FLORE
Repository istituzionale dell'Università degli Studi di Firenze

Approximate norm descent methods for constrained nonlinear systems

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:

Approximate norm descent methods for constrained nonlinear systems / Morini, Benedetta; Porcelli, Margherita; Toint Philippe, L.. - In: MATHEMATICS OF COMPUTATION. - ISSN 1088-6842. - STAMPA. - 87:(2018), pp. 1327-1351. [10.1090/mcom/3251]

Availability:

The webpage https://hdl.handle.net/2158/1079577 of the repository was last updated on 2018-02-21T12:36:51Z

Published version:

DOI: 10.1090/mcom/3251

Terms of use:

Open Access

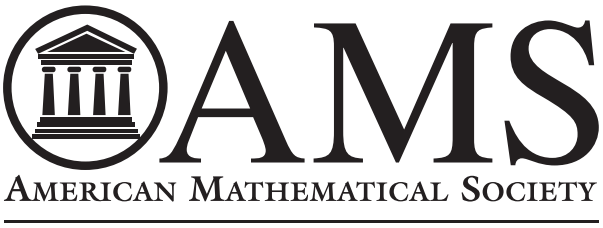
La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf)

Publisher copyright claim:

La data sopra indicata si riferisce all'ultimo aggiornamento della scheda del Repository FloRe - The above-mentioned date refers to the last update of the record in the Institutional Repository FloRe

(Article begins on next page)

23 May 2026



www.ams.org

Benedetta Morini, Margherita Porcelli, Philippe L. Toint

Approximate norm descent methods for constrained nonlinear systems

Mathematics of Computation

DOI: 10.1090/mcom/3251

Accepted Manuscript

This is a preliminary PDF of the author-produced manuscript that has been peer-reviewed and accepted for publication. It has not been copyedited, proofread, or finalized by AMS Production staff. Once the accepted manuscript has been copyedited, proofread, and finalized by AMS Production staff, the article will be published in electronic form as a "Recently Published Article" before being placed in an issue. That electronically published article will become the Version of Record.

This preliminary version is available to AMS members prior to publication of the Version of Record, and in limited cases it is also made accessible to everyone one year after the publication date of the Version of Record.

The Version of Record is accessible to everyone five years after publication in an issue.

Approximate norm descent methods for constrained nonlinear systems[‡]

Benedetta Morini[*], Margherita Porcelli[*] and Philippe L. Toint[†]

December 16, 2016

Abstract

We address the solution of convex-constrained nonlinear systems of equations where the Jacobian matrix is unavailable or its computation/storage is burdensome. In order to efficiently solve such problems, we propose a new class of algorithms which are "derivative-free" both in the computation of the search direction and in the selection of the steplength. Search directions comprise the residuals and Quasi-Newton directions while the steplength is determined by using a new linesearch strategy based on a nonmonotone approximate norm descent property of the merit function. We provide a theoretical analysis of the proposed algorithm and we discuss several conditions ensuring convergence to a solution of the constrained nonlinear system. Finally, we illustrate its numerical behaviour also in comparison with existing approaches.

Keywords: nonlinear systems of equations, bound constraints, numerical algorithms, convergence theory.
AMS Subject Classification: 65H10, 90C06, 90C56.

1 Introduction

Solving nonlinear systems of equations is an ubiquitous task in applied mathematics, and has generated considerable interest for a long time. In this paper, we focus on an important variant of this task: that of solving a nonlinear system subject to convex constraints (such as bounds). More precisely, let $F : X \to \mathbb{R}^n$ be a continuous mapping and $X \subseteq \mathbb{R}^n$ be an open set. We address the problem of finding a vector $x \in \mathbb{R}^n$ satisfying the nonlinear system with convex-constraints

$$F(x) = 0, \quad x \in \Omega, \tag{1}$$

where $\Omega \subset X$ is a convex set whose relative interior is non-empty.

[*]Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, viale G.B. Morgagni 40, 50134 Firenze, Italia. Emails: benedetta.morini@unifi.it, margherita.porcelli@unifi.it

[†]Namur Center for Complex Systems (naXys), University of Namur, 61, rue de Bruxelles, B-5000 Namur, Belgium. Email: philippe.toint@unamur.ac.be

[‡]The work of the first two authors was supported by *Gruppo Nazionale per il Calcolo Scientifico* (GNCS-INdAM) of Italy. Part of the research was conducted during a visit supported by GNCS-INdAM of the third author to the Università degli Studi di Firenze.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

The solution of problem (1) has been intensively investigated in the last years. Most of the proposed methods require the calculation of the derivatives of F and are Newton-based methods belonging to the class of affine-scaling procedures, see e.g., [3, 4, 19, 26, 29, 31]. However, such methods may become computationally expensive for medium and large scale problems, due to the evaluation cost of the Jacobian J of F, unless this matrix has structure which can be exploited. Whenever this is not the case, spectral residual methods [21, 22] and Quasi-Newton methods [6, 27] may become competitive, and implementations which do not involve derivatives at all (derivative-free algorithms) are of special interest, as exemplified by the algorithms proposed in [1, 17, 21, 23] for unconstrained problems and in [12, 20, 30] for constrained ones.

Our interest in this paper is in a class of derivative-free methods covering both spectral residual and quasi-Newton algorithms. As it turns out (and as we demonstrate in the paper), these methods can be used for relatively large problems and can be surprisingly efficient in terms of computing a solution of (1), as opposed to the easier task of computing a local minimizer of the residual

$$f(x) = \|F(x)\|_2^2. \tag{2}$$

However, it is also known that they may fail. Our objective is thus to propose an efficient algorithm which avoids some of the convergence pitfalls present in similar approaches and also to investigate conditions under which convergence to a solution of (1) can be ensured.

The algorithm developed in this paper generates feasible iterates x_k, where k is the iteration index. If F is continuous, then the residuals $\pm F(x_k)$ are used as search directions. Alternatively, if F is differentiable, search directions can be computed by using approximations B_k to the Jacobian matrices J of F at the iterates. In both cases, large savings can be obtained in the computation of the search directions compared with Newton's method. A derivative-free linesearch strategy is proposed so that, for any initial iterate, either $\|F(x_k)\|$ converges to zero or the iteration fails to do so in a small and characterized number of ways. Since the solutions of problem (1) are global minimizers of the function f and the search directions generated may be uphill directions for f, we introduce a nonmonotone *approximate norm descent condition* inspired by both the linesearch proposed by Li and Fukushima [23], and the globalization schemes for Inexact Newton methods due to Eisenstat and Walker [13].

The paper is organized as follows. Section 2 introduces the context and the PSANE method [20]. Our proposal is then developed in Section 3. We next investigate (in Section 4) some simple convergence properties of the sequences of residuals and iterates. The theoretical core of the paper is Section 5 where we discuss several conditions ensuring convergence to a solution of (1). Section 6 then illustrates the numerical properties of the proposed method and its variants, and compares it with PSANE. Some conclusions and perspectives are finally presented in Section 7.

1.1 Notations

Throughout this paper, $(x)_i$ represents the ith component of the vector x, and $\mathcal{B}(y, \delta)$ represents the closed ball with center y and radius δ. The symbol $\|\cdot\|$ denotes the Euclidean norm. The (orthogonal) projection map onto Ω is denoted with $P(\cdot)$. When discussing iterative methods for (1), the term *breakdown* refers to the case in which an iterate can not be determined. Finally, given a sequence of vectors $\{x_k\}$, for any function f, we let $f_k = f(x_k)$.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

2 Preliminaries

In this section we review both linesearch strategies which do not require directional derivatives of f and the projected derivative-free algorithm PSANE for nonlinear equations with convex constraints given in [20].

A useful contribution in global convergence of Broyden-like methods for unconstrained nonlinear systems is due to Li and Fukushima [23]. Starting from an earlier contribution by Griewank [15], they proposed a new derivative-free linesearch which is well-defined and easy to implement. At k-th iteration, given the iterate x_k and a search direction p_k, the successive iterate takes the form $x_{k+1} = x_k + \lambda p_k$, $\lambda > 0$, and satisfies

$$\|F(x_k + \lambda p_k)\| \leq (1 + \eta_k)\|F(x_k)\| - \alpha\lambda^2\|p_k\|^2, \tag{3}$$

for some constant $\alpha \in (0, 1)$ and some positive η_k. The sequence $\{\eta_k\}$ is supposed to meet the following requirement.

Assumption 2.1 *The positive sequence $\{\eta_k\}$ satisfies*

$$\sum_{k=0}^{\infty} \eta_k \leq \eta < \infty. \tag{4}$$

Due to the continuity of F, condition (3) holds for all λ sufficiently small, and it is called an approximate norm descent linesearch since

$$\|F(x_k + \lambda p_k)\| \leq (1 + \eta_k)\|F(x_k)\|. \tag{5}$$

La Cruz, Martínez and Raydan [21] then developed the derivative-free nonmonotone iterative method for unconstrained nonlinear systems named Derivative Free Spectral Algorithm for Nonlinear Equations (DF-SANE). The linesearch strategy proposed has the form

$$\phi(x_k + \lambda p_k) \leq \max_{0 \leq j \leq \min\{k, M\}} \phi(x_{k-j}) + \eta_k - \alpha\lambda^2\phi(x_k), \tag{6}$$

where $\phi(x) = \|F(x)\|^\tau$, $\tau \in \{1, 2\}$, M is a nonnegative integer. The first term on the right-hand side of (6) is responsible for the nonmonotone behaviour of ϕ, while the second term $\eta_k > 0$ guarantees that the linesearch strategy is well-defined, and the third term provides the arguments for proving global convergence. The sequence $\{\eta_k\}$ is supposed to satisfy Assumption 2.1. In [17] condition (6) is combined with a nonmonotone watchdog rule and is used with $\eta_k = 0$ for all k.

A further proposal was made by Birgin, Krejić and Martínez [5] in the context of Inexact Quasi-Newton methods for unconstrained systems. Restricting to the "exact" solution of the linear systems, the linesearch is given by

$$\|F(x_k + \lambda p_k)\| \leq (1 - \alpha\lambda)\|F(x_k)\| + \eta_k, \tag{7}$$

and, again, $\alpha \in (0, 1)$ and the sequence $\{\eta_k\}$ is supposed to satisfy Assumption 2.1. We refer to previously mentioned papers for the analysis of the resulting procedures.

In addition, La Cruz recently proposed a projected derivative-free method for the constrained nonlinear system (1), named PSANE [20]. Since the PSANE algorithm motivated the definition of our method, we restate its details.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Algorithm 2.1: The PSANE algorithm

Given $x_0 \in \Omega$, α, $\sigma \in (0,1)$, $\lambda_{\max} \in (0,1]$, $0 < \beta_{\min} < \beta_{\max} < \infty$, $\beta_0 \in [\beta_{\min}, \beta_{\max}]$, a positive sequence $\{\eta_k\}$ that satisfies (4).

