Journal of Computational and Applied Mathematics 38 (1991) 61–75
North-Holland

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

Chen, X. and T. Yamamoto, PCG methods applied to a system of nonlinear equations, Journal of Computational and Applied Mathematics 38 (1991) 61–75.

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

$$B(x_k)(x_{k+1} - x_k) = -F(x_k), \quad k \geqslant 0, \tag{1.1}$$

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

$$F(x) = Ax + g(x) = 0, \quad x \in \mathbb{R}^n, \tag{1.2}$$

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, $i=1,\dots,m$, are $m\times m$ tridiagonal symmetric matrices and A_j, $j=2,\dots,m$, are $m\times m$ diagonal. Such an equation arises from the usual discretization of the nonlinear elliptic equation

$$-\frac{\partial u}{\partial x}\left(p(x, y)\frac{\partial u}{\partial x}\right)-\frac{\partial u}{\partial y}\left(q(x, y)\frac{\partial u}{\partial y}\right)=\psi\left(x, y, u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}\right),$$
$$\text{in } \Omega=(0,1)\times(0,1)\subset\mathbb{R}^2,$$

subject to the boundary condition

$$u(x, y)=\mu(x, y), \quad \text{on } \partial\Omega,$$

where $p^*\geqslant p(x, y)\geqslant p_*>0$, $q^*\geqslant q(x, y)\geqslant q_*>0$, $x, y\in\Omega$, and ψ is a continuous function whose partial derivatives ψ_u, ψ_{u_x}, ψ_{u_y} 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 $k\geqslant 0$, let $(x_k)_i$ be the ith component of the vector x_k and $|x_k|$ be the vector with the components $|(x_k)_1|,\dots,|(x_k)_n|$. Let $k\geqslant 1$ and $\|x_k-x_{k-1}\|_\infty=|x_k-x_{k-1}|_j$ (the jth component of the vector $|x_k-x_{k-1}|$). The notations a^+ and a^- are defined by

$$a^+=\begin{cases}\dfrac{1}{a}, & a\neq 0,\\ 0, & a=0,\end{cases} \quad\text{and}\quad a^-=\begin{cases}0, & a\neq 0,\\ 1, & a=0.\end{cases}$$

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

$$\phi_1(x_k)=\operatorname{diag}\left((x_k-x_{k-1})_i^+\right),$$
$$\phi_2(x_k)=(x_k-x_{k-1})_j^{-1}\sum_{i=1}^{n}\left(e_je_i^t+e_ie_j^t\right)(x_k-x_{k-1})_i^-$$

and

$$\phi(x_k)=(\phi_1(x_k)+\phi_2(x_k))\operatorname{diag}\left((g(x_k)-g(x_{k-1}))_i\right),$$

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

$$B(x_k)(x_k-x_{k-1})=F(x_k)-F(x_{k-1}), \quad k\geqslant 1, \tag{1.3}$$

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

$$B(x_k)y=(A+\phi(x_k))y=-F(x_k), \quad k=0, 1, 2,\dots,$$

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 $k\geqslant 0$. Let $D=\operatorname{diag}(a_{ii})$, $T=\operatorname{diag}(T_i)$ (block diagonal) and L and L_c be lower triangular matrices such that

$$L+L^t=A-D \quad\text{and}\quad L_c+L_c^t=A-T.$$

Then the following matrices M are considered:

(1) $M=D$, Jacobi, (1.4)
(2) $M=T$, Block Jacobi, (1.5)

(3) $M = S_\omega = (D + \omega L)D^{-1}(D + \omega L^t)/((2-\omega)\omega)$, SSOR, (1.6)
(4) $M = C_\omega = (T + \omega L_c)T^{-1}(T + \omega L_c^t)/((2-\omega)\omega)$, Block SSOR, (1.7)
(5) $M = I$,
(6) $M = A$,
(7) $M = H$, an incomplete block Cholesky factorization of A. (1.8)

We first estimate the spectral condition number $\kappa(M^{-1}B(x_k)) = \lambda_n/\lambda_1$ $(\geqslant 1)$ with different M, where λ_1 and λ_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 λ_n/λ_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 ϵ. 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 ϕ, 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 $l \geqslant 0$:

$$\alpha_l = \frac{(r_l, q_l)}{(p_l, Bp_l)}, \qquad y_{l+1} = y_l + \alpha_l p_l, \qquad r_{l+1} = r_l + \alpha_l B p_l,$$

$$q_{l+1} = M^{-1} r_{l+1}, \qquad \beta_l = \frac{(r_{l+1}, q_{l+1})}{(r_l, q_l)}, \qquad p_{l+1} = -q_{l+1} + \beta_l p_l.$$

