Vater, N.; Borzì, A.

Article — Published Version

Convergence of a quasi-Newton method for solving systems of nonlinear underdetermined equations

Computational Optimization and Applications

Suggested Citation: Vater, N.; Borzì, A. (2024) : Convergence of a quasi-Newton method for solving systems of nonlinear underdetermined equations, Computational Optimization and Applications, ISSN 1573-2894, Springer US, New York, Vol. 91, Iss. 2, pp. 973-996, https://doi.org/10.1007/s10589-024-00606-3

Computational Optimization and Applications (2025) 91:973–996
https://doi.org/10.1007/s10589-024-00606-3

Convergence of a quasi-Newton method for solving systems of nonlinear underdetermined equations

N. Vater[1] · **A. Borzì**[1]

Received: 12 February 2024 / Accepted: 23 August 2024 / Published online: 6 September 2024

Abstract
The development and convergence analysis of a quasi-Newton method for the solution of systems of nonlinear underdetermined equations is investigated. These equations arise in many application fields, e.g., supervised learning of large overparameterised neural networks, which require the development of efficient methods with guaranteed convergence. In this paper, a new approach for the computation of the Moore–Penrose inverse of the approximate Jacobian coming from the Broyden update is presented and a semi-local convergence result for a damped quasi-Newton method is proved. The theoretical results are illustrated in detail for the case of systems of multidimensional quadratic equations, and validated in the context of eigenvalue problems and supervised learning of overparameterised neural networks.

Keywords Systems of nonlinear underdetermined equations · Nonlinear root-finding problems · Quasi-Newton methods · Least change secant update · Supervised learning

Mathematics Subject Classification 49M15 · 65H04 · 65H10 · 90C30

1 Introduction

This work is concerned with the solution of systems of nonlinear underdetermined equations represented by $F(x) = 0$, where $F : \mathbb{R}^n \to \mathbb{R}^m$, $m \leq n$. In the literature,

In Memory of Our Colleague and Friend Daniela di Serafino.

Partially supported by the BMBF-Project iDeLIVER: Intelligent MR Diagnosis of the Liver by Linking Model and Data-Driven Processes.

✉ N. Vater
nadja.vater@mathematik.uni-wuerzburg.de

A. Borzì
alfio.borzi@mathematik.uni-wuerzburg.de

[1] Institut für Mathematik, Universität Würzburg, Emil-Fischer-Strasse 30, 97074 Würzburg, Germany

the problem of determining $x^* \in \mathbb{R}^n$ such that $F(x^*) = 0$ is also referred to as a root-finding problem. These kind of problems are relevant in several application fields such as nonlinear eigenvalue problems [22], homotopy or continuation methods [24], and data fitting tasks [4] by means of supervised learning of overparameterised artificial neural networks [1, 16, 23].

The purpose of this work is to contribute to the field of methods for solving systems of nonlinear underdetermined equations by presenting an update scheme for the Moore–Penrose inverse of the approximated Jacobians coming from the Broyden update and investigating the convergence of the corresponding quasi-Newton method. Specifically, we prove a new semi-local convergence result for a damped quasi-Newton method, whose update schemes satisfies a bounded deterioration principle. From this theorem, we conclude the convergence of our quasi-Newton method with update of the approximation of the Moore–Penrose inverse of the Jacobian. In contrast to previous works, we state explicit assumptions on the residual function F at the starting point x_0 of the quasi-Newton iteration that can be easily checked for a given problem. We illustrate our results for the quadratic case and further specialise them in the case of supervised learning of overparameterised shallow neural networks. In the latter case, we also provide sufficient conditions on the number of the unknown parameters of the network for the convergence of our method to a root of an exact data fitting problem. We successfully validate our theoretical findings by results of numerical experiments, that showcase the potential of our method for the computation of eigenpairs in eigenvalue problems and for the supervised training of a multilayer neural network for the classification of the Iris data set.

