Noname manuscript No.
(will be inserted by the editor)

Convergence properties of the Broyden–like method for mixed linear–nonlinear systems of equations

Florian Mannel

Received: date / Accepted: date

Abstract We consider the Broyden–like method for a nonlinear mapping $F : \mathbb{R}^n \to \mathbb{R}^n$ that has some affine component functions, using an initial matrix B_0 that agrees with the Jacobian of F in the rows that correspond to affine components of F. We show that in this setting the iterates belong to an affine subspace and can be viewed as outcome of the Broyden–like method applied to a lower-dimensional mapping $G : \mathbb{R}^d \to \mathbb{R}^d$, where d is the dimension of the affine subspace. We use this subspace property to make some small contributions to the decades–old question of whether the Broyden–like matrices converge: First, we observe that the only available result concerning this question cannot be applied if the iterates belong to a subspace because the required uniform linear independence does not hold. By generalizing the notion of uniform linear independence to subspaces we can extend the available result to this setting. Second, we infer from the extended result that if at most one component of F is nonlinear while the others are affine and the associated $n-1$ rows of the Jacobian of F agree with those of B_0, then the Broyden–like matrices converge if the iterates converge; this holds whether the Jacobian at the root is invertible or not. In particular, this is the first time that convergence of the Broyden–like matrices is proven for $n > 1$, albeit for a special case only. Third, under the additional assumption that the Broyden–like method turns into Broyden's method after a finite number of iterations, we prove that the convergence order of iterates and matrix updates is bounded from below by $\frac{\sqrt{5}+1}{2}$ if the Jacobian at the root is invertible. If the nonlinear component of F is actually affine, we show finite convergence. We provide high-precision numerical experiments to confirm the results.

Keywords Broyden–like method · Broyden's method · convergence of Broyden–like matrices · quasi-Newton methods · uniform linear independence

Mathematics Subject Classification (2010) 49M15 · 65H10 · 65K05 · 90C30 · 90C53

F. Mannel
University of Graz
E-mail: florian.mannel@uni-graz.at

1 Introduction

This work is devoted to convergence properties of the Broyden–like method for systems of equations in which some of the equations are linear. Among others it provides the first answer to the decades–old question whether the Broyden–like matrices converge under the standard assumptions for q-superlinear convergence of the iterates, albeit for a special case only.

Given a smooth nonlinear mapping $F : \mathbb{R}^n \to \mathbb{R}^n$, Broyden's method [3] aims at finding $\bar{u} \in \mathbb{R}^n$ with

$$F(\bar{u}) = 0.$$

It is a well-established member of the class of quasi-Newton methods and shares its local q-superlinear convergence, cf. [9,15,21,23,14]. The *Broyden–like* method generalizes Broyden's method by allowing an additional parameter σ_k in the matrix update. It reads as follows.

Algorithm BL: Broyden–like method

Input: $(u^0, B_0) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$, $0 < \sigma_{\min} < \sigma_{\max} < 2$
for $k = 0, 1, 2, \dots$ **do**
  if $F(u^k) = 0$ **then** let $u^* := u^k$; STOP
  Solve $B_k s^k = -F(u^k)$ for s^k
  Let $u^{k+1} := u^k + s^k$ and $y^k := F(u^{k+1}) - F(u^k)$
  Choose $\sigma_k \in [\sigma_{\min}, \sigma_{\max}]$
  Let $B_{k+1} := B_k + \sigma_k (y^k - B_k s^k) \frac{(s^k)^T}{\|s^k\|^2}$
end
Output: u^*

For $(\sigma_k) \equiv 1$ we recover Broyden's method. An appropriate choice of σ_k ensures that B_{k+1} is invertible if B_k is invertible. In fact, by the Sherman–Morrison formula all choices but one maintain invertibility. The Broyden–like method is well-known, cf. [22], [28, Section 6] and [16, Algorithm 1].

In this work we consider Algorithm BL for mixed linear–nonlinear systems of equations. That is, there exists $J \subset \{1, \dots, n\}$ such that $F_j(u) = a_j^T u + b_j$, where $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$ for all $j \in J$. In addition, we suppose that the initial matrix B_0 agrees with the Jacobian of F in the rows that correspond to (some of) the affine components of F, i.e., $B_0^j = a_j^T$ for all $j \in J$. For $j \notin J$ the functions F_j can be nonlinear and B_0^j is not restricted. This framework includes many practically relevant systems of equations. Also, it fits two standard suggestions for the choice of B_0, which are to use $B_0 = F'(u^0)$ or a finite difference approximation of $F'(u^0)$. In the following we speak of *exact initialization* if $B_0^j = a_j^T$ for all $j \in J$.

This article is divided into four parts. In the first part we show that exact initialization ensures that the steps $(s_k)_{k\geq 1}$ stay in a subspace $\mathcal{S}$ and that they can be generated by applying Algorithm BL to a lower-dimensional mapping $G : \mathbb{R}^d \to \mathbb{R}^d$, where d is the dimension of $\mathcal{S}$. This extends results from [18].

The second part is concerned with the consequences of the first part for the convergence of the Broyden–like matrices (B_k). We point out that it is still largely open if (B_k) converges and that several renowned researchers have mentioned this issue in their works, cf. the survey articles [8, Example 5.3], [21, p. 117], [14, p. 306] and [2, p. 940]. The convergence of (B_k) is for example of interest because it is closely related to the rate of convergence of (u^k), see e.g. Lemma 2 and 3. For invertible $F'(\bar{u})$ there is only one result available: It is established in [22, Theorem 5.7] and in [17] that if the sequence of steps (s^k) is *uniformly linearly independent*, then (B_k) converges and $\lim_{k\to\infty} B_k = F'(\bar{u})$. We include the precise result as Theorem 4. Unfortunately, conditions that imply uniform linear independence of (s^k) are unknown and we are not aware of a single example—be it theoretical or numerical—in which (s^k) is actually uniformly independent. In the setting of this work, anyway, $(s^k)_{k\geq 1}$ is confined to the subspace $\mathcal{S}$ and thus violates uniform linear independence. After extending the notion of uniform linear independence to subspaces we generalize the above convergence result for (B_k) to the setting of this work, cf. Theorem 5. In doing so we also obtain a formula for the limit of (B_k).

In the third part we observe that if F has only one nonlinear component function and B_0 is initialized exactly, then the generalized convergence result from the second part implies that (B_k) converges whenever the iterates (u^k) converge, and this holds for regular and for singular $F'(\bar{u})$, cf. Corollary 2. Since the assumption of only one nonlinear component function is very restrictive, we stress that this is the first time that convergence of (B_k) is shown for $n > 1$ and invertible $F'(\bar{u})$. We will also see that even though each B_k agrees with $F'(\bar{u})$ in $n-1$ of n rows, the limit of (B_k) is generally not $F'(\bar{u})$.

We continue the third part by paying special attention to the case that $\sigma_k = 1$ for all $k \geq k_0$ and some $k_0 \geq 0$, i.e., Algorithm BL turns into Broyden's method. The result of the first part implies that in this case Broyden's method essentially reduces to the one-dimensional secant method. This yields a comprehensive characterization of the convergence of (u^k) including a lower bound for its q-order, which in turn allows us to establish significantly stronger convergence properties of (B_k) than for the Broyden–like method, cf. Theorem 6. For affine F we prove finite convergence if $\sigma_k = 1$ is selected at least once, cf. Theorem 7. The third part concludes with a brief application of the developed convergence theory to two examples from the literature.

In the last part we verify the results from the third part in numerical experiments with high precision. Among others, we find that if $F'(\bar{u})$ is invertible, then choosing $(\sigma_k)_{k\geq k_0} \equiv 1$ for some $k_0 \geq 0$ leads to much faster convergence than, e.g., $(\sigma_k) \equiv 0.99$, while this is not the case if $F'(\bar{u})$ is not invertible.

The convergence theory of Broyden's method and specific versions of the Broyden–like method are developed in, e.g., [4,22,12,16]. There is only one further result available on the convergence of the Broyden(–like) matrices besides the one mentioned above: In [19] it was recently shown for Broyden's method that if $F'(\bar{u})$ is singular with some additional structure, then $(\|B_{k+1} - B_k\|)$ converges q-linearly to zero under appropriate assumptions, so (B_k) converges.

For other quasi-Newton updates convergence results are available. We are aware of results for the SR1 update [5,11,30], for the Powell-symmetric-Broyden update [26], for the DFP and the BFGS update [13], and for the convex Broyden class excluding the DFP update [29].

This paper is organized as follows. In section 2 we collect preparatory results and we present the generalization of uniform linear independence that is useful for subspaces. In section 3 we prove the subspace property of $(s^k)_{k\geq 1}$ and show that $(s^k)_{k\geq 1}$ can be obtained by applying Algorithm BL to a suitable mapping $G : \mathbb{R}^d \to \mathbb{R}^d$. Section 4 contains the convergence results for the Broyden–like matrices and the application to examples from the literature. Section 5 presents numerical experiments and section 6 summarizes.

Notation We use $\mathbb{N} = \{1, 2, 3, \ldots\}$. For $n \in \mathbb{N}$ we set $[n] := \{1, 2, \ldots, n\}$, $[n]_0 := [n] \cup \{0\}$ and $[0] := \emptyset$. The Euclidean norm of $v \in \mathbb{R}^n$ is $\|v\|$, while $\|A\|$ is the spectral norm if $A \in \mathbb{R}^{m\times n}$. For $A \in \mathbb{R}^{m\times n}$, A^j indicates the j-th row of A, regarded as a row vector, whereas $A^{i,j} \in \mathbb{R}$ is the usual notation for entries. The span of $C \subset \mathbb{R}^n$ is indicated by $\langle C\rangle$. We will use tacitly that Algorithm BL cannot generate a step s^k satisfying $s^k = 0$. For $k \geq 0$ we define

$$E_k := B_k - F'(\bar{u}) \qquad \text{and} \qquad \hat{s}^k := \frac{s^k}{\|s^k\|},$$

where the first definition assumes that Algorithm BL has generated (B_k) and (u^k) with $\lim_{k\to\infty} u^k = \bar{u}$ for some $\bar{u}$ at which F is differentiable, while the second definition already makes sense if Algorithm BL has generated s^k. We employ the *q-order of convergence* and the *r-order of convergence* in this work. They are studied in, e.g., [25, Section 9].

2 Preliminaries

2.1 Convergence of the Broyden–like method

The main convergence result for Algorithm BL reads as follows.

Theorem 1 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable in a neighborhood of $\bar{u}$ with $F(\bar{u}) = 0$ and let $\|F'(u) - F'(\bar{u})\| \leq L\|u - \bar{u}\|^\alpha$ for all u from this neighborhood and constants $L, \alpha > 0$. Let $F'(\bar{u})$ be invertible. If Algorithm BL generates a sequence (u^k) that satisfies $\sum_k \|u^k - \bar{u}\|^\alpha < \infty$, then there holds*

$$\sum_{k=0}^{\infty} \left(\frac{\|u^{k+1} - \bar{u}\|}{\|u^k - \bar{u}\|} \right)^2 < \infty, \tag{1}$$

implying that (u^k) converges q-superlinearly to $\bar{u}$.

Moreover, there are $\delta, \varepsilon > 0$ such that for every (u^0, B_0) with $\|u^0 - \bar{u}\| \leq \delta$ and $\|B_0 - F'(\bar{u})\| \leq \varepsilon$, Algorithm BL either terminates with output $u^* = \bar{u}$ or it generates (u^k) such that all B_k are invertible and $\sum_k \|u^k - \bar{u}\|^\alpha < \infty$.

Proof This follows from [20, Theorem 1]. □

If we restrict attention to Broyden's method instead of the Broyden–like method, then a stronger result is available, namely Gay's theorem on $2n$–step q-quadratic convergence [12, Theorem 3.1]. For mixed linear–nonlinear systems with exact initialization, this result has recently been generalized.

