SPECTRAL RESIDUAL METHOD
WITHOUT GRADIENT INFORMATION FOR SOLVING
LARGE-SCALE NONLINEAR SYSTEMS OF EQUATIONS

WILLIAM LA CRUZ, JOSÉ MARIO MARTÍNEZ, AND MARCOS RAYDAN

Abstract. A fully derivative-free spectral residual method for solving large-scale nonlinear systems of equations is presented. It uses in a systematic way the residual vector as a search direction, a spectral steplength that produces a nonmonotone process and a globalization strategy that allows for this nonmonotone behavior. The global convergence analysis of the combined scheme is presented. An extensive set of numerical experiments that indicate that the new combination is competitive and frequently better than well-known Newton-Krylov methods for large-scale problems is also presented.

1. Introduction

We introduce a derivative-free nonmonotone iterative method for solving the nonlinear system of equations

\[ F(x) = 0, \]

where \( F : \mathbb{R}^n \to \mathbb{R}^n \) is a continuously differentiable mapping. We are interested in large-scale systems for which the Jacobian of \( F \) is not available or requires a prohibitive amount of storage.

Recently, La Cruz and Raydan \[17\] introduced the Spectral Algorithm for Nonlinear Equations (SANE) for solving (1). SANE uses in a systematic way the residual \( \pm F(x_k) \) as a search direction. The first trial point at each iteration is \( x_k - \sigma_k F(x_k) \), where \( \sigma_k \) is a spectral coefficient. Global convergence is guaranteed by means of a variation of the nonmonotone strategy of Grippo, Lampariello and Lucidi \[13\]. This approach requires descent directions with respect to the squared norm of the residual. As a consequence, the computation of a directional derivative, or a very good approximation of it, is necessary at every iteration.

The spectral coefficient is an appropriate Rayleigh quotient with respect to a secant approximation of the Jacobian. Spectral gradient methods for minimization were originated in the Barzilai–Borwein paper \[1\]. The properties of their method for general quadratic functions were elucidated in \[23\]. Further analysis of spectral
gradient methods can be found in [8,10,24], among others. For a review containing the more recent advances on spectral choices of the steplength for minimization problems (see [11]).

In this paper we introduce a new nonmonotone line-search technique that can be associated to the same search directions and the same initial steplengths as the SANE algorithm. In other words, the first trial point at each iteration will be the same as in SANE, but the line-search strategy will be different. The main consequence is that, in the new approach, directional derivatives are not required at all.

We also present an extensive set of numerical experiments that indicate that the new method is competitive and sometimes better than the SANE algorithm. We recall that, in [17], SANE was in turn compared favorably with several Newton–Krylov methods (see, e.g., [2,6,7]). Therefore, the new algorithm represents an encouraging low-cost scheme for solving (1).

Notation.

- \(J(x)\) will denote the Jacobian matrix of \(F\) computed at \(x\).
- For all \(x \in \mathbb{R}^n\) we denote \(g(x) = 2J(x)^T F(x) = \nabla \|F(x)\|_2^2\).
- The set of natural numbers will be denoted \(\mathbb{N} = \{0,1,2,\ldots\}\).
- If \(\{z_k\}_{k \in \mathbb{N}}\) is a sequence and \(K = \{k_1, k_2, k_3, \ldots\}\) is an infinite sequence of natural numbers such that \(k_i < k_j\) if \(i < j\), we denote \(\lim_{k \in K} z_k = \lim_{j \to \infty} z_{k_j}\).
- The symbol \(\| \cdot \|\) will always denote the Euclidian norm.
- \(B(x, \varepsilon)\) will denote the open ball with center \(x\) and radius \(\varepsilon\). That is, \(B(x, \varepsilon) = \{z \in \mathbb{R}^n \mid \|z - x\| < \varepsilon\}\).

2. The new nonmonotone line-search strategy

The best known nonmonotone line-search technique for unconstrained optimization was introduced by Grippo, Lampariello and Lucidi [13]. It has been used to globalize the spectral gradient method [24] and some of its extensions for convex constrained optimization [3,4] and nonlinear systems of equations [17]. Different nonmonotone line-search techniques, associated to Newton and quasi-Newton strategies, have been proposed for solving (1) (see [12,18]). Li and Fukushima [18] presented an interesting idea that avoids the necessity of descent directions to guarantee that each iteration is well defined. Let us briefly describe the Grippo–Lampariello–Lucidi (GLL) and the Li–Fukushima (LF) schemes.

The GLL condition can be written as follows:

\[
f(x_k + \alpha_k d_k) \leq \max_{0 \leq j \leq M-1} f(x_{k-j}) + \gamma \alpha_k \nabla f(x_k)^T d_k,
\]

where \(M\) is a nonnegative integer, \(0 < \gamma < 1\) and \(f\) is a merit function such that \(f(x) = 0\) if and only if \(\|F(x)\| = 0\).

The LF condition can be written as follows:

\[
\|F(x_k + \alpha_k d_k)\| \leq (1 + \eta_k)\|F(x_k)\| - \gamma \alpha_k^2 \|d_k\|_2^2,
\]

where \(\sum_k \eta_k \leq \eta < \infty\).

It follows that if \(\nabla f(x_k)^T d_k < 0\), then the GLL condition is satisfied for \(\alpha_k\) sufficiently close to zero and we can compute a steplength \(\alpha_k\) by using a finite backtracking process. However, when \(\nabla f(x_k)^T d_k = 0\), the existence of \(\alpha_k\) satisfying
the GLL condition is not guaranteed. Moreover, when \( d_k = \pm F(x_k), \nabla f(x_k) \) could be close to zero or zero, and then stagnation or breakdown might occur during the backtracking process.

One possible remedy is to use the LF condition. This condition does not need the computation of \( J(x_k) d_k \). Moreover, it is satisfied, if \( \alpha_k \) is small enough, independently of the choice of \( d_k \). However, since \( \eta_k \) is usually very small when \( k \) is large, the Li–Fukushima strategy generally imposes an almost monotone behavior of the merit function when \( x_k \) is close to a solution. This is not a good feature when one uses spectral gradient or spectral residual steps because, in these cases, the pure undamped methods (where \( \alpha_k = 1 \) for all \( k \)), although generally effective, are usually highly nonmonotone even in the neighborhood of an isolated solution. The reason for this is not completely understood, but the analogy with the behavior of the spectral gradient (or Barzilai–Borwein) method for minimizing convex quadratics may be useful (see [23]). In the quadratic case the spectral gradient method does not need line-search strategies for being globally convergent, but the functional values do not decrease monotonically at all. Therefore, imposing any kind of monotonicity is not convenient. Many authors, including Fletcher [11], pointed out the necessity of avoiding monotonicity requirements in the spectral framework as much as possible.

In this work we combine and extend the GLL and LF conditions to produce a robust nonmonotone line-search globalization strategy that somehow takes into account the advantages of both schemes. Roughly speaking the new descent condition can be written as

\[
 f(x_{k+1}) \leq \max_{0 \leq j \leq M-1} f(x_{k-j}) + \eta_k - \gamma \alpha_k^2 f(x_k). 
\]

The GLL term \( \max_{0 \leq j \leq M-1} f(x_{k-j}) \) is responsible for the sufficiently nonmonotone behavior of \( f(x_k) \) even when \( k \) is large. On the other hand, the presence of \( \eta_k > 0 \) guarantees that all the iterations are well defined, and the forcing term \( -\gamma \alpha_k^2 f(x_k) \) provides the arguments for proving global convergence.

We would like to mention that a similar extension that also combines the GLL with the LF conditions was presented and briefly discussed in the final remarks of [17]. However, it was neither analyzed nor tested. The present work was motivated by the need for studying the theoretical and practical properties of this globalization technique.

3. Model algorithm and convergence

We assume that \( F : \mathbb{R}^n \to \mathbb{R}^n \) has continuous partial derivatives. Let \( n_{exp} \in \{1, 2\} \). Define

\[
 f(x) = \|F(x)\|^{n_{exp}} \quad \text{for all} \quad x \in \mathbb{R}^n.
\]

Assume that \( \{\eta_k\} \) is a sequence such that \( \eta_k > 0 \) for all \( k \in \mathbb{N} \) and

\[
 \sum_{k=0}^{\infty} \eta_k = \eta < \infty.
\]

Assume that \( 0 < \gamma < 1 \) and \( 0 < \sigma_{\min} < \sigma_{\max} < \infty \). Let \( M \) be a positive integer. Let \( \tau_{\min}, \tau_{\max} \) be such that \( 0 < \tau_{\min} < \tau_{\max} < 1 \), and let \( x_0 \in \mathbb{R}^n \) be an arbitrary initial point.