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:

$$y_{l+1/2} = \omega D^{-1}\{-Ly_{l+1/2} - L^t y_l + b\} + (1-\omega)y_l,$$

$$y_{l+1} = \omega D^{-1}\{-Ly_{l+1/2} - L^t y_{l+1} + b\} + (1-\omega)y_{l+1/2},$$

(4) Block SSOR:

$$y_{l+1/2} = \omega T^{-1}\{-L_c y_{l+1/2} - L_c^t y_l + b\} + (1-\omega)y_l,$$

$$y_{l+1} = \omega T^{-1}\{-L_c y_{l+1/2} - L_c^t y_{l+1} + b\} + (1-\omega)y_{l+1/2}.$$

They can be rewritten in the form $M(y_l - y_{l+1}) = \tilde{r}_l$, where $\tilde{r}_l = Ay_l - b$ 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 Σ is the symmetric block diagonal matrix with $m \times m$ blocks Σ_i satisfying

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

we construct the matrix H as follows. Put

$$\Delta_1 = T_1, \qquad \Delta_i = T_i - A_i \Lambda_{i-1} A_i, \quad i = 2, \dots, m,$$

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

Decompose the matrices Δ_i and Λ_i:

$$\Delta_i = P_i P_i^t, \quad \Lambda_i = Q_i Q_i^t, \qquad i = 1, \dots, m,$$

where P_i and Q_i are lower bidiagonal. Put $W_i = A_i Q_{i-1}$, $i = 2, \dots, m$,

$$U^t = \begin{pmatrix} P_1 & & & \\ W_2 & P_2 & & \\ & \ddots & \ddots & \\ & & W_m & P_m \end{pmatrix}$$

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

$$Z = \begin{pmatrix} Q_1 P_1^t & & \\ & \ddots & \\ & & Q_m P_m^t \end{pmatrix} = (z_{ij}).$$

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

Here Λ_i are computed by the following method (see [8]). Let

$$\Delta_i = T_0 = \begin{pmatrix} b_1 & a_2 & & \\ a_2 & \ddots & \ddots & \\ & \ddots & \ddots & a_m \\ & & a_m & b_m \end{pmatrix}, \quad a_2, \dots, a_m \neq 0.$$

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

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

$$v_{m+1} = 0, \qquad v_m = h_2, \qquad v_i = -\frac{1}{a_{i+1}}(b_{i+1}v_{i+1} + a_{i+2}v_{i+2}), \qquad 0 \leqslant i \leqslant m-1, \tag{2.2}$$

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 $\Lambda_i = \text{trid}(\Delta_i^{-1}) = (\tau_{ij})$ is given by

$$(\tau_{ij}) = \frac{-1}{a_1 h_1 v_0} \begin{pmatrix} u_1 v_1 & u_1 v_2 & & & \\ u_1 v_2 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & u_{m-1} v_m \\ & & & u_{m-1} v_m & u_m v_m \end{pmatrix}.$$

Let

$$\alpha = \max_{2 \leqslant i \leqslant m} \frac{|b_{i-1}|}{|a_i|}, \quad \beta = \max_{3 \leqslant i \leqslant m} \frac{|a_{i-1}|}{|a_i|}, \quad \tilde{\alpha} = \max_{1 \leqslant i \leqslant m} \frac{|b_i|}{|a_i|}, \quad \tilde{\beta} = \max_{1 \leqslant i \leqslant m-1} \frac{|a_{i+1}|}{|a_i|}.$$

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*

$$|u_1| < |u_2| < \cdots < |u_m|, \qquad |v_0| > |v_1| > \cdots > |v_m|;$$

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

$$|u_i| \leqslant s t_1^{i-1} + \sigma t_2^{i-1}, \quad i = 1, 2, \ldots, m, \tag{2.3}$$

$$|v_i| \leqslant \tilde{s} \tilde{t}_1^{m-i} + \tilde{\sigma} \tilde{t}_2^{m-i}, \quad i = 0, 1, 2, \ldots, m, \tag{2.4}$$

where t_1 and t_2 are the roots of $t^2 - \alpha t - \beta = 0$, and $\tilde{t}_1$ and $\tilde{t}_2$ are the roots of $t^2 - \tilde{\alpha} t - \tilde{\beta} = 0$, which satisfy

$$-1 < t_2 < 0 < 1 < t_1, \qquad -1 < \tilde{t}_2 < 0 < 1 < \tilde{t}_1;$$

(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 $i \geqslant 3$,

$$|u_i| \geqslant \frac{1}{|a_i|}(|b_{i-1}||u_{i-1}| - |a_{i-1}||u_{i-2}|) > \frac{1}{|a_i|}(|b_{i-1}| - |a_{i-1}|)|u_{i-1}| \geqslant |u_{i-1}|,$$