For $k = 0, 1, 2, \ldots$ do

1. If $\|F(x_k)\| = 0$ stop.
2. Set $d_- = P(x_k - \beta_k F(x_k)) - x_k$ and $d_+ = P(x_k + \beta_k F(x_k)) - x_k$.
3. Choose $\lambda \in (0, \lambda_{\max}]$.
4. Repeat
   4.1 If
   $$\|F(x_k + \lambda d_-)\|^2 \leq \|F(x_k)\|^2 + \eta_k - \alpha\lambda^2\beta_k^2\|F(x_k)\|^2, \tag{8}$$
   set $\lambda_k = \lambda, d_k = d_-$ and go to Step 5.
   4.2 If
   $$\|F(x_k + \lambda d_+)\|^2 \leq \|F(x_k)\|^2 + \eta_k - \alpha\lambda^2\beta_k^2\|F(x_k)\|^2 \tag{9}$$
   set $\lambda_k = \lambda, d_k = d_+$ and go to Step 5.
   4.3 Set $\lambda = \sigma\lambda$.
5. Set $x_{k+1} = x_k + \lambda_k d_k$, $s_k = x_{k+1} - x_k$, $y_k = F(x_{k+1}) - F(x_k)$.
6. Update β_k:
   $$\begin{array}{l} \text{Set } b_k = \dfrac{s_k^T y_k}{s_k^T s_k}. \\ \text{If } \left|\dfrac{1}{b_k}\right| \in [\beta_{\min}, \beta_{\max}], \text{ set } \beta_{k+1} = \dfrac{1}{b_k}, \\ \text{else } \beta_{k+1} = \min\left[\beta_{\max}, \max\left[\beta_{\min}, \left|\dfrac{1}{b_k}\right|\right]\right]. \end{array} \tag{10}$$

One distinguishing feature of PSANE is that the computation of the search directions d_- and d_+ does not involve the solution of linear systems. The spectral coefficient $1/b_k$ formed in Step 6 is closely related to the Barzilai-Borwein's steplength [2]; it may be positive or negative, and the absolute value $|1/b_k|$ is constrained to belong to the given interval $[\beta_{\min}, \beta_{\max}]$ [21]. The iterate x_{k+1} is determined through a backtracking strategy and each repetition of Step 4 requires a number of evaluations of F between 1 and 2. It is easy to observe that each iterate x_{k+1} is feasible as it is the convex combination of the feasible points x_k and $P(x_k \pm \beta_k F(x_k))$.

Convergence properties of both $\{\|F(x_k)\|\}$ and $\{x_k\}$ have been established under Assumption 2.1. In particular, it is shown that the sequence $\{\|F(x_k)\|\}$ converges [20, Proposition 2.4] and that, if an isolated solution of (1) is a limit point of $\{x_k\}$, then the whole sequence converges to such a solution [20, Theorem 2.7].

As such, the PSANE algorithm is not without drawbacks. We first note that the acceptance conditions (8) and (9) depend on the spectral coefficient β_k such that $|\beta_k| \in [\beta_{\min}, \beta_{\max}]$ (Step 6). Since in practice $1/\beta_{\min}$ and $\beta_{\max}$ are large values, the term $\alpha\lambda^2\beta_k^2\|F(x_k)\|^2$ may become

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

either negligible for small values of $|\beta_k|$, or excessively large for big values of $|\beta_k|$. In the latter case, a large number of backtracks may be necessary to generate the new iterate x_{k+1}.

Moreover, PSANE may breakdown prematurely if an iterate x_k lies on the boundary of Ω, the step d_- has zero norm and d_- is accepted in Step 4.1. In this case, $x_{k+1} = x_k$ and therefore b_k in Step 6 is not well-defined. This can be observed when solving the nonlinear system [20, eqn (28)]

$$F(x) = \begin{pmatrix} 54 - 18x_1 + 3x_3 \\ 78 - 26x_2 + 2x_3 \\ x_3(18 - 3x_1 - 2x_2) \end{pmatrix} = 0 \quad x \in \Omega, \tag{11}$$

where Ω is the box $\{x \in \mathbb{R}^n \text{ s.t. } l \leq x \leq u\}$, $l = (0, 0, 0)^T$, $u = (4, 6, \infty)$. This system admits the unique solution $x^* = (3, 3, 0)^T$. Breakdown occurs at the starting point when running a MATLAB implementation of PSANE with the parameters declared in [20], and initial guesses $x_0^{(1)} = (0,\, 0,\, 0)^T$ and $x_0^{(2)} = (4, 6, 0)^T$.

3 The new algorithm

Building on the concepts developed above, we now introduce our new Projected Approximate Norm Descent algorithm (PAND), which builds a sequence of feasible iterates $\{x_k\}$ satisfying the approximate norm descent property (5) for all k by using the projection operator onto Ω and a linesearch strategy,

At k-th iteration, let x_k be the current feasible iterate and B_k be a suitable invertible matrix. First, the linear system

$$B_k p_k^{\mathrm{QN}} = -F(x_k), \tag{12}$$

is solved and two steps

$$p_+(p_k^{\mathrm{QN}}, \lambda) \stackrel{\mathrm{def}}{=} P(x_k + \lambda p_k^{\mathrm{QN}}) - x_k, \qquad p_-(p_k^{\mathrm{QN}}, \lambda) \stackrel{\mathrm{def}}{=} P(x_k - \lambda p_k^{\mathrm{QN}}) - x_k, \tag{13}$$

$\lambda \in (0, 1]$, are formed, see e.g. [7]. A feasible point of the form

$$x_{k+1} = x_k + p_k = x_k + p_k(\lambda),$$

is then selected by such that, for some $\alpha \in (0, 1)$, $\eta_k > 0$, and $\{\eta_k\}$ satisfying (4),

$$\|F(x_k + p_k(\lambda))\| \leq (1 - \alpha(1 + \lambda))\|F(x_k)\|, \tag{14}$$

or

$$\|F(x_k + p_k(\lambda))\| \leq (1 + \eta_k - \alpha\lambda)\|F(x_k)\|. \tag{15}$$

where $p_k(\lambda) = p_\pm(p_k^{\mathrm{QN}}, \lambda)$.

In this procedure, p_k^{QN} is a Quasi-Newton step (which explains its superscript). The matrix B_k can be chosen as a two-point approximation to the secant equation by letting

$$B_k = \beta_k^{-1} I, \qquad |\beta_k| \in [\beta_{\min}, \beta_{\max}], \tag{16}$$

with β_k given in (10) [2, 21]. Alternatively, B_k can be built by using the Broyden's update or any other secant formula, see e.g., [9, 24, 27]. The use of such matrices B_k is intended to make the computation of p_k^{QN} cheap.

The formal description of PAND method is as follows.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Algorithm 3.1: The PAND algorithm

Given $x_0 \in \Omega$, $B_0 \in \mathbb{R}^{n\times n}$ nonsingular, α, $\sigma \in (0,1)$, $\{\eta_k\}$ satisfying (4).

For $k = 0, 1, 2, \ldots$ do

1. Solve the linear system (12).
2. Set $\lambda = 1$.
3. Repeat
   - 3.1 Set $p_+ = p_+(p_k^{\mathrm{QN}}, \lambda)$ and $p_- = p_-(p_k^{\mathrm{QN}}, \lambda)$ as in (13).
   - 3.2 If $p_k(\lambda) = p_+$ satisfies (14), go to Step 4.
   - 3.3 If $p_k(\lambda) = p_-$ satisfies (14), go to Step 4.
   - 3.4 If $\|p_+\| \neq 0$ and $p_k(\lambda) = p_+$ satisfies (15), go to Step 4.
   - 3.5 If $\|p_-\| \neq 0$, and $p_k(\lambda) = p_-$ satisfies (15), go to Step 4.
   - 3.6 Otherwise set $\lambda = \sigma\,\lambda$.
4. Set $p_k = p_k(\lambda)$, $\lambda_k = \lambda$, $x_{k+1} = x_k + p_k$.
5. If $\|F(x_{k+1})\| = 0$ stop.
   Else form an invertible matrix B_{k+1}.

Trivially, $x_k + p_\pm(p_k^{\mathrm{QN}}, \lambda)$ is feasible. If F is continuously differentiable, either p_+ or p_- is a descent direction for f in (2), unless $\nabla f(x_k)^T p_+ = \nabla f(x_k)^T p_- = 0$. Thus, the use of both p_+ and p_- promotes a decrease of $\|F\|$, cfr. [21,22].

We also observe that the vector

$$v_k^{\mathrm{QN}} = P(x_k + p_k^{\mathrm{QN}}) - x_k, \tag{17}$$

is the first step tested in the PAND algorithm. From the properties of the projection map P, we may deduce that

$$\|v_k^{\mathrm{QN}}\| \leq \|p_k^{\mathrm{QN}}\|, \tag{18}$$

$$\|p_\pm\| \leq \lambda \|p_k^{\mathrm{QN}}\|, \tag{19}$$

and, by Steps 3 and 4 of PAND algorithm, we have that

$$\|p_k\| \leq \lambda_k \|p_k^{\mathrm{QN}}\|. \tag{20}$$

Acceptance of the trial steps is tested in Step 3, which terminates in a finite number of steps. Indeed, from the continuity of F and the positivity of η_k, there exists a scalar $\bar{\lambda}$ such that

$$(F(x_k + p_k(\lambda))_i^2 \leq (1 + \eta_k - \alpha\bar{\lambda})^2 (F(x_k))_i^2,$$

with $\lambda \in (0, \bar{\lambda}]$ and for $i = 1, \ldots, n$. Trivially the above inequalities imply that (15) holds for λ small enough. The number of F-evaluations performed at each loop within Step 3 is either 1 or 2.

The linesearch conditions (14) and (15) are derivative-free. The first is related to globally convergent Inexact Newton methods [13] where a sufficient decrease in $\|F\|$ is imposed at each iteration. It is tested on both p_+ and p_- in order to promote a decrease in $\|F\|$. The second

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

allows for an increase in $\|F\|$, possibly for λ small enough. We exclude the use of zero-norm steps as in this case the inequality (15) is trivially satisfied as long as $(\eta_k - \alpha\lambda) \geq 0$.

It is important to observe that inequality (14) implies (15), and the latter implies (5). Thus, the approximate norm descent condition (5) holds for all k.

As in (3), (7) and (8), the scalar η_k in (15) allows a nonmonotone behaviour of $\|F\|$. Conditions (14) and (15) however differ from (3) and (8) in two respects. Firstly, they are independent from the norm of the step used, which may be convenient whenever this norm is very large, see §2. Secondly, η_k appears as a multiplicative term for $\|F(x_k)\|$ in (15), while the impact of η_k on the value $\|F(x_k)\|^2 + \eta_k$ in (6) is unpredictable as η_k is not adjusted to reflect the size of $\|F(x_k)\|$.

