On the order of convergence of Broyden's method

Faster convergence on mixed linear–nonlinear systems of equations and a conjecture on the q-order

Florian Mannel[*]

Preprint, 11th July 2021

Abstract

We present two theoretical results and two surprising conjectures concerning convergence properties of Broyden's method for smooth nonlinear systems of equations. First, we show that when Broyden's method is applied to a nonlinear mapping $F:\mathbb{R}^n \to \mathbb{R}^n$ with $n-d$ affine component functions and the initial matrix B_0 is chosen suitably, then the generated sequence $(u^k, F(u^k), B_k)_{k\geq 1}$ can be identified with a lower-dimensional sequence that is also generated by Broyden's method. This property enables us to prove, second, that for such mixed linear–nonlinear systems of equations a proper choice of B_0 ensures $2d$–step q-quadratic convergence, which improves upon the previously known $2n$ steps. Numerical experiments of high precision confirm the faster convergence and show that it is not available if B_0 deviates from the correct choice. In addition, the experiments suggest two surprising possibilities: It seems that Broyden's method is $(2d-1)$–step q-quadratically convergent for $d>1$ and that it admits a q-order of convergence of $2^{1/(2d)}$. These conjectures are new even for $d=n$.

Keywords. Broyden's method, quasi-Newton methods, systems of nonlinear equations, local convergence, Gay's theorem, $2n$–step quadratic convergence, q-order of convergence

AMS subject classifications. 49M15, 65H10, 65K05, 90C30, 90C53

1 Introduction

Given a smooth nonlinear mapping $F:\mathbb{R}^n \to \mathbb{R}^n$, Broyden's method aims at finding a root $\bar{u}$ of F,

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

It is one of the most widely used quasi-Newton methods for systems of nonlinear equations and converges locally q-superlinearly, as was shown by Broyden, Dennis and Moré in their seminal paper [BDM73]. We state the method as Algorithm BROY.

[*] University of Graz, Heinrichstr. 36, 8010 Graz, Austria (florian.mannel@uni-graz.at).

Algorithm BROY: Broyden's method

```
Input: $(u^0, B_0) \in \mathbb{R}^n \times \mathbb{R}^{n\times n}$, $B_0$ invertible
for $k = 0, 1, 2, \ldots$ 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 $y^k := F(u^{k+1}) - F(u^k)$
    Let $B_{k+1} := B_k + (y^k - B_k s^k)\frac{(s^k)^T}{\|s^k\|^2}$
end
Output: $u^*$
```

Before discussing the content of this paper in more detail, let us outline its main contributions:

- A well-known result of Gay [Gay79, Theorem 3.1] asserts local $2n$–step q-quadratic convergence of Broyden's method under appropriate assumptions. We show under the same assumptions that if $n-d$ of the equations are affine and the corresponding $n-d$ rows of B_0 agree with the corresponding $n-d$ rows of F', then Broyden's method is locally $2d$–step q-quadratically convergent.
- We provide high-precision numerical experiments that confirm the improved convergence speed and observe that it is lost if the relevant rows of B_0 are perturbed.
- The experiments suggest that Broyden's method enjoys a q-order of convergence no smaller than $2^{1/(2d)}$. This is the first time that the q-order of Broyden's method is studied numerically, and even for $d = n$ the conjecture that a q-order larger than one may exist is novel.

The starting point of this work is the property of Algorithm BROY that if $n-d$ rows of F' are constant and the initial guess B_0 matches these rows exactly, then there exist d-dimensional subspaces $\mathcal{S} \subset \mathbb{R}^n$ and $\mathbb{R}^n_d \subset \mathbb{R}^n$ such that $(s^k)_{k\geq 1} \subset \mathcal{S}$ and $(F(u^k))_{k\geq 1} \subset \mathbb{R}^n_d$, where $\mathbb{R}^n_d := \{(y_1, \ldots, y_n)^T \in \mathbb{R}^n : y_j = 0 \ \forall j > d\}$ and where we have assumed without loss of generality that the constant rows of F' are the last $n-d$ ones. This subspace property is well-known and appears, for instance, in the classical book of Dennis and Schnabel [DS96]. To the best of the author's knowledge, however, it has not been noted before that in this situation the sequence $(u^k, F(u^k), B_k)_{k\geq 1}$ can be identified with a sequence $(w^k, G(w^k), C_k)_{k\geq 0} \subset \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d}$ that is generated by applying Broyden's method to a suitable mapping $G : \mathbb{R}^d \to \mathbb{R}^d$. We stress that many well-known quasi-Newton methods do *not* have this property, e.g., the BFGS update, cf. Remark 2.3 in Section 2. We will use it to show that $(u^k)_{k\geq 1}$ is $2d$–step q-quadratically convergent under the assumptions of Gay's theorem on $2n$–step q-quadratic convergence (which coincide with the classical assumptions for local q-superlinear convergence of Broyden's method). As a corollary we obtain that (u^k) exhibits an *r-order of convergence* [OR00, Section 9.2] no smaller than $2^{\frac{1}{2d}}$. We emphasize that no modification of Algorithm BROY is necessary to enjoy the faster convergence; it is only required to choose B_0 in such a way that it agrees with F' on the rows that correspond to affine components of F. On the other hand, it may be numerically advantageous to carry out Algorithm BROY for G instead of F, for instance because smaller linear systems have to be solved; cf. also Remark 2.6 2).

It is clear that there is an abundance of practically relevant nonlinear systems of equations that contain some linear equations; these systems are covered by the result on $2d$–step q-quadratic convergence. In addition, this result supports two standard suggestions for the choice of B_0, which are to use either $B_0 = F'(u^0)$ or a componentwise finite difference approximation of $F'(u^0)$, cf. for instance [NW06, comment after Theorem 11.5] and [Mar00, Section 10], since both choices imply that B_0 and F' agree on rows associated to affine component functions of F.

We confirm the $2d$–step convergence in numerical experiments of high precision and observe that it is lost if the rows of B_0 that correspond to affine components of F are perturbed, while perturbations in other rows have no such effect. This shows that choosing B_0 to match the constant rows of F' (if possible) will usually decrease both iteration count and runtime.

Besides confirming the $2d$–step q-quadratic convergence of Broyden's method on mixed systems of equations, the numerical experiments lead us to three conjectures: The iterates (u^k) of Broyden's method

- may converge $(2d-1)$–step q-quadratically for $d \in \{2, \ldots, n\}$ (which, if true, implies an r-order of convergence no smaller than $2^{\frac{1}{2d-1}}$),
- may exhibit a *q-order of convergence* [OR00, Section 9.1] larger than one,
- may admit the lower bound $2^{1/(2d)}$ for their q-order for $d \in \{1, \ldots, n\}$.

Even for $d = n$ (i.e., fully nonlinear systems) these conjectures are new and perhaps somewhat surprising. We therefore stress that the conjectured lower bound of $2^{1/(2n)}$ for the q-order of Broyden's method is consistent with further numerical experiments of the author, e.g. in [Man21a, Man21c, Man21b]. Also, we are aware that the $2n$–step q-quadratic convergence is derived from the $2n$–step convergence of Broyden's method on regular linear systems [Gay79, O'L95, GL81] and that it is more or less accepted that fewer than $2n$ iterations rarely suffice in the case of linear systems. On the other hand, since the present work is the first to numerically assess the $2n$–step q-quadratic convergence and the q-order of convergence of Broyden's method, no numerical evidence that contradicts the above conjectures is available.

Broyden's method is one of the most prominent quasi-Newton methods for solving nonlinear equations, and there is an ever-growing body of literature available. For the case of smooth systems of equations, the surveys [ASM14, DM77, Gri12, Mar00] cover many aspects and provide further references, so we restrict ourselves to mentioning [BDM73, Gri87, Sac86] that develop the local convergence theory of Broyden's method. Works that are too recent to be included in the surveys are, for instance, concerned with the extension of Broyden's method to constrained nonlinear systems of equations [MMP18], to set-valued mappings [ABDL14], and to single- and set-valued nonsmooth problems in infinite-dimensional spaces [MHMP13, AN18, MR20]. Broyden's method is also used in implicit deep learning, cf. [BKK19, BKK20], which further underlines that it continues to play a role nowadays. Despite the vast amount of contributions, we are not aware of works that contain the identification of $(u^k, F(u^k), B_k)_{k\geq 1}$ with $(w^k, G(w^k), C_k)_{k\geq 0}$ or that exploit the presence of affine equations to prove faster convergence. We acknowledge that the use of exact initialization of constant rows can be regarded as a special case of the least-change update theory [DW81], of structured Broyden methods [Sch70, Mar79, AC79, HK92], and of reduced quasi-Newton methods [Gil89, GL01], but in these and similar contributions it is not specified how the rate of convergence depends on the dimension of the underlying spaces, so they yield no improvement for the situation at hand.