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

$$|u_i| \leqslant \frac{|b_{i-1}|}{|a_i|} |u_{i-1}| + \frac{|a_{i-1}|}{|a_i|} |u_{i-2}| \leqslant \alpha |u_{i-1}| + \beta |u_{i-2}|.$$

The linear difference equation $z_{i+1} = \alpha z_i + \beta z_{i-1}$ has the general solution $z_i = s t_1^i + \sigma t_2^i$, where s and σ are constants, and it is a simple matter to show that

$$-1 < t_2 < 0 < 1 < t_1.$$

For example, we have

$$t_2 = \tfrac{1}{2}\left(\alpha - \sqrt{\alpha^2 + 4\beta}\right) < 0$$

and

$$|t_2| = \tfrac{1}{2}\left(\sqrt{\alpha^2 + 4\beta} - \alpha\right) \leqslant \tfrac{1}{2}\left(\sqrt{\alpha^2 + 4\alpha} - \alpha\right) \leqslant \tfrac{1}{2}\left(\sqrt{(\alpha+2)^2} - \alpha\right) = 1, \quad \text{etc.}$$

The constants s and σ can uniquely be determined by the relations

$$z_0 = s + \sigma = |u_1|, \qquad z_1 = st_1 + \sigma t_2 = |u_2|.$$

Furthermore, it can easily be shown that s, $\sigma > 0$. We now obtain by induction that $|u_i| \leqslant z_{i-1}$, $i = 1, 2, \ldots, m$. In fact, if this is true up to some $i \geqslant 2$, then

$$|u_{i+1}| \leqslant \alpha |u_i| + \beta |u_{i-1}| \leqslant \alpha z_{i-1} + \beta z_{i-2} = z_i.$$

By the same way, we can obtain $|v_0| > \cdots > |v_m|$ and (2.4). □

The following corollary justifies our procedure which approximates Δ_i^{-1} by the tridiagonal matrix Λ_i.

Corollary 2. *Suppose that the conditions of Theorem* 1 *hold. Then we have*

$$|\tau_{ij}| \geqslant |\tau_{ij+1}|, \quad \textit{for } i \leqslant j, \tag{2.5}$$

$$|\tau_{ij}| \geqslant |\tau_{ij-1}|, \quad \textit{for } j \leqslant i, \tag{2.6}$$

and

$$|\tau_{ij}| \leqslant \frac{|b_1| r}{|a_1 a_2|} R^{-3} \left(\frac{R}{r}\right)^m \frac{1}{R^{|i-j|}}, \tag{2.7}$$

where

$$r = \min_{2 \leqslant i \leqslant m-1} \left\{ \frac{|b_i| - |a_{i+1}|}{|a_i|}, \frac{|b_i| - |a_i|}{|a_{i+1}|} \right\} \geqslant 1$$

and

$$R = \max_{2 \leqslant i \leqslant m-1} \left\{ \frac{|b_i| + |a_{i+1}|}{|a_i|}, \frac{|b_i| + |a_i|}{|a_{i+1}|} \right\} \geqslant 1.$$

Proof. The inequalities (2.5) and (2.6) are direct consequences from the definition of (τ_{ij}) and the assertion (i) of Theorem 1. To prove (2.7), take h_1 arbitrarily. Then, for $u_1 = h_1$, we have

$$|u_2| = \left| -\frac{1}{a_2}(a_1 u_0 + b_1 u_1) \right| = \left| \frac{b_1}{a_2} \right| |u_1| \geqslant |u_1|.$$

If $|u_{i-1}| \geqslant |u_{i-2}|$, $i \geqslant 3$, then

$$|u_i| = \left| \frac{1}{a_1}(a_{i-1}u_{i-2} + b_{i-1}u_{i-1}) \right| \geqslant \frac{1}{|a_i|}(|b_{i-1}||u_{i-1}| - |a_{i-1}||u_{i-2}|)$$

$$\geqslant \left(\frac{|b_{i-1}| - |a_{i-1}|}{|a_i|} \right) |u_{i-1}| \geqslant r |u_{i-1}| \geqslant r^{i-2} \left| \frac{b_1}{a_2} \right| |u_1|$$

and

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

By the same way, we obtain

$$|v_i| \geqslant r^{m-i-1} \left| \frac{b_m}{a_m} \right| |v_m|$$

and

$$|v_i| \leqslant R^{m-i-1} \left| \frac{b_m}{a_m} \right| |v_m|, \quad i = 0, 1, \dots, m.$$