We now present the DF-SANE algorithm.
Algorithm DF-SANE (Derivative-free SANE).

Step 0.
Set $k \leftarrow 0$.

Step 1.
- Choose $\sigma_k$ such that $|\sigma_k| \in [\sigma_{\min}, \sigma_{\max}]$ (the spectral coefficient).
- Compute $\bar{f}_k = \max \{f(x_k), \ldots, f(x_{\max\{0,k-M+1\}})\}$.
- Set $d \leftarrow -\sigma_k F(x_k)$.
- Set $\alpha_+ \leftarrow 1$, $\alpha_- \leftarrow 1$.

Step 2.
If $f(x_k + \alpha_+ d) \leq \bar{f}_k + \eta_k - \gamma \alpha_+^2 f(x_k)$, then
Define $d_k = d$, $\alpha_k = \alpha_+$, $x_{k+1} = x_k + \alpha_k d_k$
else if $f(x_k - \alpha_- d) \leq \bar{f}_k + \eta_k - \gamma \alpha_-^2 f(x_k)$, then
Define $d_k = -d$, $\alpha_k = \alpha_-$, $x_{k+1} = x_k + \alpha_k d_k$
else
choose $\alpha_{\text{new}} \in [\tau_{\min} \alpha_+, \tau_{\max} \alpha_+], \alpha_{\text{new}} \in [\tau_{\min} \alpha_-, \tau_{\max} \alpha_-],$
replace $\alpha_+ \leftarrow \alpha_{\text{new}}$, $\alpha_- \leftarrow \alpha_{\text{new}}$
and go to Step 2.

Step 3.
If $F(x_{k+1}) = 0$, terminate the execution of the algorithm. Else, set $k \leftarrow k + 1$
and go to Step 1.

Remark. As we will see later, the coefficient $\sigma_k$ will be chosen to be an approximation
of the quotient $\| F(x_k) \|^2 \langle J(x_k) F(x_k), F(x_k) \rangle$. This quotient may be positive or negative
(or even null).

Proposition 1. The iteration is well defined.

Proof. Since $\eta_k > 0$, after a finite number of reductions of $\alpha_+$ the condition
\[ f(x_k + \alpha_+ d) \leq \bar{f}_k + \eta_k - \gamma \alpha_+^2 f(x_k) \]
necessarily holds. \qed

In the rest of this section we will prove several convergence results:

1. There exists an infinite sequence of indices $K \subset \mathbb{N}$ such that at every limit point
of the subsequence $\{x_k\}_{k \in K}$, the gradient of $\|F(x)\|^2$ is orthogonal to
the residual $F(x)$. Therefore, if $\|F(x)\|$ has bounded level sets, there exists
a limit point $x_*$ of $\{x_k\}_{k \in \mathbb{N}}$ such that
\[ \langle J(x_*) F(x_*), F(x_*) \rangle = 0. \]
2. If some limit point of $\{x_k\}_{k \in \mathbb{N}}$ is a solution of (1), then every limit point is
a solution.
3. If a limit point $x_*$ of $\{x_k\}_{k \in \mathbb{N}}$ is an isolated solution, then the whole sequence
converges to $x_*$.
4. If the initial point $x_0$ is close enough to some strongly isolated solution $x_*$, then the whole sequence converges to $x_*$. 
Only the first result was proved for the original SANE algorithm [17], although it is easy to realize that the other ones hold for SANE as well. When breakdown does not occur, under standard assumptions, Newton–Krylov methods have stronger local convergence results in the sense that linear or, sometimes, superlinear convergence can be proved.

Before we proceed with the convergence analysis, we need some preliminary definitions (see [5]). Define

\[ V_0 = f(x_0) \]

and

\[ V_k = \max\{f(x_{(k-1)M+1}), \ldots, f(x_{kM})\} \]

for all \( k = 1, 2, \ldots \).

Let \( \nu(k) \in \{(k-1)M + 1, \ldots, kM\} \) be such that, for all \( k = 1, 2, \ldots \),

\[ f(x_{\nu(k)}) = V_k. \]

Clearly,

\[ f(x_{kM+1}) \leq \max\{V_k, f(x_{kM+1})\} + \eta_{kM} - \gamma \alpha_{kM}^2 f(x_{kM}) \]

and so on.

Therefore, by an inductive argument,

\[ f(x_{kM+\ell}) \leq V_k + \sum_{j=0}^{\ell-1} \eta_{kM+j} - \gamma \alpha_{kM+\ell-1}^2 f(x_{kM+\ell-1}) \]

for all \( \ell = 0, 1, 2, \ldots, M \). Moreover,

\[ f(x_{kM+\ell}) \leq V_k + \sum_{j=0}^{\ell-1} \eta_{kM+j} \forall k, \ell \in \mathbb{N}. \]

But \( \nu(k+1) \in \{kM+1, \ldots, kM+M\} \), thus, by (4),

\[ V_{k+1} = f(x_{\nu(k+1)}) \leq V_k + \sum_{j=0}^{M-1} \eta_{kM+j} - \gamma \alpha_{\nu(k+1)-1}^2 f(x_{\nu(k+1)-1}) \]

(6)

\[ = f(x_{\nu(k)}) + \sum_{j=0}^{M-1} \eta_{kM+j} - \gamma \alpha_{\nu(k)-1}^2 f(x_{\nu(k)-1}). \]

Using (4), (5) and (6) we can prove the following propositions:

**Proposition 2.** For all \( k, \ell \in \mathbb{N} \),

\[ f(x_{kM+\ell}) \leq f(x_{\nu(k)}) + \sum_{i=\nu(k)}^{\infty} \eta_i \leq f(x_{\nu(k)}) + \eta. \]

**Proof.** Straightforward, using (5). \( \square \)

**Proposition 3.**

\[ \lim_{k \to \infty} \alpha_{\nu(k)-1}^2 f(x_{\nu(k)-1}) = 0. \]
Proof. Write the inequalities (6) for \( k = 1, 2, \ldots, L \). Observe that \( f(x_{\nu(k)}) \) occurs on the left-hand side of the \( k \)th inequality and also on the right-hand side of the \( k + 1 \)st inequality. Adding the \( L \) inequalities, we get

\[
f(x_{\nu(L+1)}) \leq f(x_{\nu(1)}) + \sum_{j=M}^{(L+1)M-1} \eta_j - \gamma \sum_{j=1}^L \alpha_{\nu(j+1)-1}^2 f(x_{\nu(j+1)-1}).
\]

Therefore, for all \( L = 1, 2, \ldots, \), we obtain

\[
\gamma \sum_{j=1}^L \alpha_{\nu(j+1)-1}^2 f(x_{\nu(j+1)-1}) \leq f(x_{\nu(1)}) + \sum_{j=M}^{(L+1)M-1} \eta_j - f(x_{\nu(L+1)}) \\
\leq f(x_{\nu(1)}) + \sum_{j=M}^{(L+1)M-1} \eta_j \\
\leq f(x_{\nu(1)}) + \eta.
\]

So, the series \( \sum_{j=1}^\infty \alpha_{\nu(j+1)-1}^2 f(x_{\nu(j+1)-1}) \) is convergent. This implies the desired result. \( \square \)

From now on we define
\[
K = \{ \nu(1) - 1, \nu(2) - 1, \nu(3) - 1, \ldots \}
\]
and
\[
K_+ = \{ \nu(1), \nu(2), \nu(3), \ldots \}.
\]
Observe that
\[
\nu(j + 1) \leq \nu(j) + 2M - 1 \text{ for all } j = 1, 2, \ldots.
\]

In Theorem 1 we prove that, at every limit point \( x_* \) of the subsequence \( \{ x_k \}_{k \in K} \) one necessarily has that \( \langle J(x_*) F(x_*), F(x_*) \rangle = \langle F(x_*), g(x_*) \rangle = 0 \). In other words the gradient of \( \| F(x) \|^2 \) at \( x_* \) is orthogonal to the residual \( F(x_*) \).