There is also a considerable amount of numerical studies available for Broyden's method. Since many of the aforementioned works contain numerical experiments and the surveys include references to a number of studies, we mention only [SH97] and the monograph [Kel95].

This paper is organized as follows. In Section 2 we present and prove the relationship between $(u^k, F(u^k), B_k)_{k\geq 1}$ and its lower-dimensional counterpart $(w^k, G(w^k), C_k)_{k\geq 0}$, and in Section 3 we use it to establish the result on local $2d$–step q-quadratic convergence. Section 4 is devoted to numerical experiments, and Section 5 provides a summary.

Notation. We use $\mathbb{N} = \{1, 2, 3, \ldots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. By $\|\cdot\|$ we indicate the Euclidean norm for vectors, respectively, the spectral norm for matrices. For $A \in \mathbb{R}^{n\times n}$, A^j indicates the j-th row of A, regarded as a row vector, while $A^{i,j} \in \mathbb{R}$ is an entry of A and A_k is a member of the matrix sequence (A_k). The linear hull of $C \subset \mathbb{R}^n$ is denoted by $\langle C\rangle$ and its orthogonal complement is denoted by $C^\perp$.

2 A subspace property of Broyden's method

The following assumption presents the setting that we are interested in.

Assumption 2.1. Let $n \in \mathbb{N}$, $d \in \{0, 1, \ldots, n\}$ and $J := \{d+1, d+2, \ldots, n\}$. 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 $u^0 \in \mathbb{R}^n$ and $B_0 \in \mathbb{R}^{n\times n}$ invertible with $B_0^j = a_j^T$ for all $j \in J$.

The first lemma describes the behavior of Algorithm BROY on affine components of F provided that B_0 is initialized with the rows of F' for these components. It appears in [DS96, Section 8.5, Exercise 10], albeit without proof.

Lemma 2.2. *Let Assumption 2.1 hold and let (u^k) and (B_k) be generated by Algorithm BROY with initial guess (u^0, B_0). Then we have for all $j \in J$ 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. Let $j \in J$. From $B_0^j s^0 = -F_j(u^0)$ and $F_j'(u^0) = a_j^T = B_0^j$ we deduce

$$F_j(u^1) = F_j(u^0 + s^0) = F_j(u^0) + F_j'(u^0)s^0 = F_j(u^0) + B_0^j s^0 = 0.$$

Since $B_0 s^0 = -F(u^0)$ implies $y^0 - B_0 s^0 = F(u^1)$, the Broyden update formula yields $B_1^j = B_0^j = a_j^T$. From $B_1^j s^1 = -F_j(u^1) = 0$ we infer that $a_j^T s^1 = 0$. This implies $F_j(u^2) = F_j(u^1 + s^1) = F_j(u^1) + F_j'(u^1)s^1 = 0$, hence $B_2^j = B_1^j = a_j^T$. By induction we confirm for all $k \geq 1$ that $B_k^j = a_j^T$, $F_j(u^k) = 0$, and $a_j^T s^k = 0$. The update formula entails $(B_{k+1} - B_k)a_j = 0$ for all $k \geq 1$, whence $B_k a_j = B_1 a_j$ for these k. □

Remark 2.3. Several other quasi-Newton methods do not have the property described in Lemma 2.2. Let us consider the BFGS method as an example. We choose $F(u) := u$ (obtained as $F = \nabla f$ for $f(u) = \frac{1}{2}\|u\|^2$), $u_0 = (1, 1)^T$ and $B_0 = \operatorname{diag}(1, 2)$, so that $B_0^1 = (1, 0) = F_1'$. It is easy to confirm that $u^1 = F(u^1) = (0, \frac{1}{2})^T$, but $B_1^1 = \frac{1}{15}(17, -4) \neq B_0^1$ and $F_1(u^2) = -\frac{2}{25} \neq 0$.

To state the main result of this section we introduce the following notation.

Definition 2.4. Let Assumption 2.1 hold. For any matrix $B \in \mathbb{R}^{n\times n}$ we set

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

We now establish the fundamental property of Broyden's method that under Assumption 2.1 the sequence $(u^k, F(u^k), B_k)_{k\geq 1}$ can be identified with a sequence $(w^k, G(w^k), C_k)_{k\geq 0}$ that is generated by applying Broyden's method to a suitable mapping G from $\mathbb{R}^d$ into $\mathbb{R}^d$.

Theorem 2.5. *Let Assumption 2.1 hold and let (u^k) and (B_k) be generated by Algorithm BROY with initial guess (u^0, B_0). Suppose that each B_k is invertible. Let $S \in \mathbb{R}^{n\times d}$ be a matrix whose columns form an orthonormal basis of $\langle \{a_j\}_{j\in J}\rangle^\perp$. Define*

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

as well as

$$C_0 := \widetilde{B}_1 S \in \mathbb{R}^{d\times d} \qquad \text{and} \qquad w^0 := 0 \in \mathbb{R}^d.$$

Then C_0 is invertible and the application of Algorithm BROY to G with initial guess (w^0, C_0) 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*

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

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 $k \geq 1$*

$$\bar{u} = u^1 + S\bar{w} \qquad \text{and} \qquad \|u^k - \bar{u}\| = \|w^{k-1} - \bar{w}\|. \tag{2}$$

Proof. Before we prove 1) and 2), let us point out that S is well-defined. This follows since the invertibility of B_0 implies that $\mathcal{A} := \langle \{a_j\}_{j\in J}\rangle$ has dimension $n-d$, hence $\mathcal{S} := \mathcal{A}^\perp$ has dimension d.

Proof of 1): We show first that the invertibility of B_k implies the invertibility of $\widetilde{B}_k S$. Let $k \geq 0$ and let $\widetilde{B}_k S v = 0$ for some $v \in \mathbb{R}^d$. For $w := Sv$ we have $\widetilde{B}_k w = 0$. Since $w \in \mathcal{S} = \mathcal{A}^\perp$ we obtain $B_k^j w = 0$ for all $j \in J$, where we used that $B_k^j = a_j^T$ for all $k \geq 0$ by Lemma 2.2. We infer that $B_k w = 0$, hence $w = 0$, thus $v = 0$, which shows that $\widetilde{B}_k S$ is invertible.

We now prove that Algorithm BROY generates sequences (w^k) and (C_k), that each C_k is invertible, and that the first and last of the three asserted equalities in (1) hold. To proceed by induction, we note that $w^0 = 0$, so $u^1 + Sw^0 = u^1$ holds. Also, $C_0 = \widetilde{B}_1 S$ by definition, so C_0 is invertible by the first part of the proof. For the induction step let $k \in \mathbb{N}$ and assume that w^{k-1} and C_{k-1} satisfying $u^k = u^1 + Sw^{k-1}$ and $C_{k-1} = \widetilde{B}_k S$ have been generated and that C_{k-1} is invertible. Since $s^k \in \mathcal{S}$ by Lemma 2.2 and since the columns of S are a basis of $\mathcal{S}$, there is a vector $\lambda \in \mathbb{R}^d$ such that $s^k = S\lambda$. The i-th equation of

$B_k s^k = -F(u^k)$ thus reads $B_k^i S\lambda = -F_i(u^1 + Sw^{k-1})$ for $i \in \{1, \ldots, n\}$, where we used the induction assumption. By definition, the first d of these equations can be expressed as $\widetilde{B}_k S\lambda = -G(w^{k-1})$. Since $C_{k-1} = \widetilde{B}_k S$ by induction assumption and C_{k-1} is regular, it follows that the linear system $C_{k-1} s_w^{k-1} = -G(w^{k-1})$ in Algorithm BROY has the unique solution $s_w^{k-1} = \lambda$, hence $w^k = w^{k-1} + s_w^{k-1}$ exists, and we obtain $s^k = S\lambda = Ss_w^{k-1} = S(w^k - w^{k-1})$. Adding u^k and using the induction assumption $u^k = u^1 + Sw^{k-1}$ this yields $u^{k+1} = u^1 + Sw^k$, which is the first equality in (1). We observe that

$$\|s^k\| = \|S\lambda\| = \|Ss_w^{k-1}\| = \|s_w^{k-1}\|, \tag{3}$$

where the last equality follows since the columns of S are orthonormal. To conclude the induction it is left to demonstrate the third equality in (1) and the invertibility of C_k. Invoking $S^T S = I \in \mathbb{R}^{d\times d}$, $C_{k-1} = \widetilde{B}_k S$, (3) and $Ss_w^{k-1} = s^k$ we infer that

$$\begin{aligned} C_k &= C_{k-1} + \frac{G(w^k) - G(w^{k-1}) - C_{k-1}s_w^{k-1}}{\|s_w^{k-1}\|^2}(s_w^{k-1})^T S^T S \\ &= \widetilde{B}_k S + \frac{\widetilde{F}(u^{k+1}) - \widetilde{F}(u^k) - \widetilde{B}_k s^k}{\|s^k\|^2}(s^k)^T S, \end{aligned} \tag{4}$$

where we used the first equality from (1) to rewrite the argument of $\widetilde{F}$ as u^{k+1}, respectively, u^k. Since $F(u^k) \neq 0$ implies $s^k \neq 0$, we conclude from (3) that C_k exists and from (4) that it satisfies

$$C_k = \widetilde{B}_k S + \left(\widetilde{B}_{k+1} - \widetilde{B}_k\right) S = \widetilde{B}_{k+1} S,$$

as desired. By the first part of the proof this representation of C_k also implies that C_k is invertible, which concludes the induction. The remaining second equality of (1) follows from the first and the definition of G.