Hence, if $i \leqslant j$, then we have

$$|\tau_{ij} = \left| -\frac{u_i v_j}{a_1 h_1 v_0} \right| \leqslant \left| \frac{1}{a_1 h_1} \cdot \frac{a_m}{r^{m-1} b_m v_m} R^{m-j-1} \frac{b_m}{a_m} v_m R^{i-2} \frac{b_1}{a_2} u_1 \right|$$

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

Furthermore, if $j \leqslant i$, then

$$|\tau_{ij}| = \left| -\frac{u_j v_i}{a_1 h_1 v_0} \right| \leqslant \left| \frac{b_1}{a_1 a_2} \right| \left(\frac{R}{r} \right)^{m-1} R^{j-i-2},$$

so that (2.7) holds. □

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 α such that $\|\phi\|_\infty \leqslant \alpha h^2 < 4(p_* + q_*)\sin^2\frac{1}{2}\pi h$, then as $h \to 0$, we have*

$$\kappa(B) \geqslant \frac{4(p_* + q_*) \sin^2\frac{1}{2}\pi(1-h) - \alpha h^2}{4(p^* + q^*) \sin^2\frac{1}{2}\pi h + \alpha h^2} \to \infty \tag{3.1}$$

and

$$\kappa(A^{-1}B) \leqslant \frac{4(p_* + q_*) \sin^2\frac{1}{2}\pi h + \alpha h^2}{4(p_* + q_*) \sin^2\frac{1}{2}\pi h - \alpha h^2} \to \frac{(p_* + q_*)\pi^2 + \alpha}{(p_* + q_*)\pi^2 - \alpha}. \tag{3.2}$$

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

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

$$\Gamma_2 = \begin{pmatrix} 2(p_*+q_*)I_m - p_*C & -q_*I_m & & & \\ -q_*I_m & 2(p_*+q_*)I_m - p_*C & & & -q_*I_m \\ & & & \ddots & \\ & & & & -q_*I_m \\ & & -q_*I_m & & 2(p_*+q_*)I_m - p_*C \end{pmatrix},$$

I_m is the $m \times m$ identity and

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

We have

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

$$\lambda_1(\Gamma_2) = 2(p_*+q_*)(1-\cos \pi h) = 4(p_*+q_*)\sin^2\tfrac{1}{2}\pi h$$

and

$$\lambda_n(\Gamma_2) = 2(p_*+q_*)(1-\cos m\pi h) = 4(p_*+q_*)\sin^2\tfrac{1}{2}\pi(1-h)$$

(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

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

and

$$(\nabla_2)_{ii} = (A - \Gamma_2)_{ii} = a_{ii} - 2(p_*+q_*) \geqslant \sum_{i \neq j} |(\nabla_2)_{ij}| > 0.$$

From

$$\frac{(Bx, x)}{(x, x)} \geqslant \frac{(\Gamma_2 x, x)}{(x, x)} + \frac{(\phi x, x)}{(x, x)} \geqslant 4(p_*+q_*)\sin^2\tfrac{1}{2}\pi h - \alpha h^2 > 0,$$

for any vector $x \neq 0$, we have that the eigenvalues of B and $A^{-1}B$ are positive and

$$\lambda_n(B) = \max_{x \neq 0} \frac{(Bx, x)}{(x, x)} \geqslant \max_{x \neq 0} \frac{(\Gamma_2 x, x)}{(x, x)} + \min_{x \neq 0} \frac{(\phi x, x)}{(x, x)} \geqslant \lambda_n(\Gamma_2) - \alpha h^2$$

and

$$\lambda_1(B) = \min_{x \neq 0} \frac{(Bx, x)}{(x, x)} \leqslant \min_{x \neq 0} \frac{(\Gamma_1 x, x)}{(x, x)} + \max_{x \neq 0} \frac{(\phi x, x)}{(x, x)} \leqslant \lambda_1(\Gamma_1) + \alpha h^2.$$

This implies that

$$\kappa(B) = \frac{\lambda_n(B)}{\lambda_1(B)} \geqslant \frac{\lambda_n(\Gamma_2) - \alpha h^2}{\lambda_1(\Gamma_1) + \alpha h^2} = \frac{4(p_* + q_*) \sin^2 \frac{1}{2}\pi(1-h) - \alpha h^2}{4(p^* + q^*) \sin^2 \frac{1}{2}\pi h + \alpha h^2} \to \infty$$
$$\text{(as } h \to 0\text{)}.$$