Theorem 1. Assume that \( \{ x_k \}_{k \in \mathbb{N}} \) is generated by the DF-SANE algorithm. Then, every limit point \( x_* \) of \( \{ x_k \}_{k \in K} \) satisfies
\[
\langle F(x_*), J(x_*)^T F(x_*) \rangle = 0.
\]

Proof. By Proposition 3 we have that
\[
\lim_{k \in K} \alpha_k^2 f(x_k) = 0.
\]
Let \( x_* \) be a limit point of \( \{ x_k \}_{k \in K} \). Let \( K_1 \subset K \) be an infinite sequence of indices such that
\[
\lim_{k \in K_1} x_k = x_*.
\]
Then, by (12),
\[
\lim_{k \in K_1} \alpha_k^2 f(x_k) = 0.
\]

If \( \{ \alpha_k \}_{k \in K_1} \) does not tend to zero, there exists an infinite sequence of indices \( K_2 \subset K_1 \) such that \( \alpha_k \) is bounded away from zero for \( k \in K_2 \). Then, by (13),
\[
\lim_{k \in K_2} f(x_k) = 0.
\]
Since \( f \) is continuous and \( \lim_{k \in K_2} x_k = x_* \), this implies that \( f(x_*) = 0 \). Thus \( F(x_*) = 0 \), and (11) holds.

So, we only need to analyze the case
\[
\lim_{k \in K_1} \alpha_k = 0.
\]
At Step 2 of Algorithm DF-SANE one tests the inequality
\[ f(x_k + \alpha_d) \leq \tilde{f}_k + \eta_k - \gamma \alpha^2 f(x_k). \]  
If (15) does not hold, the inequality
\[ f(x_k - \alpha_d) \leq \tilde{f}_k + \eta_k - \gamma \alpha^2 f(x_k) \]
is tested.

The first trial steps at (15)–(16) are $\alpha_+ = \alpha_- = 1$. By (14), there exists $k_0 \in K_1$ such that $\alpha_k < 1$ for all $k \geq k_0, k \in K_1$. For those iterations of the DF-SANE algorithm, the line-search was not immediately successful and $\alpha_+$ and $\alpha_-$ were adapted at least once. Suppose that in DF-SANE step $k$ (i.e., the step which generates $x_{k+1}$) $\alpha_+$ and $\alpha_-$ were adapted $m_k$ times in the line search process (i.e., the inequalities (15) and (16) were both violated together $m_k$ times). Let $\alpha^+_k$ and $\alpha^-_k$ be the values of $\alpha_+$ and $\alpha_-$ respectively in the last unsuccessful line search step (the last step at which (15) and (16) were violated together) in DF-SANE step $k$. Because of the choice of $\alpha_{\text{new}}$ and $\alpha_{\text{new}}$ at Step 2 of the DF-SANE algorithm we have that

\[ \alpha_k \geq \tau^{m_k} \]

for all $k \geq k_0, k \in K_1$ and so, by (14)

\[ \lim_{k \in K_1} m_k = \infty. \]

But, again by the choice of $\alpha_{\text{new}}$ and $\alpha_{\text{new}}$,

\[ \alpha_k^+ \leq \tau^{m_k-1} \]

and

\[ \alpha_k^- \leq \tau^{m_k-1}. \]

Therefore, since $\tau_{\max} < 1$,

\[ \lim_{k \in K_1} \alpha_k^+ = \lim_{k \in K_1} \alpha_k^- = 0. \]

Clearly, since $d = -\sigma_k F(x_k)$, the fact that (15) and (16) are not satisfied by $\alpha_k^+$ and $\alpha_k^-$ respectively implies that

\[ f(x_k - \alpha_k^+ \sigma_k F(x_k)) > \tilde{f}_k + \eta_k - \gamma (\alpha_k^+)^2 f(x_k) \]

and

\[ f(x_k + \alpha_k^- \sigma_k F(x_k)) > \tilde{f}_k + \eta_k - \gamma (\alpha_k^-)^2 f(x_k) \]

for all $k \in K_1, k \geq k_0$.

The inequality (17) implies that

\[ f(x_k - \alpha_k^- \sigma_k F(x_k)) > f(x_k) - \gamma (\alpha_k^+)^2 f(x_k). \]

So,

\[ f(x_k - \alpha_k^+ \sigma_k F(x_k)) - f(x_k) \geq -\gamma (\alpha_k^+)^2 f(x_k). \]

By Proposition 2, $f(x_k) \leq c \equiv f(x_0) + \eta$ for all $k \in \mathbb{N}$. Thus,

\[ f(x_k - \alpha_k^+ \sigma_k F(x_k)) - f(x_k) \geq -c \gamma (\alpha_k^+)^2. \]

Let us first consider the case $n e x p = 2$. Then, by (19),

\[ \| F(x_k - \alpha_k^+ \sigma_k F(x_k)) \|^2 - \| F(x_k) \|^2 \geq -c \gamma (\alpha_k^+)^2. \]
Now consider the case \( n_{\text{exp}} = 1 \). The subsequence \( \{x_k\}_{k \in K_1} \) is convergent and, therefore, bounded. Since \( \|F(x_k)\| \), \( \alpha_k^+ \) and \( \sigma_k \) are also bounded, we have that \( \{x_k - \alpha_k^+ \sigma_k F(x_k)\}_{k \in K_1} \) is bounded. So, by the continuity of \( F \), there exists \( c_1 > 0 \) such that

\[
\|F(x_k - \alpha_k^+ \sigma_k F(x_k))\| + \|F(x_k)\| \leq c_1 \text{ for all } k \in K_1.
\]

Multiplying both sides of (19) by \( \|F(x_k - \alpha_k^+ \sigma_k F(x_k))\| + \|F(x_k)\| \), we obtain that

\[
\|F(x_k - \alpha_k^+ \sigma_k F(x_k))\|^2 - \|F(x_k)\|^2 \geq -cc_1\gamma(\alpha_k^+)^2.
\]

Setting \( C = c \) if \( n_{\text{exp}} = 2 \) and \( C = cc_1 \) if \( n_{\text{exp}} = 1 \), we obtain that

\[
\|F(x_k - \alpha_k^+ \sigma_k F(x_k))\|^2 - \|F(x_k)\|^2 \geq -C\gamma(\alpha_k^+)^2.
\]

So,

\[
\frac{\|F(x_k - \alpha_k^+ \sigma_k F(x_k))\|^2 - \|F(x_k)\|^2}{\alpha_k^+} \geq -C\gamma\alpha_k^+.
\]

By the Mean Value Theorem, there exists \( \xi_k \in [0, 1] \) such that

\[
\langle g(x_k - \xi_k \alpha_k^+ \sigma_k F(x_k)), -\sigma_k F(x_k) \rangle \geq -C\gamma\alpha_k^+.
\]

Therefore,

\[
\sigma_k \langle g(x_k - \xi_k \alpha_k^+ \sigma_k F(x_k)), -F(x_k) \rangle \geq -C\gamma\alpha_k^+.
\]

By the definition of the algorithm we have that \( \sigma_k > 0 \) for infinitely many indices or \( \sigma_k < 0 \) for infinitely many indices. If \( \sigma_k > 0 \) for infinitely many indices \( k \in K_2 \subset K_1 \), the inequality (20) implies that, for \( k \in K_2, k \geq k_0 \),

\[
\langle g(x_k - \xi_k \alpha_k^+ \sigma_k F(x_k)), F(x_k) \rangle \leq \frac{C\gamma\alpha_k^+}{\sigma_k} \leq \frac{C\gamma\alpha_k^+}{\sigma_{\text{min}}}.
\]

Using (18) and proceeding in the same way, we obtain that, for \( k \in K_2, k \geq k_0 \),

\[
\langle g(x_k + \xi_k \alpha_k^- \sigma_k F(x_k)), F(x_k) \rangle \geq -\frac{C\gamma\alpha_k^-}{\sigma_k} \geq -\frac{C\gamma\alpha_k^-}{\sigma_{\text{min}}}
\]

for some \( \xi_k^\prime \in [0, 1] \).

Since \( \alpha_k^+ \rightarrow 0 \), \( \alpha_k^- \rightarrow 0 \), and \( \|\sigma_k F(x_k)\| \) is bounded, taking limits in (21) and (22), we obtain that

\[
\langle g(x_\ast), F(x_\ast) \rangle = 0.
\]

If \( \sigma_k < 0 \) for infinitely many indices, proceeding in an analogous way, we also deduce (23). Thus, the thesis is proved. \( \square \)

**Corollary 1.** Assume that \( \{x_k\}_{k \in \mathbb{N}} \) is generated by the DF-SANE algorithm, \( x_\ast \) is a limit point of \( \{x_k\}_{k \in K} \) and for all \( v \in \mathbb{R}^n, v \neq 0 \),

\[
\langle J(x_\ast)v, v \rangle \neq 0.
\]

Then, \( F(x_\ast) = 0 \).

