A - \Gamma_2)_{ii} = a_{ii} - 2(p^* + q^*) \geq \sum_{i \neq j} |(\nabla_2)_{ij}| > 0.$$
for any vector \( x \neq 0 \), we have that the eigenvalues of \( B \) and \( A^{-1}B \) are positive and

\[
\lambda_n(B) - \max_{x \neq 0} \left( \frac{Bx}{x} , \frac{x}{x} \right) \leq \max_{x \neq 0} \left( \frac{\Gamma_2 x}{x} , \frac{x}{x} \right) + \min_{x \neq 0} \left( \frac{\phi x}{x} , \frac{x}{x} \right) \geq \lambda_n(\Gamma_2) - \alpha h^2
\]

and

\[
\lambda_1(B) = \min_{x \neq 0} \left( \frac{Bx}{x} , \frac{x}{x} \right) \leq \min_{x \neq 0} \left( \frac{\Gamma_1 x}{x} , \frac{x}{x} \right) + \max_{x \neq 0} \left( \frac{\phi x}{x} , \frac{x}{x} \right) \leq \lambda_1(\Gamma_1) + \alpha h^2.
\]

This implies that

\[
\kappa(B) = \frac{\lambda_n(B)}{\lambda_1(B)} \geq \frac{\lambda_n(\Gamma_2) - \alpha h^2}{\lambda_1(\Gamma_1) + \alpha h^2} = \frac{4(\rho + q) \sin^2 \frac{\pi}{2}(1 - h) - \alpha h^2}{4(\rho + q) \sin^2 \frac{\pi}{2}h + \alpha h^2} \to \infty
\]

(as \( h \to 0 \)).

On the other hand, we consider \( A^{-1}B = I + A^{-1} \phi \). Since \( \|A^{-1}B\|_2 \leq 1 + \|A^{-1}\| \cdot \|\phi\|_2 \) and \( \lambda_1(\Gamma_2) \leq \lambda_1(\Gamma) \), we have

\[
\lambda_n(A^{-1}B) \leq 1 + \alpha h^2 \lambda_1^{-1}(\Gamma_2), \quad \lambda_1(A^{-1}B) \geq 1 - \alpha h^2 \lambda_1^{-1}(\Gamma_2)
\]

and

\[
\kappa(A^{-1}B) = \frac{\lambda_n(A^{-1}B)}{\lambda_1(A^{-1}B)} \leq \frac{1 + \alpha h^2 \lambda_1^{-1}(\Gamma_2)}{1 - \alpha h^2 \lambda_1^{-1}(\Gamma_2)} \to \infty
\]

Next, we consider the cases \( M = D, M = T, M = S, \) and \( M = C \). Let

\[
\delta_1 = \min_{x \neq 0} \left( \frac{(LD^{-1}L^1 + A)x}{x} , \frac{x}{x} \right), \quad \delta_2 = \min_{x \neq 0} \left( \frac{(L,T^{-1}L^1 + A)x}{x} , \frac{x}{x} \right),
\]

\[
\gamma_1 = \frac{1}{1 - \min(\delta_1, \frac{1}{2})} \quad \text{and} \quad \gamma_2 = \frac{1}{1 - \min(\delta_2, \frac{1}{2})}.
\]

Then we have the following corollary.

**Corollary 4.** Under the conditions of Theorem 3, as \( h \to 0 \), we have

(i) \( \kappa(D^{-1}B) \geq \frac{(\rho + q)}{(\rho^* + q^*)} \kappa(B) \to \infty \),

(ii) \( \kappa(T^{-1}B) \geq \frac{2q + 4p \sin^2 \frac{\pi}{2}h}{2q^* + 4p^* \sin^2 \frac{\pi}{2}(1 - h)} \kappa(B) \to \infty \),

(iii) \( \kappa(S^{-1}B) \geq \frac{F_1(\omega)(\rho + q)^2}{4(\rho^* + q^*)^2} \kappa(B) \to \infty \), if \( \omega < \gamma_1 \),