Some review work on the solvability of systems of nonlinear equations is given in [12, 13, 15]. These surveys cover a large amount of results for the case where the number of equations is at least as large as the number of unknowns, that is, $m \geq n$. In contrast, results for nonlinear underdetermined problems are much more scattered in the literature. For this latter case, the question of solvability is explicitly addressed in [19, 20], where it is shown that a sequence generated by a Newton-type method starting at a given starting vector converges to a root. To show convergence of the sequence, local properties of the residual function F in a ball around the starting vector are exploited whilst showing that all iterates stay in this ball. In particular, convergence of the iterative sequence is shown without an a priori assumption of existence of or proximity to a root. In contrast to local convergence results in which the existence of an (isolated) root is assumed, the theoretical statements without an a priori assumption on the existence of a root are called semi-local convergence results. Early works applying this Kantorovich-type analysis are [11, 19].

There exist several iterative methods for the solution of nonlinear root-finding problems, especially in the case $m = n$. In this case, a canonical choice is Newton's method, where each iteration requires assembling and solving a linear system of equations. To reduce the computational effort for each iteration, approximate versions of this method are developed [3, 8, 10], which can be roughly classified as being either a quasi-Newton method or an inexact Newton method. This categorisation refers to the part of the calculation that is being approximated, that is, the construction or the solution of the linear equation system, respectively. In this work, we consider quasi-Newton methods.

In the nonlinear underdetermined case, similar Newton-like approaches have been investigated in the literature. We refer to this scheme as Newton method also in the underdetermined case. Inexact methods are considered in, e.g., [22] and quasi-Newton methods with least change secant update are analysed in [14, 24]. Both kinds of approximate Newton methods are discussed in [2], where they are viewed as approximate Gauss-Newton methods to solve the related least squares problem $\min_{x\in\mathbb{R}^n} \frac{1}{2}\|F(x)\|^2$. To avoid the computation of the minimal-norm solution of a linear system of equations, we present an update scheme for the Moore–Penrose inverse of the approximate Jacobians obtained by the Broyden update, which is an example of a least change secant update.

While the convergence results for quasi-Newton methods in [14] are local in the sense that the existence of a root is assumed, in [2, 22, 24] semi-local convergence results for an inexact Newton method and a quasi-Newton method are proven. These results show that there exist parameters, which describe the properties of a function in a ball around the start value, such that the iterates obtained by the method stay within this ball and converge to a root. In our work, we prove a semi-local convergence result for the damped version of a quasi-Newton method in which we give explicit conditions on the residual function F, the start value x_0, and the approximation of the Jacobian that guarantee convergence of the method. The inclusion of a damping parameter allows us to obtain convergence for a larger set of starting values compared to the approach without damping. To emphasise that the update direction is multiplied with such a step size, we refer to the resulting schemes as *damped* Newton and quasi-Newton methods, respectively. Using the approximation quality of the Broyden update, we provide sufficient conditions on the residual function and on the starting value in order to prove convergence to a root for our corresponding quasi-Newton method. With this result at hand, we also demonstrate that we can obtain a convergence sequence also without including a damping factor by imposing stricter assumptions on the start value.

Recently, semi-local convergence results for gradient-type methods were considered in the context of supervised training of overparameterised shallow neural networks [17, 23], where the start parameters are typically generated randomly. In particular, in these works it is shown that the assumptions that are required for proving the given semi-local convergence results are satisfied with high probability for a random choice of the initial approximation of the network's parameters provided that the shallow network is wide enough. Motivated by these findings, we provide similar conditions on the width of the network to claim convergence of our quasi-Newton iteration.

In the next section, we discuss existence of solutions to systems of nonlinear underdetermined equations and illustrate our assumptions for the case where the residual functions are quadratic. In Sect. 3, we introduce damped quasi-Newton methods and our quasi-Newton scheme which uses updates of the Moore–Penrose inverse of the approximation of the Jacobians obtained by Broyden's method. In Sect. 4, we prove a semi-local convergence result for the damped method and conclude convergence of our quasi-Newton method without damping. The specific case of problems arising in the context of supervised training of shallow neural networks is treated in Sect. 5. In Sect. 6, we validate our theoretical findings by showing results of numerical experiments. To this end, we consider the cases of quadratic residual functions and

of supervised training of shallow artificial neural networks, which are also discussed theoretically. Additionally, we include the computation of eigenpairs in an eigenvalue problem and the supervised training of larger networks to illustrate the potential of our quasi-Newton method. A section of conclusion completes our work.