Theorem 2 *Let $n \in \mathbb{N}$, $d \in [n]_0$ and $J := [n] \setminus [d]$. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ satisfy $F_j(u) = a_j^T u + b_j$ for all $j \in J$, where $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$ for all $j \in J$. Let F be differentiable in a neighborhood of $\bar{u}$ with $F(\bar{u}) = 0$ and let $\|F'(u) - F'(\bar{u})\| \leq L\|u - \bar{u}\|$ for all u from this neighborhood and a constant $L > 0$. Let $F'(\bar{u})$ be invertible. Then there are $\delta, \varepsilon > 0$ and $C > 0$ such that for every (u^0, B_0) with $\|u^0 - \bar{u}\| \leq \delta$, $\|B_0 - F'(\bar{u})\| \leq \varepsilon$, and $B_0^j = a_j^T$ for all $j \in J$, Algorithm BL with $(\sigma_k) \equiv 1$ either terminates with output $u^* = \bar{u}$ or it generates (u^k) that satisfies (1) and*

$$\left\|u^{k+2d} - \bar{u}\right\| \leq C \left\|u^k - \bar{u}\right\|^2 \qquad \forall k \geq 1.$$

In particular, (u^k) converges q-superlinearly and with r-order at least $2^{1/(2d)}$ to $\bar{u}$ and all B_k are invertible.

Proof See [18]. □

2.2 Convergence of the Broyden–like updates

If (u^k) and the Broyden–like updates converge, then $F(\lim_{k\to\infty} u^k) = 0$.

Lemma 1 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be continuous at $\bar{u}$. Let (u^k) and (B_k) be generated by Algorithm BL. Suppose that $u^k \to \bar{u}$ and $\sup_{k\geq 0}\|B_{k+1} - B_k\| < \infty$. Then $F(\bar{u}) = 0$.*

Proof From $\sup_{k\geq 0}\|B_{k+1} - B_k\| < \infty$ we infer $\sup_{k\geq 0} \frac{\|F(u^{k+1})\|}{\|s^k\|} < \infty$. The convergence of (u^k) yields $\lim_{k\to\infty}\|s^k\| = 0$, so $\lim_{k\to\infty}\|F(u^k)\| = 0$, whence $F(\bar{u}) = 0$. □

If (u^k) and the Broyden–like matrices converge, then the convergence of (u^k) is q-superlinear.

Lemma 2 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable at $\bar{u}$ with $F'(\bar{u})$ invertible. Let (u^k) and (B_k) be generated by Algorithm BL. Suppose that $u^k \to \bar{u}$ and $\|B_{k+1} - B_k\| \to 0$ for $k \to \infty$. Then (u^k) converges q-superlinearly to $\bar{u}$.*

Proof Due to the invertibility of $F'(\bar{u})$ and $u^k \to \bar{u}$ there is $C > 0$ such that

$$\begin{aligned}\left\|u^{k+1} - \bar{u}\right\| &\leq C \left\|F(u^{k+1}) - F(\bar{u})\right\| = \frac{C}{\sigma_k} \|B_{k+1} - B_k\| \left\|s^k\right\| \\ &\leq \frac{C}{\sigma_{\min}} \|B_{k+1} - B_k\| \left(\left\|u^{k+1} - \bar{u}\right\| + \left\|u^k - \bar{u}\right\|\right)\end{aligned}$$

for all k sufficiently large. Here, we also used that $F(\bar{u}) = 0$ by Lemma 1. Subtracting $\frac{C}{\sigma_{\min}}\|B_{k+1} - B_k\|\|u^{k+1} - \bar{u}\|$ and taking the limit yields the claim. □

Next we show that convergence of (u^k) with q-order at least $\gamma > 1$ implies convergence of $(\|B_{k+1} - B_k\|)$ with r-order at least γ, cf. also [25, 9.1.8&9.2.7].

Lemma 3 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ and let (u^k) be generated by Algorithm BL. Suppose that (u^k) converges to some $\bar{u}$ and that F satisfies $\|F(u) - F(\bar{u})\| \leq L\|u - \bar{u}\|$ for all u in a neighborhood of $\bar{u}$ and some constant $L > 0$. Let $\gamma > 1$.*

1) If $F(\bar{u}) = 0$ and there is $C > 0$ such that for all k sufficiently large

$$\left\|u^{k+1} - \bar{u}\right\| \leq C \left\|u^k - \bar{u}\right\|^{\gamma} \tag{2}$$

is satisfied, then there exists $\hat{C} > 0$ such that

$$\|B_{k+1} - B_k\| \leq \hat{C} \left\|u^k - \bar{u}\right\|^{\gamma - 1} \tag{3}$$

for all sufficiently large k.

2) If $C, \hat{C} > 0$ exist such that (2) and (3) are satisfied for all sufficiently large k, then we have $F(\bar{u}) = 0$ and $\lim_{k\to\infty}\|B_{k+1} - B_k\|^{\frac{1}{p^k}} = 0$ for all $p \in [1, \gamma)$. In particular, $\sum_k \|B_{k+1} - B_k\| < \infty$ and (B_k) converges.

Proof **Proof of 1):** Since (2) implies q-superlinear convergence of (u^k), we obtain from a well-known result of Dennis and Moré that $\|u^k - \bar{u}\|/\|s^k\| \to 1$ for $k \to \infty$, cf. [7, Lemma 2.1]. The Lipschitz–type property of F at $\bar{u}$, $F(\bar{u}) = 0$ and (2) hence yield

$$\|B_{k+1} - B_k\| = \sigma_k \frac{\|F(u^{k+1}) - F(\bar{u})\|}{\|s^k\|} \leq \hat{C} \left\|u^k - \bar{u}\right\|^{\gamma - 1}$$

for all sufficiently large k and a constant $\hat{C} > 0$, which proves (3).
Proof of 2): Lemma 1 yields $F(\bar{u}) = 0$ due to (3). To prove the remaining claims it suffices to establish that

$$\lim_{k\to\infty} \left(\|B_{k+1} - B_k\|^{\frac{1}{\gamma - 1}}\right)^{\frac{1}{p^k}} = 0 \qquad \forall p \in [1, \gamma). \tag{4}$$

As (u^k) has q-order at least γ by (2), its r-order is also at least γ, cf. [25, 9.3.2], thus $\lim_{k\to\infty}\|u^k - \bar{u}\|^{\frac{1}{p^k}} = 0$ for all $p \in [1, \gamma)$, so (4) follows from (3). □

Remark 1 For Broyden's method it is unknown whether (2) holds for any $\gamma > 1$ if $n > 1$, cf. also [18]. For $n = 1$ it is known that (2) holds with γ equal to the golden mean [31]. In Theorem 6 we show that this result extends to arbitrary n provided F has $n - 1$ affine component functions and B_0 is initialized exactly.

2.3 Uniform linear independence of dimension d

The following definition is the appropriate generalization of uniform linear independence for the purposes of this paper.

Definition 1 Let $n \in \mathbb{N}$ and $d \in \mathbb{N}$. The sequence of vectors $(s^k) \subset \mathbb{R}^n \setminus \{0\}$ is called *uniformly linearly independent of dimension d* iff there exist constants $m \in \mathbb{N}$ and $\rho > 0$ such that for every sufficiently large k the set

$$\left\{s^k, s^{k+1}, \ldots, s^{k+m}\right\}$$

contains d vectors $s^{k_1}, \ldots, s^{k_d}$ such that all singular values of the matrix

$$\left(\frac{s^{k_1}}{\|s^{k_1}\|} \quad \frac{s^{k_2}}{\|s^{k_2}\|} \quad \cdots \quad \frac{s^{k_d}}{\|s^{k_d}\|}\right) \in \mathbb{R}^{n\times d}$$

are larger than ρ.

Remark 2 The usual notion of uniform linear independence, cf. [5, (AS.4)], is recovered for $d = n$. If d is not specified, then it is understood that $d = n$.

3 Behavior of the Broyden–like method on mixed systems

To conveniently state results for mixed linear–nonlinear systems of equations, we will use the following assumption.

Assumption 1 *Let $n \in \mathbb{N}$, $d \in [n]_0$ and $J := [n] \setminus [d]$. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ satisfy $F_j(u) = a_j^T u + b_j$ for all $j \in J$, where $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$ for all $j \in J$. Let $B_0 \in \mathbb{R}^{n\times n}$ satisfy $B_0^j = a_j^T$ for all $j \in J$ and suppose that B_0 is invertible.*

Remark 3 Due to $B_0^j = a_j^T$ for all $j \in J$ and the invertibility of B_0, Assumption 1 implies $\dim(\langle\{a_j\}_{j\in J}\rangle) = n - d$, hence $\dim(\langle\{a_j\}_{j\in J}\rangle^\perp) = d$.

The first result establishes basic properties of Algorithm BL under Assumption 1. It generalizes [18, Lemma 2.1].

Lemma 4 *Let Assumption 1 hold and let (u^k), (s^k) and (B_k) be generated by Algorithm BL. Then we have for each $j \in J$ and all $k \geq 1$ the identities $B_k^j = a_j^T$, $F_j(u^k) = 0$, $a_j^T s^k = 0$ and $B_k a_j = B_1 a_j$.*

Proof The proof of [18, Lemma 2.1] applies without changes. □

Under the assumptions of Lemma 4 the sequence (s^k) necessarily violates uniform linear independence except if $J = \emptyset$.

Corollary 1 *Any selection $\{s^{k_1}, \ldots, s^{k_{d+1}}\}$ of $d+1$ vectors from the sequence $(s^k)_{k\geq 1}$ of Lemma 4 is linearly dependent.*

Proof Lemma 4 yields $a_j^T s^k = 0$ for all $j \in J$ and all $k \geq 1$, thus $s^k \in \langle\{a_j\}_{j\in J}\rangle^\perp$ for all $k \geq 1$. The claim follows from $\dim(\langle\{a_j\}_{j\in J}\rangle^\perp) = d$. □

To conveniently state the next result we introduce some notation.

Definition 2 Let Assumption 1 hold. We set $\mathcal{A} := \langle \{a_j\}_{j\in J}\rangle$ and $\mathcal{S} := \mathcal{A}^\perp$. Furthermore, we let $\{\mathfrak{s}^i\}_{i\in[d]}$ be an orthonormal basis of $\mathcal{S}$ and we denote $S := \begin{pmatrix}\mathfrak{s}^1 & \dots & \mathfrak{s}^d\end{pmatrix} \in \mathbb{R}^{n\times d}$. For any matrix $B \in \mathbb{R}^{n\times n}$ we denote

$$\widetilde{B} := \begin{pmatrix} B^1 \\ \vdots \\ B^d \end{pmatrix} \in \mathbb{R}^{d\times n} \qquad \text{and similarly} \qquad \widetilde{F}(u) := \begin{pmatrix} F_1(u) \\ \vdots \\ F_d(u) \end{pmatrix}.$$

We show that under Assumption 1 the iterates $(u^k)_{k\geq 1}$ obtained by applying Algorithm BL to F can also be generated by applying it to a mapping G acting between $\mathbb{R}^d$. The following result extends [18, Theorem 2.3].