**Proof.** Straightforward, using Theorem 1. \( \square \)
As usual, we say that a matrix $A \in \mathbb{R}^{n \times n}$ is positive-definite if $\langle Av, v \rangle > 0$ for all $v \in \mathbb{R}^n, v \neq 0$. If $F(x)$ is positive-definite for all $x \in \mathbb{R}^n$, we say that the mapping $F$ is strictly monotone. If $F$ is strictly monotone or $-F$ is strictly monotone, we say that the mapping $F$ is strict. If a mapping is strict and admits a solution, its solution must be unique (see [22], Chapter 5).

**Corollary 2.** Assume that $\{x_k\}_{k \in \mathbb{N}}$ is generated by the DF-SANE algorithm and the mapping $F$ is strict. Then, every bounded subsequence of $\{x_k\}_{k \in \mathcal{K}}$ converges to the solution of (1).

**Proof.** Straightforward, using Corollary [1].

**Corollary 3.** Assume that $\{x_k\}_{k \in \mathbb{N}}$ is generated by the DF-SANE algorithm, the mapping $F$ is strict and the level set $\{x \in \mathbb{R}^n | f(x) \leq f(x_0) + \eta\}$ is bounded. Then, $\{x_k\}_{k \in \mathcal{K}}$ converges to the solution of (1).

**Proof.** Straightforward, using Corollary 2.

So far, we proved that at every limit point of $\{x_k\}_{k \in \mathcal{K}}$ the gradient of $\|F(x)\|^2$ is orthogonal to the residual $F(x_\ast)$. The case in which there exists a limit point of $\{x_k\}_{k \in \mathbb{N}}$ at which $F(x_\ast) = 0$ deserves further analysis. The theorem below shows that, when such a limit point exists, all the limit points of the sequence generated by the algorithm are solutions of the nonlinear system.

**Theorem 2.** Assume that the sequence $\{x_k\}_{k \in \mathbb{N}}$ is generated by the DF-SANE Algorithm and that there exists a limit point $x_\ast$ of $\{x_k\}_{k \in \mathbb{N}}$ such that $F(x_\ast) = 0$. Then

$$\lim_{k \to \infty} F(x_k) = 0.$$  

Consequently, $F(x)$ vanishes at every limit point of $\{x_k\}_{k \in \mathbb{N}}$.

**Proof.** Let $K_1$ be an infinite subset of $\mathbb{N}$ such that

$$\lim_{k \in K_1} x_k = x_\ast$$

and

$$(24) \quad F(x_\ast) = 0.$$  

Then,

$$\lim_{k \in K_1} F(x_k) = 0.$$  

Therefore, since $x_{k+1} = x_k \pm \alpha_k \sigma_k F(x_k)$ and $|\alpha_k \sigma_k| \leq \sigma_{\text{max}}$ for all $k \in \mathbb{N}$,

$$\lim_{k \in K_1} \|x_{k+1} - x_k\| = 0.$$  

So,

$$\lim_{k \in K_1} x_{k+1} = x_\ast.$$  

Proceeding by induction, we may prove that for all fixed $\ell \in \{0, 1, 2, \ldots, 2M - 1\}$,

$$(25) \quad \lim_{k \in K_1} x_{k+\ell} = x_\ast.$$  

Now, by (10), for all $k \in K_1$, we can choose $\mu(k) \in \{0, 1, \ldots, 2M - 1\}$ such that

$$(26) \quad k + \mu(k) \in K_+.$$
Moreover, at least one value of $\mu(k)$ must be repeated infinitely many times. So, there exists $\ell_0 \in \{0, 1, \ldots, 2M - 1\}$ such that $\mu(k) = \ell_0$ for infinitely many indices $k \in K_1$. Consider

$$K_2 = \{k + \mu(k) \mid k \in K_1 \text{ and } \mu(k) = \ell_0\}.$$ 

By (25) and (26) we have that $K_2 \subset K_+ \cap K_1$. By (25) and (26) we have that $K_2 \subset K_+ \cap K_1$. Then, by (24),

$$\lim_{k \in K_2} x_k = x_*.$$

Then, by (24),

$$\lim_{k \in K_2} F(x_k) = 0.$$ 

Since $K_2 \subset K_+$, there exists an infinite subsequence of indices $J_1$ such that

$$\lim_{j \in J_1} x_{\nu(j)} = x_*$$

and

$$\lim_{j \in J_1} f(x_{\nu(j)}) = \lim_{j \in J_1} V_j = 0.$$ 

Let us write $J_1 = \{j_1, j_2, j_3, \ldots\}$, where $j_1 < j_2 < j_3 < \cdots$ and $\lim_{i \to \infty} j_i = \infty$. By (25), we have that

$$\lim_{i \to \infty} V_{j_i} = 0.$$ 

Now, by (6) we have that for all $j \in \mathbb{N}, j > j_i$,

$$V_j \leq V_{j_i} + \sum_{\ell=M_{j_i}}^{\infty} \eta_\ell.$$

Therefore,

$$\sup_{j \geq j_i} V_j \leq V_{j_i} + \sum_{\ell=M_{j_i}}^{\infty} \eta_\ell.$$ 

By the summability of $\eta_k$,

$$\lim_{i \to \infty} \sum_{\ell=M_{j_i}}^{\infty} \eta_\ell = 0.$$ 

Then, by (29), taking limits on both sides of (30), we get

$$\lim_{i \to \infty} \sup_{j \geq j_i} V_j = 0.$$ 

Thus,

$$\lim_{j \to \infty} V_j = 0.$$ 

By the definition of $V_j$, this implies that

$$\lim_{k \to \infty} \|F(x_k)\| = \lim_{k \to \infty} f(x_k) = 0,$$

as we wanted to prove.

The second part of the proof is straightforward: if $\bar{x}$ is a limit point of $\{x_k\}_{k \in \mathbb{N}}$ there exists a subsequence $\{x_k\}_{k \in K_3}$ that converges to $\bar{x}$. By (31) and the continuity of $F$ we have that

$$F(\bar{x}) = \lim_{k \in K_3} F(x_k) = 0.$$ 

This completes the proof. □
Now we prove two theorems of local convergence type. Theorem 3 says that if an isolated solution is a limit point of \( \{x_k\} \), then the whole sequence \( x_k \) converges to this solution.

**Theorem 3.** Assume that the sequence \( \{x_k\} \) is generated by the DF-SANE algorithm and that there exists a limit point \( x^* \) of the sequence \( \{x_k\} \) such that \( F(x^*) = 0 \). Moreover, assume that there exists \( \delta > 0 \) such that \( F(x) \neq 0 \) whenever \( 0 < \|x - x^*\| \leq \delta \). Then, \( \lim_{k \to \infty} x_k = x^* \).

**Proof.** By Theorem 2 we have that

\[
\lim_{k \to \infty} F(x_k) = 0.
\]

Therefore, since \( \alpha_k \) and \( \sigma_k \) are bounded,

\[
\lim_{k \to \infty} \|x_{k+1} - x_k\| = 0.
\]

Thus, there exists \( k_1 \in \mathbb{N} \) such that

\[
\|x_{k+1} - x_k\| \leq \delta/2 \quad \text{for all} \quad k \geq k_1.
\]  

(32)

Consider the set

\[
S = \{x \in \mathbb{R}^n \mid \delta/2 \leq \|x - x^*\| \leq \delta\}.
\]

By hypothesis, \( S \) does not contain any solution of \( F(x) = 0 \). But, by Theorem 2, all the limit points of \( \{x_k\} \) are solutions of (1). Therefore, \( S \) does not contain any limit point of \( \{x_k\} \). Thus, since \( S \) is compact, it cannot contain infinitely many iterates \( x_k \). This implies that there exists \( k_2 \in \mathbb{N} \) such that

\[
x_k \notin S \quad \text{for all} \quad k \geq k_2.
\]  

(33)

Let \( k_3 \geq \max\{k_1, k_2\} \) be such that

\[
\|x_{k+1} - x^*\| \leq \delta/2.
\]  

By (32), we have

\[
\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \|x_{k+1} - x_k\| \leq \delta.
\]

But, by (33), \( x_{k+1} \notin S \), therefore, we have that

\[
\|x_{k+1} - x^*\| \leq \delta/2.
\]