Proof of 2): All claims follow readily by use of $u^k = u^1 + Sw^{k-1}$. □

Remark 2.6.

1) To illustrate Theorem 2.5, let us consider the special case that S is given by the first d columns of the $n \times n$ identity matrix. In the notation of the previous proof this corresponds to $\mathcal{S} = \mathbb{R}^n_d := \{(y_1, \ldots, y_n)^T \in \mathbb{R}^n : y_j = 0 \ \forall j > d\}$. In this setting we find that $C_{k-1}^{i,j} = B_k^{i,j}$ for $i, j \in \{1, \ldots, d\}$, i.e., C_{k-1} is the $d \times d$ submatrix of B_k that results from deleting the last $n - d$ rows and columns of B_k. Due to $(s^k)_{k\geq 1} \subset \mathbb{R}^n_d$ and $(F(u^k))_{k\geq 1} \subset \mathbb{R}^n_d$ the Broyden update affects only the entries of this submatrix of B_k for $k \geq 1$. Similarly, for $k \geq 1$ only the first d entries of u^k, respectively, of $F(u^k)$ change, and w^{k-1}, respectively, $G(w^{k-1})$ contain exactly these entries.

2) Theorem 2.5 1) shows that once u^1 is computed, all further iterates (u^k, B_k, F_k), $k \geq 2$, can be obtained by an application of Broyden's method to G with initial guess (w^0, C_0). Using G instead of F may reduce the numerical costs, for instance because the linear systems that have to be solved involve the $d \times d$ matrix C_k rather than the $n \times n$ matrix B_k. Moreover, if F is used it is possible that rounding errors destroy the properties $(s^k)_{k\geq 1} \subset \mathcal{S}$ and $(F(u^k))_{k\geq 1} \subset \mathbb{R}^n_d$, which cannot happen if G is used instead. The properties $(s^k)_{k\geq 1} \subset \mathcal{S}$ and $(F(u^k))_{k\geq 1} \subset \mathbb{R}^n_d$ are crucial to obtain the improved convergence result of this work, both in theory and practice; loosing them slows down the convergence.

3 Improved convergence for mixed linear–nonlinear systems

In this section we show that Gay's result on local $2n$–step q-quadratic convergence of Broyden's method can be improved if F has some affine component functions. This requires the notion of multi–step q-quadratic converge.

Definition 3.1. Let $(u^k) \subset \mathbb{R}^n$ with $\lim_{k\to\infty} u^k = \bar{u}$ for some $\bar{u}$. Then (u^k) is called *m–step q-quadratically convergent*, $m \in \mathbb{N}$, iff there is a constant $C > 0$ such that

$$\|u^{k+m} - \bar{u}\| \le C\|u^k - \bar{u}\|^2 \qquad \forall k \ge 0$$

is satisfied.

Remark 3.2. For $m = 1$ this is q-quadratic convergence in the usual sense.

Gay's theorem is based on the following assumption.

Assumption 3.3. Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable in a neighborhood of some $\bar{u}$ with $F(\bar{u}) = 0$. Let there be $L > 0$ such that $\|F'(u) - F'(\bar{u})\| \le L\|u - \bar{u}\|$ is satisfied for all u in that neighborhood. Let $F'(\bar{u})$ be invertible.

Gay's theorem on local $2n$–step q-quadratic convergence, cf. [Gay79, Theorem 3.1], reads as follows.

Theorem 3.4. *Let Assumption 3.3 hold. Then there exist $\delta, \varepsilon > 0$ and $C > 0$ such that for all (u^0, B_0) with $\|u^0 - \bar{u}\| \le \delta$ and $\|B_0 - F'(\bar{u})\| \le \varepsilon$, Algorithm BROY either terminates with output $u^* = \bar{u}$ or it generates a sequence (u^k) that converges to $\bar{u}$ and satisfies*

$$\|u^{k+2n} - \bar{u}\| \le C\|u^k - \bar{u}\|^2 \qquad \forall k \ge 0.$$

Remark 3.5.

1) It is well known that under the assumptions of Theorem 3.4, (u^k) is q-superlinearly convergent and the sequences $(\|B_k\|)$ and $(\|B_k^{-1}\|)$ are bounded.
2) While it is not required by Definition 3.1, we note that the constant C in Theorem 3.4 is independent of (u^0, B_0).

The following result improves Theorem 3.4 in the presence of linear equations.

Theorem 3.6. *Let Assumption 3.3 hold. Let $d \in \{0, 1, \dots, n\}$, $J := \{d+1, d+2, \dots, n\}$ and suppose that F satisfies $F_j(u) = a_j^T u + b_j$ for all $j \in J$, where $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$ for all $j \in J$. Then there exist $\delta, \varepsilon > 0$ and $C > 0$ such that for all (u^0, B_0) with $\|u^0 - \bar{u}\| \le \delta$, $\|B_0 - F'(\bar{u})\| \le \varepsilon$ and $B_0^j = a_j^T$ for all $j \in J$, Algorithm BROY either terminates with output $u^* = \bar{u}$ or it generates a sequence (u^k) that converges to $\bar{u}$ and satisfies*

$$\|u^{k+2d} - \bar{u}\| \le C\|u^k - \bar{u}\|^2 \qquad \forall k \ge 1. \tag{5}$$

Proof. If Algorithm BROY terminates with $u^* = \bar{u}$, then there is nothing to show, so we can assume that this termination does not occur. Since the assumptions of Theorem 3.4 are satisfied, it follows that Algorithm BROY successfully generates (u^k) and (B_k) and that (u^k) converges to $\bar{u}$. Since each B_k is invertible, we can apply Theorem 2.5. In particular, this endows us with a matrix S, a mapping G, sequences (w^k) and (C_k), and a point $\bar{w}$, all satisfying the properties stated in that theorem. We observe that $G(\bar{w}) = \widetilde{F}(u^1 + S\bar{w}) = \widetilde{F}(\bar{u}) = 0$ by the definition of G and (2). Using again that $u^1 + S\bar{w} = \bar{u}$, we infer that

$$\begin{aligned}\|G'(w) - G'(\bar{w})\| &= \|\widetilde{F}'(u^1 + Sw) - \widetilde{F}'(u^1 + S\bar{w})\| \\ &= \|F'(u^1 + Sw) - F'(u^1 + S\bar{w})\| \leq L\,\|S(w - \bar{w})\| = L\|w - \bar{w}\|\end{aligned}$$

for all w sufficiently close to $\bar{w}$. Thus, Theorem 3.4 is applicable to G if $G'(\bar{w}) = \widetilde{F}'(\bar{u})S$ is invertible and if the norms $\|w^0 - \bar{w}\|$ and $\|C_0 - G'(\bar{w})\|$ are sufficiently small. Once these properties are established, Theorem 3.4 yields

$$\|w^{k+2d} - \bar{w}\| \leq C\|w^k - \bar{w}\|^2 \qquad \forall k \geq 0,$$

from which (5) follows by virtue of the second identity in (2). It remains to show the aforementioned properties. By (3.5) in Gay's proof of [Gay79, Theorem 3.1] (or, alternatively, by the q-linear convergence of Broyden's method) we can assume without loss of generality that $\|u^1 - \bar{u}\| \leq \|u^0 - \bar{u}\|$. Therefore, it is evident that $\|w^0 - \bar{w}\| = \|u^1 - \bar{u}\|$ becomes sufficiently small if δ is small enough. Using that $B_1^j = a_j^T = F_j'(\bar{u})$ for all $j \in J$ by Lemma 2.2, we deduce

$$\|C_0 - G'(\bar{w})\| = \|\widetilde{B}_1 S - \widetilde{F}'(\bar{u})S\| = \|\widetilde{B}_1 - \widetilde{F}'(\bar{u})\| = \|B_1 - F'(\bar{u})\|.$$

Thus, it follows from (3.6) in the proof of [Gay79, Theorem 3.1] (or, alternatively, from the well-known bounded deterioration principle) that $\|C_0 - G'(\bar{w})\| \leq 2\varepsilon$. The invertibility of $F'(\bar{u})$ implies the invertibility of $G'(\bar{w}) = \widetilde{F}'(\bar{u})S$ verbatim as in the first part of the proof of Theorem 2.5. □

Remark 3.7. If (u^k) satisfies (5) for some $C > 0$, then it also satisfies Definition 3.1 for $m = 2d$ and the constant $\hat{C} := \max\{C, \|u^{2d} - \bar{u}\|/\|u^0 - \bar{u}\|^2\}$, so (u^k) is indeed $2d$–step q-quadratically convergent. Note, however, that while the constant C in Theorem 3.6 is independent of (u^0, B_0), this may no longer be true for $\hat{C}$.