Theorem 3 *Let Assumption 1 hold and let (u^k), (B_k) and (σ_k) be generated by Algorithm BL, where each B_k is assumed to be invertible. Define*

$$G : \mathbb{R}^d \to \mathbb{R}^d, \qquad G(w) := \widetilde{F}(u^1 + Sw)$$

as well as

$$C_0 := \widetilde{B}_1 S \in \mathbb{R}^{d\times d}, \qquad w^0 := 0 \in \mathbb{R}^d, \qquad \text{and} \qquad \tau_k := \sigma_{k+1} \quad \forall k \geq 0.$$

Then the application of Algorithm BL to G with initial guess (w^0, C_0) and updating sequence (τ_k) generates sequences (w^k) and (C_k) with the following properties:

1) Each C_k is invertible and for all $k \geq 1$ there hold

$$u^k = u^1 + Sw^{k-1}, \qquad \widetilde{F}(u^k) = G(w^{k-1}) \qquad \text{and} \qquad C_{k-1} = \widetilde{B}_k S. \tag{5}$$

2) The iterates (u^k) converge to $\bar{u} \in \mathbb{R}^n$ if and only if there is $\bar{w} \in \mathbb{R}^d$ such that (w^k) converges to $\bar{w}$. If (u^k) and (w^k) converge to $\bar{u}$ and $\bar{w}$, respectively, then we have for all $k \geq 1$

$$\bar{u} = u^1 + S\bar{w} \qquad \text{and} \qquad \left\|u^k - \bar{u}\right\| = \left\|w^{k-1} - \bar{w}\right\|. \tag{6}$$

3) The matrices (B_k) converge to $B \in \mathbb{R}^{n\times n}$ if and only if there is $C \in \mathbb{R}^{d\times d}$ such that (C_k) converges to C. If (B_k) and (C_k) converge to B and C, respectively, then we have for all $k \geq 1$

$$C = \widetilde{B}S \qquad \text{and} \qquad \|C_k - C\| = \|B_k - B\|.$$

Proof **Proof of 1):** The proof of [18, Theorem 2.3], which is for $(\sigma_k) \equiv 1$, can be used almost verbatim.
Proof of 2): We will use several times that $\|Sv\| = \|v\|$ for all $v \in \mathbb{R}^d$ because the columns of S are orthonormal.
Let (u^k) converge to $\bar{u}$. From (5) it follows that $u^n - u^m = S(w^{n-1} - w^{m-1})$ for all $n, m \geq 1$, which implies that (w^k) is a Cauchy sequence, hence convergent. Denoting the limit by $\bar{w}$ we deduce from (5) that $\bar{u} = u^1 + S\bar{w}$, which in turn

yields $\|u^k - \bar{u}\| = \|S(w^{k-1} - \bar{w})\|$, hence $\|u^k - \bar{u}\| = \|w^{k-1} - \bar{w}\|$. If (w^k) converges to $\bar{w}$, then we can argue similarly.

Proof of 3): Let (B^k) converge to B. From (5) it follows that $\|C_{n-1} - C_{m-1}\| \leq \|\widetilde{B}_n - \widetilde{B}_m\| = \|B_n - B_m\|$ for all $n, m \geq 1$, where we used that $\|S\| = 1$ and that $B_n^j - B_m^j = 0$ for all $j \in J$ due to Lemma 4. This implies that (C^k) is a Cauchy sequence, hence convergent. Denoting the limit by C we deduce from (5) that $C = \widetilde{B}S$. Let now (C^k) converge to C. We denote by $A \in \mathbb{R}^{n\times(n-d)}$ the matrix

$$A := \begin{pmatrix} \mathfrak{a}^1 & \dots & \mathfrak{a}^{n-d} \end{pmatrix},$$

where $\{\mathfrak{a}^i\}_{i\in[n-d]}$ is an orthonormal basis of $\mathcal{A}$. Furthermore, let $\hat{S} \in \mathbb{R}^{n\times n}$ be given by $\hat{S} := \begin{pmatrix} S & A \end{pmatrix}$. Since $B_k^j S = a_j^T S = 0$ and $B_k A = B_1 A$ for all $j \in J$ and all $k \geq 1$ by Lemma 4, we infer that

$$B_k \hat{S} = \left(\begin{array}{c|c} \widetilde{B}_k S & B_k A \\ 0 & \end{array}\right) = \left(\begin{array}{c|c} C_{k-1} & B_1 A \\ 0 & \end{array}\right), \tag{7}$$

where we also used the identity $\widetilde{B}_k S = C_{k-1}$ from (5). Since $\hat{S}\hat{S}^T = I$, it follows that

$$B_k = \left(\begin{array}{c|c} C_{k-1} & B_1 A \\ 0 & \end{array}\right) \hat{S}^T$$

for all $k \geq 1$. Since (C_k) converges, we see that (B_k) converges, too. Denoting the limit of (B_k) by B we conclude from (5) that $C = \widetilde{B}S$ and from (7) that $\|C_{k-1} - C\| = \|(B_k - B)\hat{S}\| = \|B_k - B\|$, where we used that $\hat{S}$ is orthogonal. □

Remark 4 Theorem 3 does not require invertibility of $F'(\bar{u})$, which allows us to derive results for singular $F'(\bar{u})$, too, cf. Theorems 6 and 7.

4 Convergence of the Broyden–like matrices

4.1 The general result

From [22, Theorem 5.7] we recall the following sufficient condition for convergence of (B_k) to $F'(\bar{u})$.

Theorem 4 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be strictly differentiable at $\bar{u}$. Let (u^k), (s^k) and (B_k) be generated by Algorithm BL. Let (u^k) converge to $\bar{u}$ and let (s^k) be uniformly linearly independent. Then $B := \lim_{k\to\infty} B_k$ exists and satisfies $B = F'(\bar{u})$. Moreover, we have $F(\bar{u}) = 0$. If, in addition, $F'(\bar{u})$ is invertible, then (u^k) converges q-superlinearly.*

Proof There are three differences to [22, Theorem 5.7]. The first is that we replaced continuous differentiability of F by strict differentiability. It is easy to verify that the proof of [22, Theorem 5.7] still holds under this weaker

assumption. The second and third difference are the statements for $F(\bar{u}) = 0$ and the q-superlinear convergence of (u^k), which we added. They follow from Lemma 1 and Lemma 2, respectively. □

Corollary 1 shows that for mixed linear–nonlinear systems with exact initialization, the uniform linear independence required in Theorem 4 does not hold. The following result extends Theorem 4 to mixed systems. We recall that the matrix S is introduced in Definition 2.

Theorem 5 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$. Let Assumption 1 hold and let (u^k), (s^k) and (B_k) be generated by Algorithm BL, where each B_k is assumed to be invertible. Let (u^k) converge to $\bar{u}$ and suppose that $w \mapsto \widetilde{F}(\bar{u}+Sw)$ is strictly differentiable at $w = 0$. Let (s^k) be uniformly linearly independent of dimension d. Then $B := \lim_{k\to\infty} B_k$ exists and satisfies $\widetilde{B}S = \widetilde{F}'(\bar{u})S$, $Ba_j = B_1 a_j$ and $B^j = a_j^T = F_j'(\bar{u})$ for all $j \in J$. Moreover, we have $F(\bar{u}) = 0$. If $\widetilde{F}'(\bar{u})S$ is invertible, then (u^k) converges q-superlinearly. If F is strictly differentiable at $\bar{u}$, then $E := \lim_{k\to\infty} E_k$ exists and satisfies $E = E_1(I - SS^T)$.*

Proof For $d = n$ we have $J = \emptyset$, $\widetilde{E} = E$ and $S \in \mathbb{R}^{n\times n}$ is orthogonal, so the result is equivalent to Theorem 4 and there is nothing to prove. For $d < n$ we begin by noting that Lemma 4 yields $B_k^j = a_j^T$ and $B_k a_j = B_1 a_j$ for all $j \in J$ and all $k \geq 1$, which carries over to $\lim_{k\to\infty} B_k$ if it exists. Next we show the existence of $\lim_{k\to\infty} B_k$. By applying Theorem 3 we obtain sequences (C_k) and (w^k) and a point $\bar{w}$ as stated in that theorem. Part 3) of that theorem shows that for convergence of (B_k) it suffices to demonstrate the convergence of (C_k). Denoting $s_w^k := w^{k+1} - w^k$ we now prove that $(s_w^k) \subset \mathbb{R}^d \setminus \{0\}$ is uniformly linearly independent (of dimension d). Indeed, using (5) we have

$$\hat{s}^k = \frac{Ss_w^{k-1}}{\|s^k\|} = \frac{S(w^k - w^{k-1})}{\|S(w^k - w^{k-1})\|} = \frac{S(w^k - w^{k-1})}{\|w^k - w^{k-1}\|}.$$

This implies that the matrix $\hat{S}^k$ appearing in the definition of uniform linear independence of dimension d of (s^k) and the matrix appearing in the definition of uniform linear independence of (s_w^k) have identical singular values, so the uniform linear independence of dimension d of (s^k) implies the uniform linear independence of (s_w^k). The uniform linear independence of (s_w^k) and the results of Theorem 3 allow us to apply Theorem 4 to G, (w^k), (s_w^k) and (C_k). This yields convergence of (C_k) to $G'(\bar{w}) = \widetilde{F}'(\bar{u})S$, which by means of Theorem 3 3) implies $\widetilde{B}S = \widetilde{F}'(\bar{u})S$. Since (B_k) converges, Lemma 1 supplies $F(\bar{u}) = 0$ and Theorem 4 implies q-superlinear convergence of (w^k), from which the q-superlinear convergence of (u^k) follows by use of (6). If F is strictly differentiable at $\bar{u}$, then the claims for B imply that E exists and satisfies $\widetilde{E}S = 0$ as well as $E^j = 0$ and $Ea_j = E_1 a_j$ for all $j \in J$. It is easy to see that these conditions are equivalent to $E = E_1(I - SS^T)$. □

Remark 5

1) If F is strictly differentiable at $\bar{u}$, then $\widetilde{F}(\bar{u}+Sw)$ is strictly differentiable at $w=0$. If $F'(\bar{u})$ is invertible, then $\widetilde{F}'(\bar{u})S$ is invertible.
2) To illustrate the conditions obtained for B let us consider the case that $\mathcal{S}=\{(s_1,s_2,\ldots,s_n)^T\in\mathbb{R}^n : s_j=0\ \forall j>d\}$. In this case we can use for S the first d columns of the $n\times n$ identity matrix. Thus, $\widetilde{B}S$ consists of the entries $B^{i,j}$, $i,j\in[d]$, and $\widetilde{B}S=\widetilde{F}'(\bar{u})S$ states that the first $d\times d$ block of B agrees with the respective block of $F'(\bar{u})$. From $Ba_j=B_1a_j$ for all $j\in J$ we obtain in addition that the entries $B^{i,j}$, $i\in[d]$, $j\in[n]\setminus[d]$, are the same as in B_1. If F is strictly differentiable at $\bar{u}$, then this implies that $B^{i,j}$, $i\in[d]$, $j\in[n]\setminus[d]$, cannot equal the respective entries of $F'(\bar{u})$ if the rank of $(E_0^{i,j})_{i\in[d],j\in[n]\setminus[d]}$ is larger than one.

4.2 The special case $d=1$

Sufficient conditions for uniform linear independence of $(s^k)\subset\mathbb{R}^n$ are unknown for Broyden's method if $n>1$ (hence also for the more general Algorithm BL). However, any sequence $(s^k)\subset\mathbb{R}^n\setminus\{0\}$ is uniformly linearly independent of dimension 1, hence Theorem 5 implies the following result.

Corollary 2 *Let $F:\mathbb{R}^n\to\mathbb{R}^n$. Let Assumption 1 hold for $d=1$ and let (u^k), (s^k) and (B_k) be generated by Algorithm BL, where each B_k is assumed to be invertible. Let (u^k) converge to $\bar{u}$ and suppose that $t\mapsto F_1(\bar{u}+t\bar{s})$ is strictly differentiable at $t=0$, where $\bar{s}:=S$. Then $B:=\lim_{k\to\infty}B_k$ exists and satisfies $B^1\bar{s}=F_1'(\bar{u})(\bar{s})$, $B^1a_j=B_1^1a_j$ and $B^j=a_j^T=F_j'(\bar{u})$ for all $j>1$. Moreover, we have $F(\bar{u})=0$. If $F_1'(\bar{u})(\bar{s})\neq 0$, then (u^k) converges q-superlinearly. If F_1 is strictly differentiable at $\bar{u}$, then $E:=\lim_{k\to\infty}E_k$ exists and satisfies $E^1=E_1^1(I-\bar{s}\bar{s}^T)$ and $E^j=0$ for all $j>1$; in particular, (B_k) converges to $F'(\bar{u})$ iff $E_1^1a_j=0$ for all $j>1$.*

Remark 6 Under the assumptions of Corollary 2 each B_k agrees with $F'(\bar{u})$ in all rows except the first and $B:=\lim_{k\to\infty}B_k$ exists, yet B will usually be different from $F'(\bar{u})$ (provided $F'(\bar{u})$ exists). If, say, $\bar{s}$ is the first canonical unit vector, then $E^1=\begin{pmatrix}0 & E_1^{1,2} & \ldots & E_1^{1,n}\end{pmatrix}$, hence $E=0$ holds iff $B_1^{1,j}=[F_1'(\bar{u})]_j$ for all $j>1$, where $[F_1'(\bar{u})]_j$ indicates the j-th component of the vector $F_1'(\bar{u})$. This also shows that if $\|E_0\|$ is large, then $\|E\|$ will usually be large, too. The numerical results in section 5 and our numerical experience from other work confirm that (B_k) will frequently not converge to $F'(\bar{u})$ and indicate that this also holds in more nonlinear settings.