Continuing this argument inductively we have that

\[
\|x_k - x^*\| \leq \delta/2 \quad \text{for all} \quad k \geq k_3.
\]  

(34)

This implies that all the limit points \( x^* \) of the sequence \( \{x_k\} \) are such that

\[
\|x^* - x^*\| \leq \delta/2.
\]

By Theorem 2 \( F(\bar{x}) = 0 \) at every limit point \( \bar{x} \) and, by the hypothesis of this theorem, the set defined by \( 0 < \|x - x^*\| \leq \delta/2 \) does not contain solutions of (1). Therefore, this set does not contain limit points. So, the only limit point \( \bar{x} \) that satisfies \( \|x - x^*\| \leq \delta/2 \) is \( x^* \). So, by (34), the sequence converges to \( x^* \), as we wanted to prove. \(\square\)
Theorem 3 is our second local convergence theorem. For proving it we need a new definition and a technical lemma. We say that $x_* \in \mathbb{R}^n$ is a strongly isolated solution of (1) if $F(x_*) = 0$ and there exists $\varepsilon > 0$ such that 

$$0 < \|x - x_*\| \leq \varepsilon \Rightarrow \langle J(x)F(x), F(x) \rangle \neq 0.$$ 

That is, in a reduced neighborhood of a strongly isolated solution the residual $F(x)$ is not orthogonal to the gradient $J(x)F(x)$. Theorem 3 says that if the initial point $x_0$ is close enough to a strongly isolated solution $x_*$, then the sequence $\{x_k\}$ converges to $x_*$. Observe that this cannot be deduced from Theorem 3 and, moreover, Theorem 3 cannot be deduced from this result either, since the strong isolation assumption is not necessary to prove that theorem.

Lemma 1. Assume that $F(x_*) = 0$, $k \in \mathbb{N}$ and $\delta > 0$. Then, there exists $\delta(k) \in (0, \delta]$ such that for any possible initial point $x_0$ such that $\|x_0 - x_*\| < \delta(k)$, the $k$th iterate computed by the DF-SANE algorithm will satisfy 

$$\|x_k - x_*\| < \delta.$$ 

(The value of $\delta(k)$ does not depend on the particular choice of $\sigma_k$ at Step 1 or on the choices of $\alpha_{+\text{ new}}, \alpha_{-\text{ new}}$ at Step 2.)

Proof. We proceed by induction. If $k = 0$ the result is trivial with $\delta(\hat{k}) = \delta$. Assume that it is true for $k = 0, 1, \ldots, \hat{k}$, and let us prove it for $\hat{k} + 1$. Let $\hat{\delta} > 0$. Observe that, by the definition of the algorithm, 

$$\|x_{\hat{k} + 1} - x_\hat{k}\| \leq \sigma_{\text{max}}\|F(x_{\hat{k}})\|.$$ 

Since $F$ is continuous and $F(x_*) = 0$, there exists 

$$\delta' \in (0, \hat{\delta}/2)$$ 

such that 

$$\|x - x_*\| < \delta' \Rightarrow \|F(x)\| < \hat{\delta}/2\sigma_{\text{max}}.$$ 

So, by (35), 

$$\|x_{\hat{k} + 1} - x_*\| < \delta' \Rightarrow \|x_{\hat{k} + 1} - x_{\hat{k}}\| < \hat{\delta}/2.$$ 

But, by the inductive hypothesis, there exists $\delta \in (0, \delta']$ such that 

$$\|x_{\hat{k} + 1} - x_{\hat{k}}\| < \delta \Rightarrow \|x_{\hat{k} + 1} - x_*\| < \delta'.$$ 

By (36), (37) and (38), if $\|x_{\hat{k} + 1} - x_*\| < \delta$, we have 

$$\|x_{\hat{k} + 1} - x_*\| \leq \|x_{\hat{k}} - x_*\| + \|x_{\hat{k} + 1} - x_{\hat{k}}\| < \delta' + \hat{\delta}/2 < \delta.$$ 

This completes the proof. 

Theorem 4. Assume that $x_*$ is a strongly isolated solution of $F(x) = 0$. Then, there exists $\delta > 0$ such that 

$$\|x_0 - x_*\| < \delta \Rightarrow \lim_{k \to \infty} x_k = x_*,$$ 

where $\{x_k\}_{k \in \mathbb{N}}$ is the sequence generated by the DF-SANE algorithm for the starting point $x_0$. 


Proof. Let \( \varepsilon > 0 \) be such that
\[
0 < \|x - x_*\| \leq \varepsilon \Rightarrow \langle J(x)F(x), F(x) \rangle \neq 0.
\]
Since \( |\sigma_i \alpha_k| \leq \sigma_{\text{max}} \), \( \|x_{k+1} - x_k\| \leq |\sigma_i \alpha_k|\|F(x_k)\| \) and \( F(x_*) = 0 \), the continuity of \( F \) implies that there exists
\[
\varepsilon_1 \in (0, \varepsilon/2]
\]
such that
\[
\|x_k - x_*\| \leq \varepsilon_1 \Rightarrow \|F(x_k) - F(x_*)\| = \|F(x_k)\| \leq \frac{\varepsilon}{2\sigma_{\text{max}}} \Rightarrow \|x_{k+1} - x_k\| < \varepsilon/2.
\]
Define
\[
C_\varepsilon = \{x \in \mathbb{R}^n \mid \varepsilon_1 \leq \|x - x_*\| \leq \varepsilon\}.
\]
Since \( C_\varepsilon \) is compact and \( f \) is continuous, \( f \) has a minimum in \( C_\varepsilon \). So, there exists \( \hat{x} \in C_\varepsilon \), \( \beta > 0 \), such that
\[
\beta \equiv f(\hat{x}) \leq f(x) \quad \text{for all} \quad x \in C_\varepsilon.
\]
Since \( f \) is continuous, the set \( \{x \in B(x_*, \varepsilon_1/2) \mid f(x) < \beta/2\} \) is an open neighborhood of \( x_* \). Therefore, there exists \( \delta \in (0, \varepsilon_1/2) \) such that
\[
\|x - x_*\| < \delta \Rightarrow f(x) < \beta/2.
\]
Let \( m \geq 1 \) be such that
\[
\sum_{i=(m-1)M+1}^{\infty} \eta_i < \beta/2.
\]
By Proposition 2 writing \( k_0 = mM \), we have that
\[
f(x_k) \leq V_m + \sum_{i=(m-1)M+1}^{\infty} \eta_i < V_m + \beta/2
\]
for all \( k \geq k_0 \).

Now, apply Lemma 1 for \( \hat{k} = (m-1)M + 1, \ldots, mM \) with \( \delta \) given by (42). It turns out that there exist \( \delta_{(m-1)M+1}, \ldots, \delta_{mM} \) such that
\[
\|x_0 - x_*\| < \delta_{(m-1)M+j} \Rightarrow \|x_{(m-1)M+j} - x_*\| \leq \delta < \varepsilon_1, \quad j = 1, 2, \ldots, M.
\]
So, taking \( \delta = \min\{\delta_{(m-1)M+1}, \ldots, \delta_{mM}\} \), we have that
\[
\|x_0 - x_*\| < \delta \Rightarrow \|x_{(m-1)M+j} - x_*\| \leq \delta < \varepsilon_1, \quad j = 1, 2, \ldots, M.
\]
In particular, since \( k_0 = mM \),
\[
\|x_0 - x_*\| < \delta \Rightarrow \|x_{k_0} - x_*\| \leq \delta < \varepsilon_1.
\]
By (42) and (44), \( f(x_{(m-1)M+j}) < \beta/2 \) and thus \( V_m < \beta/2 \). So, by (43),
\[
f(x_k) < \beta \quad \text{for all} \quad k \geq k_0.
\]
Let us prove by induction that, choosing \( \|x_0 - x_*\| < \delta \),
\[
\|x_{k_0+j} - x_*\| < \varepsilon_1
\]
for all \( j \in \mathbb{N} \). By (45), we have that (47) is true for \( j = 0 \).