Finally, the sufficient decrease condition (14) with p_+ and p_- is important for establishing theoretical results on the convergence of $\{\|F(x_k)\|\}$ to zero (see next section). Such results are valuable as convergence to stationary points of (2) cannot be obtained in our framework, cfr. [17, 20, 21, 23]. We are aware that (14) may slow the convergence of the method but the numerical experience presented in §6 shows that it does not either prevent the nonmonotone behaviour of $\|F\|$ or slow convergence down compared with PSANE.

Detailed numerical experience with PAND will be presented in §6. We only observe at this stage that the implementation of PAND (with B_k given by (16), and the PSANE parameters as used in [20]) is successful on problem (11) starting from the initial guesses $x_0^{(1)}$, $x_0^{(2)}$ given in §2: the algorithm converges to a solution in 8 and 10 F-evaluations, respectively.

4 Convergence analysis

This section is devoted to the theoretical study of the PAND algorithm. Summarizing our main results:

- We show that the sequence $\{\|F_k\|\}$ is convergent.
- We show that sequence $\{x_k\}$ is convergent and give an upper bound on the distance between x_0 and the limit point x^*.
- We investigate some conditions under which $\lim_{k\to\infty} \|F_k\| = 0$, i.e. $F(x^*) = 0$.

The following technical result shown in [23, Lemma 2.1] will be useful.

Lemma 4.1 *Let $\{\eta_k\}$ satisfy Assumption 2.1. Then $\prod_{i=0}^{k}(1+\eta_i) \leq e^{\eta}$ with $k \geq 0$.*

4.1 Analysis of the sequences $\{\|F_k\|\}$ and $\{\lambda_k\}$

We start by analyzing the asymptotic behaviour of the sequences $\{\|F_k\|\}$ and $\{\lambda_k\}$ and make a first attempt to detect both occurrences where $\lim_{k\to\infty} \|F_k\| = 0$ and where PAND method fails to solve (1). The following theorem characterizes the behaviour of $\{\|F_k\|\}$ and is valid for any continuous function F. The proof relies on inequality (5).

Theorem 4.2 *Let Assumption 2.1 hold and $\{x_k\}$ be generated by the* PAND *algorithm. Then*

(i) the sequence $\{\|F_k\|\}$ is bounded and

$$\|F_{k+1}\| \leq e^{\eta}\|F_0\|, \tag{21}$$

for all $k \geq 0$.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

(ii) The sequence $\{\|F_k\|\}$ *is convergent.*

(iii)

$$\lim_{k\to\infty} \lambda_k \|F_k\| = 0. \tag{22}$$

(iv)

$$\liminf_{k\to\infty} \lambda_k > 0 \qquad \text{implies that} \qquad \lim_{k\to\infty} \|F_k\| = 0. \tag{23}$$

(v) If (14) is satisfied for infinitely many k*, then* $\lim_{k\to\infty} \|F_k\| = 0$*.*

(vi) If $\|F_k\| \le \|F_{k+1}\|$ *for infinitely many iterations, then* $\liminf_{k\to\infty} \lambda_k = 0$*.*

(vii) If $\|F_k\| \le \|F_{k+1}\|$ *for all* k *sufficiently large, then* $\{\|F_k\|\}$ *does not converge to 0.*

Proof. (i) Applying (5) recursively, we obtain that

$$\|F_{k+1}\| \le \prod_{i=0}^{k} (1+\eta_i)\|F_0\|,$$

for all $k \ge 0$. The proof is then completed by using Lemma 4.1.
(ii) We know that any positive sequence $\{a_k\}$ satisfying

$$a_{k+1} \le (1+r_k)a_k + r_k,$$

with $r_k > 0$ and $\sum_{k=0}^{\infty} r_k < \infty$, is convergent by [10, Lemma 3.3]. Hence, since $\{\|F_k\|\}$ satisfies (5) for all k, it converges.
(iii) By (15), we have that

$$\alpha\lambda_k\|F_k\| \le (1+\eta_k)\|F_k\| - \|F_{k+1}\|. \tag{24}$$

Using $\lim_{k\to\infty} \eta_k = 0$ and the convergence of $\{\|F_k\|\}$ we obtain (22).
(iv) The implication (23) directly follows from (22).
(v) If the norm decrease (14) holds for infinitely many k, there exists a subsequence $\{\|F_{k_j}\|\}$, blue $1 \le k_0 < k_1 < \cdots$, such that

$$\|F_{k_j}\| \le (1-\alpha-\alpha\lambda_{k_j})\|F_{k_j-1}\| \le (1-\alpha)\|F_{k_j-1}\|,$$

whereas by (5)

$$\|F_{k_j-1}\| \le (1+\eta_{k_j-2})\|F_{k_j-2}\| \le \prod_{i=k_{j-1}}^{k_j-2} (1+\eta_i)\|F_{k_j-1}\|.$$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Thus,

$$\begin{aligned}
\|F_{k_j}\| &\leq (1-\alpha)\|F_{k_j-1}\| \\
&\leq (1-\alpha)\prod_{i=k_{j-1}}^{k_j-2}(1+\eta_i)\|F_{k_{j-1}}\| \\
&\leq (1-\alpha)^2\prod_{i=k_{j-1}}^{k_j-2}(1+\eta_i)\|F_{k_{j-1}-1}\| \\
&\leq \dots \\
&\leq (1-\alpha)^{j+1}\prod_{i=k_0}^{k_j-2}(1+\eta_i)\|F_{k_0-1}\| \\
&\leq (1-\alpha)^{j+1}\prod_{i=0}^{k_j-2}(1+\eta_i)\|F_0\| \\
&\leq (1-\alpha)^{j+1}e^{\eta}\|F_0\|,
\end{aligned}$$

where the last inequality follows from Lemma 4.1. Hence, $\lim_{k_j\to\infty}\|F_{k_j}\| = 0$ and the convergence of $\{\|F_k\|\}$ implies $\lim_{k\to\infty}\|F_k\| = 0$.
(vi) If $\|F_k\| \leq \|F_{k+1}\|$ for infinitely many steps then there exists a subsequence of indices $\{k_j\}$ such that

$$\|F_{k_j}\| \leq \|F_{k_j+1}\| \leq (1+\eta_{k_j}-\alpha\lambda_{k_j})\|F_{k_j}\|,$$

and this gives

$$\alpha\lambda_{k_j} \leq \eta_{k_j}.$$

Since $\lim_{k\to\infty}\eta_k = 0$, we get $\liminf_{k\to\infty}\lambda_k = 0$.
(vii) In case we have that

$$\|F_k\| \leq \|F_{k+1}\| \leq (1+\eta_k-\alpha\lambda_k)\|F_k\|,$$

for all k sufficiently large, we trivially conclude that $\{\|F_k\|\}$ does not converge to 0. □

4.2 Analysis of the sequence $\{x_k\}$

Now we analyze the sequence of iterates generated by the PAND algorithm and make the following assumption.

Assumption 4.1 *Matrices B_k^{-1} are uniformly bounded for $k \geq 0$, i.e. $\|B_k^{-1}\| \leq c_B$ for some positive scalar c_B.*

Assumption 4.1 immediately yields that the step p_k^{QN} in (12) satisfies

$$\|p_k^{\mathrm{QN}}\| \leq c_B\|F_k\|. \tag{25}$$

We observe that B_k of the form (16) is guaranteed to fulfill Assumption 4.1 as $\|B_k^{-1}\| = |\beta_k| \leq \beta_{\max}$. We start showing that $\{x_k\}$ is convergent.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Theorem 4.3 *Let Assumptions 2.1 and 4.1 hold and $\{x_k\}$ be the sequence generated by the* PAND *algorithm. Then the sequence $\{x_k\}$ is convergent and, if x^* is the limit point, then*

$$\|x_0 - x^*\| \leq c_B \left(\frac{1}{\alpha} + \frac{\eta}{\alpha} e^{\eta}\right) \|F_0\|.$$

Proof. First note that (20) and (25) yield

$$\|p_k\| \leq c_B \lambda_k \|F_k\|. \tag{26}$$

Consider $\sum_{k=0}^{\infty} \lambda_k \|F_k\|$. Using (24) and (21), we obtain that

$$\begin{aligned}
\sum_{k=0}^{\infty} \lambda_k \|F_k\| &\leq \sum_{k=0}^{\infty} \left(\frac{1+\eta_k}{\alpha}\|F_k\| - \frac{1}{\alpha}\|F_{k+1}\|\right) \\
&= \sum_{k=0}^{\infty} \frac{1}{\alpha}\left(\|F_k\| - \|F_{k+1}\|\right) + \sum_{k=0}^{\infty} \frac{\eta_k}{\alpha}\|F_k\| \\
&\leq \frac{1}{\alpha}\|F_0\| + \sum_{k=0}^{\infty} \frac{\eta_k}{\alpha} e^{\eta}\|F_0\| \\
&\leq \left(\frac{1}{\alpha} + \frac{\eta}{\alpha} e^{\eta}\right) \|F_0\|.
\end{aligned} \tag{27}$$

Then $\sum_{k=0}^{\infty} \lambda_k \|F_k\|$ is convergent since the terms $\lambda_k \|F_k\|$ are nonnegative. Moreover, by (26), we have that

$$\sum_{k=0}^{\infty} \|p_k\| < \infty. \tag{28}$$

In order to show that $\{x_k\}$ is convergent, let $m \geq \ell$ and consider

$$\|x_m - x_\ell\| \leq \sum_{k=\ell}^{m-1} \|p_k\| \leq \sum_{k=\ell}^{\infty} \|p_k\|.$$

Now,

$$\sum_{k=\ell}^{\infty} \|p_k\| = \sum_{k=0}^{\infty} \|p_k\| - \sum_{k=0}^{\ell-1} \|p_k\|$$

tends to zero as ℓ tends to infinity. Consequently, for any $\epsilon > 0$, there exists ℓ sufficiently large such that $\|x_m - x_\ell\| \leq \epsilon$ for $m \geq \ell$. This means that $\{x_k\}$ is a Cauchy sequence and hence it converges. Finally,

$$\|x_0 - x_\ell\| \leq \sum_{k=0}^{\ell-1} \|p_k\|,$$

and letting ℓ tend to infinity, we obtain that

$$\|x_0 - x^*\| \leq \sum_{k=0}^{\infty} \|p_k\| \leq c_B \sum_{k=0}^{\infty} \lambda_k \|F_k\|.$$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

The desired conclusion then follows from (27). □

Assumption 4.1 has an important consequence. The bound on $\|x_0 - x^*\|$ given above implies that if a solution $\bar{x}$ of (1) satisfies

$$\|x_0 - \bar{x}\| > c_B \left(\frac{1}{\alpha} + \frac{\eta}{\alpha} e^{\eta} \right) \|F_0\|,$$

then $\{x_k\}$ cannot converge to $\bar{x}$. Fortunately, α is typically chosen quite small in practice [9], but this remains a drawback of PAND. An analogous result to the bound (28) was established by Li and Fukushima on the steps taken in their derivative-free Broyden-like method, see [23, Theorem 2,2]. This class of methods is therefore best suited to cases where a solution is known to exist in a reasonable neighbourhood of the initial point.