2 Problem setting and characterisation of the solution

We consider the underdetermined root-finding problem:

$$\text{Find } x^* \in \mathbb{R}^n \quad \text{such that } F(x) = 0, \tag{1}$$

with a nonlinear function $F : \mathbb{R}^n \to \mathbb{R}^m$, $n \geq m$. We call F the residual function of the problem and assume that F satisfies the following assumption:

Assumption 1 Let $\Omega \subseteq \mathbb{R}^n$ be closed and convex. Assume that F is differentiable in Ω and the Jacobian of F is Lipschitz continuous in Ω with Lipschitz constant $\gamma > 0$, that is, it holds

$$\left\| F'(x) - F'(y) \right\| \leq \gamma \, \|x - y\| \quad \text{for } x, y \in \Omega. \tag{2}$$

The existence of a root of the residual function F in Ω can be proved based on properties of the function F in a neighbourhood of a suitable initial approximation (guess) $x_0 \in \Omega$ of the solution. Let $\|\cdot\|$ denote the Euclidean norm of a vector or the spectral norm of a matrix. Let $N(x_0, \varrho) := \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq \varrho\}$ denote the ball of radius $\varrho > 0$ around x_0. Specifically, we have the following result [21].

Theorem 1 *Assume that the residual function F satisfies Assumption 1 in a closed and convex region $\Omega \subseteq \mathbb{R}^n$.*

Let $x_0 \in \Omega$ and $\mu > 0$ be such that $\left\| F'(x_0)^\top h \right\| \geq \frac{2}{\mu} \|h\|$, $h \in \mathbb{R}^m$, and $N\left(x_0, \frac{1}{\mu\gamma}\right) \subseteq \Omega$. Assume that it holds

$$\|F(x_0)\| < \frac{1}{\mu^2 \gamma}. \tag{3}$$

Then there exists a solution x^* of problem (1) with $x^* \in N(x_0, \mu \|F(x_0)\|)$.

The assumption on the Jacobian $F'(x_0)$ implies that this matrix has full rank and the largest singular value of its Moore–Penrose inverse $F'(x_0)^\dagger$ is bounded by $\frac{\mu}{2}$. The Lipschitz continuity of the Jacobian function F' together with the full rank assumption implies that the Jacobian $F'(x)$ has full rank for all values x in a neighbourhood of the guess x_0. By bounding the size of this neighbourhood, we obtain a bound on the largest singular value of the Moore–Penrose inverse for all values x in this neighbourhood as follows:

Remark 1 Assume that the residual function F satisfies Assumption 1 in a closed and convex region $\Omega \subseteq \mathbb{R}^n$.

Let $x_0 \in \mathbb{R}^n$ and $\mu > 0$ be such that $\left\| F'(x_0)^\top h \right\| \geq \frac{2}{\mu} \|h\|$, $h \in \mathbb{R}^m$, and $\varrho \in \left(0, \frac{2}{\mu\gamma}\right)$. Then, for $x \in N(x_0, \varrho)$, it holds

$$\begin{aligned} \left\| F'(x)^\top h \right\| &\geq \left\| F'(x_0)^\top h \right\| - \left\| (F'(x_0) - F'(x))^\top h \right\| \geq \frac{2}{\mu} \|h\| - \gamma\varrho \|h\| \\ &= \left(\frac{2}{\mu} - \gamma\varrho \right) \|h\| . \end{aligned} \tag{4}$$

Hence, for the specific choice $\varrho = \frac{1}{\mu\gamma}$ we have

$$\left\| F'(x)^\top h \right\| \geq \frac{1}{\mu} \|h\| , \quad h \in \mathbb{R}^m$$

for all $x \in N\left(x_0, \frac{1}{\gamma\mu}\right)$, that is, it holds $\left\| F'(x)^\dagger \right\| \leq \mu$ for all $x \in N\left(x_0, \frac{1}{\gamma\mu}\right)$.

Using this property, the proof of Theorem 1 is based on a semi-local convergence result for a damped Newton method with iterates defined by $x_{k+1} = x_k - \eta F'(x_k)^\dagger F(x_k)$ starting at x_0. It is shown in [21] that it is possible to choose the step size $\eta > 0$ small enough such that all iterates stay inside the ball $N(x_0, \varrho)$, $\varrho = \frac{1}{\mu\gamma}$, and $\|F(x_{k+1})\| \leq q \|F(x_k)\|$ for $q \in (0, 1)$. Thus, the sequence (x_k) converges to a root of F. We use a similar approach to show convergence of a damped quasi-Newton method in Sect. 4.