Assume, as an inductive hypothesis, that, for some \( j \geq 1 \),
\[
\|x_{k_0+j-1} - x_*\| < \varepsilon_1.
\]
Since $\varepsilon_1 \leq \varepsilon/2$, then
\[ \|x_{k_0+j-1} - x_*\| < \varepsilon/2. \]

But, by (40),
\[ (48) \quad \|x_{k_0+j} - x_*\| \leq \|x_{k_0+j-1} - x_*\| + \|x_{k_0+j} - x_{k_0+j-1}\| < \varepsilon/2 + \varepsilon/2 = \varepsilon. \]

Since, by (40), $f(x_{k_0+j}) < \beta$, (41) and (48) imply that $\|x_{k_0+j} - x_*\| < \varepsilon_1$. This completes the inductive proof.

So, $\{x_k\}_{k \geq k_0} \subset B(x_*, \varepsilon_1)$. Therefore, all the limit points $\bar{x}$ of $\{x_k\}_{k \in \mathbb{N}}$ are such that $\|\bar{x} - x_*\| \leq \varepsilon_1 < \varepsilon$. But, by (39) and Theorem 1, the only possible limit point is $x_*$. Therefore, $\lim_{k \to \infty} x_k = x_*$, as we wanted to prove. \[\square\]

**Corollary 4.** Assume that $x_*$ is a solution of $F(x) = 0$ and assume that $J(x_*)$ is either positive definite or negative definite. Then, there exists $\delta > 0$ such that
\[ \|x_0 - x_*\| < \delta \Rightarrow \lim_{k \to \infty} x_k = x_* \]
with $\{x_k\}_{k \in \mathbb{N}}$ the sequence generated by the DF-SANE algorithm for the starting point $x_0$.

**Proof.** Using the continuity of $J$ we obtain that $x_*$ is strongly isolated. Then, the thesis follows from Theorem 4 \[\square\]

4. **Numerical results**

We implemented DF-SANE with the following parameters: $nexp = 2, \sigma_{\text{min}} = 10^{-10}, \sigma_{\text{max}} = 10^{10}, \sigma_0 = 1, \tau_{\text{min}} = 0.1, \tau_{\text{max}} = 0.5, \gamma = 10^{-4}, M = 10, \eta_k = \|F(x_0)/(1+k)^2\|$ for all $k \in \mathbb{N}$.

The spectral steplength was computed by the formula
\[ \sigma_k = \frac{\langle s_k, s_k \rangle}{\langle s_k, y_k \rangle}, \]
where $s_k = x_{k+1} - x_k$ and $y_k = F(x_{k+1}) - F(x_k)$. Observe that
\[ y_k = \int_0^1 J(x_k + ts_k)dt|s_k, \]
so $\sigma_k$ is the inverse of the Rayleigh quotient
\[ \frac{\langle \int_0^1 J(x_k + ts_k)dt|s_k, s_k \rangle}{\langle s_k, s_k \rangle}. \]

However, if $|\sigma_k| \notin [\sigma_{\text{min}}, \sigma_{\text{max}}]$, we replace the spectral coefficient by
\[ \sigma_k = \begin{cases} 
1 & \text{if } \|F(x_k)\| > 1, \\
\|F(x_k)\|^{-1} & \text{if } 10^{-5} \leq \|F(x_k)\| \leq 1, \\
10^5 & \text{if } \|F(x_k)\| < 10^{-5}.
\end{cases} \]

Since we use big values for $\sigma_{\text{max}}$ and $1/\sigma_{\text{min}}$, this replacement rarely occurs. In the few cases in which the replacement is necessary, the first trial point is $x_k - F(x_k)$ if $\|F(x_k)\| \geq 1$. If $10^{-5} \leq \|F(x_k)\| < 1$ the step $\sigma_k$ is such that the distance between $x_k$ and the first trial point is equal to 1. When $\|F(x_k)\| < 10^{-5}$ we prefer to allow the distance between $x_k$ and the trial point to be smaller, choosing for $\sigma_k$ the fixed value $10^{-5}$. 

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
For choosing $\alpha_+^\text{new}$ and $\alpha_-^\text{new}$ at Step 2, we proceed as follows. Given $\alpha_+ > 0$, we take $\alpha_+^\text{new} > 0$ as

$$\alpha_+^\text{new} = \begin{cases} 
\tau_\text{min}\alpha_+ & \text{if } \alpha_t < \tau_\text{min}\alpha_+, \\
\tau_\text{max}\alpha_+ & \text{if } \alpha_t > \tau_\text{max}\alpha_+, \\
\alpha_t & \text{otherwise}, 
\end{cases}$$

where

$$\alpha_t = \frac{\alpha^2 f(x_k)}{f(x_k + \alpha d) + (2\alpha - 1)f(x_k)}.$$  

We use similar formulae for choosing $\alpha_-^\text{new}$ as a function of $\alpha_-$, $f(x_k)$ and $f(x_k - \alpha d)$. This parabolic model is similar to the one described in [15, pp. 142–143], in which the Jacobian at $x_k$ is replaced by the identity matrix (see also [9]).

We also implemented SANE [17] with the following parameters: $\gamma = 10^{-4}$, $\varepsilon = 10^{-8}$, $\sigma_1 = 0.1$, $\sigma_2 = 0.5$, $\alpha_0 = 1$, $M = 10$, and

$$\delta = \begin{cases} 
1 & \text{if } \|F(x_k)\| > 1, \\
\|F(x_k)\|^{-1} & \text{if } 10^{-5} \leq \|F(x_k)\| \leq 1, \\
10^5 & \text{if } \|F(x_k)\| < 10^{-5}.
\end{cases}$$

Both in SANE and DF-SANE we stop the process when

$$\frac{\|F(x_k)\|}{\sqrt{n}} \leq e_a + e_r \frac{\|F(x_0)\|}{\sqrt{n}},$$

where $e_a = 10^{-5}$ and $e_r = 10^{-4}$.

We ran SANE and DF-SANE using a set of large-scale test problems. The first twenty (1–20) test problems are fully described in [17] or references therein. The complete set of test problems is described in our expanded report [16].

The numerical results are shown in Tables 1 and 2. We report only one failure, denoted by the symbol (*), when running problem 18 with $n = 100$. In that case, DF-SANE fails because it generates a sequence that converges to a point $\bar{x}$ at which $F(\bar{x})^T g(\bar{x}) = 0$, but $F(\bar{x}) \neq 0$ and $g(\bar{x}) \neq 0$.

The results from Tables 1 and 2 are summarized in Table 3. In Table 3 we compare the performance (the number of problems for which each method is a winner with respect to the number of iterations, function evaluations and computer time) between SANE and DF-SANE. In Tables 1 and 2 we report the problem number and the dimension of the problem (Function(n)), the number of iterations (IT), the number of function evaluations (including the additional functional evaluations that SANE uses for approximating directional derivatives) (FE), the number of backtrackings (BK), and the CPU time in seconds (T). In SANE it is necessary to evaluate the directional derivative $\langle F(x_k), J(x_k)^T F(x_k) \rangle$ at each iteration. Since we assume that the Jacobian is not easily available, we use the fact that

$$\langle F(x_k), J(x_k)^T F(x_k) \rangle = \langle J(x_k)F(x_k), F(x_k) \rangle$$

and the approximation

$$J(x_k)F(x_k) \approx \frac{F(x_k + tF(x_k)) - F(x_k)}{t},$$

where $t > 0$ is a small parameter. Therefore, computing the approximate directional derivative involves an additional function evaluation (included in FE) at each iteration.
Table 1. SANE vs. DF-SANE for the first set of test problems.