We conclude our analysis of the convergence of $\{x_k\}$ by considering the case where the limit point of $\{x_k\}$ solves (1) and lies in the interior of Ω. Part of our results is obtained under the well-known Dennis-Moré condition [9] and the following assumption.

Assumption 4.2 *F is continuously differentiable on Ω and the Jacobian J is Lipschitz continuous on Ω and satisfies*

$$\|J(x) - J(y)\| \leq 2L\|x - y\|, \quad \forall x, y \in \Omega.$$

Lemma 4.4 *Let Assumptions 2.1, 4.1 hold, and $\{x_k\}$ be the sequence generated by the* PAND *algorithm. Suppose that the limit point x^* of $\{x_k\}$ is such that $x^* \in int(\Omega)$ and $F(x^*) = 0$. Then the following conclusions hold.*

i) For k sufficiently large it holds $p_+(p_k^{\mathrm{QN}}, 1) = p_k^{\mathrm{QN}}$ and $x_k + p_\pm(p_k^{\mathrm{QN}}, \lambda) \in int(\Omega)$ for all $\lambda \in (0, 1]$.

ii) If Assumption 4.2 holds, $J(x^*)$ is nonsingular, and

$$\lim_{k \to \infty} \frac{\|E_k p_k^{\mathrm{QN}}\|}{\|p_k^{\mathrm{QN}}\|} = 0, \tag{29}$$

with $E_k = B_k - J(x^*)$, then $\{x_k\}$ converges to x^* superlinearly.

Proof. (i) Since $x^* \in int(\Omega)$, there exist $\rho^* > 0$ such that $\mathcal{B}(x^*, \rho) \subset int(\Omega)$ for $\rho \in (0, \rho^*)$. Since $\{x_k\}$ converges to x^*, we know that $x_k \in \mathcal{B}(x^*, \rho)$ for all k sufficiently large. From (25), $\|p_k^{\mathrm{QN}}\|$ tends to 0, and for k large enough

$$\|x^* - (x_k + p_k^{\mathrm{QN}})\| \leq \|x^* - x_k\| + \|p_k^{\mathrm{QN}}\| \leq \rho + \|p_k^{\mathrm{QN}}\| < \rho^*.$$

Thus, we have that $x_k + p_k^{\mathrm{QN}} \in int(\Omega)$ and $p_+(p_k^{\mathrm{QN}}, 1) = p_k^{\mathrm{QN}}$ by (13). Further, (19) yields that $x_k + p_\pm \in int(\Omega)$ for all $\lambda \in (0, 1]$.

(ii) See [9, Chapter 8]. □

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

5 Ensuring the convergence of $\{\|F_k\|\}$ to zero

In Theorem 4.2 we pointed out one case where the PAND algorithm solves problem (1), i.e., $\{\|F_k\|\}$ converges to zero. In this section we complete our theoretical analysis of the PAND algorithm by detecting further occurrences where $\{\|F_k\|\}$ converges to zero. We address this issue considering the use of both spectral residual steps and more general Quasi-Newton steps. In order to interpret the results given, it is again useful to remember that α is typically quite small [9].

We start by recalling a simple observation.

Lemma 5.1 *Let f defined in (2) be continuously differentiable. For $p_k = \pm\lambda_k\beta_k F_k$, it holds*

$$\begin{aligned} f(x_{k+1}) &= f(x_k) \pm 2\lambda_k\beta_k \int_0^1 F_k^T J(x_k + tp_k)F_k\, dt + \\ &\quad \pm 2\lambda_k\beta_k \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k)F_k\, dt. \end{aligned} \tag{30}$$

Proof. Using [9, Lemma 4.1.2], we have that

$$\begin{aligned} f(x_{k+1}) &= f(x_k) + \int_0^1 \nabla f(x_k + tp_k)^T p_k\, dt \\ &= f(x_k) \pm 2\lambda_k\beta_k \int_0^1 F(x_k + tp_k)^T J(x_k + tp_k)F_k\, dt, \end{aligned}$$

from which (30) follows. □

Under specific assumptions on the Jacobian J at the limit point x^* of $\{x_k\}$, the next two theorems analyze the acceptance of the spectral residual steps $p_k = \pm\lambda_k\beta_k F_k$, $|\beta_k| \in (\beta_{\min}, \beta_{\max})$ for k large enough. Our first result concerns the case when $J_S(x^*)$, the symmetric part* of $J(x^*)$, is positive (negative) definite and ensures that $\lim_{k\to\infty} \|F_k\| = 0$ when the 2-norm condition number of $J_S(x^*)$ is of order $O\left(\alpha^{-1}\right)$. The notation G_k is used for the "average Jacobian" matrix along the step p_k, defined by

$$G_k \stackrel{\text{def}}{=} \int_0^1 J(x_k + tp_k)\, dt, \tag{31}$$

while $(G_S)_k$ denotes the average matrix associated to J_S along the step p_k, defined by

$$(G_S)_k \stackrel{\text{def}}{=} \int_0^1 J_S(x_k + tp_k)\, dt. \tag{32}$$

Theorem 5.2 *Let Assumptions 2.1, 4.1, 4.2 hold and $\{x_k\}$ be the sequence generated by the* PAND *algorithm with B_k given by (16). Suppose that for k sufficiently large, the steps taken have the form $p_k = \pm\lambda_k\beta_k F_k$, $|\beta_k| \in (\beta_{\min}, \beta_{\max})$. Moreover assume that the symmetric part J_S of J is positive (negative) definite at the limit point x^* of $\{x_k\}$, and that the 2-norm condition number $\kappa(J_S(x^*))$ satisfies*

$$\kappa(J_S(x^*)) < \frac{\gamma}{\alpha}, \tag{33}$$

for some $\gamma \in (0,1)$, and $\alpha \in (0,1)$ as in (14)-(15). Then $F(x^*) = 0$.

*We recall here that the symmetric part A_S of any matrix A is defined as $A_S = (A + A^T)/2$. It holds $v^T Av = v^T A_S v$ for any vector v.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Proof. Without loss of generality, let us assume that $J_S(x^*)$ is positive definite. Then $J(x^*)$ is nonsingular and by (10) and (31) we get

$$\beta_k = \frac{\|p_{k-1}\|^2}{p_{k-1}^T(F_k - F_{k-1})} = \frac{\|p_{k-1}\|^2}{p_{k-1}^T \int_0^1 J(x_{k-1} + tp_{k-1})p_{k-1}\,dt} = \frac{\|p_{k-1}\|^2}{p_{k-1}^T \int_0^1 J_S(x_{k-1} + tp_{k-1})p_{k-1}\,dt},$$

i.e., by (32)

$$\beta_k = \frac{\|p_{k-1}\|^2}{p_{k-1}^T (G_S)_{k-1} p_{k-1}}.$$

Moreover, since $F_k^T G_k F_k = F_k^T (G_S)_k F_k$, using Lemma 5.1, we have that

$$\begin{aligned} f(x_{k+1}) &= f(x_k) \pm 2\lambda_k\beta_k \frac{F_k^T (G_S)_k F_k}{\|F_k\|^2} f(x_k) + \\ &\quad \pm 2\lambda_k\beta_k \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k) F_k\,dt. \end{aligned} \tag{34}$$

Now, continuity implies that there exists a scalar $\rho > 0$ sufficiently small such that, for all $y \in \mathcal{B}(x^*, \rho)$,

$$\sigma_{\min}(J_S(y)) \geq (1-\epsilon)\sigma_{\min}(J_S(x^*)) \;\text{ and }\; \sigma_{\max}(J_S(y)) \leq (1+\epsilon)\sigma_{\max}(J_S(x^*)), \tag{35}$$

and

$$\sigma_{\max}(J(y)) \leq (1+\epsilon)\sigma_{\max}(J(x^*)), \tag{36}$$

with $\epsilon \in (0,1)$ given by

$$\epsilon \stackrel{\text{def}}{=} \frac{1-\gamma}{1+\gamma}. \tag{37}$$

Moreover, the convergence of the sequence $\{x_k\}$ implies that $x_{k-1} + tp_{k-1}$ and $x_k + tp_k$ both belong to $\mathcal{B}(x^*, \rho)$ for large enough k and all $t \in [0,1]$. As a consequence, we deduce that, for k sufficiently large,

$$\min\left[\sigma_{\min}((G_S)_k), \sigma_{\min}((G_S)_{k-1})\right] \geq (1-\epsilon)\sigma_{\min}(J_S(x^*)), \tag{38}$$

and

$$\max\left[\sigma_{\max}((G_S)_k), \sigma_{\max}((G_S)_{k-1})\right] \leq (1+\epsilon)\sigma_{\max}(J_S(x^*)). \tag{39}$$

This in turn implies that, for k sufficiently large, $\beta_k > 0$ lies in the interval

$$\beta_k \in \left[\frac{1}{\sigma_{\max}((G_S)_{k-1})}, \frac{1}{\sigma_{\min}((G_S)_{k-1})}\right], \tag{40}$$

and that

$$\beta_k \frac{F_k^T (G_S)_k F_k}{\|F_k\|^2} \in \left[\frac{\sigma_{\min}((G_S)_k)}{\sigma_{\max}((G_S)_{k-1})}, \frac{\sigma_{\max}((G_S)_k)}{\sigma_{\min}((G_S)_{k-1})}\right] \subseteq \left[\frac{1-\epsilon}{1+\epsilon}\left(\frac{\sigma_{\min}(J_S(x^*))}{\sigma_{\max}(J_S(x^*))}\right), \frac{1+\epsilon}{1-\epsilon}\left(\frac{\sigma_{\max}(J_S(x^*))}{\sigma_{\min}(J_S(x^*))}\right)\right],$$

which yields

$$\beta_k \frac{F_k^T (G_S)_k F_k}{\|F_k\|^2} \geq \frac{\gamma}{\kappa(J_S(x^*))}. \tag{41}$$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Consider $p_k = -\lambda_k \beta_k F_k$. The inequality (34) implies that

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - 2\lambda_k\beta_k \frac{F_k^T (G_S)_k F_k}{\|F_k\|^2} f(x_k) + \\ &\quad + 2\lambda_k\beta_k \left| \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k) F_k \, dt \right|, \end{aligned} \tag{42}$$

in which the last absolute value can be written

$$\left| \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k) F_k \, dt \right| = \left| \int_0^1 \left(\int_0^1 J(x_k + \zeta\, tp_k) tp_k \, d\zeta \right) J(x_k + tp_k) F_k \, dt \right|,$$