To illustrate the application of Theorem 1, we consider the case where each component F_i of the residual function F is quadratic; see also [21].

Quadratic case Let the residual function be defined component-wise by

$$F_i(x) := \frac{1}{2} (A_i x, x) + (b_i, x) + c_i , \quad i = 1, \ldots, m, \tag{5}$$

with symmetric matrices $A_i = A_i^\top \in \mathbb{R}^{n\times n}$, vectors $b_i \in \mathbb{R}^n$ and scalar values $c_i \in \mathbb{R}$ for $i = 1, \ldots, m$.

The gradients of these quadratic functions are given by

$$\nabla F_i(x) = A_i x + b_i , \quad i = 1, \ldots, m. \tag{6}$$

Thus, the Jacobian has the form

$$F'(x) = \begin{pmatrix} x^\top A_1 + b_1^\top \\ \vdots \\ x^\top A_m + b_m^\top \end{pmatrix} \in \mathbb{R}^{m\times n}, \tag{7}$$

which is Lipschitz continuous on any closed and convex set $\Omega \subseteq \mathbb{R}^n$ with Lipschitz constant given by [18]

$$\gamma = \sqrt{\sum_{i=1}^{m} \|A_i\|^2}. \tag{8}$$

For $x_0 = 0$, it holds $\|F(x_0)\| = \|c\|$, where $c \in \mathbb{R}^m$ is the vector with entries c_i, and $F'(x_0) = B$, where $B \in \mathbb{R}^{m\times n}$ with the rows given by the vectors $b_1, \ldots, b_m$. If B has full rank with $\left\|B^\top h\right\| \geq \frac{2}{\mu}\|h\|$, $h \in \mathbb{R}^m$, and $\|c\| < \frac{1}{\mu^2\gamma}$, then the corresponding root-finding problem has a solution according to Theorem 1.

As a specific problem of this form, we consider $F : \mathbb{R}^3 \to \mathbb{R}^2$ defined by

$$F(x) = \begin{pmatrix} x_1^2 + (x_2 - 1)^2 + x_3^2 - \frac{15}{16} \\ (x_1 - 1)^2 + x_2^2 + x_3^2 - 1 \end{pmatrix}, \tag{9}$$

where $A_1 = A_2 = 2I \in \mathbb{R}^{3\times 3}$, with I denoting the identity matrix, and

$$b_1 = (0, -2, 0)^\top, \quad b_2 = (-2, 0, 0)^\top, \quad c_1 = \frac{1}{16}, \quad \text{and } c_2 = 0.$$

For this case, we have $\gamma = \sqrt{\|A_1\|^2 + \|A_2\|^2} = 2\sqrt{2}$,

$$F'(0) = B = \begin{pmatrix} 0 & -2 & 0 \\ -2 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad \|F(0)\| = \|c\| = \frac{1}{16}.$$

The smallest singular value of B is 2, hence $\left\|B^\top h\right\| \geq \frac{2}{\mu}\|h\|$, $h \in \mathbb{R}^2$, holds with $\mu = 1$.

It holds $\|F(0)\| = \frac{1}{16} < \frac{1}{2\sqrt{2}} = \frac{1}{\mu^2\gamma}$. Thus, there exists a solution x^* of the root-finding problem with $\|x^*\| < \frac{1}{16} = \mu\,\|F(0)\|$ according to Theorem 1. In fact, the problem has a root $x^* = \left(\frac{31-\sqrt{959}}{64}, \frac{33-\sqrt{959}}{64}, 0\right)^\top$ with $\|x^*\| < \frac{\sqrt{10}}{64} < \frac{1}{16} = \mu\,\|F(0)\|$.

Next, we introduce damped quasi-Newton methods and present an update scheme for the Moore–Penrose inverse of the approximate Jacobians coming from the Broyden update. Further, we investigate their convergence to a solution of our system of nonlinear underdetermined equations. Specifically, we prove a semi-local convergence result for damped quasi-Newton methods with similar assumptions on the problem as in Theorem 1 and give sufficient conditions for the convergence of the quasi-Newton method corresponding to our update scheme of the Moore–Penrose inverse of the approximate Jacobians.