(iv) \( \kappa(C^{-1}B) \geq \frac{F_2(\omega)(2q + 4p \sin^2 \frac{\pi}{2}h)^2}{16(\rho^* + q^*)^2} \kappa(B) \to \infty \), if \( \omega < \gamma_2 \),
where

\[ F_1(\omega) = \begin{cases} \omega^2 \delta_1 + (1 - \omega), & 0 < \omega \leq 1, \\ \omega \delta_1 + (1 - \omega), & 1 \leq \omega < \gamma_1, \end{cases} \quad F_2(\omega) = \begin{cases} \omega^2 \delta_2 + (1 - \omega), & 0 < \omega \leq 1, \\ \omega \delta_2 + (1 - \omega), & 1 \leq \omega < \gamma_2. \end{cases} \]

Furthermore,

\[ \lambda_1(D^{-1}A) \leq \delta_1 \leq \frac{(p^* + q^*)^2}{(p^* + q^* + p_* + q_*)^2} + \lambda_1(D^{-1}A), \]

\[ \lambda_1(T^{-1}A) \leq \delta_2 \leq \frac{(q^*)^2}{(2q_* + 4p_\ast \sin^2 \frac{1}{2} \pi h)^2} + \lambda_1(T^{-1}A) \]

and

\[ \min_{0 < \omega \leq \gamma_1} F_1(\omega) = \begin{cases} F_1(\gamma_1), & \text{if } \delta_1 \leq \delta^*, \\ F_1\left(\frac{1}{2\delta_1}\right), & \text{if } \delta_1 \geq \delta^*, \end{cases} \quad (3.4) \]

\[ \min_{0 < \omega \leq \gamma_2} F_2(\omega) = \begin{cases} F_2(\gamma_2), & \text{if } \delta_2 \leq \delta^*, \\ F_2\left(\frac{1}{2\delta_2}\right), & \text{if } \delta_2 \geq \delta^*, \end{cases} \]

where \( \delta^* = \frac{1}{2} + 1/(2\sqrt{2}) \).

**Proof.** Since

\[ \kappa(M^{-1}B) = \frac{\max\{(Bx, x)/(Mx, x)\}}{\min\{(Bx, x)/(Mx, x)\}}, \]

we have

\[ \kappa(M^{-1}B) \geq \frac{\kappa(B)}{\kappa(M)}. \]

Hence, to prove (i)-(iv), it suffices to estimate the lower bounds of \( \lambda_1(D) \), \( \lambda_1(T) \), \( \lambda_1(S_\omega) \) and \( \lambda_1(C_\omega) \) and the upper bounds of \( \lambda_n(D) \), \( \lambda_n(T) \), \( \lambda_n(S_\omega) \) and \( \lambda_n(C_\omega) \). We obtain the following:

(i) \( \lambda_1(D) \geq 2(p_* + q_\ast) \) and \( \lambda_n(D) \leq 2(p^* + q^*) \).

(ii) \( \lambda_1(T) \geq 2q_* + 4p_\ast \sin^2 \frac{1}{2} \pi h \) and \( \lambda_n(T) \leq 2q^* + 4p^* \sin^2 \frac{1}{2} \pi (1 - h) \).

(iii) The matrix \( S_\omega \) can be expressed as

\[ S_\omega = \frac{\omega A + (1 - \omega) D + \omega^2 LD^{-1}L^t}{(2 - \omega) \omega}. \]

Hence

\[ (S_\omega x, x) \geq \begin{cases} \omega^2 \left(\frac{(LD^{-1}L^t + A)x, x}{(Dx, x)} + (1 - \omega)\right) \frac{(Dx, x)}{(2 - \omega) \omega}, & \text{for } \omega \leq 1, \\ \omega \left(\frac{(LD^{-1}L^t + A)x, x}{(Dx, x)} + (1 - \omega)\right) \frac{(Dx, x)}{(2 - \omega) \omega}, & \text{for } \omega \geq 1, \end{cases} \]
Furthermore, we have from (1.6),

\[ \lambda_n(S_\omega) \leq \|S_\omega\|_\infty \leq \frac{\|D + \omega L\|_\infty \|D + \omega L^T\|_\infty \|D^{-1}\|_\infty}{(2 - \omega)\omega} \leq 8 \frac{(p_* + q_*)^2}{(p_* + q_*)(2 - \omega)\omega}. \]

This proves (iii).