We now focus on Broyden's method, where $(\sigma_k)\equiv 1$. In fact, it is enough if $\sigma_k=1$ for all k sufficiently large. For this case we can strengthen the findings of Corollary 2 in several ways, for instance by providing orders of convergence for (u^k) and $(\|B_{k+1}-B_k\|)$. These results are derived by exploiting the fact that if

$\sigma_k = 1$ for a $k \in \mathbb{N}$, then s^{k+1} and thus u^{k+2} can also be generated by the one-dimensional secant method, cf. the proof of Theorem 6 1). Correspondingly, let us first argue for the one-dimensional case.

Lemma 5 *Let $G : \mathbb{R} \to \mathbb{R}$. Let (w^k), (s_w^k) and (C_k) be generated by Algorithm BL applied to G, using an update sequence (τ_k) that satisfies*

$$\lim_{k\to\infty} \frac{\tau_{k+1}}{\tau_k} = 1.$$

Let (w^k) converge to $\bar{w}$ with $G(\bar{w}) = 0$. For $k \geq 0$, respectively, $k \geq 1$ define

$$q_k^G := \frac{|w^{k+1} - \bar{w}|}{|w^k - \bar{w}|} \qquad \text{and} \qquad Q_k^G := \frac{|C_{k+1} - C_k|}{|C_k - C_{k-1}|}.$$

Then the following statements hold:

1) Let G be differentiable at $\bar{w}$ with $G'(\bar{w}) \neq 0$. Let $\varphi := \frac{1+\sqrt{5}}{2}$ and suppose that

$$\lim_{k\to\infty} \frac{|w^{k+1} - \bar{w}|}{|w^k - \bar{w}|^\varphi} \tag{8}$$

exists. Then we have

$$\lim_{k\to\infty} \frac{Q_k^G}{q_{k-2}^G} = 1.$$

If, in addition, $\lim_{k\to\infty} \tau_k = 1$ is satisfied, then there holds

$$\lim_{k\to\infty} \frac{|C_{k+1} - C_k|}{|C_k - C_{k-1}|^\varphi} = |G'(\bar{w})|^{1-\varphi}.$$

2) Let $m_0 \in \mathbb{N}$, $\kappa \in (0,1)$ and $\hat{\kappa} > 0$. Let G be $m_0 + 1$ times differentiable at $\bar{w}$. Let $G^{(m)}(\bar{w}) = 0$ for all $m \in [m_0]$ and $G^{(m_0+1)}(\bar{w}) \neq 0$. Suppose that

$$\lim_{k\to\infty} q_k^G = \kappa \qquad \text{and} \qquad \lim_{k\to\infty} \frac{|s_w^k|}{|w^k - \bar{w}|} = \hat{\kappa}$$

are satisfied. Then we have

$$\lim_{k\to\infty} Q_k^G = \kappa^{m_0}.$$

Proof **Proof of 1):** Using $G(\bar{w}) = 0$ we find

$$\begin{aligned}\frac{\tau_{k-1}}{\tau_k} \cdot \frac{|C_{k+1} - C_k|}{|C_k - C_{k-1}|} &= \frac{|G(w^{k+1})||s_w^{k-1}|}{|s_w^k||G(w^k)|} \\ &= \frac{|G'(\bar{w})(w^{k+1} - \bar{w}) + o(|w^{k+1} - \bar{w}|)||s_w^{k-1}|}{|s_w^k||G'(\bar{w})(w^k - \bar{w}) + o(|w^k - \bar{w}|)|}\end{aligned}$$

for all $k \geq 1$. As (8) implies that (w^k) converges q-superlinearly, a well-known lemma of Dennis and Moré, cf. [7, Lemma 2.1], yields $\lim_{k\to\infty} \frac{|s_w^k|}{|w^k-\bar{w}|} = 1$. Therefore, we have

$$\begin{aligned}\lim_{k\to\infty} \frac{Q_k^G}{q_{k-2}^G} &= \lim_{k\to\infty} \frac{|C_{k+1}-C_k|}{|C_k-C_{k-1}|} \frac{|w^{k-2}-\bar{w}|}{|w^{k-1}-\bar{w}|} \\ &= \lim_{k\to\infty} \frac{|G'(\bar{w})||w^{k+1}-\bar{w}||w^{k-1}-\bar{w}|}{|w^k-\bar{w}||G'(\bar{w})||w^k-\bar{w}|} \frac{|w^{k-2}-\bar{w}|}{|w^{k-1}-\bar{w}|} \\ &= \lim_{k\to\infty} \frac{|w^{k+1}-\bar{w}||w^{k-2}-\bar{w}|}{|w^k-\bar{w}|^2},\end{aligned}$$

provided the latter limit exists. By applying (8) multiple times we obtain

$$\lim_{k\to\infty} \frac{|w^{k+1}-\bar{w}||w^{k-2}-\bar{w}|}{|w^k-\bar{w}|^2} = \lim_{k\to\infty} \mu^{\varphi-1-\frac{1}{\varphi}} |w^{k-1}-\bar{w}|^{\varphi^2-2\varphi+\frac{1}{\varphi}} = 1,$$

where $\mu \in [0,\infty)$ denotes the limit from (8) and where we used the identities $\varphi^2 - 2\varphi + \frac{1}{\varphi} = -\varphi + 1 + \frac{1}{\varphi} = \varphi - 1 - \frac{1}{\varphi} = 0$ that follow from $\varphi^2 - \varphi - 1 = 0$. Similar considerations show that

$$\lim_{k\to\infty} \frac{|C_{k+1}-C_k|}{|C_k-C_{k-1}|^\varphi} = \bar{\mu} \lim_{k\to\infty} \frac{|w^{k+1}-\bar{w}|}{|w^k-\bar{w}|} \cdot \frac{|w^{k-1}-\bar{w}|^\varphi}{|w^k-\bar{w}|^\varphi} = \bar{\mu}$$

for $\bar{\mu} := |G'(\bar{w})|^{1-\varphi}$, where we used (8) to obtain the final equality.

Proof of 2): Let us prove the claim for $m_0 = 1$; it is readily generalized to arbitrary $m_0 \geq 1$. Taylor expansion around $\bar{w}$ together with $G(\bar{w}) = 0$ implies by similar arguments as in the proof of 1) that

$$\begin{aligned}&\lim_{k\to\infty} \frac{|G(w^{k+1})|}{|G(w^k)|} \\ &= \lim_{k\to\infty} \frac{|G'(\bar{w})(w^{k+1}-\bar{w}) + \frac{1}{2}G''(\bar{w})(w^{k+1}-\bar{w})^2 + o(|w^{k+1}-\bar{w}|^2)|}{|G'(\bar{w})(w^k-\bar{w}) + \frac{1}{2}G''(\bar{w})(w^k-\bar{w})^2 + o(|w^k-\bar{w}|^2)|} \\ &= \lim_{k\to\infty} \frac{|G''(\bar{w})|}{|G''(\bar{w})|} \cdot \frac{|w^{k+1}-\bar{w}|^2}{|w^k-\bar{w}|^2} = \kappa^2 = \kappa^{m_0+1}.\end{aligned}$$

By assumption we have $\hat{\kappa} = \lim_{k\to\infty} \frac{|s_w^k|}{|w^k-\bar{w}|} > 0$, hence

$$\lim_{k\to\infty} \frac{|s_w^{k-1}|}{|s_w^k|} = \lim_{k\to\infty} \frac{\hat{\kappa}|w^{k-1}-\bar{w}|}{\hat{\kappa}|w^k-\bar{w}|} = \frac{1}{\kappa}.$$

By definition there holds for all $k \geq 1$

$$\frac{\tau_{k-1}}{\tau_k} \cdot Q_k^G = \frac{|G(w^{k+1})|}{|G(w^k)|} \cdot \frac{|s_w^{k-1}|}{|s_w^k|}.$$

Taking the limit for $k \to \infty$ yields the claim. □

We now provide a detailed description of the convergence behavior of Algorithm 1 with $\sigma_k = 1$ for all large k and $d = 1$, where F has $n-1$ affine component functions $F_2, \dots, F_n$. We first present a result for nonlinear F_1 and then deal with affine F_1.

Theorem 6 *Let Assumption 1 hold for $d = 1$ and let (u^k), (s^k) and (B_k) be generated by Algorithm BL, with each B_k invertible. Suppose that $\sigma_k = 1$ for all k large enough and that (u^k) converges to some $\bar{u}$. Set $\bar{s} := S$ and define*

$$q_k := \frac{\left\|u^{k+1} - \bar{u}\right\|}{\|u^k - \bar{u}\|} \qquad \text{and} \qquad Q_k := \frac{\|B_{k+1} - B_k\|}{\|B_k - B_{k-1}\|}$$

for all $k \geq 0$, respectively, $k \geq 1$. Then the following statements hold:

1) Let $t \mapsto F_1(\bar{u}+t\bar{s})$ be twice differentiable near $t = 0$ with $t \mapsto F_1''(\bar{u}+t\bar{s})(\bar{s},\bar{s})$ continuous at $t = 0$ and $F_1'(\bar{u})(\bar{s}) \neq 0$. Then we have

$$\limsup_{k\to\infty} \frac{\left\|u^{k+1} - \bar{u}\right\|}{\|u^k - \bar{u}\|^{\varphi}} \leq \left|\frac{F_1''(\bar{u})(\bar{s},\bar{s})}{2F_1'(\bar{u})(\bar{s})}\right|^{\frac{1}{\varphi}}, \tag{9}$$

where $\varphi := \frac{1+\sqrt{5}}{2}$. For all $p \in [1, \varphi)$ there holds

$$\lim_{k\to\infty} \|B_{k+1} - B_k\|^{\frac{1}{p^k}} = 0. \tag{10}$$

If, in addition, $F_1''(\bar{u})(\bar{s},\bar{s}) \neq 0$, then (9) holds with equality and $\limsup$ replaced by $\lim$, and we have

$$\lim_{k\to\infty} \frac{\|B_{k+1} - B_k\|}{\|B_k - B_{k-1}\|^{\varphi}} = |F_1'(\bar{u})(\bar{s})|^{1-\varphi} \qquad \text{and} \qquad \lim_{k\to\infty} \frac{Q_k}{q_{k-2}} = 1. \tag{11}$$

2) Let $m_0 \in \mathbb{N}$ and denote by $\kappa \in (0,1)$ the unique root of the polynomial $x^{m_0+1} + x^{m_0} - 1$ in $(0,1)$. Let $t \mapsto F_1(\bar{u} + t\bar{s})$ be $m_0 + 1$ times differentiable near $t = 0$ with its (m_0+1)-th derivative continuous at $t = 0$. If $F_1^{(m)}(\bar{u})(\bar{s},\dots,\bar{s}) = 0$ for all $m \in [m_0]$ and $F_1^{(m_0+1)}(\bar{u})(\bar{s},\dots,\bar{s}) \neq 0$, then

$$\lim_{k\to\infty} q_k = \kappa \qquad \text{and} \qquad \lim_{k\to\infty} Q_k = \kappa^{m_0}.$$