3 Damped quasi-Newton methods

The iterative method which is employed in [19] is a damped Newton method with iterates $x_{k+1} = x_k - \eta F'(x_k)^\dagger F(x_k)$, which is also called normal flow algorithm.

This iteration corresponds to Newton's method in the case $m = n$ and $\eta = 1$. For both the normal flow algorithm and Newton's method, it is necessary to compute the Jacobian F' and the solution of a linear system of equations in each iteration. For large problems, these computations are rather expensive, and in quasi-Newton methods the computation of the Jacobian $F'(x_k)$ is replaced by a suitable approximation B_k in each iteration, which should be cheaper to compute than the Jacobian. Let $B_0 \in \mathbb{R}^{m\times n}$ be an approximation of the Jacobian $F'(x_0)$ at the start value $x_0 \in \mathbb{R}^n$. An effective way to obtain an approximation to $F'(x_k)$ is by computing the approximation B_k as a low-rank update of the predecessor B_{k-1}. We can write such an algorithm as follows:

$$x_{k+1} = x_k - \eta\, B_k^{\dagger} F(x_k), \quad B_{k+1} = U(x_k, x_{k+1}, B_k), \tag{10}$$

where $U : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is an update function and $\eta > 0$ is the damping factor. We say that (x_k, B_k) is the sequence of iterates obtained by the damped quasi-Newton method (10).

For the case $m = n$, there exist a variety of algorithms of this type; see, e.g., [9]. Notice that the popular updates BFGS and SR1 are not suitable to update our rectangular matrix B_k as they apply only to square symmetric matrices. For underdetermined problems mainly quasi-Newton methods with least change secant updates are discussed; see, e.g., [14, 24]. We introduce least change secant updates in Appendix A. Several generalisations of well-known quasi-Newton methods for the root finding problem with $m = n$ have been considered for the underdetermined case [14]. However, the convergence results proven for these methods are local in the sense that they assume the existence of a root and proximity of the start value to the root. In [24], in addition to quasi-Newton methods for underdetermined root-finding problems, also approaches tailored to problems arising in homotopy or continuation methods are discussed. In these cases, semi-local convergence results are provided for both the general and the problem-specific quasi-Newton method by assuming that the update function U satisfies a bounded deterioration property. For our new result for the damped version of the quasi-Newton method, we assume that our update function U satisfies a variation of the bounded deterioration principle used in [24] as follows:

Assumption 2 Let $\Omega \subseteq \mathbb{R}^n$ be closed and convex. Assume that the update function $U : \Omega \times \Omega \times \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ satisfies the bounded deterioration property [24] with a constant $\alpha > 0$, that is,

$$\left\| U(x, x+s, B) - F'(x+s) \right\| \le \left\| B - F'(x) \right\| + \alpha \left\| s \right\|, \tag{11}$$

for $x \in \Omega$, $B \in \mathbb{R}^{m\times n}$ and $s \in \mathcal{R}(B^{\top})$ such that $x + s \in \Omega$.

The iterative application of this property with $s_k = -\eta B_k^{\dagger} F(x_k)$ allows us to bound $\left\| B_k - F'(x_k) \right\|$ by the sum of the initial deterioration $\left\| B_0 - F'(x_0) \right\|$ and the sum over $\|s_\ell\|$, $\ell = 0, \ldots, k-1$. If $F'(x_k)$ has full rank and this bound is small enough, we can conclude that B_k has full rank as well. This reasoning will be crucial in our convergence proof.

As a specific update scheme suitable for rectangular matrices in the class of least change secant updates, we consider the first Broyden method introduced in [14], where

$U(x, x+s, B) := B + \frac{(F(x+s)-F(x)-Bs)s^\top}{s^\top s}$. We have

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)s_k^\top}{s_k^\top s_k}, \tag{12}$$

where $s_k = x_{k+1} - x_k$ and $y_k = F(x_{k+1}) - F(x_k)$.

