Journal of Computational and Applied Mathematics 40 (1992) 73–89
North-Holland

CAM 1102

Nonlinear CG-like iterative methods *

A.T. Chronopoulos

Department of Computer Science, University of Minnesota, Minneapolis, MN 55455, United States

Received 8 November 1990
Revised 27 March 1991

Abstract

Chronopoulos, A.T., Nonlinear CG-like iterative methods, Journal of Computational and Applied Mathematics 40 (1992) 73–89.

A nonlinear conjugate gradient method has been introduced and analyzed by J.W. Daniel. This method applies to nonlinear operators with symmetric Jacobians. Orthomin(1) is an iterative method which applies to nonsymmetric and definite linear systems. In this article we generalize Orthomin(1) to a method which applies directly to nonlinear operator equations. Each iteration of the new method requires the solution of a scalar nonlinear equation. Under conditions that the Hessian is uniformly bounded away from zero and the Jacobian is uniformly positive definite the new method is proved to converge to a globally unique solution. Error bounds and local convergence results are also obtained. Numerical experiments on solving nonlinear operator equations arising in the discretization of nonlinear elliptic partial differential equations are presented.

Keywords: Nonlinear algebraic systems, iterative methods, Orthomin.

1. Introduction

Nonlinear systems of equations often arise when solving initial- or boundary value problems in ordinary or partial differential equations. We consider the nonlinear system of equations

$$F(x) = 0, \tag{1.1}$$

where $F(x)$ is a nonlinear operator from a real Euclidean space of dimension N or Hilbert space into itself. The Newton method coupled with direct linear system solvers is an efficient way to solve such nonlinear systems when the dimension of the Jacobian is small. When the Jacobian is large and sparse, some kind of iterative method may be used. This can be a nonlinear iteration (for example, functional iteration for contractive operators), or an inexact Newton method. In an inexact Newton method the solution of the resulting linear systems is

Correspondence to: Prof. A.T. Chronopoulos, Department of Computer Science, University of Minnesota, 200 Union Street S.E., Minneapolis, MN 55455, United States.

* The research was partially supported by NSF under grant CCR-8722260. The research was also partially supported by the U.S. Department of Energy while the author was at the University of Illinois. The Minnesota Supercomputing Institute provided time on the CRAY-2.

approximated by a linear iterative method. The following are typical steps in an inexact Newton method for solving the nonlinear system (1.1).

(1) Choose x_0.
For $j = 0$ **Until** Convergence **do**
(2) Solve iteratively: $F'(x_n)\,\Delta_n = -F(x_n)$;
(3) $x_{n+1} = x_n + \Delta_n$;
EndFor.

Step (2) often consists of an inner linear solver iteration. If the linear iterative method is a Krylov subspace method (e.g., conjugate gradient, Chebyshev), then the Jacobian is only required for performing Jacobian times vector operations. The explicit computation of the Jacobian requires additional sparse storage and computation time. Efficient methods to compute directly sparse Jacobians have been proposed [19]. Alternatively, the Jacobian times vector operation can be approximated [4,16] using the following divided difference:

$$F'(x_0)v \equiv \frac{F(x_0 + \epsilon v) - F(x_0)}{\epsilon}. \tag{1.2}$$

A very important question is how to terminate the inner and outer iterations in an inexact Newton algorithm. Let the outer iteration terminate if

$$\| F(x_n) \| \leqslant \epsilon, \tag{1.3}$$

where ϵ is a given error tolerance. The following two possibilities for terminating the inner iteration and retaining convergence for an inexact Newton algorithm have been studied. Chan and Jackson [5] proved that the inexact Newton method converges if the Jacobian is a uniformly definite operator in the neighborhood of the solution and the inner iteration stopping criterion is

$$\| F(x_n) + F'(x_n)\,\Delta_n \| \leqslant \epsilon. \tag{1.4}$$

Dembo et al. [10] proved that if we choose $0 < \eta_n \leqslant t < 1$, for all n, and if the inner iteration stopping criterion is

$$\| F(x_n) + F'(x_n)\,\Delta_n \| \leqslant \eta_n \| F(x_n) \|, \tag{1.5}$$

then the convergence of the inexact Newton algorithm is locally at least linear. Thus, we can choose $\eta_n \equiv \eta < 1$ and obtain linear convergence rate in an inexact Newton algorithm. Superlinear or quadratic convergence for the outer Newton iteration can be obtained if the linear residual norm is $o(\| F(x_n) \|)$ or $O(\| F(x_n) \|^2)$, respectively. For quadratic convergence the Jacobian $F'(x)$ needs to be locally Lipschitz continuous. We note that several inner iterations may still be required to obtain convergence (linear or higher rate) with this stopping criterion.

Nonlinear steepest descent methods for the minimal residual and normal equations have been studied by many authors (cf. [22,23]). Fletcher and Reeves [15], and Daniel [7] have obtained a nonlinear conjugate gradient method which converges if the Jacobian is symmetric and uniformly positive definite. These nonlinear methods reduce to the standard conjugate gradient methods for linear systems. These methods are based on exact line search at each iteration and thus must solve a scalar nonlinear minimization problem in order to determine the steplengths. Several authors have suggested inexact line search and have given conditions

under which these methods would still converge [1,12,14]. This is done to avoid solving exactly the scalar minimization problem whose derivative evaluation involves evaluation of the nonlinear operator.

In the last two decades many Krylov subspace iterative methods have been derived for solving nonsymmetric linear systems of equations. These methods are generalizations of the conjugate gradients methods for symmetric and positive definite linear systems. Some outstanding examples are the generalized conjugate residual method (GCR), Orthomin(k) [13,25] and the generalized conjugate gradient method (GCG) [2].