Proof **Proof of 1):** From Theorem 3 we obtain $G : \mathbb{R} \to \mathbb{R}$, (w^k), (C_k) and $\bar{w}$ as stated in that theorem. We let $s_w^k := w^{k+1} - w^k$ for all $k \geq 0$. Due to $C_k s_w^k (s_w^k)^T / |s_w^k|^2 = C_k$ we have $C_{k+1} = (G(w^{k+1}) - G(w^k))/s_w^k$ if $\sigma_k = 1$ and thus Algorithm BL for G agrees with the one-dimensional secant method for all sufficiently large k. As $(G(w^{k+1}) - G(w^k))/s_w^k \to G'(\bar{w})$ for $k \to \infty$, we obtain the convergence of (C_k), thus $G(\bar{w}) = 0$ by Lemma 1. Furthermore, there holds $G'(\bar{w}) = \widetilde{F}'(\bar{u})S = F_1'(\bar{u})(\bar{s}) \neq 0$. Since (w^k) converges to $\bar{w}$ with $G(\bar{w}) = 0$ and $G'(\bar{w}) \neq 0$, classical results for the secant method, cf. [31, (6)], yield that if $G''(\bar{w}) \neq 0$, then

$$\lim_{k\to\infty} \frac{|w^k - \bar{w}|}{|w^{k-1} - \bar{w}|^{\varphi}} = \left|\frac{G''(\bar{w})}{2G'(\bar{w})}\right|^{\frac{1}{\varphi}},$$

which by use of (5) is readily transformed into (9) with equality and lim sup replaced by lim. Similarly for (9). The r-order (10) follows from Lemma 3 using that $F(\bar{u}) = 0$ due to Corollary 2. Since $Q_k^G = Q_{k+1}$ and $q_{k-2}^G = q_{k-1}$ by (5) and (6), Lemma 5 1) yields (11).

Proof of 2): We argue only for $m_0 = 1$. It follows from Corollary 2 that $F(\bar{u}) = 0$. It is a standard result for the one-dimensional secant method, cf. [10, Section 2.2.2], that $\lim_{k\to\infty} q_k^G = \kappa$, hence $\lim_{k\to\infty} q_k = \kappa$, too. The claim on (Q_k) follows via (Q_k^G) from Lemma 5 2) if we can show that there is $\hat{\kappa} > 0$ such that

$$\lim_{k\to\infty} \frac{|s_w^k|}{|w^k - \bar{w}|} = \hat{\kappa}.$$

Using $G'(\bar{w}) = 0$, $G''(\bar{w}) \neq 0$, and $\lim_{k\to\infty} q_k^G = \kappa$, elementary considerations show that there is an index k_0 such that $(w^k - \bar{w})_{k\geq k_0}$ converges to zero without changing signs. For sufficiently large k we thus have

$$\left|s_w^k\right| = \left|(w^{k+1} - \bar{w}) - (w^k - \bar{w})\right| = (1 - q_k^G)\left|w^k - \bar{w}\right|,$$

hence the desired limit exists with $\hat{\kappa} = 1 - \kappa > 0$. □

Remark 7

1) If $F'(\bar{u})$ is invertible, then $F_1'(\bar{u})(\bar{s}) \neq 0$. Indeed, since $\bar{s} \in \mathcal{S}$ and since $F_j'(\bar{u}) = a_j^T \in \mathcal{A} = \mathcal{S}^\perp$ for all $j > 1$, we have $F_j'(\bar{u})(\bar{s}) = 0$ for all $j > 1$, hence $F_1'(\bar{u})(\bar{s}) = 0$ would imply $F'(\bar{u})(\bar{s}) = 0$.
2) (9) and (10) show that (u^k), respectively, $(\|B_{k+1} - B_k\|)$ have q-order, respectively, r-order no less than φ. If $F_1''(\bar{u})(\bar{s}, \bar{s}) \neq 0$, then the additional part of 1) implies that both (u^k) and $(\|B_{k+1} - B_k\|)$ have q-order and r-order φ, cf. [25, 9.3.3]. For (u^k), the q-order φ improves the best available result, which is the 2–step q-quadratic convergence ensured by Theorem 2 for $d = 1$. Moreover, the example in section 4.3.2 shows that if $F_1''(\bar{u})(\bar{s}, \bar{s}) = 0$, then it is possible to have a higher q-order than φ.
3) For $m_0 = 1$, Theorem 6 2) is related to the results in [6,19].
4) Corollary 2 is valid under the assumptions of Theorem 6, so in 1) and 2) we also have $F(\bar{u}) = 0$ and B satisfies the conditions from that corollary.

In the affine setting Algorithm BL terminates after finitely many steps, provided a root exists and $\sigma_k = 1$ for at least one k (if the Jacobian is regular). More precisely, we have the following result.

Theorem 7 *Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be affine. Let Assumption 1 hold for $d = 1$ and let (u^k), (s^k) and (B_k) be generated by Algorithm BL, with each B_k invertible. Then the following statements hold:*

1) Let F' be invertible. Then F has a unique root $\bar{u}$. If there is an index $k \geq 1$ with $\sigma_k = 1$, then $u^{k+2} = \bar{u}$ (provided $u^j \neq \bar{u}$ for all $j \in [k+1]_0$). If there is no $k \geq 0$ such that $u^k = \bar{u}$, then (u^k) converges to $\bar{u}$ and satisfies (1).
2) Let F' be singular. If F has a root, then $F(u^0) = 0$ or $F(u^1) = 0$. If F does not have a root, then the algorithm generates a diverging sequence (u^k) such that $F(u^k) = (\omega, 0, \ldots, 0)^T$ for all $k \geq 1$ and some $\omega \neq 0$.

Proof **Proof of 1):** From [22, Theorem 3.2] we know that for affine F with invertible F', Algorithm BL converges q-superlinearly for any u^0 if all B_k are invertible and the algorithm does not terminate with output $u^* = \bar{u}$. (Since $d = 1$, it is also not difficult to establish this directly.) Theorem 1 now yields (1). Corollary 2 yields the convergence of (B_k). It remains to prove that if $\sigma_k = 1$ and $F(u^{k+1}) \neq 0$, then $F(u^{k+2}) = 0$. Since $F_j(u^k) = 0$ for all $j > 1$ and all $k \geq 1$ by Lemma 4, we have to show that $F_1(u^{k+2}) = 0$. Similar as in the proof of Theorem 6 we use Theorem 3 to obtain $\{w^j\}_{j=0}^{k+1}$ and $\{C_j\}_{j=0}^{k+1}$ by applying Algorithm BL to the affine function $G : \mathbb{R} \to \mathbb{R}$, $G(w) := F_1(u^1 + w\bar{s})$, where $\bar{s} := S$. In view of (5) we have to show that $G(w^{k+1}) = 0$. From $\tau_{k-1} = \sigma_k = 1$ it follows that $C_k = (G(w^k) - G(w^{k-1}))/(w^k - w^{k-1}) = G'$. Using $C_k(w^{k+1} - w^k) = -G(w^k)$ we find $G(w^{k+1}) = G(w^k) + G' \cdot (w^{k+1} - w^k) = G(w^k) - G(w^k) = 0$, hence $F(u^{k+2}) = 0$.
Proof of 2): Defining $A := F'$ we note that A has rank $n-1$ since $A\bar{s} = 0$ and since $n-1$ rows of A agree with the invertible B_0. Thus, A^1 can be expressed as a linear combination of $\{A^j\}_{j=2}^n$. Since F has a root and since $F_j(u^1) = 0$ for all $j > 1$ by Lemma 4, it readily follows that $F_1(u^1) = 0$, whence $F(u^1) = 0$. Now suppose that F does not have a root. By applying Theorem 3 again, we obtain that $G' = A\bar{s} = 0$, hence G is constant, say $G \equiv \omega$ for some $\omega \in \mathbb{R}$. Since F has no root, we must have $\omega \neq 0$. Since G is constant, there holds $F_1(u^k) = G(w^{k-1}) = \omega$ for all $k \geq 1$. The sequence (u^k) cannot be convergent because Corollary 2 would entail that the limit point is a root of F. □

Remark 8

1) The starting point u^0 is arbitrary in Theorem 7.
2) The finite convergence in Theorem 7 1) is related to the $2n$–step convergence of Broyden's method for regular linear systems [12,24]. Indeed, in the proof of Thm. 7 1) we can replace the computation for showing $G(w^{k+1}) = 0$ by an application of the $2n$–step convergence to G using that due to $\tau_{k-1} = 1$, s_w^{k-1} and s_w^k are the Broyden steps for initial (w^{k-1}, C_{k-1}).
3) If in Theorem 7 1), Algorithm BL does not terminate with $u^* = \bar{u}$, then $\lim_{k\to\infty} E_k$ exists and satisfies the conditions from Corollary 2.

4.3 Application to two examples from the literature

We illustrate some of our findings on two examples from the literature. The second example also hints at two extensions.

4.3.1 An example by Dennis and Schnabel

In [9, Example 8.1.3] and [9, Lemma 8.2.7] it is shown that for

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \qquad F(u) = \begin{pmatrix} u_1 + u_2 - 3 \\ u_1^2 + u_2^2 - 9 \end{pmatrix}$$

with root $\bar{u} = (0,3)^T$ the initial data

$$u^0 = \begin{pmatrix} 1 \\ 5 \end{pmatrix} \qquad \text{and} \qquad B_0 = F'(u^0) = \begin{pmatrix} 1 & 1 \\ 2 & 10 \end{pmatrix}$$

yields sequences (u^k) and (B_k) with $u^k \to \bar{u}$ for $k \to \infty$ and

$$B_1 = \begin{pmatrix} 1 & 1 \\ 0.375 & 8.625 \end{pmatrix}, \qquad B := \lim_{k\to\infty} B_k = \begin{pmatrix} 1 & 1 \\ 1.5 & 7.5 \end{pmatrix}, \qquad F'(\bar{u}) = \begin{pmatrix} 1 & 1 \\ 0 & 6 \end{pmatrix}.$$

The affine component F_1 has coefficient vector $a_1 = (1,1)^T$, so $\mathcal{S} = \langle\{a_1\}\rangle^\perp = \{t\bar{s} : t \in \mathbb{R}\}$ with $\bar{s} := \frac{1}{\sqrt{2}}(1,-1)^T$. Theorem 3 yields that $(s^k)_{k\geq 1} \subset \mathcal{S}$ and $(F_1(u^k))_{k\geq 1} \equiv 0$. Of course, this can also be verified directly, cf. also [9, Example 8.1.3&Lemma 8.2.7]. In agreement with Theorem 5 and Corollary 2 there holds $\widetilde{B}S = B^2\bar{s} = -3\sqrt{2} = \widetilde{F}'(\bar{u})S$, $B^1 = B_0^1$ and $B(1,1)^T = B_1(1,1)^T$. (From B_1, $F'(\bar{u})$ and $\bar{s}$ we can actually determine the limit B.) Because of $F_2'(\bar{u})\bar{s} \neq 0 \neq F_2''(\bar{u})(\bar{s},\bar{s})$, Theorem 6 1) yields q-order φ for (u^k) and $(\|B_{k+1} - B_k\|)$ as well as the validity of (11).