Theorem 3.6 yields the following bound for the r-order of convergence [OR00, Section 9.2] of Broyden's method.

Corollary 3.8. *Any sequence (u^k) that satisfies (5) and converges to $\bar{u}$ admits an r-order of convergence of at least $p_0 := \sqrt[2d]{2}$ if $d \geq 1$.*

Proof. Without loss of generality we can assume that the constant C appearing in (5) satisfies $C \geq 1$. Let $D := 2d$. Since (u^k) converges to $\bar{u}$, there is $T \in \mathbb{N}$ such that $\|u^k - \bar{u}\| < 1$ for all $k \geq T$. By induction it follows from (5) that

$$\|u^{t+jD} - \bar{u}\| \leq C^{(2^j-1)} \cdot \|u^t - \bar{u}\|^{(2^j)}$$

for all $t \in \{T, T+1, \dots, T+D-1\}$ and all $j \in \mathbb{N}_0$. As $p_0^{t+jD} = 2^{\frac{t}{D}+j}$ and $C \geq 1$, this readily yields

$$\|u^{t+jD} - \bar{u}\|^{\frac{1}{p_0^{t+jD}}} \leq \|u^t - \bar{u}\|^{\frac{1}{2^{\frac{t}{D}}}} =: \alpha_t < 1$$

for all t and j as before. Setting $\alpha := \max_{T \leq t \leq T+D-1} \alpha_t$ it follows that

$$\|u^k - \bar{u}\|^{\frac{1}{p_0^k}} \leq \alpha < 1$$

for all $k \geq T$, which proves the claim. □

Remark 3.9. Corollary 3.8 shows that in the setting of Theorem 3.6 the r-order of convergence of Broyden's method is bounded from below by $2^{1/(2d)}$ for $d \geq 1$. In addition, the numerical experiments in Section 4 suggest that for $d > 1$, Broyden's method may be $(2d-1)$–step q-quadratically convergent, which would imply that the r-order is no smaller than $2^{1/(2d-1)}$. While the exact r-order of Broyden's method remains unknown, cf. also [Gri12, pp. 308–310], we mention that the exact r-order is known for the *adjoint Broyden method*, cf. [GSW08].

4 Numerical experiments

We provide numerical results to verify the $2d$–step q-quadratic convergence established in Theorem 3.6. We also assess the q-order of convergence of Broyden's method, which has not been done before. We will see that the numerical results are consistent with the following conjectures C1 and C2:

$$\boxed{\text{C1: } \textit{Broyden's method has q-order at least } 2^{1/(2d)} \textit{ for } d \geq 1.}$$

$$\boxed{\text{C2: } \textit{Broyden's method is } (2d-1)\textit{–step q-quadratically convergent for } d \geq 2.}$$

Here, as before, we have supposed that $n-d$ of n equations are linear. Both conjectures are new even for $d = n$ and, in fact, the existence of a nontrivial q-order for Broyden's method has not been proposed before.

In Section 4.1 we present the design of the experiments, Section 4.2 contains the examples and results.

4.1 Design of the experiments

4.1.1 Implementation and accuracy

The experiments are carried out in Matlab 2017a using its *variable precision arithmetic (vpa)* with a precision of 1000 digits. We replace the termination criterion $F(u^k) = 0$ in Algorithm BROY by $\|F(u^k)\| \leq 10^{-320}$. The rather small residual tolerance of 10^{-320} ensures that the number of iterations is large enough to meaningfully assess asymptotic properties such as the q-order of convergence. Of course, the small residual tolerance necessitates the usage of sufficiently many digits. By $\bar{k} \geq 0$ we denote the final value of k in Algorithm BROY.

4.1.2 Known solution and random initialization

All examples have an explicitly available solution $\bar{u}$ and the initial guesses are made in such a way that convergence to $\bar{u}$ takes place in all runs. In all examples $F'(\bar{u})$ is invertible. The initial point u^0 is generated using MATLAB'S function `rand`, which produces uniformly distributed random numbers, and satisfies $u^0 \in \bar{u} + [-10^{-3}, 10^{-3}]^n$. For B_0 we choose $B_0 = F'(u^0) + \hat{\alpha}\|F'(u^0)\|R$, where $\hat{\alpha} \in \{0, 10^{-30}, 10^{-10}, 10^{-3}\}$ and where $R \in \mathbb{R}^{n\times n}$ is a random matrix whose entries belong to $[-1, 1]$, but that has a particular structure in which many entries are zero. Specifically, we use two schemes for R: Either R affects only those rows of B_0 that correspond to nonlinear components of F, in which case $R^j = 0$ for all $j \in J$ (cf. Assumption 2.1 for notation), or it affects only those rows that correspond to affine components, in which case $R^j = 0$ for all $j \in \{1, \dots, d\}$. In the first case we want to demonstrate that the perturbation has essentially no effect, so we allow R to be nonzero in the entire rows, i.e., for each $j \in \{1, \dots, d\}$ the row R^j is randomly drawn from $[-1, 1]^n$. In the second case the aim is to show that even minimal perturbations significantly decrease the order of convergence, so we modify only one entry of B_0 per row, i.e., for each $j \in J$ all entries of R^j except one are zero. The nonzero entry is taken to be a random number in $[-1, 1]$ and its position within the row is random, too. We denote $\hat{\alpha} = \hat{\alpha}_n$ (nonlinear) in the first case and $\hat{\alpha} = \hat{\alpha}_l$ (linear) in the second.

4.1.3 Quantities of interest

Let (u^k) be generated by Algorithm BROY. For $k \geq 0$ we define

$$F_k := F(u^k), \qquad \rho_k^m := \frac{\log(\|u^k - \bar{u}\|)}{\log(\|u^{k-m} - \bar{u}\|)} \qquad \text{and} \qquad C_k^m := \frac{\|u^k - \bar{u}\|}{\|u^{k-m} - \bar{u}\|^2},$$

where $m \in \{1, \dots, 2n\}$ will be specified for each example. Whenever any of these quantities is undefined we set it to -1; e.g., $\rho_k^m := -1$ for all $k \in \{0, \dots, m-1\}$.

4.1.4 Single runs and cumulative runs

We perform *single runs* and *cumulative runs*. For single runs we display the quantities of interest during the course of Algorithm BROY. For cumulative runs we perform 10000 single runs with varying initial data as described in Section 4.1.2. For each of the 10000 single runs we compute

$$\hat{\rho}_m^j := \min_{k_0(j) \leq k \leq \bar{k}(j)} \rho_k^m \qquad \text{and} \qquad \hat{C}_m^j := \max_{k_0(j) \leq k \leq \bar{k}(j)} C_k^m,$$

where $j \in I := \{1, \dots, 10000\}$ indicates the respective single run and we always use the value $k_0(j) := \lfloor 0.75\bar{k}(j) \rfloor$. As outcome of a cumulative run we display

$$\rho_m^- := \min_{j\in I} \hat{\rho}_m^j, \qquad \rho_m^+ := \max_{j\in I} \hat{\rho}_m^j,$$

$$C_m^- := \min_{j\in I} \hat{C}_m^j, \qquad C_m^+ := \max_{j\in I} \hat{C}_m^j$$

for several values of m. In case of m–step q-quadratic convergence we expect $(\hat{C}_m^j)_j$ to be bounded from above (due to the uniformity of C discussed in Remarks 3.5 and 3.7) and $(\hat{\rho}_m^j)_j$