In this article we undertake the task of deriving a nonlinear generalization of Orthomin(1). This new method is called *Nonlinear Orthomin*(1). This method consists of an iteration which requires computation of a nonlinear steplength. It coincides with the linear Orthomin(1) if the operator equation is linear. We prove global and local convergence results for this new method. We also provide asymptotic residual error bounds. We compare the Nonlinear Orthomin(1) in terms of performance to the inexact Newton-Orthomin(1) algorithms with stopping criteria (1.3), (1.4) and (1.3)–(1.5). We present two test problems. These are operator equations arising in the discretization of nonlinear elliptic partial differential equations. The Nonlinear Orthomin(1) demonstrated superior performance to the inexact Newton-Orthomin(1).

In Section 2, we review the Orthomin method. In Section 3, we derive a nonlinear extension to Orthomin(1). Under assumptions on the Jacobian and Hessian of the nonlinear systems we show that this method converges to a globally unique solution. In Section 4, we prove local convergence results and give asymptotic residual error bounds. In Section 5, we describe practical details in implementing the Newton-Orthomin(1) and Nonlinear Orthomin(1) methods. We also describe the preconditioned Nonlinear Orthomin(1) method with right constant operator preconditioning. In Section 6, we show some numerical experiments for nonlinear systems arising from the discretization of linear boundary value problems in PDEs. In Section 7, we draw conclusions and describe future work on this subject.

2. The Orthomin method

In this section, we review the Orthomin(k) method [13,25]. Let us consider the system of linear equations

$$Ax = f, \tag{2.1}$$

where A is a large and sparse matrix of order N. Direct methods may be inefficient for solving this problem because of the large amount of work and storage involved. Iterative methods can be used to obtain an approximate solution.

Assume that A is Symmetric Positive Definite (SPD). Then the Conjugate Gradient method (CG) applies. Solving a SPD linear system by use of the CG method is equivalent to minimizing a quadratic error functional

$$E(x) = (x-h)^{\mathrm{T}} A(x-h),$$

where $h = A^{-1}f$ is the solution of the system. In infinite precision arithmetic the exact solution is reached in at most N iterations. The conjugate residual (CR) method is a variant of the

conjugate gradient method in which the residual norm $E(x) = \| Ax - f \|^2$ is minimized at every iteration.

The generalized conjugate residual (GCR) [13] is an extension of CR which applies to nonsymmetric systems provided that the symmetric part of the matrix $\frac{1}{2}(A^T + A)$ is positive definite. This method also terminates (in infinite arithmetic) in at most N steps. However, the storage requirements increase at every step. Vinsome [25] proposed the Orthomin(k), as a practical version of GCR, where the latter has the drawback of keeping all the previous direction vectors. Let Pr denote the approximate inverse operator of A in a right preconditioning for the system (2.1).

Algorithm 2.1. Orthomin(k).

(0) Choose x_0.

(1) Compute $r_0 = f - Ax_0$.

(2) $p_0 = \mathrm{Pr}\ r_0$.

For $n = 0$ **Step** 1 **Until** Convergence **Do**

(3) $$c_n = \frac{(r_n, Ap_n)}{(Ap_n, Ap_n)};$$

(4) $$x_{n+1} = x_n + c_n p_n;$$

(5) $$r_{n+1} = r_n - c_n Ap_n;$$

(6) $$p_{n+1} = \mathrm{Pr}\ r_{n+1} + \sum_{j=j_n}^{n} b_j^n p_j, \quad \text{where } b_j^n = -\frac{(A\ \mathrm{Pr}\ r_{n+1}, Ap_j)}{(Ap_j, Ap_j)},\ j \leqslant n;$$

(7) $$Ap_{n+1} = A\ \mathrm{Pr}\ r_{n+1} + \sum_{j=j_n}^{n} b_j^n Ap_j, \quad \text{where } j_n = \min(0, n-k+1);$$

Endfor.

Eisenstat et al. [13] proved that Orthomin(k), $k > 0$, converges. Note that if the matrix is symmetric, Orthomin(1) is the CR method. Orthomin applied with right preconditioning minimizes the residual norm of the unpreconditioned system, while left preconditioning minimizes a preconditioned residual norm.

3. The Nonlinear Orthomin(1) method

In this section, we generalize the Orthomin(1) iteration to a nonlinear iteration which requires the solution of a scalar equation to determine the steplength. We then prove a global convergence result under assumptions that the Hessian and the Jacobian are uniformly bounded and the Jacobian is uniformly definite.

Let $F(x)$ be an operator mapping of the Euclidean space $\mathbb{R}^n$ (or, even more generally, a real Hilbert space) into itself. The notation $F'(x)$ and $F''(x)$ will be used to denote the Fréchet and Gâteaux derivatives, respectively. Also, for simplicity F_n' and F_n'' will denote $F'(x_n)$ and

$F''(x_n)$, respectively. We seek to solve iteratively the nonlinear system of equations $F(x) = 0$. In the linear case, $F(x) = Ax - b$ and $F'(x) = A$.

Assume that $F'(x)$ and $F''(x)$ exist at all x and that there exist scalars $0 < m \leqslant M$, $0 < B$, independent of x, so that the following conditions are satisfied for any vectors x and v:

$$m \| v \|^2 \leqslant (F'(x)v, v) \leqslant M \| v \|^2, \tag{3.1a}$$

$$\| F''(x) \| \leqslant B, \tag{3.1b}$$

$$m^2 \| v \|^2 \leqslant \left(\left(F'(x)^{\mathrm{T}} F'(x) \right) v, v \right) \leqslant M^2 \| v \|^2. \tag{3.1c}$$

Remark 3.0. The rightmost inequality in (3.1a) and the leftmost inequality in (3.1c) are can be derived from the remaining inequalities. To see this we use the Schwarz inequality and the rightmost inequality in (3.1c) to obtain the rightmost inequality in (3.1a). We also use the Schwarz inequality and the leftmost inequality in (3.1a) to obtain the leftmost inequality in (3.1c).

Condition (3.1a) states that the symmetric part of the Jacobian is uniformly positive definite. This stems from the identity $(F'(x)v, v) = (\frac{1}{2}[F'(x) + F'(x)^{\mathrm{T}}]v, v)$. Note that (3.1c) is satisfied if for example the Jacobian and its inverse are bounded: $\| F'(x) \| < M$ and $\| F'(x)^{-1} \| < 1/m$ imply $m^2 \| v \|^2 \leqslant (F'(x)^{\mathrm{T}} F'(x)v, v) \leqslant M^2 \| v \|^2$.