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

$$\lambda_n(A^{-1}B) \leqslant 1 + \alpha h^2 \lambda_1^{-1}(\Gamma_2), \qquad \lambda_1(A^{-1}B) \geqslant 1 - \alpha h^2 \lambda_1^{-1}(\Gamma_2)$$

and

$$\kappa(A^{-1}B) = \frac{\lambda_n(A^{-1}B)}{\lambda_1(A^{-1}B)} \leqslant \frac{1 + \alpha h^2 \lambda_1^{-1}(\Gamma_2)}{1 - \alpha h^2 \lambda_1^{-1}(\Gamma_2)}$$
$$= \frac{1 + \alpha h^2 / 4(p_* + q_*) \sin^2 \frac{1}{2}\pi h}{1 - \alpha h^2 / 4(p_* + q_*) \sin^2 \frac{1}{2}\pi h} \to \frac{(p_* + q_*)\pi^2 + \alpha}{(p_* + q_*)\pi^2 - \alpha} \quad \text{(as } h \to 0\text{)}. \qquad \square$$

Next, we consider the cases $M = D$, $M = T$, $M = S_\omega$, and $M = C_\omega$. Let

$$\delta_1 = \min_{x \neq 0} \frac{((LD^{-1}L^t + A)x, x)}{(Dx, x)}, \qquad \delta_2 = \min_{x \neq 0} \frac{((L_c T^{-1} L_c^t + A)x, x)}{(Tx, x)},$$

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

Then we have the following corollary.

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

(i) $\kappa(D^{-1}B) \geqslant \dfrac{(p_* + q_*)}{(p^* + q^*)} \kappa(B) \to \infty,$

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

(iii) $\kappa(S_\omega^{-1}B) \geqslant \dfrac{F_1(\omega)(p_* + q_*)^2}{4(p^* + q^*)^2} \kappa(B) \to \infty,$ *if* $\omega < \gamma_1$,

(iv) $\kappa(C_\omega^{-1}B) \geqslant \dfrac{F_2(\omega)(2q_* + 4p_* \sin^2 \frac{1}{2}\pi h)^2}{16(p^* + q^*)^2} \kappa(B) \to \infty,$ *if* $\omega < \gamma_2$, *where*

$$F_1(\omega)=\begin{cases}\omega^2\delta_1+(1-\omega), & 0<\omega\leqslant 1,\\ \omega\delta_1+(1-\omega), & 1\leqslant\omega<\gamma_1,\end{cases}\qquad F_2(\omega)=\begin{cases}\omega^2\delta_2+(1-\omega), & 0<\omega\leqslant 1,\\ \omega\delta_2+(1-\omega), & 1\leqslant\omega<\gamma_2.\end{cases}$$

Furthermore,

$$\lambda_1(D^{-1}A)\leqslant\delta_1\leqslant\frac{(p^*+q^*)^2}{(p^*+q^*+p_*+q_*)^2}+\lambda_1(D^{-1}A), \tag{3.3}$$

$$\lambda_1(T^{-1}A)\leqslant\delta_2\leqslant\frac{(q^*)^2}{(2q_*+4p_*\sin^2\frac{1}{2}\pi h)^2}+\lambda_1(T^{-1}A)$$

and

$$\min_{0<\omega\leqslant\gamma_1}F_1(\omega)=\begin{cases}F_1(\gamma_1), & \text{if } \delta_1\leqslant\delta^*,\\ F_1\left(\dfrac{1}{2\delta_1}\right), & \text{if } \delta_1\geqslant\delta^*,\end{cases} \tag{3.4}$$

$$\min_{0<\omega\leqslant\gamma_2}F_2(\omega)=\begin{cases}F_2(\gamma_2), & \text{if } \delta_2\leqslant\delta^*,\\ F_2\left(\dfrac{1}{2\delta_2}\right), & \text{if } \delta_2\geqslant\delta^*,\end{cases}$$

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

$$\kappa(M^{-1}B)\geqslant\frac{\kappa(B)}{\kappa(M)}.$$

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) $\lambda_1(D)\geqslant 2(p_*+q_*)$ and $\lambda_n(D)\leqslant 2(p^*+q^*)$.