Now we shall prove (3.3). We first observe that

\[ \lambda_1(D^{-1}A) = \left( \frac{Ae_i, e_i}{De_i, e_i} \right) = 1. \]

If \( L^' x \neq 0 \), then

\[ \frac{(LD^{-1}L^T + A)x, x}{(Dx, x)} > \frac{(Ax, x)}{(Dx, x)} \geq \lambda_1(D^{-1}A). \]

If \( L^' x = 0 \), then

\[ \frac{(LD^{-1}L^T + A)x, x}{(Dx, x)} = \frac{(Dx, x)}{(Dx, x)} = 1. \]

Hence \( \delta_1 \geq \lambda_1(D^{-1}A) \). We have also

\[ \delta_1 \leq \max \left( \frac{LD^{-1}L^T x}{Dx, x} \right) + \min \left( \frac{Ax, x}{Dx, x} \right) = \lambda_n(D^{-1}LD^{-1}L^T) + \lambda_1(D^{-1}A) \]

\[ \leq \|D^{-1}L\|_\infty \|D^{-1}L^T\|_\infty + \lambda_1(D^{-1}A) \leq \frac{(p_* + q_*)^2}{(p_* + q_* + p_* + q_*)^2} + \lambda_1(D^{-1}A). \]

Next we shall prove (3.4). If \( \delta_1 \leq \frac{1}{2} \), then \( F_1(\omega) \) is strictly decreasing and \( \min_{\omega} F_1(\omega) = F_1(\gamma_1) = 0 \). If \( \delta_1 > \frac{1}{2} \), then there is a unique zero \( \omega \) of \( F_1(\omega) \), \( 0 < \omega < 1 \). Furthermore,

\[ F_1\left( \frac{1}{2\delta_1} \right) = \frac{4\delta_1 - 1}{4\delta_1} \leq F_1(2) = 2\delta_1 - 1, \quad \text{if} \ \delta_1 \geq \delta^*. \]

Part (iv) may be proved in the same way as in the proof of (iii). \( \square \)

**Remark 5.** Axelsson and Barker gave an upper bound for \( \kappa(S_\omega^{-1}A) \) in [1]. Their results are stated as follows. Let

\[ \mu = \max_{x \neq 0} \left( \frac{Dx, x}{Ax, x} \right), \quad \delta = \max_{x \neq 0} \left( \frac{(LD^{-1}L^T - \frac{1}{4}D)x, x}{(Ax, x)} \right), \]

and

\[ G(\omega) = \frac{1 + \left( \frac{(2 - \omega)^2}{4\omega} \right) \mu + \omega \delta}{2 - \omega}. \]

Then, \( \delta \gg \frac{1}{4}, \ \lambda_n(S_\omega^{-1}A) < 1, \ \lambda_1(S_\omega^{-1}A) > 1/G(\omega) \) and \( \kappa(S_\omega^{-1}A) \leq G(\omega) \). Furthermore,

\[ \min_{0 < \omega < \omega_2} G(\omega) = G(\omega^*) = \sqrt{\frac{1}{2} + \delta} \mu + \frac{1}{2} \leq \sqrt{\frac{1}{2} + \delta} \mu + \frac{1}{2} \leq \sqrt{\frac{1}{2} + \delta} \kappa(A) + \frac{1}{2}, \]