Under assumptions (3.1) we consider the following nonlinear generalization of Orthomin(1).

Algorithm 3.1. Nonlinear Orthomin(1).

(0) Choose x_0.

(1) $p_0 = r_0 = -F(x_0)$.

For $n = 0$ **Until** Convergence **Do**

(2) Select the smallest $c_n > 0$ to solve $\min_{c>0} \| F(x_n + cp_n) \|$;

(3) $x_{n+1} = x_n + c_n p_n$;

(4) $r_{n+1} = -F(x_{n+1})$;

(5) $p_{n+1} = r_{n+1} + b_n p_n$, where $b_n = -\dfrac{(F'_{n+1} r_{n+1}, F'_{n+1} p_n)}{\| F'_{n+1} p_n \|^2}$;

EndFor.

The selection of scalars c_n and b_n guarantees that the two orthogonality conditions

$$(r_n, F'_n p_{n-1}) = 0 \tag{3.2}$$

and

$$(F'_n p_n, F'_n p_{n-1}) = 0 \tag{3.3}$$

hold. Under the assumptions (3.1) the following lemma holds true for Algorithm 3.1.

Lemma 3.2. *Let* $\{r_n\}$ *be the nonlinear residuals and* $\{p_n\}$ *be the direction vectors in Algorithm* 3.1; *then the following identities hold true*:

(i) $(r_n, F'_n p_n) = (r_n, F'_n r_n)$;

(ii) $(F_n'p_n, F_n'r_n) = \|F_n'p_n\|^2$;

(iii) $(F_n'p_n, p_n) = (F_n'r_n, r_n) + b_{n-1}^2(F_n'p_{n-1}, p_{n-1})$;

(iv) $\|F_n'r_n\|^2 = \|F_n'p_n\|^2 + b_{n-1}^2\|F_n'p_{n-1}\|^2$;

(v) $\|r_n\| \leq \left(\frac{M}{m}\right)^{1/2} \|p_n\|$;

(vi) $\|p_n\| \leq \frac{M}{m}\|r_n\|$;

(vii) $\|r_{n+1}\| \leq \|r_n\|$.

Proof. The orthogonality relations (3.2) and (3.3) combined with step (5) of Algorithm 3.1 imply (i)–(iv). Equality (iii) is used in proving inequality (v) as follows:

$$m\|r_n\|^2 \leq |(r_n, F_n'r_n)| \leq |(p_n, F_n'p_n)| \leq \|p_n\| \, \|F_n'p_n\| \leq M\|p_n\|^2.$$

Equality (iv) is used in proving inequality (vi) as follows:

$$m^2\|p_n\|^2 \leq \|F_n'p_n\|^2 \leq \|F_n'r_n\|^2 \leq M^2\|r_n\|^2.$$

Inequality (vii) follows from the definition of c_n. □

Remark 3.3. Let $f_n(c)$ denote the scalar function $\frac{1}{2}\|F(x_n + cp_n)\|^2$. Its first and second derivatives are given by

$$f_n'(c) = (F(x_n + cp_n), F'(x_n + cp_n)p_n), \tag{3.4}$$

$$f_n''(c) = ((F''(x_n + cp_n)p_n, p_n), F(x_n + cp_n)) + \|F'(x + cp_n)p_n\|^2. \tag{3.5}$$

The following upper and lower bounds on $f_n''(c)$ can be computed from (3.5), the assumptions (3.1) and Lemma 3.2:

$$\|p_n\|^2(m^2 - B\|r_0\|) \leq m^2\|p_n\|^2 - B\|p_n\|^2\|r_n\| \leq f_n''(c), \tag{3.6}$$

$$f_n''(c) \leq \|p_n\|^2(M^2 + B\|r_n\|) \leq \|p_n\|^2(M^2 + B\|r_0\|). \tag{3.7}$$

We next prove that if assumptions (3.1) hold, then the nonlinear Orthomin(1) iteration converges to a globally unique solution.

Theorem 3.4. *Under the assumptions* (3.1) *on the nonlinear operator* $F(x)$ *the sequence* x_n *generated by Algorithm* 3.1 *is well-defined for any* x_0, *it converges to a unique solution* x^* *of the nonlinear system* $F(x) = 0$ *and*

$$\|x_n - x^*\| < \frac{1}{m}\|F(x_n)\|.$$

Proof. The proof is divided in four parts.

Firstly, we prove the existence of c_n in step (2) of Algorithm 3.1. The derivative of the real function f_n at zero

$$f_n'(0) = -(r_n, F_n'r_n)$$

is negative because of assumptions (3.1). So there exists $c > 0$ such that $\| F(x_n + cp_n) \| < \| r_n \|$. We must prove that there is a $c > 0$ such that $f_n(0) \leqslant f_n(c)$. This would imply that there exists a $c_n > 0$ where $f_n(c)$ assumes a local minimum. We use the value theorem for the operator $F(x)$ to obtain the following equation:

$$(F(y) - F(x), (y - x)) = (F'(z)(y - x), (y - x)).$$

Combining this with condition (3.1a) we obtain the inequality

$$m \| y - x \|^2 \leqslant \| F(y) - F(x) \| \, \| y - x \|.$$

This inequality implies that

$$m \| y - x \| \leqslant \| F(y) - F(x) \|. \tag{3.8}$$

For $x = x_n$ and $y = x_n + cp_n$ we conclude that $F(y)$ grows unbounded for $c \to \infty$. This proves that there is a $c > 0$ such that $f_n(0) \leqslant f_n(c)$.