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

(iii) The matrix S_ω can be expressed as

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

Hence

$$(S_\omega x,x)\geqslant\begin{cases}\left[\omega^2\dfrac{((LD^{-1}L^{t}+A)x,x)}{(Dx,x)}+(1-\omega)\right]\dfrac{(Dx,x)}{(2-\omega)\omega}, & \text{for } \omega\leqslant 1,\\ \left[\omega\dfrac{((LD^{-1}L^{t}+A)x,x)}{(Dx,x)}+(1-\omega)\right]\dfrac{(Dx,x)}{(2-\omega)\omega}, & \text{for } \omega\geqslant 1,\end{cases}$$

and

$$\lambda_1(S_\omega) = \min_{\|x\|=1} (S_\omega x, x) \geqslant 2F_1(\omega)\frac{p_* + q_*}{(2-\omega)\omega}.$$

Furthermore, we have from (1.6),

$$\lambda_n(S_\omega) \leqslant \| S_\omega \|_\infty \leqslant \frac{\| D + \omega L \|_\infty \| D + \omega L^t \|_\infty \| D^{-1} \|_\infty}{(2-\omega)\omega} \leqslant 8\frac{(p^* + q^*)^2}{(p_* + q_*)(2-\omega)\omega}.$$

This proves (iii).

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

$$\lambda_1(D^{-1}A) \leqslant \frac{(Ae_i, e_i)}{(De_i, e_i)} = 1.$$

If $L^t x \neq 0$, then

$$\frac{((LD^{-1}L^t + A)x, x)}{(Dx, x)} > \frac{(Ax, x)}{(Dx, x)} \geqslant \lambda_1(D^{-1}A).$$

If $L^t x = 0$, then

$$\frac{((LD^{-1}L^t + A)x, x)}{(Dx, x)} = \frac{(Dx, x)}{(Dx, x)} = 1.$$

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

$$\delta_1 \leqslant \max\frac{(LD^{-1}L^t x, x)}{(Dx, x)} + \min\frac{(Ax, x)}{(Dx, x)} = \lambda_n(D^{-1}LD^{-1}L^t) + \lambda_1(D^{-1}A)$$

$$\leqslant \| D^{-1}L \|_\infty \| D^{-1}L^t \|_\infty + \lambda_1(D^{-1}A) \leqslant \frac{(p^* + q^*)^2}{(p^* + q^* + p_* + q_*)^2} + \lambda_1(D^{-1}A).$$

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

$$F_1\left(\frac{1}{2\delta_1}\right) = \frac{4\delta_1 - 1}{4\delta_1} \leqslant F_1(2) = 2\delta_1 - 1, \quad \text{if } \delta_1 \geqslant \delta^*.$$

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

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

$$\mu = \max_{x\neq 0} \frac{(Dx, x)}{(Ax, x)}, \qquad \delta = \max_{x\neq 0} \frac{((LD^{-1}L^t - \frac{1}{4}D)x, x)}{(Ax, x)},$$

and

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

Then, $\delta \geqslant -\frac{1}{4}$, $\lambda_n(S_\omega^{-1}A) \leqslant 1$, $\lambda_1(S_\omega^{-1}A) \geqslant 1/G(\omega)$ and $\kappa(S_\omega^{-1}A) \leqslant G(\omega)$. Furthermore,

$$\min_{0<\omega<2} G(\omega) = G(\omega^*) = \sqrt{(\tfrac{1}{2} + \delta)\mu} + \tfrac{1}{2} \leqslant \sqrt{(\tfrac{1}{2} + \delta)\kappa(A)} + \tfrac{1}{2},$$

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

They further proved that δ is bounded ($\delta \leqslant 0$) if

$$\|D^{-1/2}LD^{-1/2}\|_\infty \leqslant \tfrac{1}{2} \quad \text{and} \quad \|D^{-1/2}L^{\mathrm{t}}D^{-1/2}\|_\infty \leqslant \tfrac{1}{2}. \tag{3.5}$$

By using their results, we obtain

$$\kappa(S_\omega^{-1}B) \leqslant G(\omega)\frac{\lambda_1(S_\omega)+\alpha h^2}{\lambda_1(S_\omega)-\alpha h^2 G(\omega)},$$

since

$$\kappa(S_\omega^{-1}B) \leqslant \frac{\lambda_n(S_\omega^{-1}A)+\alpha h^2\lambda_n(S_\omega^{-1})}{\lambda_1(S_\omega^{-1}A)-\alpha h^2\lambda_n(S_\omega^{-1})}$$

and $\lambda_n(S_\omega^{-1}) = 1/\lambda_1(S_\omega)$. Hence under the assumptions (3.5), $\kappa(S_{\omega^*}^{-1}B) = \mathrm{O}(\sqrt{\kappa(A)})$, 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 $\mathrm{O}(\sqrt{\kappa(B)})$, i.e., $\mathrm{O}(\sqrt{n})$.