4.3.2 An example by Dennis and Moré

In [8, Example 5.3] Dennis and Moré consider Broyden's method for

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \qquad F(u) = \begin{pmatrix} u_1 \\ u_2 + u_2^3 \end{pmatrix}$$

with root $\bar{u} = (0,0)^T$ and note that for any $\delta, \epsilon \in \mathbb{R}$ the initial data

$$u^0 = \begin{pmatrix} 0 \\ \epsilon \end{pmatrix} \qquad \text{and} \qquad B_0 = \begin{pmatrix} 1+\delta & 0 \\ 0 & 1 \end{pmatrix} \tag{12}$$

yields a sequence (B_k) with $B_k^{1,1} = 1 + \delta$ for all $k \geq 0$. Hence, the incorrect entry $1 + \delta$ is never corrected (assuming $\delta \neq 0$), preventing convergence of (B_k) to $F'(\bar{u})$. According to [8], "The above example points out that one of the disadvantages of Broyden's method is that it is not self-correcting. In particular, B_k depends upon each B_j with $j < k$ and thus it may retain information which is irrelevant or even harmful.". It is well-known that the BFGS method is self-correcting, cf. e.g. [27,1].

We show that the iterates (u^k) converge rapidly despite the incorrect entry $1+\delta$ in all B_k. The affine component F_1 has coefficient vector $a_1 = (1,0)^T$, thus $\mathcal{S} = \langle\{a_1\}\rangle^\perp = \{(0,t)^T : t \in \mathbb{R}\}$. We set $\bar{s} := (0,1)^T$ and observe $(s^k)_{k\geq 0} \subset \mathcal{S}$ as well as $(F_1(u^k))_{k\geq 0} \equiv 0$. It is not difficult to see that Theorem 3 and, in turn, Theorem 6 1) apply, even though Assumption 1 is not satisfied in this example. Theorem 6 1) implies that if (u^k) converges to $\bar{u}$, then it has a q-order no smaller than φ and $(\|B_{k+1} - B_k\|)$ goes to zero with r-order no smaller than φ. The fast convergence is enabled by the fact that Broyden's method effectively reduces to the one-dimensional secant method. It should also be noted that (B_k) converges to $F'(\bar{u})$ in $\mathcal{S}$, i.e., $(B_k - F'(\bar{u}))S \to 0$,

cf. Corollary 2. Furthermore, since $B_0 S = 1$ correctly approximates the affine part of F_2 and since F_2 does not contain a quadratic part, it can be shown that $(\|B_{k+1} - B_k\|)$ has q-order 2, which implies that (u^k) has q-order 2, too. The numerical experiments confirm the q-order 2, cf. section 5.2.2.

5 Numerical experiments

We use numerical examples to verify Corollary 2 and Theorems 6 and 7. We first present the design of the experiments and then provide the examples and results.

5.1 Design of the experiments

5.1.1 Implementation and accuracy

We use the *variable precision arithmetic (vpa)* of MATLAB 2020B. Unless stated otherwise, we work with a precision of 10000 digits and replace the termination criterion $F(u^k) = 0$ in Algorithm BL by $\|F(u^k)\| \leq 10^{-5000}$. By $\bar{k}$ we denote the final value of k.

5.1.2 Known solution and random initialization

All examples have root $\bar{u} = 0$ and the experiments are set up in such a way that convergence to $\bar{u}$ takes place in all runs except possibly a handful, that are discarded. Except in the second example, the initial guess (u^0, B_0) is randomly generated using MATLAB'S function `rand` to satisfy $u^0 \in [-\alpha, \alpha]^n$ and $B_0 = F'(u^0) + \hat{\alpha}\|F'(u^0)\|R$. Here, $R \in \mathbb{R}^{n\times n}$ is a matrix with $R^j = 0$ for all $j > 1$ and the entries in R^1 randomly drawn from $[-1, 1]$. The values of $\alpha \in [10^{-3}, 1000]$ and $\hat{\alpha} \in [0, 1000]$ will be specified within each example.

5.1.3 Quantities of interest

To display the course of Algorithm BL we use the norm of $F_k := F(u^k)$, the error $\|E_k\|$, the quotients q_k and Q_k introduced in Theorem 6, and furthermore

$$\beta_k := \|B_k - B_{k-1}\|, \qquad C_k^u := \frac{\|u^k - \bar{u}\|}{\|u^{k-1} - \bar{u}\|^\varphi}, \qquad C_k^B := \frac{\|B_k - B_{k-1}\|}{\|B_{k-1} - B_{k-2}\|^\varphi},$$

as well as

$$\mathcal{R}_k^B := \log\big(\|B_k - B_{k-1}\|^{-1}\big)^{\frac{1}{k}}$$

and

$$\mathcal{Q}_k^u := \frac{\log(\|u^k - \bar{u}\|)}{\log(\|u^{k-1} - \bar{u}\|)}, \qquad \mathcal{Q}_k^B := \frac{\log(\|B_k - B_{k-1}\|)}{\log(\|B_{k-1} - B_{k-2}\|)}.$$

We note that $\mathcal{Q}_k^u$ and $\mathcal{Q}_k^B$ approximate the q-order of convergence while $\mathcal{R}_k^B$ approximates the r-order. Whenever any of these quantities is undefined we set it to -1; e.g., $\beta_0 := -1$. We will use these quantities to confirm that (B_k) converges, cf. Corollary 2, and to assess the convergence order of (u^k) and $(\|B_{k+1}-B_k\|)$, cf. Theorem 6. We are also interested in whether $\|E_k\| \to 0$, i.e., whether (B_k) converges to the true Jacobian $F'(\bar{u})$, cf. for instance Remark 6.

5.1.4 Single runs and cumulative runs

We use *single runs* and *cumulative runs*. For single runs we display the quantities of interest during the course of the algorithm. A cumulative run consists of 1000 single runs with initial data varying according to section 5.1.2, unless stated otherwise. Let us briefly describe the aggregated quantities that we use to assess cumulative runs. For instance, to gauge the q-order of $(\|B_{k+1} - B_k\|)$ we compute for each single run of a cumulative run the number

$$\mathcal{Q}_j := \min_{k_0(j)\le k\le \bar{k}(j)} \mathcal{Q}_k^B,$$

where $j \in [1000]$ indicates the respective single run and we consistently use $k_0(j) := \min\{100, \lfloor 0.75\bar{k}(j)\rfloor\}$. As outcome of the cumulative run we display

$$\mathcal{Q}_B^- := \min_{j\in[1000]} \mathcal{Q}_j \qquad \text{and} \qquad \mathcal{Q}_B^+ := \max_{j\in[1000]} \mathcal{Q}_j.$$

If the stronger conditions in Theorem 6 1) hold, then $\mathcal{Q}_B^-$ and $\mathcal{Q}_B^+$ should both be close to the golden mean φ. If the convergence is of lower order in any of the 1000 single runs, then we expect $\mathcal{Q}_B^-$ to be smaller than φ.

In the same way as just presented for $\mathcal{Q}_B^-$ and $\mathcal{Q}_B^+$, we derive $\|E\|^-$, $\|E\|^+$, q^-, q^+, $\mathcal{Q}_u^-$, $\mathcal{Q}_u^+$, β^-, β^+, Q^-, Q^+, $\mathcal{R}_u^-$ and $\mathcal{R}_u^+$ from the respective quantities used in single runs. In addition, we use

$$\|F\|^- := \min_{j\in[1000]} \|F(u^{\bar{k}(j)})\| \qquad \text{and} \qquad \|F\|^+ := \max_{j\in[1000]} \|F(u^{\bar{k}(j)})\|.$$

To keep the tables for cumulative runs of a reasonable size we will omit some of these quantities, but what is omitted varies from example to example.

5.2 Numerical examples

5.2.1 Example 1

To verify the results of Theorem 6 1) we consider $F : \mathbb{R}^{10} \to \mathbb{R}^{10}$ given by

$$F(u) = \begin{pmatrix} u_1 \cdot \left[\prod_{j=2}^{10} \left(u_j + (-1)^j\right)\right] \\ Au \end{pmatrix},$$

Table 1 Example 1: Single run with $\hat{\alpha} = 0$, i.e., $B_0 = F'(\bar{u})$

k	$\|\|F_k\|\|$	q_k	C_k^u	$\mathcal{Q}_k^u$	β_k	Q_k	C_k^B	$\mathcal{R}_k^B$	$\mathcal{Q}_k^B$	$\|\|E_k\|\|$
0	2.8e-3	-1	-1	-1	-1	-1	-1	-1	-1	3.0e-3
1	1.2e-6	5.6e-3	0.27	1.82	6.7e-4	-1	-1	2.7	-1	2.7e-3
2	1.1e-9	8.7e-4	1.1	1.61	1.0e-4	0.16	14	2.09	1.25	2.7e-3
3	4.8e-15	4.5e-6	0.43	1.66	5.4e-7	5.2e-3	1.5	1.95	1.57	2.7e-3
4	1.8e-23	3.9e-9	0.75	1.63	4.7e-10	8.7e-4	6.5	1.85	1.49	2.7e-3
5	3.2e-37	1.7e-14	0.53	1.63	2.1e-15	4.5e-6	2.6	1.8	1.57	2.7e-3
6	2.2e-59	6.8e-23	0.65	1.62	8.1e-24	3.9e-9	4.6	1.76	1.57	2.7e-3
7	2.6e-95	1.2e-36	0.57	1.62	1.4e-37	1.7e-14	3.2	1.74	1.6	2.7e-3
8	2.0e-153	7.9e-59	0.62	1.62	9.6e-60	6.8e-23	4.0	1.73	1.6	2.7e-3
9	1.9e-247	9.3e-95	0.59	1.62	1.1e-95	1.2e-36	3.5	1.71	1.61	2.7e-3
10	1.4e-399	7.4e-153	0.61	1.62	8.9e-154	7.9e-59	3.8	1.7	1.61	2.7e-3
11	9.6e-646	6.9e-247	0.6	1.62	8.3e-248	9.3e-95	3.6	1.7	1.61	2.7e-3
12	4.9e-1044	5.1e-399	0.6	1.62	6.1e-400	7.4e-153	3.7	1.69	1.62	2.7e-3
13	1.7e-1688	3.5e-645	0.6	1.62	4.2e-646	6.9e-247	3.7	1.68	1.62	2.7e-3
14	3.1e-2731	1.8e-1043	0.6	1.62	2.2e-1044	5.1e-399	3.7	1.68	1.62	2.7e-3
15	2.0e-4418	6.3e-1688	0.6	1.62	7.6e-1689	3.5e-645	3.7	1.68	1.62	2.7e-3
16	2.2e-7148	1.1e-2730	0.6	1.62	1.4e-2731	1.8e-1043	3.7	1.67	1.62	2.7e-3

where $A \in \mathbb{R}^{9\times 10}$ is a random matrix with entries in $[-1, 1]$ that is changed after each of the 1000 single runs of the cumulative run. The randomly generated A is only accepted if the resulting $F'(\bar{u})$ is invertible. We use $\alpha = 0.001$ in this example. A single and a cumulative run with $(\sigma_k) \equiv 1$ and $\hat{\alpha} = 0$ are displayed in Tables 1 and 2. The results agree with Theorem 6 1). For instance, it is apparent that (u^k) and $(\|B_{k+1} - B_k\|)$ converge with q-order $\varphi \approx 1.618$ and that $\lim_{k\to\infty} \frac{Q_k}{q_{k-2}} = 1$ (since A is random we expect $F_1''(\bar{u})(\bar{s}, \bar{s}) \neq 0$). Table 2 also shows results for a cumulative run with $(\sigma_k) \equiv 1$ and $\hat{\alpha} = 0.1$. In accordance with Theorem 6 1), deviating from the choice $B_0 = F'(u^0)$ does not affect the q-order of convergence. Next we keep $\hat{\alpha} = 0.1$ and let $\sigma_k = 0.5$ for $k \leq 3$ and $(\sigma_k)_{k\geq 4} \equiv 1$. Theorem 6 1) predicts that this choice of (σ_k) maintains q-order φ for (u^k) and $(\|B_{k+1} - B_k\|)$, and Table 2 confirms this.