Secondly, we obtain a lower bound on the steplength c_n. Taylor's expansion gives $f_n'(c_n) = 0 = f_n'(0) + c_n f_n''(\bar{c}_n)$, where $\bar{c}_n = t_n c_n$ for some t in (0, 1). We solve for c_n. We then use the upper bound in equality (3.7) and Lemma 3.2(vi) to obtain

$$\frac{m^3}{M^2(M^2 + B \| r_0 \|)} \leqslant \frac{m \| r_n \|^2}{\| p_n \|^2 (M^2 + B \| r_0 \|)} \leqslant \frac{(r_n, F_n' r_n)}{\| p_n \|^2 (M^2 + B \| r_n \|)} \leqslant c_n. \tag{3.9}$$

Thirdly, we prove that the sequence of residual norms decreases to zero. For $\bar{c} = tc$ for some t in [0, 1] we have

$$f_n(c) = \tfrac{1}{2} \| r_n \|^2 - c(r_n, F_n' r_n) + \tfrac{1}{2} c^2 f_n''(\bar{c}).$$

To obtain the upper bound on $f_n(c)$ we use (3.1a), (3.7) and Lemma 3.2(vi):

$$f_n(c) \leqslant \left[\tfrac{1}{2} - mc + \tfrac{1}{2} c^2 (B \| r_n \| + M^2) \frac{M^2}{m^2} \right] \| r_n \|^2.$$

Now by inserting

$$c = \frac{m^3}{M^2(M^2 + B \| r_0 \|)}$$

we obtain the following bound on the residual error:

$$\tfrac{1}{2} \| r_{n+1} \|^2 = f_n(c_n) \leqslant f_n(c) \leqslant \frac{1}{2} \left[1 - \frac{m^3}{M^2(M^2 + B \| r_0 \|)} \right] \| r_n \|^2.$$

Since M is the dominant term in the bound expression it follows that $\| r_n \| \to 0$.

Finally, we prove that the sequence of iterates converges to a unique solution of the nonlinear operator equation. By use of (3.8) with $x = x_n$ and $y = x_{n+k}$ we obtain that the sequence x_n is a Cauchy sequence. Thus it converges to x^* and $F(x^*) = 0$. The uniqueness and the error bound inequality in the theorem statement follow from (3.8) with $x = x_n$ and $y = x^*$. □

4. Asymptotic steplength estimates and error bounds

In this section, we obtain asymptotic estimates of the steplengths c_n near the solution. We then obtain residual error bounds and prove a local convergence result.

Firstly, we prove a lemma giving bounds on the nonlinear steplengths.

Lemma 4.1. *We consider assumptions* (3.1) *and the additional condition on the initial residual* $\|r_0\| < m^2/(2B)$. *Let* c_n *and* p_n *be as in Theorem* 3.4; *then the inequalities*

$$\frac{m^3}{M^4+m^2} \leqslant c_n \leqslant \frac{2M^{3/2}}{m} \tag{4.1}$$

and

$$\|c_n p_n\| \leqslant 2M\|r_n\| \tag{4.2}$$

hold.

Proof. To prove the rightmost inequality in (4.1) we follow the proof of (3.9). We solve $f_n'(c_n) = 0 = f_n'(0) + c_n f_n''(\bar{c}_n)$ for c_n. Then we use the lower bound in (3.6) and Lemma 3.2(i) to obtain

$$c_n \leqslant \frac{(r_n, F_n' r_n)}{\|p_n\|^2(m^2 - B\|r_0\|)} \leqslant \frac{\|r_n\|\,\|F_n' p_n\|}{\|p_n\|^2(m^2 - B\|r_0\|)}. \tag{4.3}$$

Now, using (3.1c), Lemma 3.2(v) and the condition on the initial residual $\|r_0\| < m^2/(2B)$, we obtain

$$c_n \leqslant \frac{M\|r_n\|}{\|p_n\|(m^2 - B\|r_0\|)} \leqslant 2\frac{M^{3/2}}{m}.$$

The leftmost inequality in (4.1) is proved from (3.9) and the condition on the initial residual. The inequality (4.2) follows from (4.3) and (3.1c). □

We next obtain asymptotic estimates of the steplengths c_n.

Proposition 4.2.

$$\frac{(F_n' r_n, r_n)}{\|F_n p_n\|^2}\frac{1}{(1+\epsilon_n)} \leqslant c_n \leqslant \frac{(F_n' r_n, r_n)}{\|F_n' p_n\|^2}\frac{1}{(1-\epsilon_n)}, \quad \text{where } \epsilon_n = O(\|r_n\|).$$

Proof. We will prove only the rightmost inequality. The leftmost inequality is proved similarly. The derivative of f_n by use of (3.4) and expansion into second-order derivative terms becomes

$$f_n'(c) = \left(\left[F_n' p_n + c(F_{n_1}'' p_n, p_n)\right], \left[F(x_n) + cF_n' p_n + \tfrac{1}{2}c^2(F_{n_2}'' p_n, p_n)\right]\right),$$

where $(F_{n_i}'' p_n, p_n) = (F''(x_n + ct_i p_n)p_n, p_n)$ for some t_i with $0 < t_i < 1$ and $i = 1, 2$. Now developing the inner product and using our assumptions (3.1) and Lemma 3.2 we obtain

$$f_n'(c) \leqslant -(F_n' r_n, r_n) + c\|F_n' p_n\|^2 + cB\|p_n\|^2\left(\tfrac{3}{2}c\|F_n' p_n\| + \|r_n\| + \tfrac{1}{2}c^2 B\|p_n\|^2\right).$$

Thus, for $c = c_n$ we obtain

$$(F_n' r_n, r_n) \leqslant c_n \| F_n' p_n \|^2 + c_n \| F_n' p_n \|^2 \epsilon_n,$$

where

$$\epsilon_n = \frac{\| p_n \|^2 B}{\| F_n' p_n \|^2} \left[\tfrac{3}{2} c_n \| F_n' p_n \| + \| r_n \| + \tfrac{1}{2} c_n^2 B \| p_n \|^2 \right].$$

Using condition (3.1c) and inequality (4.2) in Lemma 4.1 we obtain

$$\epsilon_n \leqslant \epsilon \| r_n \|,$$

where

$$\epsilon = \frac{B}{m^2} \left[3M^2 + 1 + 2M^2 B \| r_0 \| \right]. \qquad \square$$

We next prove a lemma relating the norms of two successive residuals in Algorithm 3.1.