If we store and update the matrices B_k, we have to solve a system of equations of size $m \times n$ in every iteration to obtain the update direction $B_k^\dagger F(x_k)$. Instead, we follow the common approach to store and update the inverse of the approximation matrix, which is in our case the Moore–Penrose inverse $H_k := B_k^\dagger$. Having H_k at hand, the update direction $B_k^\dagger F(x_k) = H_k F(x_k)$ can be computed by a matrix–vector multiplication instead of solving a linear system of equations. Similar to the update of H_k in the case $m = n$, see [7], we find that the update formula for H_k is given by

$$H_{k+1} = H_k + \frac{(s_k - H_k y_k)\, s_k^\top H_k}{s_k^\top H_k y_k}, \tag{13}$$

where $H_0 = B_0^\dagger$, see Appendix B. Hence, our quasi-Newton method with Broyden update and step size $\eta = 1$ reads as follows:

$$x_{k+1} = x_k - H_k F(x_k), \quad H_{k+1} = H_k + \frac{(s_k - H_k y_k)\, s_k^\top H_k}{s_k^\top H_k y_k}, \tag{14}$$

where $s_k = x_{k+1} - x_k$ and $y_k = F(x_{k+1}) - F(x_k)$. We say that (x_k, H_k) is the sequence of iterates obtained by the quasi-Newton method with Broyden update. Similar to our quasi-Newton method with arbitrary update function from (10) a step size $\eta \neq 1$ can be used in the method to obtain a damped scheme. In the next section, we show that for a given starting vector $x_0 \in \mathbb{R}^n$, satisfying certain assumptions, we can find a range $(0, \overline{\eta})$ such that the damped quasi-Newton scheme with step size $\eta \in (0, \overline{\eta})$ converges. Additionally, in Corollary 1, we demonstrate, that if we fix $\eta = 1$, then the starting vector needs to satisfy stronger assumptions in order to guarantee convergence of the quasi-Newton scheme.

4 Semi-local convergence of the quasi-Newton method

In this section, we discuss the convergence of our damped quasi-Newton method to a root of our nonlinear residual function. The solutions of such underdetermined systems of equations are not necessarily isolated. Hence, we cannot single out a specific solution as the limit of our iterative sequence. Instead of assuming proximity of our starting point to a solution, we assume that our starting point x_0 has properties similar to the assumptions for x_0 in Theorem 1, that is, it holds $\|F(x_0)\| < \epsilon$ and $\left\|F'(x_0)^\top h\right\| \geq \frac{2}{\mu}\|h\|$, $h \in \mathbb{R}^m$ for some specific $\epsilon > 0$. While the proof of Theorem 1 is based on showing the convergence of a damped Newton method, we extend this result to

obtain convergence of our damped quasi-Newton method. To this end, we assume that the initial approximation B_0 is sufficiently close to the Jacobian $F'(x_0)$, that is $\|B_0 - F'(x_0)\| = \delta < \frac{1}{4\mu}$ and that the update function satisfies the bounded deterioration property given in Assumption 2. Our main result is the following theorem.

Theorem 2 *Assume that there exists a closed and convex set $\Omega \subseteq \mathbb{R}^n$ such that F and U satisfy Assumption 1 with $\gamma > 0$ and Assumption 2 with $\alpha > 0$ in Ω.*

Let $x_0 \in \Omega$ and $\mu > 0$ such that $\|F'(x_0)^\top h\| \geq \frac{2}{\mu}\|h\|$, $h \in \mathbb{R}^m$, $N(x_0, \varrho) \subseteq \Omega$, where $\varrho = \frac{1}{\gamma\mu}$, and choose $B_0 \in \mathbb{R}^{m\times n}$ such that $\|B_0 - F'(x_0)\| =: \delta < \frac{1}{4\mu}$. Assume that it holds

$$\|F(x_0)\| < \frac{1}{2\mu^2} \min\left\{\frac{1}{\gamma}, \frac{1-4\mu\delta}{4\alpha}\right\}. \tag{15}$$

Let (x_k, B_k) be generated by the damped quasi-Newton method (10) starting at (x_0, B_0) with update function U and step size $\eta \in (0, \overline{\eta})$, where

$$\overline{\eta} = \min\left\{\frac{3}{2}, \left(1 - 2\mu^2\|F(x_0)\| \max\left\{\gamma, \frac{4\alpha}{1-4\mu\delta}\right\}\right)\frac{3}{4\gamma\mu^2\|F(x_0)\|}\right\}. \tag{16}$$

Then it holds $x_k \in N(x_0, \varrho)$ and $F(\lim_{k\to\infty} x_k) = 0$.