$\zeta \in [0,1]$. Again $x_k + \zeta\, tp_k \in \mathcal{B}(x^*, \rho)$ for $t, \zeta \in [0,1]$. Thus, proceeding as above and using the form $p_k = -\lambda_k \beta_k F_k$, we deduce that

$$\begin{aligned} \left| \int_0^1 (F(x_k + tp_k) - F_k)^T J(x_k + tp_k) F_k \, dt \right| &\leq \int_0^1 t\lambda_k |\beta_k| \max_{z \in \mathcal{B}(x^*,\rho)} \|J(z)\|^2 \|F_k\|^2 \, dt \\ &= \frac{1}{2} \lambda_k |\beta_k| \max_{z \in \mathcal{B}(x^*,\rho)} \sigma_{\max}(J(z))^2 \|F_k\|^2. \end{aligned} \tag{43}$$

Combining this expression with (41), (42), (40), (35), (36) and (37) we obtain that, for k sufficiently large,

$$\begin{aligned} f(x_{k+1}) &\leq \left(1 - 2\lambda_k\beta_k \frac{F_k^T (G_S)_k F_k}{\|F_k\|^2} + \lambda_k^2 \beta_k^2 \max_{z \in \mathcal{B}(x^*,\rho)} \sigma_{\max}(J(z))^2 \right) f(x_k) \\ &\leq \left(1 - 2\frac{\gamma}{\kappa(J_S(x^*))} \lambda_k + \frac{1}{\gamma^2} \left[\frac{\sigma_{\max}(J(x^*))}{\sigma_{\min}(J_S(x^*))} \right]^2 \lambda_k^2 \right) f(x_k). \end{aligned}$$

Thus, for k sufficiently large, the linesearch condition (15) holds for any λ such that

$$1 - \frac{2\gamma}{\kappa(J(x^*))} \lambda + \frac{1}{\gamma^2} \left[\frac{\sigma_{\max}(J(x^*))}{\sigma_{\min}(J_S(x^*))} \right]^2 \lambda^2 \leq (1 - \alpha\lambda)^2,$$

i.e., such that

$$\kappa_2 \lambda^2 + 2\kappa_1 \lambda \stackrel{\text{def}}{=} \left(\frac{1}{\gamma^2} \left[\frac{\sigma_{\max}(J(x^*))}{\sigma_{\min}(J_S(x^*))} \right]^2 - \alpha^2 \right) \lambda^2 + 2 \left(\alpha - \frac{\gamma}{\kappa(J(x^*))} \right) \lambda \leq 0. \tag{44}$$

By definition of J_S, $\|J_S(x^*)\| \leq \|J(x^*)\|$. Then,

$$\frac{\sigma_{\max}(J(x^*))}{\sigma_{\min}(J_S(x^*))} \geq \kappa(J_S(x^*)),$$

and $\kappa_2 > 0$ since α and γ belong to $(0,1)$. This implies that (44) is satisfied for a sufficiently small and positive λ, since (33) gives $\kappa_1 < 0$ and (15) is satisfied (for k large enough) if $\lambda \leq \lambda_* \stackrel{\text{def}}{=} -2\,\kappa_1/\kappa_2$. The mechanism of Step 3.6 of the PAND algorithm then guarantees that, for k sufficiently large, the loop in Step 3 terminates with $\lambda_k \geq \min\{1, \sigma\lambda_*\}$, and λ_*

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

independent of k. As a consequence, $\liminf_{k\to\infty} \lambda_k > 0$ and (23) allows us to conclude the proof. □

Convergence of $\{\|F_k\|\}$ to zero was also obtained in [22] by assuming the positive (negative) definiteness of J_S for all x in the lower level set $\{x : 0 \le f(x) \le f(x_0)\}$.

In the next theorem, we analyze the acceptance of the spectral residual step under the assumption that J is strongly diagonally dominant and the diagonal entries have constant sign. We use the following notation:

$$\zeta_i(x) \stackrel{\text{def}}{=} \frac{1}{|(J(x))_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |(J(x))_{ij}| \quad i = 1, \ldots, n, \tag{45}$$

$$m(x) \stackrel{\text{def}}{=} \min_{1\le i\le n} (J(x))_{ii}, \qquad M(x) \stackrel{\text{def}}{=} \max_{1\le i\le n} (J(x))_{ii}, \tag{46}$$

$$\widetilde{m}(x) \stackrel{\text{def}}{=} \min_{1\le i\le n} |(J(x))_{ii}|, \qquad \widetilde{M}(x) \stackrel{\text{def}}{=} \max_{1\le i\le n} |(J(x))_{ii}|. \tag{47}$$

Observe that all this quantities only depend on the Jacobian matrix at x. The value of $\zeta_i(x)$ measure the degree of diagonal dominance of the i-th row of $J(x)$, $m(x)$ and $M(x)$ measure the signed range of its diagonal elements while $\widetilde{m}(x)$ and $\widetilde{M}(x)$ measure the diagonals' absolute values' range. If $J(x)$ has positive diagonal entries, then $\widetilde{m}(x) = m(x) = |m(x)|$ and $\widetilde{M}(x) = M(x) = |M(x)|$. If the diagonal elements are negative, then $\widetilde{m}(x) = -M(x) = |M(x)|$ and $\widetilde{M}(x) = -m(x) = |m(x)|$. The conditions used are

$$\max\left[\frac{\widetilde{M}(x^*)}{|m(x^*)|}, \frac{\widetilde{M}(x^*)}{|M(x^*)|}\right] \sum_{i=1}^{n} \zeta_i(x^*) \le \frac{1-\nu}{1+\nu}, \tag{48}$$

and

$$\frac{\widetilde{M}(x^*)}{\widetilde{m}(x^*) < \left(\frac{\nu}{2-\nu}\right)\left(\frac{1-\nu}{1+\nu}\right)\frac{1}{\alpha}}, \tag{49}$$

where $\nu \in (0,1)$ and $\alpha \in (0,1)$ is the constant in (14)-(15). Such conditions are satisfied by matrices which are close to being diagonal and have a condition number of order α^{-1}. In fact, for decreasing values of $\max_{1\le i\le n} \zeta_i$, the ratio $\widetilde{M}/\widetilde{m}$ approaches $\kappa(J(x^*))$ and (49) implies a bound on such a condition number in terms of α^{-1}. For example, if $\nu = 1/2$, the right-hand side of (48) is $1/3$ and that of (49) is $1/9\alpha$.

Theorem 5.3 *Let Assumptions 2.1, 4.1, 4.2 hold and $\{x_k\}$ be the sequence generated by the* PAND *algorithm with B_k given by (16). Suppose that for k sufficiently large, the steps taken have the form $p_k = \pm\lambda_k\beta_k F_k$, $|\beta_k| \in (\beta_{\min}, \beta_{\max})$, and that $J(x^*)$ is nonsingular, where x^* is the limit point of $\{x_k\}$. Suppose in addition that $J(x^*)$ has diagonal entries of constant sign and satisfies (48) and (49), for some $\nu \in (0,1)$ and $\alpha \in (0,1)$ being the constant in (14)-(15). Then $F(x^*) = 0$.*

Proof. Because $J(x)$ is continuous, there exists a $\rho > 0$ such that

$$\max\left[\frac{\widetilde{M}(x)}{|m(x)|}, \frac{\widetilde{M}(x)}{|M(x)|}\right] \sum_{i=1}^{n} \zeta_i(x) < 1-\nu, \tag{50}$$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

and

$$\widetilde{M}(x) \leq (1+\nu)\widetilde{M}(x^*) \quad \text{and} \quad \widetilde{m}(x) \geq (1-\nu)\widetilde{m}(x^*),$$

for all $x \in \mathcal{B}(x^*, \rho)$. Moreover, for k sufficiently large x_{k-1} and x_k belong to $\mathcal{B}(x^*, \rho)$ and the same holds for $x_k + tp_k$, $x_{k-1} + tp_{k-1}$, $t \in [0,1]$. Hence (50) holds for $J(x_k)$ and $J(x_k + tp_k)$ for all k sufficiently large and all $t \in [0,1]$. As a consequence, we obtain that, for sufficiently large k,

$$\max\left[\frac{\widetilde{M}_k}{|m_k|}, \frac{\widetilde{M}_k}{|M_k|}\right] \sum_{i=1}^{n} \zeta_{i,k} \leq 1-\nu, \tag{51}$$

and

$$\widetilde{M}_k \leq (1+\nu)\widetilde{M}(x^*) \quad \text{and} \quad \widetilde{m}_k \geq (1-\nu)\widetilde{m}(x^*),$$

where $\zeta_{i,k}$, M_k, m_k, $\widetilde{m}_k$ and $\widetilde{M}_k$ are defined as in (45)-(47) using the average Jacobian G_k instead of $J(x)$.
As in the previous theorem, the steps used have the form $p_k = \pm\lambda_k\beta_k F_k$ and we analyze (30). As for $F_k^T G_k F_k$, $t \in [0,1]$, with G_k as in (31), we have that

$$F_k^T G_k F_k = \sum_{i=1}^{n} (F_k)_i \left[(G_k)_{ii}(F_k)_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} (G_k)_{ij}(F_k)_j \right].$$

Then, for i fixed,

$$\begin{aligned}
(G_k)_{ii}(F_k)_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{n} (G_k)_{ij}(F_k)_i(F_k)_j &\geq (G_k)_{ii}(F_k)_i^2 - \sum_{\substack{j=1 \\ j \neq i}}^{n} |(G_k)_{ij}|\,|(F_k)_i|\,|(F_k)_j| \\
&\geq (G_k)_{ii}(F_k)_i^2 - \sum_{\substack{j=1 \\ j \neq i}}^{n} |(G_k)_{ij}|\,\|F_k\|_\infty^2 \\
&\geq (G_k)_{ii}(F_k)_i^2 - \zeta_{i,k}|(G_k)_{ii}|\,\|F_k\|^2.
\end{aligned}$$

The entries of $(G_k)_{ii}$ are of constant sign. If they are positive, by using (51) we obtain that, for k large enough,

$$F_k^T G_k F_k \geq m_k \left(1 - \sum_{i=1}^{n} \zeta_{i,k} \frac{\widetilde{M}_k}{|m_k|}\right) \|F_k\|^2 \geq m_k \nu \|F_k\|^2.$$

Similarly, for k large enough,

$$\begin{aligned}
(G_k)_{ii}(F_k)_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{n} (G_k)_{ij}(F_k)_i(F_k)_j &\leq (G_k)_{ii}(F_k)_i^2 + \sum_{\substack{j=1 \\ j \neq i}}^{n} |(G_k)_{ij}|\,|(F_k)_i|\,|(F_k)_j| \\
&\leq (G_k)_{ii}(F_k)_i^2 + \zeta_{i,k}|(G_k)_{ii}|\,\|F_k\|^2,
\end{aligned}$$

and

$$F_k^T G_k F_k \leq M_k \left(1 + \sum_{i=1}^{n} \zeta_{i,k} \frac{\widetilde{M}_k}{|M_k|}\right) \leq M_k(2-\nu)\|F_k\|^2,$$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

where the last inequality again follows from (51). Proceeding analogously when the entries of $(G_k)_{ii}$ are negative, we have

$$m_k(2-\nu)\|F_k\|^2 \leq F_k^T G_k F_k \leq M_k \nu \|F_k\|^2.$$

Hence, for sufficiently large k the scalars $F_k^T G_k F_k$ have constant sign and

$$|F_k^T G_k F_k| \geq \nu\, \widetilde{m}_k \|F_k\|^2 \geq (1-\nu)\nu\, \widetilde{m}(x^*)\|F_k\|^2. \tag{52}$$