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 $\mathrm{O}(\kappa(B)) = \mathrm{O}(n) = \mathrm{O}(h^{-2})$, if $\omega < \gamma_1$. Furthermore we remark that $\gamma_1 < \omega^*$ if $\lambda_1(D^{-1}A) \leqslant \frac{1}{4}(2-\sqrt{2})$. In fact, under the assumptions (3.5), we have $\delta_1 \leqslant \frac{1}{4} + \lambda_1(D^{-1}A)$ so that $\gamma_1 \leqslant 4/(3-4\lambda_1(D^{-1}A))$. On the other hand, observing that $\lambda_n(A^{-1}D)^{-1} = \lambda_1(D^{-1}A) \leqslant 1$, we have

$$\begin{aligned}\omega^* &> \frac{2\sqrt{\lambda_n(A^{-1}D)}}{\sqrt{\lambda_n(A^{-1}D)}+\sqrt{2}} \geqslant \frac{2\lambda_n(A^{-1}D)}{\lambda_n(A^{-1}D)+\sqrt{2}}\\ &= \frac{2}{1+\lambda_1(D^{-1}A)\sqrt{2}} \geqslant \frac{4}{3-4\lambda_1(D^{-1}A)} \geqslant \gamma_1.\end{aligned}$$

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 $\kappa(C_\omega^{-1}A) \leqslant G(\omega)$, where μ and δ in $G(\omega)$ are replaced by

$$\mu = \max_{x\neq 0}\frac{(Tx, x)}{(Ax, x)}, \qquad \delta = \max_{x \neq 0}\frac{((L_cT^{-1}L_c^{\mathrm{t}} - \frac{1}{4}T)x, x)}{(Ax, x)}.$$

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,	$k \geqslant 0,\ l \geqslant 0$;
(2) $M = T$:	$5n$,	$k = 0,\ l = 0$,
	$3n$,	otherwise;
(3) $M = S_\omega$:	$7n$,	$k \geqslant 0,\ l \geqslant 0$;
(4) $M = C_\omega$:	$13n - 2m$,	$k = 0,\ l = 0$,
	$11n - 2m$,	otherwise;
(5) $M = I$:	0,	$k \geqslant 0,\ l \geqslant 0$;
(6) $M = A$:	$(2m+1)n + \frac{1}{2}n(n-1) + \frac{1}{6}m(7n+5)$,	$k = 0,\ l = 0$,
	$(2m+1)n$,	otherwise;
(7) $M = H$:	$19n$,	$k = 0,\ l = 0$,
	$6n$,	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 $(x_0)_i = 20(-1)^i$, $1 \leqslant i \leqslant n$, and employ the stopping criteria $\| r_l \|_2 \leqslant 10^{-6}$, $\| F(x_{k+1}) \|_\infty / \| F(x_0) \|_\infty \leqslant 10^{-5}$. 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 1
Total computing time (sec.)

n	D	T	S_1	S_{ω^*}	C_{ω^*}	I	A	H
9	0.17	0.20	0.22	0.23	0.23	0.13	0.18	0.23
49	1.43	1.80	1.65	1.53	1.70	1.37	1.33	1.70
225	10.38	11.23	9.05	7.53	7.88	8.93	11.62	8.20
961	88.25	89.48	64.75	44.23	47.22	76.53	149.85	51.43
3969	806.75	683.83	463.37	248.12	261.95	672.02	2226.67	347.83

Table 2
Number of iterations ($k[l_1, l_2, \ldots, l_k]$)

n	D	T	S_1	S_{ω^*}	C_{ω^*}	I	A	H
9	3[3,3,3]	3[4,4,4]	3[4,4,3]	3[4,4,3]	3[4,3,3]	3[3,3,3]	4[1,2,1,2]	3[6,5,4]
49	4[9,9,8,8]	4[12,9,8,8]	4[9,6,5,5]	4[8,6,5,4]	4[7,5,4,4]	4[9,9,9,9]	4[1,2,2,2]	4[9,7,6,5]
225	3[25,18,18]	3[24,17,15]	3[15,10,8]	3[12,8,7]	3[11,6,6]	3[18,19,19]	3[1,2,2]	3[14,9,7]
961	3[51,37,37]	3[46,33,28]	3[27,17,14]	3[17,12,10]	3[15,9,8]	3[38,39,39]	3[2,2,2]	3[22,13,11]
3969	3[104,74,73]	3[90,55,54]	3[50,28,27]	3[24,16,13]	3[21,12,11]	3[75,77,78]	3[2,2,2]	3[38,22,18]

Table 3
Total computing time (sec.)