Proof By induction, we prove that the iterates (x_ℓ, B_ℓ) satisfy the following properties

$$x_\ell \in N(x_0, \varrho), \tag{17}$$

$$\|B_\ell - F'(x_\ell)\| < \frac{1}{4\mu}, \tag{18}$$

$$\|F(x_\ell)\| \leq q^\ell \|F(x_0)\| \text{ with } q := 1 - \eta\left(\frac{2}{3} - \frac{8}{9}\eta\gamma\mu^2\|F(x_0)\|\right). \tag{19}$$

From the bound on $\|F(x_0)\|$ in (15), we have $\overline{\eta} < \frac{3}{4\gamma\mu^2\|F(x_0)\|}$. Hence, it holds $\frac{2}{3} - \frac{8}{9}\eta\gamma\mu^2\|F(x_0)\| \in \left(0, \frac{2}{3}\right)$. These bounds together with (15) immediately imply $q < 1$ and with $\eta < \bar{\eta} \leq \frac{2}{3}$, it holds $q > 1 - \frac{2}{3}\eta > 0$. Thus, we have $q \in (0, 1)$, that is, the sequence $(\|F(x_k)\|)$ is monotonically decreasing in the sense that $\|F(x_\ell)\| \leq q^\ell\|F(x_0)\|$ can be shown to hold for all $\ell = 0, 1, \ldots$.

From the assumptions above, we know that (x_0, B_0) satisfies properties (17), (18), and (19). Let $k \in \mathbb{N}$ and assume the properties (17), (18), and (19) hold for all $\ell = 0, \ldots, k$. We show that these properties hold also for $\ell = k + 1$.

For $\ell = 0, \ldots, k$, we have

$$\begin{aligned}
\left\|B_\ell^\top h\right\| &\geq \left\|F'(x_\ell)^\top h\right\| - \left\|(F'(x_\ell) - B_\ell)^\top h\right\| \\
&\geq \left\|F'(x_0)^\top h\right\| - \left\|(F'(x_0) - F'(x_\ell))^\top h\right\| - \frac{1}{4\mu}\|h\| \\
&\geq \left(\frac{2}{\mu} - \gamma\varrho - \frac{1}{4\mu}\right)\|h\| = \frac{3}{4\mu}\|h\|,
\end{aligned} \tag{20}$$

where we have used the triangle inequality to obtain the first inequality, the triangle inequality and (18) to reach the second inequality, and the bound $\left\| F'(x_0)^\top h \right\| \geq \frac{2}{\mu} \|h\|$ together with the Lipschitz continuity of F' and $x_\ell \in N(x_0, \varrho)$ to obtain the last inequality. Finally, the choice $\varrho = \frac{1}{\gamma\mu}$ gives the last expression. Thus, we obtain $\left\| B_\ell^\dagger F(x_\ell) \right\| \leq \frac{4\mu}{3} \|F(x_\ell)\|$ for $\ell = 0, \ldots, k$.

Additionally, for $\eta \in (0, \overline{\eta})$ and the choice of $\overline{\eta}$ in (16), it holds

$$\begin{aligned} \frac{\eta}{1-q} = \frac{1}{\frac{2}{3} - \frac{8}{9}\eta\gamma\mu^2 \|F(x_0)\|} &< \frac{1}{\frac{2}{3} - \frac{8}{9}\overline{\eta}\gamma\mu^2 \|F(x_0)\|} \\ &= \frac{1}{\frac{2}{3} - \frac{2}{3}\left(1 - 2\mu^2 \|F(x_0)\| \max\left\{\gamma, \frac{4\alpha}{1-4\mu\delta}\right\}\right)} \\ &= \frac{3}{4\mu^2 \|F(x_0)\|} \min\left\{\frac{1}{\gamma}, \frac{1-4\mu\delta}{4\alpha}\right\}. \end{aligned} \tag{21}$$

By using $\left\| B_\ell^\dagger F(x_\ell) \right\| \leq \frac{4\mu}{3} \|F(x_\ell)\|$ for $\ell \leq k$ combined with (19), the closed form of the geometric series $\sum_{\ell=0}^{\infty} q^\ell = \frac{1}{1-q}$, and the upper bound (21), we obtain

$$\begin{aligned} \sum_{\ell=0}^{k} \|x_{\ell+1} - x_\ell\| \leq \sum_{\ell=0}^{k} \eta \left\| B_\ell^\dagger F(x_\ell) \right\| \leq \sum_{\ell=0}^{k} \eta \frac{4\mu}{3} q^\ell \|F(x_0)\| &\leq \frac{4\eta\mu \|F(x_0)\|}{3(1-q)} \\ &< \frac{1}{\mu} \min\left\{\frac{1}{\gamma}, \frac{1-4\mu\delta}{4\alpha}\right\}. \end{aligned} \tag{22}$$