Moreover,

$$\beta_k = \frac{\|p_{k-1}\|^2}{p_{k-1}^T(F_k - F_{k-1})} = \frac{\|p_{k-1}\|^2}{p_{k-1}^T \int_0^1 J(x_{k-1}+tp_{k-1})p_{k-1}\,dt} = \frac{\|p_{k-1}\|^2}{p_{k-1}^T G_{k-1} p_{k-1}}.$$

Thus, using similar arguments, for k large enough the scalars β_k have the same sign as $F_k^T G_k F_k$, and

$$|\beta_k| \in \left[\frac{1}{(2-\nu)\widetilde{M}_{k-1}}, \frac{1}{\nu\widetilde{m}_{k-1}}\right] \subseteq \left[\frac{1}{(1+\nu)(2-\nu)\widetilde{M}(x^*)}, \frac{1}{(1-\nu)\nu\widetilde{m}(x^*)}\right]. \tag{53}$$

Consequently,

$$\beta_k F_k^T G_k F_k = |\beta_k F_k^T G_k F_k| \geq \left(\frac{\nu}{2-\nu}\right) \frac{(1-\nu)\widetilde{m}(x^*)}{(1+\nu)\widetilde{M}(x^*)} \|F_k\|^2, \tag{54}$$

for k large enough. Now, without loss of generality, consider $p_k = -\lambda_k \beta_k F_k$. Lemma 5.1 then gives that

$$f(x_{k+1}) \leq f_k - 2\lambda_k\beta_k F_k^T G_k F_k + 2\lambda_k|\beta_k| \left| \int_0^1 (F(x_k+tp_k) - F_k)^T J(x_k+tp_k)F_k\,dt \right|, \tag{55}$$

and the last absolute value above satisfies (43). Denoting the diagonal and off diagonal part of a matrix as $\mathrm{diag}(\cdot)$ and $\mathrm{off}(\cdot)$ respectively, and using $\|J(z)\| \leq \|\mathrm{diag}(J(z))\| + \|\mathrm{off}(J(z))\|$ we obtain that

$$\|J(z)\| \leq \widetilde{M}(z) + \sqrt{n}\|\mathrm{off}(J(z))\|_\infty \leq \widetilde{M}(z)\left(1 + \sqrt{n}\max_{1\leq i\leq n}\zeta_i(z)\right).$$

Using this bound, (43), (53) and the fact that (50) implies that $\zeta_i(z) \leq 1$ in $\mathcal{B}(x^*,\rho)$, we deduce that, for k large enough,

$$\begin{aligned}
\left|\int_0^1 (F(x_k+tp_k) - F_k)^T J(x_k+tp_k)F_k\,dt\right| & \\
&\leq \frac{1}{2}\lambda_k|\beta_k| \max_{z\in\mathcal{B}(x^*,\rho)} \|J(z)\|^2\|F_k\|^2 \\
&\leq \frac{1}{2}\lambda_k|\beta_k|\|F_k\|^2 \max_{z\in\mathcal{B}(x^*,\rho)} \left[\widetilde{M}(z)\left(1+\sqrt{n}\max_{1\leq i\leq n}\zeta_i(z)\right)\right]^2 \\
&\leq \frac{(1+\sqrt{n})^2}{2\nu}\,\frac{((1+\nu)\widetilde{M}(x^*))^2}{(1-\nu)\widetilde{m}(x^*)}\lambda_k\|F_k\|^2.
\end{aligned}$$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

This bound, (55) and (53) then imply that

$$f(x_{k+1}) \;\leq\; \left[1-\left(\frac{2\nu}{2-\nu}\right)\frac{(1-\nu)\widetilde{m}(x^*)}{(1+\nu)\widetilde{M}(x^*)}\,\lambda_k+\frac{(1+\sqrt{n})^2}{\nu^2}\frac{((1+\nu)\widetilde{M}(x^*))^2}{(1-\nu)^2\widetilde{m}(x^*)^2}\,\lambda_k^2\right] f(x_k).$$

The linesearch condition (15) thus holds for k large enough and for any λ such that

$$1-\left(\frac{2\nu}{2-\nu}\right)\frac{(1-\nu)\widetilde{m}(x^*)}{(1+\nu)\widetilde{M}(x^*)}\,\lambda+\frac{(1+\sqrt{n})^2}{\nu^2}\frac{(1+\nu)^2\widetilde{M}(x^*)^2}{(1-\nu)^2\widetilde{m}(x^*)^2}\,\lambda^2\leq(1-\alpha\lambda)^2,$$

that is such that

$$\left(\frac{(1+\sqrt{n})^2}{\nu^2}\frac{(1+\nu)^2\widetilde{M}(x^*)^2}{(1-\nu)^2\widetilde{m}(x^*)^2}-\alpha^2\right)\lambda^2+2\left(\alpha-\left(\frac{\nu}{2-\nu}\right)\frac{(1-\nu)\widetilde{m}(x^*)}{(1+\nu)\widetilde{M}(x^*)}\right)\lambda\leq 0. \qquad (56)$$

Again,

$$\kappa_2 \stackrel{\rm def}{=} \frac{(1+\sqrt{n})^2}{\nu^2}\frac{(1+\nu)^2\widetilde{M}(x^*)^2}{(1-\nu)^2\widetilde{m}(x^*)^2}-\alpha^2>0,$$

by (47) and the fact that α and ν belong to $(0,1)$, and

$$\kappa_1 \stackrel{\rm def}{=} \alpha-\left(\frac{\nu}{2-\nu}\right)\frac{(1-\nu)\widetilde{m}(x^*)}{(1+\nu)\widetilde{M}(x^*)}<0$$

by (49). Thus (15) holds for $\lambda\leq\lambda_* \stackrel{\rm def}{=} -2\kappa_1/\kappa_2$ and all k sufficiently large, $\liminf_{k\to\infty}\lambda_k\geq\min[1,\sigma\lambda_*]>0$ and (23) finally allows us to deduce that $F(x^*)=0$. □

We conclude our investigation of some cases where the PAND algorithm can be proved to converge to a solution by showing that $\{\|F_k\|\}$ converges to zero if the limit point x^* lies in the interior of Ω and the step $p_k^{\rm QN}$ in (12) is, eventually, an Inexact Newton step.

Theorem 5.4 *Let Assumptions 2.1, 4.1 and 4.2 hold and $\{x_k\}$ be generated by the* PAND *algorithm. If the limit point x^* of $\{x_k\}$ is such that $x^*\in int(\Omega)$ and the step $p_k^{\rm QN}$ in (12) satisfies*

$$\|J_kp_k^{\rm QN}+F_k\|=\tau_k\|F_k\|,\qquad \tau_k\leq\tau_{\max}<1-\alpha, \qquad (57)$$

for all k sufficiently large, then $\lim_{k\to\infty}\|F_k\|=0$.

Proof. Let $\rho^*>0$ and $\rho\in(0,\rho^*)$ such that $\mathcal{B}(x^*,\rho)\subset int(\Omega)$. Since $\{x_k\}$ converges to x^*, we have $x_k\in\mathcal{B}\left(x^*,\frac{\rho}{2}\right)$ for all k sufficiently large. Suppose k is large enough so that $x_k\in\mathcal{B}\left(x^*,\frac{\rho}{2}\right)$. Then, possibly for λ small enough, $P(x_k+\lambda p_k^{\rm QN})-x_k=\lambda p_k^{\rm QN}$, i.e., $x_k+\lambda p_k^{\rm QN}$ belongs to the interior of Ω. In particular, if $\|\lambda p_k^{\rm QN}\|=\frac{\rho}{2}$, then $x_k+\lambda p_k^{\rm QN}\in\mathcal{B}(x^*,\rho)$. By using (21) and (25), and setting $\underline{\lambda}=\dfrac{\rho}{2c_Be^{\eta}\|F_0\|}$, independent of k, we get that equation $\|\lambda p_k^{\rm QN}\|=\frac{\rho}{2}$ is satisfied for some $\lambda\geq\underline{\lambda}$, namely $x_k+\lambda p_k^{\rm QN}$ belongs to the interior of Ω for some λ uniformly bounded away from zero.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Now we show that $\lambda p_k^{\rm QN}$ satisfies (15) for some λ uniformly bounded away from zero , which implies our claim. Since Assumption 4.2 holds, we know that

$$\begin{aligned} F(x_k + \lambda p_k^{\rm QN}) &= F(x_k) + \int_0^1 J(x_k + t\lambda p_k^{\rm QN})\lambda p_k^{\rm QN} dt \\ &= (1-\lambda)F(x_k) + \lambda(J(x_k)p_k^{\rm QN} + F(x_k)) + \\ &\quad \int_0^1 (J(x_k + t\lambda p_k^{\rm QN}) - J(x_k))\lambda p_k^{\rm QN} dt, \end{aligned}$$

see [9, Lemma 4.1.9]. Hence, using (57), the bounds (25) and (21) we obtain

$$\begin{aligned} \|F(x_k + \lambda p_k^{\rm QN})\| &\leq (1-\lambda)\|F_k\| + \lambda\tau_k\|F_k\| + L\lambda^2\|p_k^{\rm QN}\|^2 \\ &\leq (1 - \lambda + \lambda\tau_{\max})\|F_k\| + L\lambda^2 c_B^2\|F_k\|^2 \\ &\leq (1 - \lambda + \lambda\tau_{\max} + L\lambda^2\, c_B^2\, e^{\eta}\|F_0\|)\|F_k\|. \end{aligned}$$

Now observe that if

$$1 - \lambda + \lambda\tau_{\max} + L\lambda^2\, c_B^2\, e^{\eta}\|F_0\| \leq 1 - \alpha\lambda,$$

then (15) is satisfied and the step is accepted. In particular, if

$$\lambda \leq \lambda_* \stackrel{\rm def}{=} \frac{1 - \alpha - \tau_{\max}}{L\, c_B^2\, e^{\eta}\|F_0\|},$$

then (15) is fulfilled for all k sufficiently large. Now, considering Step 3.6 of PAND algorithm, we conclude that the repeat-loop at Step 3 terminates with $\lambda \geq \min\{1, \sigma\lambda_*\}$, λ_* independent of k. Combining this bound with $\lambda \geq \underline{\lambda}$, we get $\liminf_{k\to\infty} \lambda_k > 0$ and (23) implies $\lim_{k\to\infty} \|F_k\| = 0$. □

6 Numerical experiments

In this section we present the results of some numerical experiments conducted with different implementations of the PAND algorithm. Our goal is to test its behaviour in terms of robustness and computational cost and to compare it with PSANE [20].