Lemma 4.3. *Under the assumptions of Theorem* 3.4 *we obtain the following asymptotic difference of the residual error* $E(x_n) = \| r_n \|^2$ *of two successive iterates*:

$$E(x_{n+1}) - E(x_n) = -c_n (r_n, F_n' r_n) + O\left(\| r_n \|^3 \right).$$

Proof. From the identity $E(x_{n+1}) - E(x_n) = (r_{n+1} - r_n, r_{n+1}) - (r_n, r_n - r_{n+1})$ by expanding in Taylor's series around x_{n+1} and x_n, respectively the first and second term we obtain

$$E(x_{n+1}) - E(x_n) = \left(\left[F_{n+1}' (-c_n p_n) + \tfrac{1}{2} c_n^2 (F''_{\overline{n+1}} p_n, p_n) \right], r_{n+1} \right)$$
$$- \left(r_n, \left[c_n F_n' p_n + \tfrac{1}{2} c_n^2 (F''_{\bar{n}} p_n, p_n) \right] \right),$$

where $\overline{n+1} = x_{n+1} + t c_n p_n$ and $\bar{n} = x_n + s c_n p_n$ for some t, s in (0, 1). Since r_{n+1} is orthogonal to $F_{n+1}' p_n$ and by using Lemma 3.2(i) and Lemma 4.1 the above expression gets reduced to

$$-c_n (r_n, F_n' r_n) + \tfrac{1}{2} c_n^2 \left[\left((F''_{\overline{n+1}} p_n, p_n), r_{n+1} \right) - \left(r_n, (F''_{\bar{n}} p_n, p_n) \right) \right].$$

Now the difference in the square brackets is easily seen to be bounded by $2BM^2 \| r_n \|^3$ using assumptions (3.1b) and Lemma 4.1. $\square$

Now, we use the previous lemma to obtain an asymptotic residual error bound for iterates in Algorithm 3.1.

Proposition 4.4. *Under the assumptions of Theorem* 3.4 *we obtain the following inequality on the residual errors*:

$$E(x_{n+1}) \leqslant E(x_n) d_n,$$

where $d_n = [1 - m^2/M^2 + \sigma_n]$ *and* $\sigma_n = 2M^2 \| r_n \|$.

Proof. By Lemma 4.3 we need to have an estimate of $c_n(r_n, F_n'r_n)$. Using conditions (3.1), Lemma 3.2 and Proposition 4.2 we prove the following inequality:

$$c_n(r_n, F_n'r_n) \geqslant \frac{(r_n, F_n'r_n)^2}{\|F_n'p_n\|^2}\frac{1}{(1+\epsilon_n)} \geqslant \frac{(r_n, F_n'r_n)^2}{\|F_n'r_n\|^2}\frac{1}{(1+\epsilon_n)} \geqslant \frac{m^2}{M^2}\frac{1}{(1+\epsilon_n)}\|r_n\|^2.$$

Now using Lemma 4.3 we obtain

$$E(x_{n+1}) \leqslant \|r_n\|^2\left[1-\frac{m^2}{M^2}\frac{1}{(1+\epsilon_n)}\right]+O(\|r_n\|^3) \leqslant E(x_n)\left[1-\frac{m^2}{M^2}\right]+O(\|r_n\|^3).$$

The last term in this inequality is

$$E(x_n)\left[1-\frac{m^2}{M^2}+\sigma_n\right], \quad \text{where } \sigma_n = 2BM^2\|r_n\|. \qquad \square$$

In the following theorem we prove local convergence for the Nonlinear Orthomin(1) iteration and give an error bound estimate. Under assumptions (3.1) and an additional assumption on the initial residual norm it is proved that the iteration remains inside a ball centered at x_0 with the appropriate choice of a radius δ_0.

Theorem 4.5. *Let x_0 be selected such that*

$$\|r_0\| < \frac{1}{2BM^2} \quad \textit{and} \quad d_0^2 = \left(1-\frac{m^2}{M^2}\right)+\sigma_0 < 1,$$

where $\sigma_2 = 2BM^2\|r_n\|$. Let δ_0 denote $2M^2/(1-d_0)\|r_0\|$; the sequence x_n generated by Algorithm 3.1 remains in the ball $B(x_0, \delta_0)$ and it converges to x^*, which is the unique solution of $F(x)=0$. Furthermore, $d_n^2 = (1-m^2/M^2)+\sigma_n$ decreases to $1-m^2/M^2$ and $\|x_n - x^*\| \leqslant \delta_0 d_0 d_1 \cdots d_{n-1}$.

Proof. From Lemma 4.1 we obtain $\|c_0p_0\| \leqslant 2M^2\|r_0\| \leqslant \delta_0$. So, $x_1 = x_0 + c_0p_0$ is in $B(x_0, \delta_0)$.

Let δ_1 denote $2M^2\|r_1\|/(1-d_1)$. We will prove that $B(x_1, \delta_1) \subset B(x_0, \delta_0)$. This follows from the following inequality if we prove that the rightmost term is bounded by δ_0:

$$\|x-x_0\| \leqslant \|x-x_1\| + \|x_1-x_0\| \leqslant \delta_1 + 2M^2\|r_0\|. \tag{4.4}$$

Proposition 4.4 implies that $\|r_1\| \leqslant d_0\|r_0\|$. This inequality and $d_1 < d_0$ imply that

$$\frac{\|r_1\|}{1-d_1} + \|r_0\| = \left[\frac{d_0}{1-d_1}+1\right]\|r_0\| \leqslant \left[\frac{d_0}{1-d_0}+1\right]\|r_0\| = \frac{1}{1-d_0}\|r_0\|.$$

This proves that the last term in (4.4) is less that δ_0.