In contrast, if we choose $\hat{\alpha} = 0$ and $(\sigma_k) \equiv 0.99$, then the order of convergence drops significantly and the same holds for $(\sigma_k) \equiv 1 - (k + 2)^{-4}$, cf. Table 2. In fact, except for some special cases it can be shown that (u^k) can only converge with q-order greater than one if $\sigma_k \to 1$ fast enough. In particular, for $(\sigma_k) \equiv 0.99$ and $(\sigma_k) \equiv 1 - (k + 2)^{-4}$, both (u^k) and $(\|B_{k+1} - B_k\|)$ have q-order 1. To confirm this for $(\sigma_k) \equiv 1 - (k + 2)^{-4}$, we repeat the cumulative run with a higher precision of 100000 digits, using $\|F(u^k)\| \leq 10^{-50000}$ as termination criterion and only 100 single runs instead of 1000. We view the results in Table 2 as being in line with q-order 1. In any case, it is apparent that for $(\sigma_k) \equiv 0.99$ and $(\sigma_k) \equiv 1 - (k + 2)^{-4}$ the q-order of convergence is not φ anymore and that $(\|B_{k+1} - B_k\|)$ converges to zero at least q-linearly for all choices of (σ_k), hence (B_k) converges, which validates Corollary 2. The values of $\|E\|^-$ show that (B_k) never converges to $F'(\bar{u})$.

Table 2 Example 1: Cumulative runs with $\hat{\alpha} = 0$ (1st row), $\hat{\alpha} = 0.1$ (2nd row), $\hat{\alpha} = 0.1$ and $\sigma_{0,1,2,3} = 0.5$ (3rd), $\hat{\alpha} = 0$ and $(\sigma_k) \equiv 0.99$ (4th), $\hat{\alpha} = 0$ and $\sigma_k = 1 - (k+2)^{-4}$ (5th), $\hat{\alpha} = 0$ and $\sigma_k = 1 - (k+2)^{-4}$ with higher precision (6th)

$\|\|F\|\|^+$	q^-	q^+	$\mathcal{Q}_u^-$	$\mathcal{Q}_u^+$	β^+	Q^-	Q^+	$\mathcal{R}_B^-$	$\mathcal{R}_B^+$	$\mathcal{Q}_B^-$	$\mathcal{Q}_B^+$	$\|\|E\|\|^-$	$\|\|E\|\|^+$
5e-5001	2e-451	5e-203	1.62	1.62	3e-205	7e-173	6e-78	1.56	1.73	1.59	1.62	7e-5	6e-3
2e-5003	4e-447	3e-173	1.62	1.62	1e-173	3e-171	1e-66	1.46	1.68	1.59	1.62	0.22	0.82
4e-5002	5e-279	2e-68	1.62	1.62	7e-70	5e-107	1e-26	1.31	1.60	1.58	1.62	0.23	0.82
8e-5001	4e-107	5e-105	1.03	1.03	3e-105	0.01	0.01	1.08	1.08	1.01	1.01	2e-4	5e-3
1e-5048	1e-170	2e-166	1.05	1.05	6e-168	6e-7	6e-7	1.14	1.14	1.03	1.03	2e-4	6e-3
4e-50012	2e-611	1e-603	1.02	1.02	5e-604	1e-8	1e-8	1.06	1.06	1.01	1.01	4e-4	4e-3

5.2.2 Example 2

We provide results for the example by Dennis and Moré discussed in section 4.3.2, which concerns Broyden's method, so $(\sigma_k) \equiv 1$. A single run is displayed in Table 3 and four cumulative runs in Table 4. For the single run and the first cumulative run we use (u^0, B_0) that satisfy (12) with randomly generated $\delta, \epsilon \in [-0.5, 0.5]$. The results confirm that, as argued in section 4.3.2, both (u^k) and $(\|B_{k+1} - B_k\|)$ have q-order 2. Because of $F_2''(\bar{u}) = 0$, this does not contradict Theorem 6 1).

In the second cumulative run we let $u^0 = (\epsilon_1, \epsilon_2)^T$ for random numbers $\epsilon_1, \epsilon_2 \in [-0.5, 0.5]$, while keeping B_0 as in (12) with $\delta \in [-0.5, 0.5]$. Due to $\epsilon_1 \neq 0$ we cannot expect (s^k) to belong to a one-dimensional subspace, hence Theorem 6 does not apply anymore. Correspondingly, the second row in Table 4 shows that (u^k) does not attain the q-order φ, but suggests that the q-order may still have a lower bound larger than 1. This view is further encouraged by the fact that the r-order of $(\|B_{k+1} - B_k\|)$ seems to admit such a lower bound, too, which is a necessary condition for (u^k) to have a q-order, cf. Lemma 3. To investigate the potential q-order of (u^k) further, we repeat the cumulative run at a higher precision using $\|F(u^k)\| \leq 10^{-100000}$ as termination criterion and 400 single runs. The results are contained in Table 4 and support the existence of a q-order larger than one for (u^k).

In the third cumulative run, whose results are depicted in the last row of Table 4, we keep the choice $u^0 = (\epsilon_1, \epsilon_2)^T$ from the second cumulative run, but use $B_0 = F'(u^0)$ as initial, so that $B_0^1 = F_1'(u^0)$ and hence Assumption 1 holds. In turn, Theorem 6 1) applies, which ensures a q-order, respectively, r-order no smaller than φ for (u^k) and $(\|B_{k+1} - B_k\|)$, respectively. It can be argued in the same way as in section 4.3.2 that both sequences actually converge with q-order 2. Table 4 confirms this q-order.

The values of $\|E\|^-$ in Table 4 show that (B_k) never converges to $F'(\bar{u})$. Yet, since $(\|B_{k+1} - B_k\|)$ declines quickly, the convergence of (u^k) is still rapid.

Table 3 Example 2: Single run with initial data of the form (12)

k	$\|\|F_k\|\|$	q_k	C_k^u	$\mathcal{Q}_k^u$	β_k	Q_k	C_k^B	$\mathcal{R}_k^B$	$\mathcal{Q}_k^B$	$\|\|E_k\|\|$
0	0.42	-1	-1	-1	-1	-1	-1	-1	-1	0.413
1	0.052	0.14	0.26	3.0	0.12	-1	-1	1.45	-1	0.413
2	5.5e-3	0.11	0.66	1.76	0.12	0.98	3.6	1.29	1.01	0.413
3	1.6e-5	3.0e-3	0.074	2.12	3.0e-3	0.03	0.09	1.55	2.74	0.413
4	5.1e-10	3.1e-5	0.028	1.94	3.1e-5	0.01	0.38	1.6	1.79	0.413
5	1.4e-19	2.7e-10	1.5e-4	2.03	2.7e-10	8.8e-6	5.4e-3	1.67	2.12	0.413
6	3.5e-38	2.6e-19	1.2e-7	1.99	2.6e-19	9.5e-10	7.7e-4	1.71	1.94	0.413
7	6.6e-76	1.9e-38	2.6e-15	2.01	1.9e-38	7.3e-20	2.3e-8	1.75	2.03	0.413
8	8.2e-151	1.2e-75	3.6e-29	2.0	1.2e-75	6.6e-38	1.4e-14	1.77	1.99	0.413
9	3.6e-301	4.4e-151	2.5e-58	2.0	4.4e-151	3.5e-76	7.0e-30	1.79	2.01	0.413
10	2.4e-601	6.7e-301	3.3e-115	2.0	6.7e-301	1.5e-150	1.3e-57	1.81	2.0	0.413
11	3.1e-1202	1.3e-601	2.1e-230	2.0	1.3e-601	1.9e-301	6.3e-116	1.83	2.0	0.413
12	1.8e-2403	5.9e-1202	2.2e-459	2.0	5.9e-1202	4.5e-601	1.1e-229	1.84	2.0	0.413
13	1.8e-4806	9.8e-2404	9.2e-919	2.0	9.8e-2404	1.7e-1202	4.2e-460	1.85	2.0	0.413
14	6.1e-9612	3.4e-4806	4.4e-1836	2.0	3.4e-4806	3.4e-2403	4.7e-918	1.86	2.0	0.413

Table 4 Example 2: Cumulative runs with initial data of the form (12) (first row), random u_0 without exact initialization (2nd), random u_0 without exact initialization with higher precision (3rd), random u^0 with exact initialization (4th)

$\|\|F\|\|^-$	$\|\|F\|\|^+$	q^+	$\mathcal{Q}_u^-$	$\mathcal{Q}_u^+$	β^+	Q^-	Q^+	$\mathcal{R}_B^-$	$\mathcal{R}_B^+$	$\mathcal{Q}_B^-$	$\mathcal{Q}_B^+$	$\|\|E\|\|^-$	$\|\|E\|\|^+$
2e-9996	9e-5005	3e-157	1.99	2.00	3e-157	3e-312	3e-79	1.76	2.24	1.98	2.00	1e-3	0.50
1e-7571	1e-5007	2e-136	1.29	1.45	2e-136	3e-144	1.00	1.38	1.57	1.00	1.44	1e-5	0.25
2e-145241	5e-100356	3e-1428	1.37	1.45	3e-1428	1e-1102	3e-213	1.40	1.53	1.16	1.44	1e-5	0.25
2e-9996	4e-5008	2e-157	1.99	2.00	2e-157	1e-315	2e-79	1.73	2.13	1.97	2.00	2e-8	0.20

Table 5 Example 3 a) and b): Two cumulative runs in a) with $(\sigma_k) \equiv 1$ (top) and $(\sigma_k) \equiv 0.99$ (below top) and in b) with $(\sigma_k) \equiv 1$ (above bottom) and $(\sigma_k) \equiv 0.99$ (bottom)

$\|\|F\|\|^-$	$\|\|F\|\|^+$	q^-	q^+	β^-	β^+	Q^-	Q^+	$\|\|E\|\|^-$	$\|\|E\|\|^+$	ι^-	ι^+
4e-501	1e-500	0.618	0.618	6e-230	9e-230	0.618	0.618	0.02	0.09	1180	1200
4e-501	1e-500	0.620	0.620	4e-230	7e-230	0.620	0.620	0.02	0.09	1180	1200
4e-501	9e-501	0.755	0.755	2e-309	4e-309	0.570	0.570	0.03	0.08	1340	1360
4e-501	1e-500	0.756	0.756	2e-309	3e-309	0.572	0.572	0.03	0.08	1350	1370

5.2.3 Example 3 a)

We turn to Theorem 6 2), where $F'(\bar{u})$ is singular. Let

$$F : \mathbb{R}^3 \to \mathbb{R}^3, \qquad F(u) = \begin{pmatrix} u_2^2 - 2u_3^3 \\ u_1 + u_2 + u_3 \\ 5u_1 \end{pmatrix}.$$

Because of $\mathcal{A}^\perp = \langle \{(0,1,-1)^T\} \rangle$ we have $\bar{s} = \frac{1}{\sqrt{2}}(0,1,-1)^T$, hence $F_1'(0) = 0$ and $F_1''(0)(\bar{s},\bar{s}) = 2 \neq 0$, which implies $\lim_{k\to\infty} q_k = \lim_{k\to\infty} Q_k = \frac{\sqrt{5}-1}{2} \approx 0.618$ for the choice $(\sigma_k) \equiv 1$ that we consider first. We use $\alpha = \hat{\alpha} = 0.01$ in this example. The results of a cumulative run with $(\sigma_k) \equiv 1$ are displayed in Table 5 and are in perfect agreement with Theorem 6 2). Table 5 also provides results for $(\sigma_k) \equiv 0.99$, which are similar to those for $(\sigma_k) \equiv 1$. Moreover, it features ι^- and ι^+, which denote the minimal, respectively, maximal number of iterations of all single runs within a cumulative run. As in the previous examples we consistently find $B_k \not\to F'(\bar{u})$.