ϵ	D	T	S_1	S_{ω^*}	C_{ω^*}	I	A	H
$5.0 \cdot 10^{-7}$	*	*	*	*	*	*	13.65	*
$1.0 \cdot 10^{-6}$	*	*	*	*	*	*	13.65	*
$2.5 \cdot 10^{-6}$	*	14.30	11.32	*	*	11.80	13.65	10.08
$5.0 \cdot 10^{-6}$	10.40	11.25	9.05	7.52	9.73	8.87	11.63	8.23
$7.5 \cdot 10^{-6}$	10.40	11.30	9.05	7.53	7.87	8.87	11.62	8.28

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

Table 4
Number of iterations ($k[l_1, l_2, \ldots, l_k]$)

ϵ	D	T	S_1	S_{ω^*}	C_{ω^*}	I	A	H
$5.0 \cdot 10^{-7}$	*	*	*	*	*	*	4[1,2,2,2]	*
$1.0 \cdot 10^{-6}$	*	*	*	*	*	*	4[1,2,2,2]	*
$2.5 \cdot 10^{-6}$	*	4[24,17,15,15]	4[15,10,8,8]	*	*	4[18,19,19,19]	4[1,2,2,2]	4[14,9,7,7]
$5.0 \cdot 10^{-6}$	3[25,18,18]	3[24,17,15]	3[15,10,8]	3[12,8,7]	4[11,6,6,5]	3[18,19,19]	3[1,2,2]	3[14,9,7
$7.5 \cdot 10^{-6}$	3[25,18,18]	3[24,17,15]	3[15,10,8]	3[12,8,7]	3[11,6,6]	3[18,19,19]	3[1,2,2]	3[14,9,7]

Table 5
Upper and lower bounds for $\kappa(A^{-1}B)$ and $\kappa(B)$

	9	49	225	961	3969
$\kappa(A^{-1}B) \leqslant$	1.1127	1.1082	1.1077	1.1068	1.1068
$\kappa(B) \geqslant$	5.4826	23.9917	98.0526	394.3027	1579.305

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 ith iteration.

Now, we change the value ϵ for the stopping criterion $\| F(x_k) \|_\infty / \| F(x_0) \|_\infty \leqslant \epsilon$ 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 $M = C_{\omega^*}$ is faster than the others and we roughly conclude that

$$\kappa(B) \geqslant \kappa(D^{-1}B) \geqslant \kappa(T^{-1}b) \geqslant \kappa(S_1^{-1}B) \geqslant \kappa(H^{-1}B) \geqslant \kappa(S_{\omega^*}^{-1}B) \geqslant \kappa(C_{\omega^*}^{-1}B)$$
$$\geqslant \kappa(A^{-1}B).$$

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

$$T(A^{-1}B) \geqslant T(D^{-1}B) \geqslant T(T^{-1}B) \geqslant T(B) \geqslant T(S_1^{-1}B) \geqslant T(H^{-1}B) \geqslant T(C_{\omega^*}^{-1}B)$$
$$\geqslant T(S_{\omega^*}^{-1}B),$$

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

[1] O. Axelsson and V.A. Barker, *Finite Element Solution of Boundary Value Problems* (Academic Press, New York, 1984).

[2] X. Chen, On the convergence of Broyden-like methods for nonlinear equations with nondifferentiable terms, *Ann. Inst. Statist. Math.* **42** (1990) 387–401.

[3] X. Chen and T. Yamamoto, Convergence domains of certain iterative methods for solving nonlinear equations, *Numer. Funct. Anal. Optim.* **10** (1&2) (1989) 37–48.

[4] P. Concus, G.H. Golub and G. Meurant, Block preconditioning for the conjugate gradient method, *SIAM J. Sci. Statist. Comput.* **6** (1985) 220–252.

[5] R.T. Gregory and D.L. Karney, *A Collection of Matrices for Testing Computational Algorithms* (Wiley/Interscience, New York, 1969).

[6] J.C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations* (Wadsworth & Brooks/Cole, Pacific Grove, CA, 1990).

[7] T. Yamamoto and X. Chen, Ball-convergence theorems and error estimates for certain iterative methods for nonlinear equations, *Japan J. Appl. Math.* **7** (1990) 131–143.

[8] T. Yamamoto and Y. Ikebe, Inversion of band matrices, *Linear Algebra Appl.* **24** (1979) 105–111.