where \( \omega^* = 2\sqrt{\mu}/(\sqrt{\mu} + 2\sqrt{\frac{1}{2} + \delta}) \).
They further proved that $\delta$ is bounded ($\delta \leq 0$) if
\[
\left\| D^{-1/2}LD^{-1/2} \right\|_\infty \leq \frac{1}{2} \quad \text{and} \quad \left\| D^{-1/2}L'D^{-1/2} \right\|_\infty \leq \frac{1}{2}.
\]
By using their results, we obtain
\[
\kappa(S^{-1}_\omega B) \leq G(\omega) \frac{\lambda_1(S_\omega) + \alpha h^2}{\lambda_1(S_\omega) - \alpha h^2 G(\omega)},
\]
and $\lambda_\eta(S_\omega^{-1}) = 1/\lambda_1(S_\omega)$. Hence under the assumptions (3.5), $\kappa(S_\omega^{-1}B) = O(\sqrt[\omega]{\kappa(A)})$, and observing (3.1) and (3.2) we see that $\kappa(A)$ and $\kappa(B)$ have the same order, so that $\kappa(S_\omega^{-1}B)$ is $O(\sqrt[\omega]{\kappa(B)})$, i.e., $O(\sqrt{n})$.

The lower bound for $\kappa(S_\omega^{-1}B)$ in (iii) of Corollary 4, together with (3.1), implies that $\kappa(S_\omega^{-1}B)$ is at least $O(\kappa(B)) - O(n) - O(h^{-2})$, if $\omega < \gamma_1$. Furthermore we remark that $\gamma_1 < \omega^*$ if $\lambda_1(D^{-1}A) \leq \frac{1}{2}(2 - \sqrt{2})$. In fact, under the assumptions (3.5), we have $\delta_1 \leq \frac{1}{4} + \lambda_1(D^{-1}A)$ so that $\gamma_1 \leq 4/(3 - 4\lambda_1(D^{-1}A))$. On the other hand, observing that $\lambda_\eta(A^{-1}D)^{-1} = \lambda_1(D^{-1}A) \leq 1$, we have
\[
\omega^* > \frac{2\sqrt{\lambda_n(A^{-1}D)}}{\sqrt{\lambda_n(A^{-1}D)} + \sqrt{2}} \geq \frac{2\lambda_n(A^{-1}D)}{\lambda_n(A^{-1}D) + \sqrt{2}} \geq \frac{4}{3 - 4\lambda_1(D^{-1}A)} \geq \gamma_1.
\]

If the results are applied to the preconditioned Block SSOR, then corresponding estimates can be obtained by replacing $D$ and $L$ by $T$ and $L_c$, respectively. For example, we have $\kappa(C_\omega^{-1}A) \leq G(\omega)$, where $\mu$ and $\delta$ in $G(\omega)$ are replaced by
\[
\mu = \max_{x \neq 0} \frac{(Tx, x)}{(Ax, x)}, \quad \delta = \max_{x \neq 0} \frac{((L_cT^{-1}L_c^{-1} - 1/4 T)x, x)}{(Ax, x)}.
\]

**Remark 6.** Now we count the number of multiplication for solving the linear equations $My = b$ in the PCG method with different preconditioners. The results are as follows:

1. $M = D$: \( n, \quad k \geq 0, \quad l \geq 0; \)
2. $M = T$: \( 5n, \quad k = 0, \quad l = 0, \)
   \( 3n, \quad \text{otherwise}; \)
3. $M = S_\omega$: \( 7n, \quad k \geq 0, \quad l \geq 0; \)
4. $M = C_\omega$: \( 13n - 2m, \quad k = 0, \quad l = 0, \)
   \( 11n - 2m, \quad \text{otherwise}; \)
5. $M = I$: \( 0, \quad k \geq 0, \quad l \geq 0; \)
6. $M = A$: \( (2m + 1)n + \frac{1}{2}n(n - 1) + \frac{1}{2}m(7n + 5), \quad k = 0, \quad l = 0, \)
   \( (2m + 1)n, \quad \text{otherwise}; \)
7. $M = H$: \( 19n, \quad k = 0, \quad l = 0, \)
   \( 6n, \quad \text{otherwise}. \)
4. A numerical example

Example 7. Consider the Dirichlet problem

$$-\Delta u - |u| = -2(x(x - 1) + y(y - 1)) - |xy(x - 1)(y - 1) - 0.025|,$$

$$x, y \in (0, 1),$$

$$u(0, t) = u(t, 0) = u(1, t) = u(t, 1) = -0.025, \quad t \in [0, 1].$$

This problem has a solution $u(x, y) = xy(x - 1)(y - 1) - 0.025$.