6.1 The problem sets

We considered two sets of problems: the first comprises small and medium-size smooth nonlinear systems with box constraints from a variety of applications; the second is made of semismooth systems with nonnegative constraints which reformulate well-known nonlinear complementarity problems from the literature.

6.1.1 Bound-constrained nonlinear systems

We selected 14 constrained nonlinear systems listed in Table 1 along with their description and dimension. The convex set Ω in (1) is the n-dimensional box $\{x \in \mathbb{R}^n \ \text{ s.t. } \ l \leq x \leq u\}$, where $l \in (\mathbb{R} \cup -\infty)^n$, $u \in (\mathbb{R} \cup \infty)^n$, and the inequalities are meant component-wise. Therefore, the projection map is given by $P(x) = \max[l, \min[x, u]]$.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Pb#	Name and Source	n
1	Himmelblau function [14, 14.1.1]	2
2	Equilibrium Combustion [14, 14.1.2]	5
3	Bullard-Biegler system [14, 14.1.3]	2
4	Ferraris-Tronconi system [14, 14.1.4]	2
5	Brown's almost linear system [14, 14.1.5]	5
6	Robot kinematics problem [14, 14.1.6]	8
7	Series of CSTRs, $R = .945$ [14, 14.1.8]	2
8	Series of CSTRs, $R = .990$ [14, 14.1.8]	2
9	Chandrasekar's H-equation, $c = 0.9999$ [23, Problem 6]	1000
10	Problem 74 [25]	1000
11	Problem 77 [25]	2000
12	Trigonometric function [22, Test 8]	2000
13	Function 15 [22, Problem 15]	2000
14	Zero Jacobian function [22, Problem 19]	2000

Table 1: Bound-constrained nonlinear system.

The first 8 problems have been frequently used as a test set and are fully described in [14]; their dimension is small. The remaining problems have variable dimension and their Jacobian matrices cannot be formed at a low computational cost by finite difference procedures for sparse matrices such as [8]. Hence, computing $B_k = J(x_k)$ by finite differences is expensive and solving (12) with such B_k's cannot take advantage of sparse/structured linear algebra solvers. As for the definition of Ω, it is the positive orthant for problem 9; $l = (0, \ldots, 0)^T$, $u = (10, \ldots, 10)^T$ in problems 10, 11 and 14; $l = (5, \ldots, 5)^T$, $u = (15, \ldots, 15)^T$ in problem 12, $l = (-10, \ldots, -10)^T$, $u = (0, \ldots, 0)^T$ in problem 13.

All problems were run starting from three different initial guesses x_0 given by

$$(x_0)_i = \begin{cases} l_i + \gamma(u_i - l_i)/2 & \text{if } l_i > -\infty \text{ and } u_i < \infty, \qquad \gamma = 1, 2, 3, \\ l_i + \gamma\, 10^\gamma & \text{if } l_i > -\infty \text{ and } u_i = \infty, \qquad \gamma = 0, 1, 2. \end{cases}$$

6.1.2 Nonlinear complementarity problems

We consider the nonlinear complementarity problems listed in Table 4 and defined as

$$G(x)^T x = 0, \quad x \geq 0,\ G(x) \geq 0,$$

where $G : \mathbb{R}^n \to \mathbb{R}^n$ is continuous differentiable. Following [20], we solved the following nonlinear systems with nonnegative constraints

$$F(x) = \min[x, G(x)] = 0, \quad x \in \Omega,$$

where is the positive orthant and the function F is continuous but not everywhere differentiable. All run were started using $x_0 = 10^\gamma$ with $\gamma = 0, 1, 2$.

6.2 Implementation issues and numerical results

All the tested algorithms have been implemented in MATLAB and run using MATLAB R2015A version on a Intel(R) Core(TM) i5-6600K CPU @3.50 GHz x 4, 16.0 GB RAM.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

Pb#	Name and Source	n
15	Kojima-Shindo's problem [11]	3
16	Josephy's problem [11]	4
17	Mathiensen's problem [11]	4
18	Harker's Nash-Cournot-5 problem [18]	5
19	Harker's Nash-Cournot-10 problem [18]	10
20	Pang and Murphy's Nash-Cournot-5 problem [28]	5
21	Pang and Murphy's Nash-Cournot-10 problem [28]	10

Table 2: Nonlinear complementarity problems.

The main implementation issues are as follows. Two rules for choosing matrices B_k were implemented. The former corresponds to the choice made in PSANE, i.e. $B_k = \beta_k^{-1} I$ with β_k given in (10), and the resulting implementation is named PAND-SR (PAND algorithm with Spectral Residual step). The latter consists in starting from a given B_0 and generating matrices B_k by using the Broyden's formula

$$B_{k+1} = B_k + \frac{(y_k - B_k p_k) p_k^T}{p_k^T p_k}, \tag{58}$$

where $y_k = F(x_{k+1}) - F(x_k)$, see [6]. The resulting implementation is named PAND-BR (PAND algorithm with BRoyden step).

In order to perform a fully derivative-free implementation of PAND-BR, our default choice for B_0 was the identity matrix. The current matrix B_k was refreshed and set equal to the identity matrix every 30 iterations and whenever $\|v_k^{\mathrm{QN}}\| = 0$.

The linear systems (12) arising in PAND-BR were solved via QR factorizations. Specifically, given the QR factorization of B_k, the QR factorization of B_{k+1} was formed by the rank one update (58) using the MATLAB function `qrupdate`.

The parameters used in PAND-SR and PAND-BR were set equal to those declared in PSANE, i.e. $\beta_{\min} = 10^{-30}$, $\beta_{\max} = 10^{30}$, $\beta_0 = 1$, $\alpha = 10^{-4}$, $\sigma = 0.5$, $\eta_k = 0.99^k(100 + \|F(x_0)\|^2)$, $k \geq 0$. This allows comparing PAND-SR , PAND-BR and PSANE in terms of their distinctive features, i.e. definition of the search directions and linesearch strategy. Following [20], PSANE was tested using $\lambda_{\max} = 1$, and $\lambda = \lambda_{\max}$ in Step 3 of Algorithm 2.1.

Tables 3 and 4 collect the results obtained with PSANE, PAND-SR and PAND-BR. The problem number refers to Tables 1 and 2 and the scalar γ is associated to the starting point.We report the number of iterations (`It`) and F-evaluations (`Fe`) performed on successful runs, i.e., runs where the criterion

$$\|F_k\| \leq 10^{-6}, \tag{59}$$

was met within a maximum number of iterations (`maxIt`) and function evaluation (`maxFe`) equal to 10^5 as in [20]. For the remaining runs, we eventually stopped the iterations on the base of the behaviour of λ_k and $\|F_k\|$: the symbol F_λ indicates that λ has been reduced 40 times by a factor σ in the linesearch strategy; the symbol $\mathrm{F_i}$ indicates that

$$\|F_{k+1}\| > (1 - \alpha)\|F_k\|,$$

occurred for 50 iterations consecutively, i.e., repeatedly $\|F_k\|$ either increased or slightly decreased. We remark that the occurrences F_λ and $\mathrm{F_i}$ are suggested by the convergence

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

properties of the PAND algorithm presented in Theorem 4.2. For large values of k, the first occurrence may indicate that $\{\lambda_k\}$ is converging to zero while the second occurrence may indicate that $\{\|F_k\|\}$ does not converge to zero. Breakdowns in PSANE, described in §2, are denoted as $\mathrm{F_b}$.

The reported results show that on a total of 63 tests, PSANE and PAND-SR fail 22 and 9 times respectively, while PAND-BR solves all the tests. Most of the failures in PSANE are due to a breakdown ($\mathrm{F_b}$); on successful runs the performance of PSANE is quite similar to that of PAND-SR algorithm. In several runs where PAND-SR is successful, its computational cost is comparable to that of PAND-BR procedure in terms of F-evaluations, but the former is more efficient as it does not require forming and solving linear systems. This fact is shown in Table 5 where we report the CPU times of the methods under analysis on problems with dimension larger than or equal to 1000. On the other hand, the version of PAND based on Broyden matrices is more robust as it allows to avoid failures of the spectral residual procedures on Problems 2, 6 and 17.

Finally, Figure 1 shows the nonmonotone behaviour of $\|F_k\|$ observed in two runs performed and is representative of the tests presented.

7 Conclusion

We have proposed a new class of derivative-free methods for the solution of constrained nonlinear systems which combines the use of simple search directions with a new suitable approximate norm linesearch. The methods are suitable for both continuous and/or differentiable nonlinear systems and their convergence properties have been studied in both cases. In particular, we have focused on methods based on spectral residual steps (PAND-SR) and Quasi-Newton directions (PAND-BR). These methods exhibit good numerical performance on relatively large problems. PAND-SR has turned out to be very efficient and competitive with PAND-BR; on the other hand PAND-BR has solved a larger set of problems than PAND-SR.

References

[1] M. Ahookhosh, K. Amini, S. Bahrami, *Two derivative-free projection approaches for systems of large-scale nonlinear monotone equations*, Numer. Algorithms, 64 (2013), 1, pp 21-42.

[2] J. Barzilai, J. Borwein, *Two-point step size gradient*, IMA J. Numer. Anal., 8 (1988), pp. 141-148.

[3] S. Bellavia, M. Macconi, B. Morini, *An affine scaling trust-region approach to bound-constrained nonlinear systems*, Appl. Numer. Math., 44 (2003), pp. 257-280.

[4] S. Bellavia, S. Pieraccini, *On affine-scaling inexact dogleg methods for bound-constrained nonlinear systems*, Optim. Method. Softw., 30(2015), pp. 276-300.

[5] E.G. Birgin, N. Krejić, J.M. Martínez, *Globally convergent inexact quasi-Newton methods for solving nonlinear equations*, Numer. Algorithms, 32 (2003), pp. 249–260.