<table>
<thead>
<tr>
<th>Function(n)</th>
<th>SANE</th>
<th>DF-SANE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(1000)</td>
<td>5 10 0 .010</td>
<td>5 5 0 .000</td>
</tr>
<tr>
<td>1(10000)</td>
<td>2 4 0 .060</td>
<td>2 2 0 .050</td>
</tr>
<tr>
<td>2(500)</td>
<td>6 14 1 .010</td>
<td>11 11 0 .000</td>
</tr>
<tr>
<td>2(2000)</td>
<td>2 7 1 .010</td>
<td>11 11 0 .030</td>
</tr>
<tr>
<td>3(100)</td>
<td>5 10 0 .010</td>
<td>5 5 0 .000</td>
</tr>
<tr>
<td>3(500)</td>
<td>1 2 0 .000</td>
<td>1 1 0 .010</td>
</tr>
<tr>
<td>4(99)</td>
<td>130 335 69 .060</td>
<td>99 289 66 .060</td>
</tr>
<tr>
<td>4(999)</td>
<td>130 335 69 .011</td>
<td>101 325 71 .011</td>
</tr>
<tr>
<td>5(9)</td>
<td>23 59 12 .000</td>
<td>42 68 12 .000</td>
</tr>
<tr>
<td>5(49)</td>
<td>552 1942 424 .070</td>
<td>732 2958 660 .130</td>
</tr>
<tr>
<td>6(100)</td>
<td>2 5 1 .000</td>
<td>3 3 0 .000</td>
</tr>
<tr>
<td>6(10000)</td>
<td>2 5 1 .040</td>
<td>3 3 0 .060</td>
</tr>
<tr>
<td>7(100)</td>
<td>23 49 2 .010</td>
<td>23 29 2 .000</td>
</tr>
<tr>
<td>7(10000)</td>
<td>23 49 2 .581</td>
<td>23 29 2 .511</td>
</tr>
<tr>
<td>8(1000)</td>
<td>1 2 0 .000</td>
<td>1 1 0 .000</td>
</tr>
<tr>
<td>8(10000)</td>
<td>1 2 0 .030</td>
<td>1 1 0 .030</td>
</tr>
<tr>
<td>9(100)</td>
<td>6 12 0 .040</td>
<td>6 6 0 .020</td>
</tr>
<tr>
<td>9(1000)</td>
<td>6 12 0 3.826</td>
<td>6 6 0 2.063</td>
</tr>
<tr>
<td>10(100)</td>
<td>1 8 1 .000</td>
<td>2 12 1 .000</td>
</tr>
<tr>
<td>10(500)</td>
<td>1 8 1 .010</td>
<td>2 12 1 .010</td>
</tr>
<tr>
<td>11(99)</td>
<td>11 34 4 .000</td>
<td>17 49 7 .000</td>
</tr>
<tr>
<td>11(399)</td>
<td>11 34 4 .020</td>
<td>17 49 7 .030</td>
</tr>
<tr>
<td>12(1000)</td>
<td>6 14 2 .040</td>
<td>30 62 12 .180</td>
</tr>
<tr>
<td>12(10000)</td>
<td>5 12 2 .421</td>
<td>23 59 11 .073</td>
</tr>
<tr>
<td>13(100)</td>
<td>3 8 1 .000</td>
<td>3 7 1 .010</td>
</tr>
<tr>
<td>13(1000)</td>
<td>4 10 1 .020</td>
<td>4 8 1 .010</td>
</tr>
<tr>
<td>14(2500)</td>
<td>11 25 1 .210</td>
<td>11 17 1 .160</td>
</tr>
<tr>
<td>14(10000)</td>
<td>12 28 1 .082</td>
<td>12 20 1 .871</td>
</tr>
<tr>
<td>15(5000)</td>
<td>5 10 0 .060</td>
<td>5 5 0 .050</td>
</tr>
<tr>
<td>15(50000)</td>
<td>5 10 0 .230</td>
<td>5 5 0 .180</td>
</tr>
<tr>
<td>16(500)</td>
<td>14 29 1 .000</td>
<td>14 16 1 .010</td>
</tr>
<tr>
<td>16(2000)</td>
<td>16 32 0 .010</td>
<td>16 16 0 .010</td>
</tr>
<tr>
<td>17(100)</td>
<td>9 19 1 .010</td>
<td>9 11 1 .000</td>
</tr>
<tr>
<td>17(1000)</td>
<td>7 15 1 .030</td>
<td>7 9 1 .030</td>
</tr>
<tr>
<td>18(50)</td>
<td>24 50 2 .010</td>
<td>19 21 1 .000</td>
</tr>
<tr>
<td>18(100)</td>
<td>24 49 1 .000</td>
<td>* * *</td>
</tr>
<tr>
<td>19(1000)</td>
<td>5 10 0 .010</td>
<td>5 5 0 .010</td>
</tr>
<tr>
<td>19(5000)</td>
<td>5 10 0 .771</td>
<td>5 5 0 .611</td>
</tr>
<tr>
<td>20(100)</td>
<td>32 67 2 .010</td>
<td>40 42 1 .010</td>
</tr>
<tr>
<td>20(1000)</td>
<td>51 117 9 .100</td>
<td>44 62 5 .070</td>
</tr>
<tr>
<td>21(399)</td>
<td>4 9 1 .010</td>
<td>5 7 1 .000</td>
</tr>
<tr>
<td>21(9999)</td>
<td>4 9 1 .200</td>
<td>5 7 1 .190</td>
</tr>
<tr>
<td>22(1000)</td>
<td>1 2 0 .000</td>
<td>1 2 0 .000</td>
</tr>
<tr>
<td>22(15000)</td>
<td>1 2 0 .030</td>
<td>1 2 0 .040</td>
</tr>
</tbody>
</table>

Our results indicate that the new fully derivative-free scheme DF-SANE is competitive with the SANE algorithm, which in turn is quite frequently preferable to the Newton-Krylov methods: Newton - GMRES, Newton - BiCGSTAB, and Newton - TFQMR (see [17] for comparisons). In particular, when comparing SANE with Newton - GMRES (which was the Krylov-like method with the best performance in [17]) the summary results shown in Table 4 were obtained.
Table 2. SANE vs. DF-SANE for the second set of test problems.