Table 1: Example 1 a): Single run with $\hat{\alpha} = 0$

k	$\|\|F_k\|\|$	ρ_k^1	ρ_k^4	ρ_k^3	ρ_k^2	C_k^4	C_k^3	C_k^2
0	4.4e-3	-1	-1	-1	-1	-1	-1	-1
1	2.4e-6	2.08	-1	-1	-1	-1	-1	-1
2	2.4e-9	1.44	-1	-1	3.01	-1	-1	1.1e-3
3	6.4e-13	1.38	-1	4.15	1.99	-1	4.6e-7	1.2
4	2.2e-16	1.28	5.32	2.55	1.77	1.6e-10	4.1e-4	111
5	2.2e-22	1.38	3.53	2.44	1.77	4.1e-10	1.1e-4	600
6	3.4e-31	1.41	3.44	2.49	1.94	1.7e-13	9.2e-7	7.6
7	2.7e-41	1.33	3.32	2.59	1.87	7.2e-17	6.0e-10	600
8	6.0e-52	1.26	3.27	2.36	1.68	1.3e-20	1.4e-8	5.7e9
9	2.7e-67	1.30	3.07	2.18	1.64	6.0e-24	2.5e-6	4.2e14
10	6.2e-91	1.35	2.96	2.22	1.76	5.9e-30	9.8e-10	1.9e12
11	5.8e-120	1.32	2.94	2.33	1.79	9.1e-39	1.8e-17	9.2e13
12	2.3e-151	1.26	2.94	2.26	1.67	7.1e-49	3.6e-18	6.5e29
13	1.0e-192	1.27	2.88	2.13	1.61	1.6e-59	2.9e-12	3.4e46
14	2.5e-255	1.33	2.82	2.13	1.69	7.1e-75	8.2e-17	5.3e46
15	5.0e-337	1.32	2.82	2.23	1.75	1.7e-98	1.1e-35	5.5e47

Table 2: Example 1 a): Cumulative runs with $\hat{\alpha} = 0$ (top) and $\hat{\alpha}_n = 10^{-3}$ (bottom)

ρ_1^-	ρ_1^+	ρ_4^-	ρ_4^+	ρ_3^-	ρ_3^+	ρ_2^-	ρ_2^+	C_4^-	C_4^+	C_3^-	C_3^+	C_2^-	C_2^+
1.20	1.29	2.76	2.90	1.99	2.21	1.50	1.69	1e-50	2e-36	3e-20	27.0	2e44	9e106
1.20	1.30	2.73	2.92	1.99	2.22	1.50	1.69	1e-53	4e-31	3e-20	71.8	9e43	3e111

to be bounded from below by (approximately) 2, and this should be reflected in C_m^+ and ρ_m^-, respectively. Correspondingly, C_{m+1}^-, C_{m+1}^+, respectively, C_{m-1}^-, C_{m-1}^+ should be indicative of null sequences, respectively, unbounded growth, while ρ_{m+1}^-, ρ_{m+1}^+, respectively, ρ_{m-1}^-, ρ_{m-1}^+ should be clearly above, respectively, below 2. If $\liminf_{k\to\infty} \rho_k^1 \geq \rho$, then the q-order of (u^k) is no smaller than ρ. The r-order is at least as large as the q-order, cf. [OR00, 9.3.3.]. We view ρ_1^- as lower bound for the q- and r-order of Algorithm BROY and we expect the actual q-order to belong to the interval $[\rho_1^-, \rho_1^+]$, while the r-order may be larger.

4.2 Numerical examples

4.2.1 Example 1 a)

The first example is [DS96, Example 8.2.6]. Let

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \qquad F(u) = \begin{pmatrix} u_1^2 + u_2^2 - 2 \\ e^{u_1 - 1} + u_2^3 - 2 \end{pmatrix}.$$

The mapping F has the root $\bar{u} = (1,1)^T$. Since F does not have affine component functions, Theorem 3.6 and Corollary 3.8 assert 4–step q-quadratic convergence and an r-order no smaller than $2^{1/4} \approx 1.189$. Table 1 displays the numerical outcome of a single run with $\hat{\alpha} = 0$, while Table 2 shows the data from the cumulative runs with $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$. The data suggest that the iterates converge 3–step q-quadratically rather than 4–step, which fits with conjecture C2. The worst-case estimate for the q- and r-order is $\rho_1^- \approx 1.20$, which confirms the proven lower bound $2^{1/4}$ for the r-order and is in line with conjecture C1 that $2^{1/4}$ may also be a lower bound for the q-order. Comparing the first and second row in Table 2 shows that perturbing the rows of B_0 that correspond to nonlinear components of F has essentially no effect on the rate of convergence. In Example 1 b) we will see that this is very different if

Table 3: Example 1 b): Single run with $\hat{\alpha} = 0$

k	$\|\|F_k\|\|$	ρ_k^3	ρ_k^2	ρ_k^1	C_k^3	C_k^2	C_k^1
0	5.2e-3	-1	-1	-1	-1	-1	-1
1	3.0e-6	-1	-1	1.97	-1	-1	1.2
2	7.0e-9	-1	2.88	1.46	-1	2.9e-3	1.1e3
3	1.8e-14	4.81	2.45	1.67	7.6e-9	2.9e-3	533.0
4	1.1e-22	3.9	2.66	1.59	1.7e-11	3.3e-6	4.7e5
5	1.8e-36	4.32	2.58	1.62	5.2e-20	7.5e-9	2.0e8
6	1.7e-58	4.17	2.62	1.61	7.4e-31	2.0e-14	7.7e13
7	2.7e-94	4.24	2.61	1.62	3.1e-50	1.2e-22	1.3e22
8	4.2e-152	4.22	2.62	1.62	1.8e-80	1.9e-36	7.9e35
9	9.9e-246	4.23	2.62	1.62	4.6e-130	1.9e-58	8.1e57
10	3.6e-397	4.23	2.62	1.62	6.9e-210	3.0e-94	5.2e93

Table 4: Example 1 b): Cumulative runs with $\hat{\alpha} = 0$ (top), $\hat{\alpha}_n = 10^{-3}$ (second to top), $\hat{\alpha}_l = 10^{-3}$ (third to top), $\hat{\alpha}_l = 10^{-10}$ (second to last), $\hat{\alpha}_l = 10^{-30}$ (last)

ρ_3^-	ρ_3^+	ρ_2^-	ρ_2^+	ρ_1^-	ρ_1^+	C_3^-	C_3^+	C_2^-	C_2^+	C_1^-	C_1^+
4.18	4.23	2.60	2.62	1.61	1.62	2e-65	4e-40	1e-29	2e-18	4e75	2e122
4.11	4.23	2.59	2.62	1.61	1.62	2e-65	1e-39	1e-29	5e-18	4e75	2e122
1.99	2.21	1.51	1.69	1.20	1.30	5e-20	23.1	2e43	2e102	1e99	3e246
2.04	2.25	1.53	1.71	1.21	1.30	2e-25	4e-5	2e40	2e81	1e99	2e223
2.00	2.27	1.37	1.63	1.16	1.24	4e-29	3.05	1e52	1e98	1e99	8e217

rows are perturbed that correspond to affine components. The iteration numbers range from 14 to 16.

4.2.2 Example 1 b)

We linearize the first component function of F from Example 1 a) and obtain

$$F : \mathbb{R}^2 \to \mathbb{R}^2, \qquad F(u) = \begin{pmatrix} 2u_1 + 2u_2 - 4 \\ e^{u_1 - 1} + u_2^3 - 2 \end{pmatrix},$$

with unchanged root $\bar{u} = (1, 1)^T$. From Theorem 3.6 we expect 2–step q-quadratic convergence if $B_0^1 = (2\ 2)$, resulting in a lower bound of 1.41 for the r-order. Table 3 shows a single run with $\hat{\alpha} = 0$, while Table 4 provides the data from the cumulative runs conducted with $\hat{\alpha} = 0$, $\hat{\alpha}_n = 10^{-3}$ and $\hat{\alpha}_l \in \{10^{-30}, 10^{-10}, 10^{-3}\}$. The results are very clear: For $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$ the fact that only one component function of F is not affine induces a reduction of Algorithm BROY to one dimension, so its convergence rate is the same as that of the one-dimensional secant method, i.e., convergence with exact q- and r-order $\frac{1+\sqrt{5}}{2} \approx 1.618$, cf. [VZ92, Man21a]. This implies that the error decays faster than 2–step q-quadratically, which can also be seen from C_2^- and C_2^+. In contrast, even a deviation of $\hat{\alpha}_l = 10^{-30}$ in only one entry of $B_0^1 = (2\ 2)$ slows down the convergence to 3–step q-quadratic, which is the same as in the fully nonlinear Example 1 a), cf. Table 2. While this fits well with conjecture C2, we notice that for $\hat{\alpha}_l = 10^{-30}$ the worst-case estimate $\rho_1^- = 1.16$ of the q-order is somewhat smaller than our conjecture C1 of $2^{1/4} \approx 1.189$; on the other hand, $[\rho_1^-, \rho_1^+] = [1.16, 1.24]$ comfortably includes 1.189. The iteration numbers vary between 9 and 10 if $B_0^1 = (2\ 2)$ and between 10 and 16 otherwise.