Since $\|r_{n-1}\| \leqslant \|r_n\|$ and $d_{n+1} \leqslant d_n$, we can prove by induction that the iterate x_n generated in Algorithm 3.1 satisfies the hypothesis of this theorem.

Now $\|r_n\|^2 \leqslant d_{n-1}^2 \cdots d_0^2\|r_0\|^2 \leqslant d_0^{2n}\|r_0\|^2$ implies that the sequence of residuals r_n converges to 0. Also,

$$\|x_{n+k}-x_n\| \leqslant 2M^2\sum_{j=n}^{n+k-1}\|r_j\| \leqslant \frac{2M^2}{m^2}\|r_0\|\sum_{j=n}^{n+k-1} d_0d_1\cdots d_{j-1}$$

proves the sequence of iterates x_n converges to $x^* \in B(x_0, \delta_0)$ and $F(x^*) = 0$. When k approaches infinity, the inequality becomes

$$\| x^* - x_n \| \leqslant C_0 d_0 \cdots d_{n-1},$$

for some constant C_0. □

5. Implementation details

We next describe four algorithms based on Newton-Orthomin(1) and Nonlinear Orthomin(1) with different choices of stopping criteria. We then present the right preconditioning of the nonlinear system (1.1) with constant preconditioning matrix.

(1) *Newton-Orthomin*(1). This is the inexact Newton method with stopping criteria (1.3), (1.4) (described in the Introduction) for the outer and inner iterations, respectively.

(2) *Nonlinear Orthomin*(1). This is Algorithm 3.1. The algorithm halts when the norm of the nonlinear residual $F(x_n)$ is less than a tolerance ϵ.

(3) *Restarted Newton-Orthomin*(1). This is the inexact Newton method with stopping criteria (1.3)–(1.5) (described in the Introduction) for the outer and inner iterations, respectively.

(4) *Restarted Nonlinear Orthomin*(1). The algorithm halts when the norm of the nonlinear residual $F(x_n)$ is less than a tolerance ϵ. However, it restarts setting $x_0 = x_n$ when

$$\| F(x_n) \| \leqslant \eta_n \| F(x_0) \|.$$

Algorithms (1) and (2) are nonlinear extensions of the linear Orthomin(1) Algorithm 2.1. This means that if the problem is linear, algorithm (1) will perform only one outer iteration. It also will perform the same number of inner iterations as algorithm (2). Algorithms (3) and (4) are restarted algorithms. If the problem is linear, then the restarted algorithms (3) and (4) generate the same iterates.

In the implementation, algorithms (2) and (4) are based on modification of Algorithm 3.1 in order to avoid exact line searches. The steplength to Proposition 4.2 can be approximated by the formula

$$c_n \approx \frac{(r_n, F_n' r_n)}{\| F_n' p_n \|^2}.$$

Also, we approximate

$$b_n \approx -\frac{(F_n' r_n, F_n' p_n)}{\| F_n' p_n \|^2}$$

in order to save computational work. The only vector product computed is $F_n' r_n$ and this is used to approximate $F_n' p_n \approx F_n' r_n + b_{n-1} F_{n-1}' p_{n-1}$. This is generally expected to lead to a stable method because the Jacobian does not vary much in a single iteration.

We briefly discuss here the right preconditioning of Algorithm 3.1 with a constant preconditioner Pr. The matrix Pr is assumed to be an approximation to the inverse of the Jacobian $F'(x)$ (i.e., $F'(x)\,\mathrm{Pr} \approx I$). The problem $F(x) = 0$ is transformed to $G(y) = F(\mathrm{Pr}\ y) = 0$, where $y = \mathrm{Pr}^{-1} x$. Note that the Jacobian of the transformed problem is $G'(y) = F'(x)\,\mathrm{Pr}$. Now, it is

easy to check that Algorithm 3.1 applied to $G(y)=0$ yielding approximants y can be transformed to yield approximants x_n. The only changes required are in steps (1) and (5). These steps become (1') and (5'), respectively:

(1') $p_0 = \mathrm{Pr}\, r_0 = -F(x_0)$;

(5') $p_{n+1} = \mathrm{Pr}\, r_{n+1} + b_n p_n$, where $b_n = -\dfrac{(F'_{n+1}\, \mathrm{Pr}\, r_{n+1},\, F'_{n+1} p_n)}{\|F'_{n+1} p_n\|^2}$.

Most of the computational work is attributed to Jacobian ($F'(x)$) evaluations, Jacobian times vectors and functions ($F(x)$) evaluations. Algorithms (1) and (3) require one Jacobian, one function evaluation per outer iteration and one Jacobian times vector operation per inner iteration. Algorithms (2) and (4) require one Jacobian, one function evaluation and one matrix-vector product per iteration. If the Jacobian times vector operation is approximated as in (1.2), then no Jacobian evaluation is required. By using (1.2) algorithms (1) and (3) require one function evaluation per outer and one function evaluation per inner iteration. However, algorithms (2) and (4) require two function evaluations per iteration. For the preconditioned methods one matrix times vector product per iteration (by the matrix Pr) must be added to the work for all algorithms.

6. Numerical tests

In this section, we present two nonlinear elliptic partial differential equation problems whose discretization gives rise to nonlinear algebraic systems with nonsymmetric Jacobians. We use the four algorithms described in the previous section. We compare their execution times on a CRAY-2 vector computer.

Problem 6.1. Let us consider the nonlinear differential equation

$$-\Delta u + \beta u_x + \gamma u^3 = f(x, y), \tag{6.1}$$

where u and f are defined on the unit square $[0, 1]\times[0, 1]$. We determine $f(x, y)$ so that $u(x,y) = e^{x^2+y^2}$ is the solution of (6.1) with inhomogeneous Dirichlet conditions imposed on the boundary. We discretize the problem using a central difference approximation for Δ and upwind first-order approximation for u_x. Thus we obtain a system of nonlinear equations $F(x)=0$ of order $N=n_x^2$ with n_x being the number of interior grid points in each direction. The same problem with $\beta=0$ and $\gamma=1$ was used in [18].