<table>
<thead>
<tr>
<th>Function(n)</th>
<th>IT</th>
<th>FE</th>
<th>BK</th>
<th>T</th>
<th>IT</th>
<th>FE</th>
<th>BK</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>23(500)</td>
<td>1</td>
<td>10</td>
<td>1</td>
<td>.000</td>
<td>2</td>
<td>18</td>
<td>1</td>
<td>.010</td>
</tr>
<tr>
<td>23(1000)</td>
<td>1</td>
<td>11</td>
<td>1</td>
<td>.000</td>
<td>2</td>
<td>20</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>24(500)</td>
<td>25</td>
<td>54</td>
<td>4</td>
<td>.300</td>
<td>54</td>
<td>109</td>
<td>18</td>
<td>.070</td>
</tr>
<tr>
<td>24(1000)</td>
<td>265</td>
<td>915</td>
<td>159</td>
<td>.951</td>
<td>17</td>
<td>25</td>
<td>3</td>
<td>.030</td>
</tr>
<tr>
<td>25(100)</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>.000</td>
<td>2</td>
<td>6</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>25(500)</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>.000</td>
<td>3</td>
<td>9</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>26(1000)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.000</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>26(100)</td>
<td>10</td>
<td>20</td>
<td>0</td>
<td>.260</td>
<td>10</td>
<td>10</td>
<td>0</td>
<td>.140</td>
</tr>
<tr>
<td>27(100)</td>
<td>11</td>
<td>22</td>
<td>0</td>
<td>1.072</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>0.561</td>
</tr>
<tr>
<td>28(100)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.000</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>28(1000)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.000</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>29(100)</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>.010</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>.000</td>
</tr>
<tr>
<td>29(1000)</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>.010</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>.010</td>
</tr>
<tr>
<td>30(99)</td>
<td>18</td>
<td>39</td>
<td>3</td>
<td>.791</td>
<td>11</td>
<td>16</td>
<td>2</td>
<td>.951</td>
</tr>
<tr>
<td>30(9999)</td>
<td>18</td>
<td>39</td>
<td>3</td>
<td>.791</td>
<td>11</td>
<td>16</td>
<td>2</td>
<td>.411</td>
</tr>
<tr>
<td>31(1000)</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>.030</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>.020</td>
</tr>
<tr>
<td>31(5000)</td>
<td>4</td>
<td>9</td>
<td>0</td>
<td>.160</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>.130</td>
</tr>
<tr>
<td>32(500)</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>.010</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>.010</td>
</tr>
<tr>
<td>32(1000)</td>
<td>6</td>
<td>12</td>
<td>0</td>
<td>.020</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>.020</td>
</tr>
<tr>
<td>33(1000)</td>
<td>3</td>
<td>20</td>
<td>2</td>
<td>.500</td>
<td>37</td>
<td>50</td>
<td>3</td>
<td>.120</td>
</tr>
<tr>
<td>33(5000)</td>
<td>3</td>
<td>22</td>
<td>1</td>
<td>.270</td>
<td>4</td>
<td>16</td>
<td>2</td>
<td>.230</td>
</tr>
<tr>
<td>34(1000)</td>
<td>22</td>
<td>52</td>
<td>4</td>
<td>.110</td>
<td>78</td>
<td>155</td>
<td>26</td>
<td>.381</td>
</tr>
<tr>
<td>34(5000)</td>
<td>12</td>
<td>27</td>
<td>1</td>
<td>.361</td>
<td>12</td>
<td>18</td>
<td>1</td>
<td>.280</td>
</tr>
<tr>
<td>35(1000)</td>
<td>21</td>
<td>45</td>
<td>2</td>
<td>.180</td>
<td>21</td>
<td>27</td>
<td>2</td>
<td>.110</td>
</tr>
<tr>
<td>35(5000)</td>
<td>29</td>
<td>63</td>
<td>3</td>
<td>1.402</td>
<td>38</td>
<td>48</td>
<td>3</td>
<td>1.202</td>
</tr>
<tr>
<td>36(1000)</td>
<td>21</td>
<td>45</td>
<td>2</td>
<td>.270</td>
<td>28</td>
<td>34</td>
<td>2</td>
<td>.210</td>
</tr>
<tr>
<td>36(5000)</td>
<td>44</td>
<td>96</td>
<td>7</td>
<td>2.954</td>
<td>26</td>
<td>36</td>
<td>4</td>
<td>1.272</td>
</tr>
<tr>
<td>37(1000)</td>
<td>23</td>
<td>49</td>
<td>2</td>
<td>.010</td>
<td>26</td>
<td>38</td>
<td>5</td>
<td>.010</td>
</tr>
<tr>
<td>37(5000)</td>
<td>23</td>
<td>49</td>
<td>2</td>
<td>.140</td>
<td>26</td>
<td>38</td>
<td>5</td>
<td>.210</td>
</tr>
<tr>
<td>38(1000)</td>
<td>19</td>
<td>40</td>
<td>2</td>
<td>.050</td>
<td>25</td>
<td>30</td>
<td>2</td>
<td>.040</td>
</tr>
<tr>
<td>38(5000)</td>
<td>19</td>
<td>40</td>
<td>2</td>
<td>.320</td>
<td>25</td>
<td>30</td>
<td>2</td>
<td>.340</td>
</tr>
<tr>
<td>39(1000)</td>
<td>55</td>
<td>126</td>
<td>13</td>
<td>.160</td>
<td>14</td>
<td>20</td>
<td>1</td>
<td>.030</td>
</tr>
<tr>
<td>39(5000)</td>
<td>55</td>
<td>126</td>
<td>13</td>
<td>1.041</td>
<td>14</td>
<td>20</td>
<td>1</td>
<td>1.210</td>
</tr>
<tr>
<td>40(1000)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.000</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>40(5000)</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>.020</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>.020</td>
</tr>
<tr>
<td>41(500)</td>
<td>7</td>
<td>15</td>
<td>1</td>
<td>.010</td>
<td>7</td>
<td>9</td>
<td>1</td>
<td>.010</td>
</tr>
<tr>
<td>41(1000)</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>.010</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>.000</td>
</tr>
<tr>
<td>42(1000)</td>
<td>110</td>
<td>268</td>
<td>45</td>
<td>.190</td>
<td>173</td>
<td>412</td>
<td>85</td>
<td>.330</td>
</tr>
<tr>
<td>42(5000)</td>
<td>110</td>
<td>268</td>
<td>45</td>
<td>1.392</td>
<td>173</td>
<td>412</td>
<td>85</td>
<td>2.654</td>
</tr>
<tr>
<td>43(100)</td>
<td>80</td>
<td>175</td>
<td>11</td>
<td>.010</td>
<td>86</td>
<td>108</td>
<td>9</td>
<td>.010</td>
</tr>
<tr>
<td>43(500)</td>
<td>488</td>
<td>1704</td>
<td>348</td>
<td>.601</td>
<td>586</td>
<td>1162</td>
<td>193</td>
<td>.451</td>
</tr>
<tr>
<td>44(1000)</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>.020</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>.030</td>
</tr>
<tr>
<td>44(5000)</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>.100</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>.090</td>
</tr>
</tbody>
</table>

Table 3. Winners with respect to iterations, evaluations and time.

<table>
<thead>
<tr>
<th>Method</th>
<th>IT</th>
<th>FE</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>SANE</td>
<td>37</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>DF-SANE</td>
<td>10</td>
<td>64</td>
<td>38</td>
</tr>
<tr>
<td>Undecided</td>
<td>41</td>
<td>5</td>
<td>30</td>
</tr>
</tbody>
</table>

License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use
Table 4. Winners with respect to iterations, evaluations and time between SANE and Newton-GMRES reported in [17].

<table>
<thead>
<tr>
<th>Method</th>
<th>IT</th>
<th>FE</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newton-GMRES</td>
<td>51</td>
<td>9</td>
<td>19</td>
</tr>
<tr>
<td>SANE</td>
<td>9</td>
<td>51</td>
<td>41</td>
</tr>
</tbody>
</table>

5. Conclusions

The algorithm presented in this paper may be considered a damped quasi-Newton method for solving nonlinear systems (see [9, 20]). The iterations are

\[ x_{k+1} = x_k - \alpha_k B_k^{-1} F(x_k), \]

where the Jacobian approximation \( B_k \) has the very simple form

\[ B_k = \sigma_k I. \]

In most cases,

\[ \sigma_k = \frac{\langle s_k, s_k \rangle}{\langle y_k, s_k \rangle}. \]

Due to the simplicity of the Jacobian approximation, the method is very easy to implement, memory requirements are minimal and, so, its use for solving large-scale nonlinear systems is attractive.

In [17] it was shown that, perhaps surprisingly, a procedure that obeys the scheme (50)–(52) behaves reasonably well for solving a number of classical nonlinear systems, most of them coming from discretization of boundary value problems. However, the algorithm introduced in [17] is not completely satisfactory in the sense that a directional derivative estimate is needed in order to ensure convergence and even well-definiteness of each iteration. In the present research we overcome that difficulty introducing the method DF-SANE, which does not need directional derivatives at all.

Our theoretical results are obtained without using the specific formula of \( \sigma_k \) employed in our experiments. However, the method does not behave well for every choice of \( \sigma_k \). Therefore, much has to be said, from the theoretical point of view, to explain the behavior of algorithms associated to the safeguarded spectral choice of the steplength used here. In particular, although the theoretical properties of the Barzilai–Borwein method for minimizing convex quadratics are now well understood (see [23]), nothing is known about the properties of the spectral residual method for solving nonsymmetric linear systems. Global and local convergence theorems, as the ones presented in this paper, smooth the path for proving results on the order of convergence. Nevertheless, it is necessary to understand what happens in the linear case first.

Since the spectral residual method is a quasi-Newton method where the Jacobian approximation is a multiple of the identity matrix, the best behavior of this method must be expected when true Jacobians are close to matrices of that type. The analogy with the Barzilai-Borwein method allows one to conjecture in which (more general) situations the method should behave well. If the Jacobians are close to symmetric matrices with clustered eigenvalues (see [21]), a good behavior of the...
Barzilai–Borwein method can be predicted, and, so, we also predict a fine behavior of the spectral residual method. Very likely, in many practical situations the performance of the method should be improved using some kind of preconditioning that transforms the Jacobian on a matrix with a small number of clusters of eigenvalues. So, with respect to preconditioning features, the situation is analogous to the one of Krylov-subspace methods. Preconditioned spectral gradient methods for minimization were introduced in [19].

Our first set of experiments are discretization of boundary value problems. In general, the Jacobians are positive definite, so that the mappings \( F \) are generally monotone, or even strictly monotone. According to the corollaries of Theorem 1 this favors the behavior of DF-SANE, but also favors the behavior of almost every nonlinear system solver. Since we are not using preconditioning at all, in general eigenvalues are not clustered. In some problems the Jacobians are well conditioned and in some other problems they are not. Moreover, in some problems the Jacobian is singular at the solution. In principle ill-conditioning adversely affects both spectral methods as Krylov subspace methods.

The second set of 22 problems does not show special characteristics from the point of view of positive definiteness or conditioning. Moreover, some of these problems have many solutions. In principle, we do not have strong reasons to predict a good behavior of DF-SANE, therefore the rather robust and efficient performance of the new algorithm for solving these problems is a pleasantly surprising fact that needs theoretical explanation.

We would like to finish pointing out that a different modification of SANE, which uses watchdog techniques and coordinate search, has been recently proposed in [14].

ACKNOWLEDGMENTS

We are grateful to Raúl Vignau for his careful reading of the first draft of this paper. We are also indebted to two anonymous referees for many suggestions which greatly improved the quality and presentation of this paper.

REFERENCES


**License or copyright restrictions may apply to redistribution; see http://www.ams.org/journal-terms-of-use**