Table 5: Example 1 c): Single run with $\hat{\alpha} = 0$

k	$\|\|F_k\|\|$	ρ_k^1	ρ_k^4	ρ_k^3	ρ_k^2	C_k^4	C_k^3	C_k^2
0	5.1e-3	-1	-1	-1	-1	-1	-1	-1
1	2.4e-6	2.09	-1	-1	-1	-1	-1	-1
2	3.5e-9	1.44	-1	-1	3.01	-1	-1	1.2e-3
3	9.2e-13	1.43	-1	4.31	2.06	-1	2.1e-7	0.41
4	4.4e-16	1.27	5.45	2.61	1.81	1.0e-10	2.0e-4	45.0
5	2.6e-22	1.39	3.64	2.52	1.77	1.2e-10	2.7e-5	855.0
6	4.5e-31	1.40	3.53	2.47	1.95	4.5e-14	1.4e-6	6.2
7	1.4e-43	1.41	3.47	2.74	1.97	4.6e-19	2.0e-12	5.6
8	1.8e-56	1.30	3.56	2.55	1.83	2.4e-25	6.9e-13	2.4e5
9	7.0e-72	1.27	3.25	2.33	1.65	2.7e-28	9.4e-11	9.3e14
10	1.2e-95	1.33	3.10	2.20	1.70	1.6e-34	1.6e-9	1.1e17
11	2.1e-129	1.35	2.98	2.30	1.80	2.8e-43	1.8e-17	1.2e14
12	1.0e-167	1.30	2.98	2.34	1.76	8.8e-56	5.6e-25	1.9e23
13	1.9e-211	1.26	2.95	2.21	1.64	1.1e-68	3.6e-21	1.2e47
14	2.3e-274	1.30	2.87	2.12	1.64	4.3e-84	1.4e-16	6.2e60
15	1.2e-365	1.33	2.83	2.18	1.73	7.4e-108	3.2e-31	8.5e56

Table 6: Example 1 c): Cumulative runs with $\hat{\alpha} = 0$ (top) and $\hat{\alpha}_n = 10^{-3}$ (bottom)

ρ_1^-	ρ_1^+	ρ_4^-	ρ_4^+	ρ_3^-	ρ_3^+	ρ_2^-	ρ_2^+	C_4^-	C_4^+	C_3^-	C_3^+	C_2^-	C_2^+
1.20	1.29	2.70	2.96	1.99	2.21	1.51	1.68	2e-51	2e-34	6e-21	24.1	7e44	4e104
1.20	1.30	2.76	2.93	1.99	2.23	1.50	1.69	1e-51	5e-31	7e-20	20.5	4e41	9e110

4.2.3 Example 1 c)

We modify Example 1 a) by inserting an additional equation. Let

$$F : \mathbb{R}^3 \to \mathbb{R}^3, \qquad F(u) = \begin{pmatrix} u_1^2 + u_2^2 - 2 \\ e^{u_1 - 1} + u_2^3 - 2 \\ u_1 + u_2 - 2u_3 \end{pmatrix}.$$

The mapping F has the root $\bar{u} = (1,1,1)^T$. Theorem 3.6 implies 4–step q-quadratic convergence for $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$, yielding an r-order of at least 1.189, whereas conjectures C1 and C2 predict that the latter is also the q-order and that 3 steps are sufficient for q-quadratic convergence. Table 5 shows a single run with $\hat{\alpha} = 0$, while Table 6 and 7 provide the data from the cumulative runs conducted with $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$, respectively, $\hat{\alpha}_l \in \{10^{-30}, 10^{-10}, 10^{-3}\}$. The results in Table 6 are similar to those from Example 1 a) in Table 2 and confirm the 3–step q-quadratic convergence and the q-order of 1.19. Table 7 shows that perturbations in any entry of the third row of B_0 have a strong detrimental effect as they induce a rate between 4–step and 5–step q-quadratic convergence. This convergence is, however, consistent with conjecture C2 for $d = n$, and so is the worst-case estimate $\rho_1^- = 1.13$ with conjecture C1. The iteration numbers range from 13 to 16 for $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$, respectively, from 15 to 20 otherwise.

Table 7: Example 1 c): Cumulative runs with $\hat{\alpha}_l = 10^{-3}/10^{-10}/10^{-30}$ (top/middle/bottom)

ρ_1^-	ρ_1^+	ρ_5^-	ρ_5^+	ρ_4^-	ρ_4^+	ρ_3^-	ρ_3^+	C_5^-	C_5^+	C_4^-	C_4^+	C_3^-	C_3^+
1.13	1.20	2.25	2.57	1.84	2.10	1.54	1.74	8e-43	5e-17	3e-12	4e24	8e39	3e103
1.14	1.20	2.29	2.66	1.87	2.17	1.56	1.76	9e-50	2e-24	1e-16	1e24	3e37	2e97
1.13	1.20	2.43	2.84	1.93	2.20	1.57	1.82	2e-82	4e-54	2e-31	2e11	2e31	4e79

Table 8: Example 2: Single run with $\hat{\alpha} = 0$

k	$\|\|F_k\|\|$	ρ_k^1	ρ_k^4	ρ_k^3	ρ_k^2	C_k^4	C_k^3	C_k^2
0	0.01	-1	-1	-1	-1	-1	-1	-1
1	9.3e-7	1.61	-1	-1	-1	-1	-1	-1
2	3.0e-8	1.33	-1	-1	2.13	-1	-1	0.43
3	1.4e-11	1.57	-1	3.35	2.08	-1	1.6e-4	0.42
4	1.6e-14	1.31	4.37	2.72	2.05	2.1e-7	5.5e-4	0.49
5	2.2e-17	1.23	3.35	2.53	1.61	7.8e-7	6.9e-4	5.1e3
6	2.8e-22	1.32	3.34	2.12	1.63	9.0e-9	0.066	3.8e4
7	1.1e-30	1.42	3.01	2.31	1.88	2.7e-10	1.5e-4	77
8	4.1e-39	1.30	3.00	2.43	1.84	5.5e-13	2.8e-7	1.6e3
9	1.2e-47	1.23	2.99	2.27	1.60	8.2e-16	4.8e-6	3.0e11
10	5.2e-61	1.29	2.93	2.07	1.59	2.1e-19	0.013	9.9e14
11	1.1e-82	1.37	2.83	2.18	1.77	2.7e-24	2.1e-7	2.4e10
12	2.2e-108	1.32	2.88	2.34	1.81	4.2e-33	4.8e-16	2.6e11
13	6.9e-134	1.24	2.90	2.24	1.64	1.5e-41	8.2e-15	1.9e29
14	9.7e-169	1.26	2.83	2.07	1.57	1.1e-49	2.6e-6	6.3e45
15	1.8e-225	1.34	2.77	2.10	1.70	4.9e-63	1.2e-11	1.2e40
16	1.6e-298	1.33	2.79	2.25	1.78	1.0e-84	1.0e-33	5.4e36
17	3.2e-375	1.26	2.83	2.24	1.67	2.1e-110	1.1e-40	3.0e73

Table 9: Example 2: Cumulative runs with $\hat{\alpha} = 0$ (top) and $\hat{\alpha}_n = 10^{-3}$ (bottom)

ρ_1^-	ρ_1^+	ρ_4^-	ρ_4^+	ρ_3^-	ρ_3^+	ρ_2^-	ρ_2^+	C_4^-	C_4^+	C_3^-	C_3^+	C_2^-	C_2^+
1.20	1.29	2.66	2.88	1.97	2.19	1.50	1.68	3e-48	6e-24	3e-18	314.0	3e44	9e110
1.19	1.29	2.64	2.90	1.94	2.19	1.47	1.68	3e-38	7e-24	1e-14	4e4	2e45	8e111

4.2.4 Example 2

Let $F : \mathbb{R}^4 \to \mathbb{R}^4$ be given by

$$F(u) = \begin{pmatrix} \sin(u_1)\cos(u_2) + u_3^3 - u_4^2 \\ e^{u_2+u_3} - (u_4+1)^2 \\ 10u_1 + u_2 - u_3 + 0.1u_4 \\ 2u_1 - u_2 + 5u_3 - 3u_4 \end{pmatrix}.$$

The mapping F has the root $\bar{u} = 0$. The developed theory guarantees 4–step q-quadratic convergence and an r-order no smaller than 1.189 provided B_0 is chosen appropriately. Table 8 shows a single run with $\hat{\alpha} = 0$, while Tables 9 and 10 provide the data from the cumulative runs conducted with $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$, respectively, with $\hat{\alpha}_l \in \{10^{-30}, 10^{-10}, 10^{-3}\}$. Table 9 displays 3–step q-quadratic convergence and $\rho_1^- \geq 1.19$, both of which are in line with conjectures C1 and C2. Table 10 indicates that if B_0 is perturbed in the third and fourth row, then the convergence lies somewhere between 5– and 6–step q-quadratic. The values $\rho_1^- = 1.08/1.09/1.1$ in Table 10 fit with the prediction $2^{1/8} \approx 1.091$ obtained from conjecture C1 for $d = n$. The iteration numbers range from 14 to 19 for $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$, respectively, from 17 to 26 otherwise.