[6] C.G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comput., 19 (1965), pp. 577-593.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

		PSANE		PAND-SR		PAND-BR	
Pb#	γ	It	Fe	It	Fe	It	Fe
1	1	11	14	12	15	14	18
	2	11	18	12	16	11	14
	3	14	25	17	23	14	20
2	1	F_λ		F_λ		284	433
	2	F_λ		F_λ		54	80
	3	F_λ		F_λ		119	180
3	1	F_b		26	41	14	19
	2	F_b		192	319	58	88
	3	F_b		1090	1817	1581	2568
4	1	25	26	27	46	10	12
	2	31	32	24	42	106	164
	3	F_b		23	39	28	39
5	1	26	27	26	34	13	15
	2	26	27	26	35	12	15
	2.5	19	22	26	35	11	13
6	1	F_i		F_i		144	234
	2	F_λ		F_i		46	69
	3	F_b		F_i		44	62
7	1	F_b		642	2427	51	79
	2	F_λ		430	849	546	1316
	3	F_b		825	1426	659	1098
8	1	13	14	9	13	6	9
	2	12	14	12	16	7	10
	3	11	13	11	14	8	11
9	0	30	31	30	41	13	14
	1	60	61	122	192	15	16
	2	37	38	37	50	15	16
10	1	14	15	14	16	13	14
	2	20	25	18	24	43	50
	3	29	31	15	19	16	19
11	1	F_b		8	11	34	61
	2	F_b		7	10	15	19
	3	F_b		6	9	16	20
12	1	92	99	21	24	2937	6911
	2	56	59	24	27	2736	6506
	3	372	645	29	35	1728	4858
13	1	115	116	243	375	418	596
	2	511	521	447	745	497	711
	3	71	72	221	356	412	586
14	1	F_b		19	22	2	4
	2	F_b		20	23	2	4
	3	F_b		20	23	2	4

Table 3: Computational results obtained with PSANE, PAND-SR and PAND-BR algorithms on bound-constrained nonlinear systems

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

		PSANE		PAND-SR		PAND-BR	
Pb#	γ	It	Fe	It	Fe	It	Fe
15	0	181	371	75	108	15	20
	1	110	111	110	167	22	32
	2	29	30	29	39	30	40
16	0	24	25	24	33	14	18
	1	22	23	22	28	19	24
	2	21	22	21	26	15	18
17	0	F_λ		F_λ		9	15
	1	F_i		F_λ		45	63
	2	F_i		F_λ		41	60
18	0	6	11	1	3	1	3
	1	20	30	24	39	26	35
	2	1	2	1	2	1	2
19	0	3	4	3	4	6	8
	1	63	72	97	132	260	375
	2	2	5	4	6	4	6
20	0	21	22	21	23	22	23
	1	16	17	16	17	15	16
	2	17	18	17	19	15	16
21	0	24	25	24	27	46	52
	1	27	28	27	34	40	43
	2	22	23	22	26	49	57

Table 4: Computational results obtained with PSANE, PAND-SR and PAND-BR algorithms on nonlinear complementarity problems.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

		Execution time		
Pb#	γ	PSANE	PAND-SR	PAND-BR
9	0	0.20	0.26	0.27
	1	0.38	1.23	0.36
	2	0.24	0.32	0.32
10	1	0.09	0.09	0.31
	2	0.15	0.14	1.00
	3	0.18	0.11	0.42
11	1	F_b	0.01	2.24
	2	F_b	0.01	1.04
	3	F_b	0.01	0.96
12	1	0.08	0.02	182.80
	2	0.04	0.02	170.73
	3	0.50	0.02	108.98
13	1	0.06	0.20	25.73
	2	0.28	0.40	30.26
	3	0.04	0.16	25.23
14	1	F_b	0.01	0.08
	2	F_b	0.01	0.08
	3	F_b	0.01	0.08

Table 5: CPU time (in seconds) obtained with PSANE, PAND-SR and PAND-BR algorithms on bound-constrained nonlinear systems.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

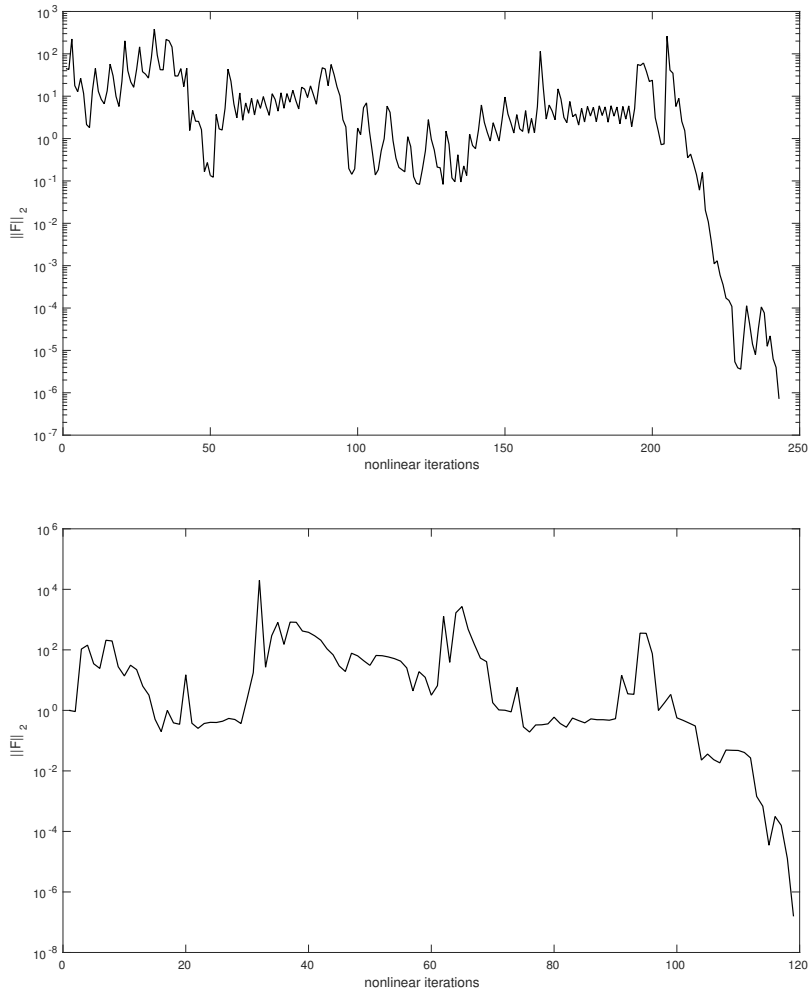


Figure 1: Norm of F on a log scale against the number of iterations. Top: problem 13 solved by PAND-SR, $\gamma = 1$. Bottom: problem 2 solved by PAND-BR, $\gamma = 3$

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

[7] A.R. Conn, N.I.M. Gould, P.L. Toint, *Testing a class of methods for solving minimization problems with simple bounds on the variables*, Math. Comput., 60 (1988), pp. 399-430.

[8] A.R. Curtis, M.J.D. Powell, J.K. Reid, *On the estimation of sparse Jacobian matrices*, J. Inst. Maths Applics (1974), 13, pp. 117-119.

[9] J.E. Dennis, R.B. Schnabel, *Numerical methods for unconstrained optimization and nonlinear equations*, Prentice Hall, Englewood Cliffs, NJ, 1983.

[10] J.E. Dennis, J. Moré, *A characterization of superlinear convergence and its application to Quasi-Newton methods*, Math. Comput., 28 (1974), pp. 549-560.

[11] S.P. Dirkse, M.C. Ferris, *MCPLIB: A Collection of Nonlinear Mixed Complementary Problems*, Technical Report, Computer Sciences Department, University of Wisconsin, Madison, Wisconsin, 1994.

[12] N. Echebest, M.L. Schuverdt, R.P. Vignau, *A derivative-free method for solving box-constrained underdetermined nonlinear systems of equations*, Appl. Math. Comput., 219 (2012), 6, pp. 3198-3208.

[13] S.C. Eisenstat, H.F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim. 4 (1994), pp. 393-422.

[14] C.A. Floudas, P.M. Pardalos, C.S. Adjiman, W.R. Esposito, Z.H. Gumus, S.T. Harding, J.L. Klepeis, C.A. Meyer, C.A. Schweiger, *Handbook of test problems in local and global optimization*, Kluwer Academic Publishers, Nonconvex Optimization and its Applications, 33, 1999.

[15] A. Griewank, *The global convergence of Broyden-like methods with a suitable line search*, J. Aust. Math. Soci., Series B 28, 1 (1986), pp. 75-92.

[16] G.H. Golub, C.F. van Loan, *Matrix Computation*, The Johns Hopkins University Press, Baltimore, 1983.

[17] L. Grippo, M. Sciandrone, *Nonmonotone derivative-free methods for nonlinear equations*, Comput. Optim. Appl., 27 (2007), pp. 297-328.

[18] P.T. Harker, *Accelerating the convergence of the diagonalization and projection algorithms for finite-dimensional variational inequalities*, Math. Program. 41 (1988), pp. 29-59.

[19] C. Kanzow, N. Yamashita, M. Fukushima, *Levenberg-Marquardt methods with strong local convergence properties for solving nonlinear equations with convex constraints*, J. Comput. Appl. Math., 172 (2004), pp. 375-397.

[20] W. La Cruz, *A projected derivative-free algorithm for nonlinear equations with convex constraints*, Optim. Method. Softw., 29 (2014), pp. 24-41.

[21] W. La Cruz, J.M. Martínez, M. Raydan, *Spectral residual method without gradient information for solving large-scale nonlinear systems of equations*, Math. Comput., 18 (2006), pp. 1429-1448.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.

[22] W. La Cruz, M. Raydan, *Nonmonotone spectral methods for large-scale nonlinear systems*, Optim. Method. Softw., 18 (2003), pp. 583-599.

[23] D.H. Li, M. Fukushima, *A derivative-free line search and global convergence of Broyden-like method for nonlinear equations*, Optim. Method. Softw., 13 (2000), pp. 181-201.

[24] L. Lukšan, J. Vlček, *Computational experience with globally convergent descent methods for large sparse systems of nonlinear equations*, Optim. Method. Softw., 8 (1998), pp. 201-223.

[25] L. Lukšan, J. Vlček, *Test Problems for Unconstrained Optimization*, Tech. Rep. V-897, ICS AS CR, November 2003.

[26] M. Macconi, B. Morini, M. Porcelli, *Trust-region quadratic methods for nonlinear systems of mixed equalities and inequalities*, Appl. Numer. Math., 59 (2009), pp. 859-876.

[27] J.M. Martínez, *Practical Quasi-Newton methods for solving nonlinear systems*, J. Comput. Appl. Math., 124 (2000),pp. 97-121.

[28] F.H. Murphy, H.D. Sherali, and A.L. Soyster, *A mathematical programming approach for determining oligopolistic market equilibrium*, Math. Program. 24 (1982), pp. 92-106.

[29] M. Porcelli, *On the convergence of an inexact Gauss-Newton trust-region method for nonlinear least-squares problems with simple bounds*, Optim. Lett., 7:3 (2013), pp. 447–465.

[30] P. Wang, D. Zhu, *An inexact derivative-free Levenberg-Marquardt method for linear inequality constrained nonlinear systems under local error bound conditions*, Appl. Math. Comput., 282 (2016), pp. 32-52.

[31] D. Zhu, *An affine scaling trust-region algorithm with interior backtracking technique for solving bound-constrained nonlinear systems*, J. Comput. Appl. Math, 184 (2005), pp. 343-361.

This is a pre-publication version of this article, which may differ from the final published version. Copyright restrictions may apply.