We first discretize the problem by the standard five-point difference formula, and obtain a system of nonlinear algebraic equations. Next, we solve the system by the quasi-Newton iteration (1.1) and (1.3) combined with the PCG method, with preconditioners given in Section 2. We choose the initial values $(x_0)_i = 20(-1)^i$, $1 \leq i \leq n$, and employ the stopping criteria $\|r_i\|_2 \leq 10^{-6}$, $\|F(x_{k+1})\|_\infty / \|F(x_0)\|_\infty \leq 10^{-5}$. Total computing times are shown in Table 1, together with the number of iterations in Table 2, where $h$: square mesh size ($h = 1/(m + 1)$); $n$: interior

<table>
<thead>
<tr>
<th>$n$</th>
<th>$D$</th>
<th>$T$</th>
<th>$S_1$</th>
<th>$S_2^*$</th>
<th>$C_\omega^*$</th>
<th>$I$</th>
<th>$A$</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.17</td>
<td>0.20</td>
<td>0.22</td>
<td>0.23</td>
<td>0.23</td>
<td>0.13</td>
<td>0.18</td>
<td>0.23</td>
</tr>
<tr>
<td>49</td>
<td>1.43</td>
<td>1.80</td>
<td>1.65</td>
<td>1.53</td>
<td>1.70</td>
<td>1.37</td>
<td>1.33</td>
<td>1.70</td>
</tr>
<tr>
<td>225</td>
<td>10.38</td>
<td>11.23</td>
<td>9.05</td>
<td>7.53</td>
<td>7.88</td>
<td>8.93</td>
<td>11.62</td>
<td>8.20</td>
</tr>
<tr>
<td>961</td>
<td>88.25</td>
<td>89.48</td>
<td>64.75</td>
<td>44.23</td>
<td>47.22</td>
<td>76.53</td>
<td>149.85</td>
<td>51.43</td>
</tr>
<tr>
<td>3969</td>
<td>806.75</td>
<td>683.83</td>
<td>463.37</td>
<td>248.12</td>
<td>261.95</td>
<td>672.02</td>
<td>2226.67</td>
<td>347.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$n$</th>
<th>$D$</th>
<th>$T$</th>
<th>$S_1$</th>
<th>$S_2^*$</th>
<th>$C_\omega^*$</th>
<th>$I$</th>
<th>$A$</th>
<th>$H$</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>$\epsilon$</th>
<th>$D$</th>
<th>$T$</th>
<th>$S_1$</th>
<th>$S_2^*$</th>
<th>$C_\omega^*$</th>
<th>$I$</th>
<th>$A$</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0 * 10^{-7}</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>13.65</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>1.0 * 10^{-6}</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>13.65</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2.5 * 10^{-6}</td>
<td>*</td>
<td>14.30</td>
<td>11.32</td>
<td>*</td>
<td>*</td>
<td>11.80</td>
<td>13.65</td>
<td>10.08</td>
</tr>
<tr>
<td>5.0 * 10^{-6}</td>
<td>10.40</td>
<td>11.25</td>
<td>9.05</td>
<td>7.52</td>
<td>9.73</td>
<td>8.87</td>
<td>11.63</td>
<td>8.23</td>
</tr>
<tr>
<td>7.5 * 10^{-6}</td>
<td>10.40</td>
<td>11.30</td>
<td>9.05</td>
<td>7.53</td>
<td>7.87</td>
<td>8.87</td>
<td>11.62</td>
<td>8.28</td>
</tr>
</tbody>
</table>