Table 10: Example 2: Cumulative runs with $\hat{\alpha}_l = 10^{-3}/10^{-10}/10^{-30}$ (top/middle/bottom)

ρ_1^-	ρ_1^+	ρ_6^-	ρ_6^+	ρ_5^-	ρ_5^+	ρ_4^-	ρ_4^+	C_6^-	C_6^+	C_5^-	C_5^+	C_4^-	C_4^+
1.08	1.19	2.07	3.02	1.78	2.49	1.55	2.07	3e-49	3e-7	7e-33	8e42	5e-9	6e100
1.09	1.19	2.19	3.15	1.83	2.60	1.55	2.11	7e-68	3e-15	2e-46	3e31	1e-11	5e98
1.10	1.18	2.34	3.59	1.93	2.86	1.61	2.21	4.e-101	4e-54	8e-75	5e13	1e-32	4e81

Table 11: Example 3: Single run with $\hat{\alpha} = 0$

k	$\|\|F_k\|\|$	ρ_k^1	ρ_k^5	ρ_k^4	ρ_k^3	C_k^5	C_k^4	C_k^3
0	4.6e-3	-1	-1	-1	-1	-1	-1	-1
1	6.0e-7	2.28	-1	-1	-1	-1	-1	-1
2	2.5e-11	1.75	-1	-1	-1	-1	-1	-1
3	7.7e-15	1.34	-1	-1	5.36	-1	-1	1.1e-9
4	3.5e-18	1.21	-1	6.48	2.85	-1	1.1e-12	7.1e-6
5	9.1e-22	1.20	7.75	3.41	1.94	4.5e-16	2.9e-9	4.2
6	5.3e-26	1.20	4.10	2.34	1.74	1.7e-13	2.4e-4	4.9e3
7	7.6e-36	1.39	3.26	2.43	2.01	3.5e-14	7.1e-7	0.67
8	1.3e-46	1.31	3.18	2.63	2.20	1.3e-17	1.2e-11	7.2e-5
9	2.6e-56	1.21	3.19	2.67	2.22	2.3e-21	1.4e-14	4.1e-6
10	4.0e-66	1.18	3.14	2.61	1.87	2.2e-24	6.4e-16	3.1e4
11	6.4e-77	1.17	3.04	2.18	1.67	1.0e-26	4.9e-7	1.6e15
12	5.9e-91	1.19	2.59	1.97	1.63	4.5e-21	14.0	3.9e20
13	7.1e-115	1.27	2.50	2.06	1.75	1.7e-23	4.7e-4	1.9e16
14	4.9e-145	1.26	2.61	2.21	1.90	3.3e-34	1.3e-14	5.3e7
15	4.5e-175	1.21	2.68	2.29	1.94	1.2e-44	4.9e-23	5.8e5
16	1.3e-205	1.18	2.70	2.28	1.80	1.4e-53	1.7e-25	1.2e23
17	3.7e-237	1.15	2.63	2.07	1.64	4.8e-57	3.3e-9	6.7e51
18	5.1e-276	1.16	2.42	1.91	1.58	4.5e-48	9.3e12	1.1e73
19	6.8e-338	1.23	2.34	1.94	1.65	1.2e-49	1.5e11	1.7e72

Table 12: Example 3: Cumulative runs with $\hat{\alpha} = 0$ (top) and $\hat{\alpha}_n = 10^{-3}$ (bottom)

ρ_1^-	ρ_1^+	ρ_5^-	ρ_5^+	ρ_4^-	ρ_4^+	ρ_3^-	ρ_3^+	C_5^-	C_5^+	C_4^-	C_4^+	C_3^-	C_3^+
1.13	1.20	2.29	2.60	1.85	2.12	1.56	1.74	3e-45	2e-21	3e-12	3e29	6e32	6e100
1.13	1.19	2.24	2.51	1.83	2.08	1.53	1.73	9e-37	5e-18	7e-8	7e18	3e37	2e94

4.2.5 Example 3

Let $F : \mathbb{R}^{10} \to \mathbb{R}^{10}$ be given by

$$F(u) = \begin{pmatrix} u_1 - u_3^2 + u_5 u_6 u_7 - (u_8 + 1)(u_9 - 1) - 1 \\ u_1 + 0.5 \ln(1 + u_9^2) - 2\exp(u_{10}) + 2 \\ u_2 + 0.5 \ln(1 + u_8^2) - \exp(u_{10}) + 1 \\ u_1 + u_2 + 2u_3 + u_4 + u_5 + u_6 - 3u_7 - 2u_8 + u_{10} \\ u_2 - 4u_3 + 3u_5 - u_7 - u_{10} \\ -2u_3 + 0.1u_7 + 0.3u_9 \\ u_1 + u_3 - 10u_2 - 5u_4 - u_6 - u_8 \\ 2u_1 + 2u_3 + 2u_5 + 2u_7 + u_9 \\ u_6 - u_7 + 2u_9 \\ 2u_3 + 2u_8 + 2u_9 + u_{10} \end{pmatrix}.$$

The mapping F has the root $\bar{u} = 0$. We expect no more than 6 steps for q-quadratic convergence if $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$ as well as an r-order no smaller than 1.122. Table 11 displays a single run with $\hat{\alpha} = 0$ and Table 12 provides the data from the cumulative runs conducted with $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$. Five steps are sufficient for quadratic convergence and $\rho_1^- = 1.13$ is compatible with the conjectured lower bound 1.122 for the q-order. Table 13 provides the data for $\hat{\alpha}_l \in \{10^{-30}, 10^{-10}, 10^{-3}\}$, but in contrast to previous experiments we have only perturbed three of the seven rows of B_0 that correspond to affine components of F, so Theorem 3.6 and Corollary 3.8 ensure 12–step q-quadratic convergence and an r-order of at least 1.06, while C1 and C2 predict 11 steps and a q-order of 1.06. Table 13 confirms the q-order and indicates that 6 to 8 steps are enough for quadratic convergence depending on the magnitude of the perturbation. The iteration numbers range from 17 to 23 for $\hat{\alpha} = 0$ and $\hat{\alpha}_n = 10^{-3}$, respectively, from 19 to 33 otherwise.

Table 13: Example 3: Cumulative runs with $\hat{\alpha}_l = 10^{-3}/10^{-10}/10^{-30}$ (top/middle/bottom)

ρ_1^-	ρ_1^+	ρ_8^-	ρ_8^+	ρ_7^-	ρ_7^+	ρ_6^-	ρ_6^+	C_8^-	C_8^+	C_7^-	C_7^+	C_6^-	C_6^+
1.06	1.14	1.96	3.15	1.76	2.77	1.60	2.36	1e-63	1100	4e-48	2e21	7e-30	3e54
1.06	1.15	2.24	3.30	1.96	2.91	1.74	2.49	1e-75	2e-16	2e-57	1644	6e-36	1e25
1.08	1.14	3.01	3.84	2.49	3.29	2.04	2.72	8e-119	6e-77	2e-94	3e-65	1e-63	2e-7

Table 14: Example 4: Cumulative runs with $\hat{\alpha} = 0$

ρ_1^-	ρ_1^+	ρ_4^-	ρ_4^+	ρ_3^-	ρ_3^+	ρ_2^-	ρ_2^+	C_4^-	C_4^+	C_3^-	C_3^+	C_2^-	C_2^+
1.20	1.30	2.77	2.93	1.98	2.23	1.50	1.71	3e-51	2e-34	3e-21	46.3	2e44	3e104

4.2.6 Example 4

As final example we consider $F : \mathbb{R}^6 \to \mathbb{R}^6$ given by

$$F(u) = \begin{pmatrix} u_1 u_2 u_3 u_4 + (u_5 - 1)(u_6 + 1) + 1 \\ e^{\sum_{j=1}^{6} u_j} - 1 \\ Au \end{pmatrix},$$

where $A \in \mathbb{R}^{4\times 6}$ is a random matrix with entries in $[-1, 1]$ that is changed after each of the 10000 runs of the cumulative run. The root of F is $\bar{u} = 0$ and A is chosen such that $F'(\bar{u})$ is invertible. Theorem 3.6 ensures 4–step q-quadratic convergence for $\hat{\alpha} = 0$ and this is clearly confirmed in Table 14, that actually suggests 3–step q-quadratic convergence. In accordance with conjecture C1 the worst-case estimate $\rho_1^- = 1.20$ of the q-order is slightly larger than $2^{1/4}$. The iteration numbers lie between 14 and 16. In passing, let us point out that the values depicted in Table 14 are quite similar to those in Table 2, which illustrates the key point of Theorem 3.6 that d determines the behavior of Broyden's method rather than n.