Table 6 Example 3 b): Single run with $(\sigma_k) \equiv 1$

k	$\|\|F_k\|\|$	q_k	β_k	Q_k	$\|\|E_k\|\|$
0	0.023	-1	-1	-1	0.0704
1	6.1e-7	0.815	7.0e-5	-1	0.0704
2	6.2e-7	1.0	0.03	424	0.0639
3	1.8e-7	0.666	6.5e-5	0.002	0.0639
4	9.0e-8	0.79	7.7e-5	1.18	0.0639
5	3.7e-8	0.742	3.2e-5	0.42	0.0639
6	1.6e-8	0.76	2.1e-5	0.638	0.0639
7	6.8e-9	0.753	1.1e-5	0.545	0.0639
8	3.0e-9	0.756	6.5e-6	0.579	0.0639
9	1.3e-9	0.755	3.7e-6	0.566	0.0639
10	5.5e-10	0.755	2.1e-6	0.571	0.0639
⋮	⋮	⋮	⋮	⋮	⋮
500	1.6e-189	0.755	4.4e-126	0.57	0.0639
501	7.1e-190	0.755	2.5e-126	0.57	0.0639
502	3.0e-190	0.755	1.4e-126	0.57	0.0639
503	1.3e-190	0.755	8.1e-127	0.57	0.0639
⋮	⋮	⋮	⋮	⋮	⋮
1000	1.1e-372	0.755	3.3e-248	0.57	0.0639
1001	4.6e-373	0.755	1.9e-248	0.57	0.0639
1002	2.0e-373	0.755	1.1e-248	0.57	0.0639
⋮	⋮	⋮	⋮	⋮	⋮
1348	3.4e-500	0.755	3.3e-333	0.57	0.0639
1349	1.5e-500	0.755	1.9e-333	0.57	0.0639
1350	6.3e-501	0.755	1.1e-333	0.57	0.0639

5.2.4 Example 3 b)

We change F_1 in example 3 a), using $F_1(u) = u_2^3 - 2u_3^3$ instead. This results in $F_1'(0) = 0$, $F_1''(0)(\bar{s}, \bar{s}) = 0$ and $F_1'''(0)(\bar{s}, \bar{s}, \bar{s}) \neq 0$, so Theorem 6 2) implies $\lim_{k\to\infty} q_k \approx 0.755$ and $\lim_{k\to\infty} Q_k \approx 0.570$. Table 5 confirms this for $(\sigma_k) \equiv 1$ and shows that the choice $(\sigma_k) \equiv 0.99$ induces only marginal changes. Overall, example 3 exhibits a remarkably uniform convergence behavior of iterates and matrix updates, as evidenced, for instance, by the fact that $q^- = q^+$ and $Q^- = Q^+$. Table 6 exemplifies this for example 3 b) in a single run with $(\sigma_k) \equiv 1$. Since this uniformity is characteristic for singular $F'(\bar{u})$ of rank $n - 1$, cf. also [19], we used $\|F(u^k)\| \leq 10^{-500}$ as termination criterion in example 3 and the cumulative runs consisted of 100 single runs.

5.2.5 Example 4

To verify Theorem 7 1) we consider $F(u) = Au$, where $A \in \mathbb{R}^{10\times 10}$ is an invertible random matrix with entries in $[-1000, 1000]$ that is changed after each single run of the cumulative run. We choose $\alpha = \hat{\alpha} = 1000$. In the first cumulative run we use $\sigma_4 = 1$, $\sigma_k = 0.1$ otherwise. Theorem 7 1) guarantees $F(u^6) = 0$ if $F(u^k) \neq 0$ for $0 \leq k \leq 5$. Table 7 shows that $\iota^- = \iota^+ = 6$, so

Table 7 Example 4: Cumulative runs with $\sigma_{0,1,2,3,5} = 0.1$ and $\sigma_4 = 1$ (top), with $(\sigma_k) \equiv 1 - (k+2)^{-4}$ (middle), with $(\sigma_k) \equiv 1 - (k+2)^{-4}$ and higher precision (bottom)

$\|\|F\|\|^-$	$\|\|F\|\|^+$	q^+	$\mathcal{Q}_u^-$	$\mathcal{Q}_u^+$	β^+	Q^-	Q^+	$\mathcal{Q}_B^-$	$\mathcal{Q}_B^+$	$\|\|E\|\|^-$	$\|\|E\|\|^+$	ι^-	ι^+
3e-10001	1e-9995	1.04	-	-	-	9.00	9.00	-	-	2322	8388	6	6
2e-5251	1e-5001	3e-160	1.05	1.05	4e-162	5e-7	6e-7	1.03	1.03	2320	8060	49	50
2e-50881	1e-50008	1e-604	1.02	1.02	9e-605	1e-8	1e-8	1.01	1.01	3200	7340	130	131

all runs use exactly 6 steps. On a side note we remark that $Q^- = Q^+ = 9$ can easily be proven. The second experiment displayed in Table 7 uses $(\sigma_k) \equiv 1 - (k+2)^{-4}$. The outcome is in line with Theorem 7 1) that asserts global q-superlinear, but not finite convergence for this choice of (σ_k), as well as convergence of (B_k). As in example 1 it can be shown that the q-order of (u^k) and $(\|B_{k+1} - B_k\|)$ is 1. To verify this we repeat the cumulative run with $(\sigma_k) \equiv 1-(k+2)^{-4}$, using a precision of 100000 digits and $\|F(u^k)\| \leq 10^{-50000}$ as termination criterion, but only 100 single runs. The result in Table 7 is in line with q-orders of 1. Despite the fact that all B_k agree with A on $n-1$ of n rows, the difference between B_k and A in the last 25% of iterations is large in norm, which, however, does not prevent finite convergence if $\sigma_k = 1$ for at least one $k \geq 1$; cf. Theorem 7 and Remark 6.

6 Summary

We have shown that, up to a translation, the iterates of the Broyden–like method for mixed linear–nonlinear systems of equations can be obtained by applying the Broyden–like method to a lower-dimensional mapping, provided that the rows of the initial matrix agree with the rows of the Jacobian for (some of) the linear equations. We have used this subspace property to extend a sufficient condition for convergence of the Broyden–like matrices. For the special case that at most one equation is nonlinear we have concluded that the Broyden–like matrices converge whenever the iterates converge. For Broyden's method we could, in addition, quantify how fast iterates and updates converge, respectively, prove finite convergence if the system is linear. We verified the results in high-precision numerical experiments.

References

1. Al-Baali, M.: Extra updates for the BFGS method. Optim. Methods Softw. **13**(3), 159–179 (2000). DOI 10.1080/10556780008805781
2. Al-Baali, M., Spedicato, E., Maggioni, F.: Broyden's quasi-Newton methods for a nonlinear system of equations and unconstrained optimization: a review and open problems. Optim. Methods Softw. **29**(5), 937–954 (2014). DOI 10.1080/10556788.2013.856909
3. Broyden, C.: A class of methods for solving nonlinear simultaneous equations. Math. Comput. **19**, 577–593 (1965). DOI 10.2307/2003941
4. Broyden, C., Dennis, J., More, J.J.: On the local and superlinear convergence of quasi-Newton methods. J. Inst. Math. Appl. **12**, 223–245 (1973). DOI 10.1093/imamat/12.3.223

5. Conn, A.R., Gould, N.I.M., Toint, P.L.: Convergence of quasi-Newton matrices generated by the symmetric rank one update. Math. Program. **50**(2 (A)), 177–195 (1991). DOI 10.1007/BF01594934
6. Decker, D.W., Kelley, C.T.: Broyden's method for a class of problems having singular Jacobian at the root. SIAM J. Numer. Anal. **22**, 566–574 (1985). DOI 10.1137/0722034
7. Dennis, J., More, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**, 549–560 (1974). DOI 10.2307/2005926
8. Dennis, J., More, J.J.: Quasi-Newton methods, motivation and theory. SIAM Rev. **19**, 46–89 (1977). DOI 10.1137/1019005
9. Dennis, J., Schnabel, R.B.: Numerical methods for unconstrained optimization and nonlinear equations. Repr. Philadelphia, PA: SIAM (1996). DOI 10.1137/1.9781611971200
10. Díez, P.: A note on the convergence of the secant method for simple and multiple roots. Appl. Math. Lett. **16**(8), 1211–1215 (2003). DOI 10.1016/S0893-9659(03)90119-4
11. Fayez Khalfan, H., Byrd, R.H., Schnabel, R.B.: A theoretical and experimental study of the symmetric rank-one update. SIAM J. Optim. **3**(1), 1–24 (1993). DOI 10.1137/0803001
12. Gay, D.M.: Some convergence properties of Broyden's method. SIAM J. Numer. Anal. **16**, 623–630 (1979). DOI 10.1137/0716047
13. Ge, R., Powell, M.J.D.: The convergence of variable metric matrices in unconstrained optimization. Math. Program. **27**, 123–143 (1983). DOI 10.1007/BF02591941
14. Griewank, A.: Broyden updating, the good and the bad! Doc. Math. (Bielefeld) pp. 301–315 (2012). URL https://www.emis.de/journals/DMJDMV/vol-ismp/45_griewank-andreas-broyden.pdf
15. Kelley, C.: Iterative methods for linear and nonlinear equations. Philadelphia, PA: SIAM (1995). DOI 10.1137/1.9781611970944
16. Li, D., Fukushima, M.: A derivative-free line search and global convergence of Broyden-like method for nonlinear equations. Optim. Methods Softw. **13**(3), 181–201 (2000). DOI 10.1080/10556780008805782
17. Li, D., Zeng, J., Zhou, S.: Convergence of Broyden-like matrix. Appl. Math. Lett. **11**(5), 35–37 (1998). DOI 10.1016/S0893-9659(98)00076-7
18. Mannel, F.: On the $2n$–step q-quadratic convergence and the q-order of Broyden's method. Submitted (2020). URL https://imsc.uni-graz.at/mannel/Broy2n.pdf
19. Mannel, F.: On the convergence of Broyden's method and of the Broyden matrices for a class of singular problems. Submitted (2020). URL https://imsc.uni-graz.at/mannel/CGB_sing.pdf
20. Mannel, F.: On the convergence rate of Broyden–like methods. In preparation (2021)
21. Martínez, J.M.: Practical quasi-Newton methods for solving nonlinear systems. J. Comput. Appl. Math. **124**(1-2), 97–121 (2000). DOI 10.1016/S0377-0427(00)00434-9
22. More, J., Trangenstein, J.: On the global convergence of Broyden's method. Math. Comput. **30**, 523–540 (1976). DOI 10.2307/2005323
23. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer Series in Operations Research and Financial Engineering. Springer, New York (2006). DOI 10.1007/978-0-387-40065-5
24. O'Leary, D.P.: Why Broyden's nonsymmetric method terminates on linear equations. SIAM J. Optim. **5**(2), 231–235 (1995). DOI 10.1137/0805012
25. Ortega, J.M., Rheinboldt, W.C.: Iterative solution of nonlinear equations in several variables, vol. 30. Philadelphia, PA: SIAM (2000). DOI 10.1137/1.9780898719468
26. Powell, M.J.D.: A New Algorithm for Unconstrained Optimization. In: J. Rosen, O. Mangasarian, K. Ritter (eds.) Nonlinear Programming, pp. 31–65. Academic Press (1970). DOI 10.1016/B978-0-12-597050-1.50006-3
27. Powell, M.J.D.: How bad are the BFGS and DFP methods when the objective function is quadratic? Math. Program. **34**, 34–47 (1986). DOI 10.1007/BF01582161
28. Sachs, E.: Convergence rates of quasi-newton algorithms for some nonsmooth optimization problems. SIAM J. Control Optim. **23**, 401–418 (1985). DOI 10.1137/0323026
29. Stoer, J.: The convergence of matrices generated by rank-2 methods from the restricted β-class of Broyden. Numer. Math. **44**, 37–52 (1984). DOI 10.1007/BF01389753
30. Sun, L.: The convergence of quasi-Newton matrices generated by the self-scaling symmetric rank one update. Indian J. Pure Appl. Math. **29**(1), 51–58 (1998)
31. Vianello, M., Zanovello, R.: On the superlinear convergence of the secant method. Am. Math. Mon. **99**(8), 758–761 (1992). DOI 10.2307/2324244