* Iteration diverged. $\omega^*$ are chosen based on Remark 5, where $\delta = 0.$
Table 4

Number of iterations \((k[l_l, l_2, \ldots, l_k])\)

<table>
<thead>
<tr>
<th>(\varepsilon)</th>
<th>(D)</th>
<th>(T)</th>
<th>(S_1)</th>
<th>(S_{a^*})</th>
<th>(C_{a^*})</th>
<th>(I)</th>
<th>(A)</th>
<th>(H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.0 \times 10^{-7}</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>4[1,2,2,2]</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>1.0 \times 10^{-6}</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>4[1,2,2,2]</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>2.5 \times 10^{-6}</td>
<td>*</td>
<td>4[24,17,15,19]</td>
<td>4[15,10,8,8]</td>
<td>*</td>
<td>*</td>
<td>4[18,19,19,19]</td>
<td>4[12,2,2,2]</td>
<td>4[14,9,7,7]</td>
</tr>
<tr>
<td>7.5 \times 10^{-6}</td>
<td>3[25,18,18]</td>
<td>3[24,17,15]</td>
<td>3[15,10,8]</td>
<td>3[12,8,7]</td>
<td>3[11,6,6]</td>
<td>3[18,19,19]</td>
<td>3[1,2,2]</td>
<td>3[14,9,7]</td>
</tr>
</tbody>
</table>

Table 5

Upper and lower bounds for \(\kappa(A^{-1}B)\) and \(\kappa(B)\)

<table>
<thead>
<tr>
<th></th>
<th>9</th>
<th>49</th>
<th>225</th>
<th>961</th>
<th>3969</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\kappa(A^{-1}B))</td>
<td>1.1127</td>
<td>1.1082</td>
<td>1.1077</td>
<td>1.1068</td>
<td>1.1068</td>
</tr>
<tr>
<td>(\kappa(B))</td>
<td>5.4826</td>
<td>23.9917</td>
<td>98.0526</td>
<td>394.3027</td>
<td>1579.305</td>
</tr>
</tbody>
</table>

mesh number \((n = m \times m\) and \(h = 1/(\sqrt{n} + 1))\); \(k\): number of the iterations for the quasi-Newton method; \(l_i\): iterative number of the PCG method at the \(i\)th iteration.

Now, we change the value \(\varepsilon\) for the stopping criterion \(\|F(x_k)\|_{\infty}/\|F(x_0)\|_{\infty} \leq \varepsilon\) to solve equation (1.2) in \(\mathbb{R}^{225}\). Total computing times are shown in Table 3, together with the number of iterations in Table 4.

According to Theorem 3, we give in Table 5 upper and lower bounds for \(\kappa(A^{-1}B)\) and \(\kappa(B)\), respectively.

Remark 8. From Table 2, Theorem 3 and Corollary 4, we see that convergence speed of the PCG method with preconditioner \(M = A\) or \(M = C_{a^*}\) is faster than the others and we roughly conclude that

\[
\kappa(B) \geq \kappa(D^{-1}B) \geq \kappa(T^{-1}B) \geq \kappa(S_1^{-1}B) \geq \kappa(H^{-1}B) \geq \kappa(S_{a^*}^{-1}B) \geq \kappa(C_{a^*}^{-1}B)
\]

\[
\geq \kappa(A^{-1}B).
\]

However, from Remark 6 and Table 1, we observe that if the stopping constant \(\varepsilon\) is not so small, then

\[
T(A^{-1}B) \geq T(D^{-1}B) \geq T(T^{-1}B) \geq T(B) \geq T(S_1^{-1}B) \geq T(H^{-1}B) \geq T(C_{a^*}^{-1}B)
\]

\[
\geq T(S_{a^*}^{-1}B),
\]

for larger \(n\), where \(T(P^{-1}B)\) stands for the computing time for solving (1.2) by the iteration (1.1) with the preconditioner \(P\).

Computations were carried out on the Apollo DOMAIN 3000 (single precision) at the Department of Mathematics, Ehime University.
References