5 Summary

We have demonstrated that the local convergence of Broyden's method improves from $2n$–step q-quadratic to $2d$–step q-quadratic if $F : \mathbb{R}^n \to \mathbb{R}^n$ has $n-d$ affine component functions and the corresponding $n-d$ rows of B_0 match those of the Jacobian of F. We have confirmed the faster convergence in numerical experiments and observed that it is stable under perturbations of the d rows of B_0 associated to nonlinear component functions of F, but not under perturbations of the remaining $n-d$ rows. Based on the numerical results we have proposed the conjectures that Broyden's method enjoys $(2d-1)$–step q-quadratic convergence for $d \in \{2, \dots, n\}$ and admits a q-order of convergence that is bounded from below by $2^{1/(2d)}$.

References

[ABDL14] F.J. Aragón Artacho, A. Belyakov, A.L. Dontchev, and M. López. Local convergence of quasi-Newton methods under metric regularity. *Comput. Optim. Appl.*, 58(1):225–247, 2014. doi:10.1007/s10589-013-9615-y.

[AC79] J.H. Avila and P. Concus. Update methods for highly structured systems of nonlinear equations. *SIAM J. Numer. Anal.*, 16:260–269, 1979. doi:10.1137/0716019.

[AN18] S. Adly and H.V. Ngai. Quasi-Newton methods for solving nonsmooth equations: Generalized Dennis-Moré theorem and Broyden's update. *J. Convex Anal.*, 25(4):1075–1104, 2018.

[ASM14] M. Al-Baali, E. Spedicato, and F. Maggioni. 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.

[BDM73] C.G. Broyden, J.E. Dennis, and J.J. More. 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.

[BKK19] S. Bai, J.Z. Kolter, and V. Koltun. Deep equilibrium models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 690–701. Curran Associates, Inc., 2019. URL: https://papers.nips.cc/paper/8358-deep-equilibrium-models.pdf.

[BKK20] S. Bai, J.Z. Kolter, and V. Koltun. Multiscale deep equilibrium models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5238–5250. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/3812f9a59b634c2a9c574610eaba5bed-Paper.pdf.

[DM77] J.E. Dennis and J.J. More. Quasi-Newton methods, motivation and theory. *SIAM Rev.*, 19:46–89, 1977. doi:10.1137/1019005.

[DS96] J.E. Dennis and R.B. Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations.* Philadelphia, PA: SIAM, repr. edition, 1996. doi:10.1137/1.9781611971200.

[DW81] J.E. Dennis and H.F. Walker. Convergence theorems for least-change secant update methods. *SIAM J. Numer. Anal.*, 18:949–987, 1981. doi:10.1137/0718067.

[Gay79] D.M. Gay. Some convergence properties of Broyden's method. *SIAM J. Numer. Anal.*, 16:623–630, 1979. doi:10.1137/0716047.

[Gil89] J.Ch. Gilbert. On the local and global convergence of a reduced quasi-Newton method. *Optimization*, 20(4):421–450, 1989. doi:10.1080/02331938908843462.

[GL81] R.R. Gerber and F.T. Luk. A generalized Broyden's method for solving simultaneous linear equations. *SIAM J. Numer. Anal.*, 18:882–890, 1981. doi:10.1137/0718061.

[GL01] P.E. Gill and M.W. Leonard. Reduced-Hessian quasi-Newton methods for unconstrained optimization. *SIAM J. Optim.*, 12(1):209–237, 2001. doi:10.1137/S1052623400307950.

[Gri87] A. Griewank. The local convergence of Broyden-like methods on lipschitzian problems in Hilbert spaces. *SIAM J. Numer. Anal.*, 24(3):684–705, 1987. doi:10.1137/0724045.

[Gri12] A. Griewank. Broyden updating, the good and the bad! *Doc. Math. (Bielefeld)*, pages 301–315, 2012. URL: https://www.emis.de/journals/DMJDMV/vol-ismp/45_griewank-andreas-broyden.pdf.

[GSW08] A. Griewank, S. Schlenkrich, and A. Walther. Optimal r-order of an adjoint Broyden method without the assumption of linearly independent steps. *Optim. Methods Softw.*, 23(2):215–225, 2008. doi:10.1080/10556780701766549.

[HK92] D.M. Hwang and C.T. Kelley. Convergence of Broyden's method in Banach spaces. *SIAM J. Optim.*, 2(3):505–532, 1992. doi:10.1137/0802025.

[Kel95] C.T. Kelley. *Iterative methods for linear and nonlinear equations*. Philadelphia, PA: SIAM, 1995. doi:10.1137/1.9781611970944.

[Man21a] F. Mannel. Convergence properties of the Broyden–like method for mixed linear–nonlinear systems of equations. *Numer. Algorithms*, Online First:1–29, 2021. doi:10.1007/s11075-020-01060-y.

[Man21b] F. Mannel. On the convergence of the Broyden–like method and the Broyden–like matrices. *Submitted*, 2021. URL: https://imsc.uni-graz.at/mannel/CBLM.pdf.

[Man21c] F. Mannel. On the convergence of the Broyden matrices for a class of singular problems. *Submitted*, 2021. URL: https://imsc.uni-graz.at/mannel/CGB_sing.pdf.

[Mar79] E. Marwil. Convergence results for Schubert's method for solving sparse nonlinear equations. *SIAM J. Numer. Anal.*, 16:588–604, 1979. doi:10.1137/0716044.

[Mar00] J.M. Martínez. 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.

[MHMP13] P.Q. Muoi, D.N. Hào, P. Maass, and M. Pidcock. Semismooth Newton and quasi-Newton methods in weighted ℓ^1-regularization. *J. Inverse Ill-Posed Probl.*, 21(5):665–693, 2013. doi:10.1515/jip-2013-0031.

[MMP18] L. Marini, B. Morini, and M. Porcelli. Quasi-Newton methods for constrained nonlinear systems: complexity analysis and applications. *Comput. Optim. Appl.*, 71(1):147–170, 2018. doi:10.1007/s10589-018-9980-7.

[MR20] F. Mannel and A. Rund. A hybrid semismooth quasi-Newton method for nonsmooth optimal control with PDEs. *Optim. Eng.*, Online First:1–39, 2020. doi:10.1007/s11081-020-09523-w.

[NW06] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2nd edition, 2006. doi:10.1007/978-0-387-40065-5.

[O'L95] D.P. O'Leary. Why Broyden's nonsymmetric method terminates on linear equations. *SIAM J. Optim.*, 5(2):231–235, 1995. doi:10.1137/0805012.

[OR00] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables*, volume 30. Philadelphia, PA: SIAM, 2000. doi:10.1137/1.9780898719468.

[Sac86] E.W. Sachs. Broyden's method in Hilbert space. *Math. Program.*, 35:71–82, 1986. doi:10.1007/BF01589442.

[Sch70] L.K. Schubert. Modification of a quasi-Newton method for nonlinear equations with a sparse Jacobian. *Math. Comput.*, 24:27–30, 1970. doi:10.2307/2004874.

[SH97] E. Spedicato and Z. Huang. Numerical experience with Newton-like methods for nonlinear algebraic systems. *Computing*, 58(1):69–89, 1997. doi:10.1007/BF02684472.

[VZ92] M. Vianello and R. Zanovello. On the superlinear convergence of the secant method. *Am. Math. Mon.*, 99(8):758–761, 1992. doi:10.2307/2324244.