Problem 6.2. Let us consider the nonlinear differential equation

$$-\Delta u + \beta u_x + \gamma\, e^u = f(x, y), \tag{6.2}$$

where u and f are defined on the unit square $[0, 1]\times[0, 1]$. Again we determine $f(x, y)$ so that $u(x, y) = e^{x^2+y^2}$ is the solution of (6.2) with inhomogeneous Dirichlet conditions imposed on the boundary. This is a simplified form of the Bratu problem [17]. It is known that for $\gamma \geqslant 0$ there exists a unique solution to the problem. We discretize this problem in the same way as Problem 6.1.

Table 1
Problem 6.1. Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 10$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-ITT	Times
16	27	0.0243	4	58	16	0.0427
32	44	0.0662	4	109	32	0.1372
64	77	0.2505	4	215	64	0.6200
128	151	1.6035	5	525	128	4.8985
160	183	2.9454	5	661	160	9.1898
200	220	5.4893	5	852	200	22.4173

Table 2
Problem 6.1. Restarted Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 10$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	25	0.0222	15	28	2	0.0259
32	43	0.0622	16	40	4	0.0595
64	84	0.2804	19	86	8	0.2776
128	171	1.9352	21	172	16	1.7147
160	202	3.3169	21	205	20	3.1165
200	235	6.0235	22	263	24	5.8935

Table 3
Problem 6.1. Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 30$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	26	0.0251	4	50	16	0.0461
32	52	0.0786	4	98	32	0.1323
64	113	0.3911	4	197	64	0.5515
128	280	3.0074	4	392	128	3.6663
160	379	6.1082	4	500	160	6.9018
200	535	16.3794	4	644	200	16.7861

The constants γ and β are used to control the nonlinearity of the problems and the nonsymmetry of the Jacobians. We considered $\gamma = 1$, $\beta = 10$ or 30. The Jacobian of this problem for small enough mesh size can be proven to satisfy conditions (3.1) locally in $\mathbb{R}^N$. Thus nonlinear Orthomin(1) can be used to solve the nonlinear discretized problems. The Jacobian is closer to symmetry in the case of $\beta = 10$. In the case of $\beta = 30$, although the Jacobian has a more significant skew symmetric part, it is also closer to linearity.

For the test problems described here the linear part of the Jacobian was computed once initially and it was stored in five diagonals. The nonlinear part consisting of a single diagonal was computed explicitly whenever needed. It was computed once in each outer iteration of algorithms (1) and (3). However it was computed once in every iteration of algorithms (2) and (4).

In Tables 1–8, we show numerical results on solving Problems 6.1 and 6.2 using the algorithms (1)–(4) with ILU(0) vectorizable preconditioning [6,21,24]. Right preconditioning

Table 4
Problem 6.1. Restarted Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 30$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	23	0.0215	15	20	3	0.01809
32	41	0.0626	15	36	5	0.0511
64	78	0.2827	17	68	11	0.2175
128	109	1.2906	21	173	25	1.7402
160	141	2.3649	21	168	27	2.5152
200	197	5.066	21	217	42	5.2995

Table 5
Problem 6.2. Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 10$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	23	0.0194	4	54	16	0.0443
32	38	0.0494	4	108	32	0.1299
64	73	0.2072	4	209	64	0.5091
128	135	1.2548	4	400	128	3.0540
160	158	2.1547	4	418	160	5.8677
200	233	5.2345	4	682	200	15.9824

Table 6
Problem 6.2. Restarted Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 10$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	24	0.0210	14	26	2	0.0225
32	41	0.0555	17	42	4	0.0543
64	84	0.2535	19	88	7	0.2401
128	164	1.5348	20	167	14	1.3579
160	189	2.6093	21	197	18	2.3018
200	233	6.1126	21	261	22	7.0066

Table 7
Problem 6.2. Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 30$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	26	0.0242	4	47	16	0.0432
32	50	0.0705	4	97	32	0.1197
64	109	0.3297	4	186	64	0.4676
128	264	2.4841	4	371	128	2.8146
160	367	5.2559	4	454	160	5.1785
200	509	15.4821	4	551	21	12.3918

Table 8
Problem 6.2. Restarted Nonlinear Orthomin(1) and Newton-Orthomin(1) ($\gamma = 1$, $\beta = 30$)

$\sqrt{N}$	NIT	Ntimes	OIT	IIT	MAX-IIT	Times
16	20	0.0190	15	21	3	0.0183
32	35	0.0474	15	35	5	0.0445
64	66	0.1968	18	82	11	0.2135
128	134	1.2832	19	135	23	1.0743
160	161	2.2292	20	137	28	1.5894
200	157	3.3931	20	200	40	3.5133

was used with the ILU(0) preconditioning obtained from the discrete linear part of the differential operators. The error tolerance was $\epsilon = 10^{-6}$. The initial vector was chosen to be the average of the solution in the four vertices of the square domain $x_0 = \frac{1}{4}[1 + 2e^1 + e^2]$. The stopping criterion parameter η was chosen $\eta = \frac{1}{2}$. This choice was observed to be best compared to $\eta = 0.1, \ldots, 0.9$. Varying η between iterations was not considered.

The flags used in the tables are defined as follows. Firstly, for algorithms (1) and (3): OIT = number of outer (Newton) iterations, IIT = total number of inner iterations, MAX-IIT = maximum number of inner linear iterations in a single Newton step. The maximum number of linear iterations allowed was set at nx. Secondly, for algorithms (2) and (4): NIT = number of nonlinear iterations. The execution times (Ntimes for algorithms (2), (4) and Times for algorithms (1), (3)) are given in seconds on a single vector processor of CRAY-2 at the University of Minnesota. Although the computations were not performed in a single-user mode, the load of the machine was low. Several runs were performed and timings with variations 0.001 seconds were observed.

7. Conclusions and future work

We have presented and analyzed a nonlinear iterative method for solving nonlinear algebraic systems of equations. This method is a generalization of existing methods for linear systems of equations. We show that under certain uniform assumptions on the Jacobians and Hessians the method is guaranteed to converge globally to a unique solution. We also prove local convergence results and give asymptotic error bound estimates. These results extend the work of other authors [8,15] to deriving nonlinear methods for nonsymmetric Jacobians.

It is well known that Krylov subspace iterative methods for linear systems require the formation of a small-dimensional subspace on which they project to obtain an approximation to the solution. What we suggest here is the inversion of the Newton linear iterative method to a nonlinear iteration Newton method.

The test results show that the Nonlinear Orthomin(1) algorithms are in most cases more efficient than the Newton-Orthomin(1) algorithms. The number of (inner) iterations can be used as a rough indicator to explain the difference in efficiency of the methods. However, each iteration of the Nonlinear Orthomin(1) is more costly than the corresponding inner iterations of Newton-Orthomin(1). For $\beta = 10$ the Jacobians are close to symmetric. This explains the superiority of the Nonlinear Orthomin(1) algorithm. For $\beta = 30$ the Jacobians are sufficiently

nonsymmetric but closer to linear than for $\beta = 10$. For $\beta = 10$, Nonlinear Orthomin(1) is slower than the restarted Newton-Orthomin(1) or the restarted Nonlinear Orthomin(1). In this case a method based on GCR [13] or GCG [2] with more vectors in storage must be used for the linear problems. We did not consider nonlinear extensions of these methods here. In general the restarted methods gave unexpected number of iterations as a function of nx. The Newton-Orthomin(1) showed the worst performance in all cases.

The method presented here requires the solution of a scalar nonlinear equation. In a future publication we will investigate inexact line search approaches similar to [1,8,12] for determining the parameters in these nonlinear methods.

Acknowledgements

The author thanks the anonymous referees whose comments helped enhance significantly the quality of presentation of this article. He also thanks Dr. C.W. Gear for his help and advice.

References

[1] M. Al-Baali, Descent property and global convergence of the Fletcher–Reeves method with inexact line searches, *IMA J. Numer. Anal.* **5** (1985) 121–124.
[2] O. Axelsson, A generalized conjugate gradient, least square method, *Numer. Math.* **51** (1987) 209–227.
[3] P.N. Brown, A local convergence theory for combined inexact-Newton/finite-difference projection methods, *SIAM J. Numer. Anal.* **24** (1987) 610–638.
[4] P.N. Brown and Y. Saad, Hybrid Krylov methods for nonlinear systems of equations, *SIAM J. Sci. Statist. Comput.* **11** (3) (1990) 450–481.
[5] T. Chan and K. Jackson, The use of iterative linear-equation solvers in codes for large systems of stiff IVPs for ODEs, *SIAM J. Sci. Statist. Comput.* **7** (1986) 378–417.
[6] A.T. Chronopoulos and C.W. Gear, On the efficient implementation of preconditioned s-step conjugate gradient methods on multiprocessors with memory hierarchy, *Parallel Comput.* **11** (1) (1989) 37–53.
[7] J.W. Daniel. The conjugate gradient method for linear and nonlinear operator equations, *SIAM J. Numer. Anal.* **4** (1967) 10–26.
[8] J.W. Daniel, Convergence of the conjugate gradient method with computationally convenient modifications, *Numer. Math.* **10** (1967) 125–131.
[9] J.W. Daniel, *The Approximate Minimization of Functionals* (Prentice-Hall, Englewood Cliffs, NJ, 1971).
[10] R.S. Dembo, S.C. Eisenstat and T. Steighaug, Inexact Newton methods, *SIAM J. Numer. Anal.* **19** (1982) 400–408.
[11] J.E. Dennis and J.J. More, Quasi-Newton methods, motivation and theory, *SIAM Rev.* **19** (1) (1977) 46–89.
[12] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations* (Prentice-Hall, Engewood Cliffs, NJ, 1983).
[13] S.C. Eisenstat, H.C. Elman and M.H. Schultz, Variational iterative methods for nonsymmetric systems of linear equations, *SIAM J. Numer. Anal* **20** (1983) 345–357.
[14] R. Fletcher, *Practical Methods in Optimization, Vol. 2, Unconstrained Optimization* (Wiley, Chichester, 1980).
[15] R. Fletcher and C.M. Reeves, Function minimization by conjugate gradients, *Comput. J.* **7** (1964) 149–154.
[16] W. Gear and Y. Saad, Iterative solution of linear equations in ODE codes, *SIAM J. Sci. Statist. Comput.* **4** (1983) 583–601.
[17] R. Glowinski, H.B. Keller and L. Reinhart, Continuation conjugate gradient methods for the least squares solution of nonlinear boundary value problems, *SIAM J. Sci. Statist. Comput.* **6** (1985) 793–832.
[18] G.H. Golub and R. Kannan, Convergence of a two stage Richardson process for nonlinear equations, *BIT* (1986) 209–216.

[19] A. Griewank, On automatic differentiation, in: M. Iri and K. Tanabe, Eds., *Mathematical Programming* (Kluwer Academic Publishers, Dordrecht, 1989).
[20] M. Hestenes and E. Stiefel, Methods of conjugate gradients for solving linear systems, *J. Res. Nat. Bur. Standards* **49** (1952) 409–436.
[21] S. Ma and A.T. Chronopoulos, Implementation of iterative methods for large sparse nonsymmetric linear systems on parallel vector computers, *Internat. J. Supercomput.* **4** (4) (1990) 9–24.
[22] M.Z. Nashed, The convergence of the method of steepest descent for nonlinear equations with variational or quasi-variational operators, *J. Math. Mech.* **13** (1964) 765–794.
[23] T.L. Saaty, *Modern Nonlinear Equations* (Dover, New York, 1981).
[24] H. van der Vorst, A vectorizable variant of some ICCG methods, *SIAM J. Sci. Statist. Comput.* **3** (3) (1982) 350–356.
[25] P. Vinsome, An iterative method for solving sparse sets of simultaneous equations, Society of Petroleum Engineers of AIME, SPE 5729, 1976.