We use this upper bound to prove $x_{k+1} \in N(x_0, \varrho)$ and $\left\| B_{k+1} - F'(x_{k+1}) \right\| < \frac{1}{4\mu}$.

To show $x_{k+1} \in N(x_0, \varrho)$, we compute

$$\|x_{k+1} - x_0\| \leq \sum_{\ell=0}^{k} \|x_{\ell+1} - x_\ell\| < \frac{1}{\mu} \min\left\{\frac{1}{\gamma}, \frac{1-4\mu\delta}{4\alpha}\right\} \leq \frac{1}{\mu\gamma} = \varrho,$$

by using the triangle inequality and (22).

From the bounded deterioration principle (11), we have

$$\left\| B_{\ell+1} - F'(x_{\ell+1}) \right\| \leq \left\| B_\ell - F'(x_\ell) \right\| + \alpha \|x_{\ell+1} - x_\ell\|,$$

for $\ell = 0, \ldots, k$. Thus, by repeated application of this principle and the bound (22), for B_{k+1}, we have

$$\begin{aligned} \left\| B_{k+1} - F'(x_{k+1}) \right\| &\leq \left\| B_0 - F'(x_0) \right\| + \alpha \sum_{\ell=0}^{k} \|x_{\ell+1} - x_\ell\| \\ &< \delta + \frac{\alpha}{\mu} \min\left\{\frac{1}{\gamma}, \frac{1-4\mu\delta}{4\alpha}\right\} \leq \delta + \frac{1-4\mu\delta}{4\mu} = \frac{1}{4\mu}. \end{aligned}$$

Finally, it holds

$$
\begin{aligned}
\|F(x_{k+1})\| &= \left\| F(x_k) + \int_0^1 F'(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k)dt \right\| \\
&= \left\| F(x_k) - \eta B_k B_k^\dagger F(x_k) - \int_0^1 (F'(x_k + t(x_{k+1} - x_k)) - B_k)\eta B_k^\dagger F(x_k)dt \right\| \\
&\leq \left\| F(x_k) - \eta B_k B_k^\dagger F(x_k) \right\| \\
&\quad + \left\| \int_0^1 (F'(x_k + t(x_{k+1} - x_k)) - F'(x_k) + F'(x_k) - B_k)\eta B_k^\dagger F(x_k)dt \right\| \\
&\leq (1 - \eta)\|F(x_k)\| + \eta \int_0^1 \left\| F'(x_k + t(x_{k+1} - x_k)) - F'(x_k) \right\| \left\| B_k^\dagger F(x_k) \right\| dt \\
&\quad + \eta \int_0^1 \left\| F'(x_k) - B_k \right\| \left\| B_k^\dagger F(x_k) \right\| dt \\
&\leq (1 - \eta)\|F(x_k)\| + \eta \int_0^1 \gamma t \left\| x_{k+1} - x_k \right\| \left\| B_k^\dagger F(x_k) \right\| dt \\
&\quad + \eta \int_0^1 \left\| F'(x_k) - B_k \right\| \left\| B_k^\dagger F(x_k) \right\| dt \\
&\leq (1 - \eta)\|F(x_k)\| + \frac{\eta^2 \gamma}{2} \left\| B_k^\dagger F(x_k) \right\|^2 + \eta \frac{1}{4\mu} \frac{4\mu}{3} \|F(x_k)\| \\
&\leq \left(1 - \frac{2}{3}\eta\right) \|F(x_k)\| + \frac{8}{9}\eta^2 \gamma \mu^2 \|F(x_k)\|^2 ,
\end{aligned}
$$

where we used $B_k B_k^\dagger = I$, the update formula yielding $\|x_{k+1} - x_k\| = \eta \left\| B_k^\dagger F(x_k) \right\|$, the Lipschitz continuity of F', the bounds $\left\| B_k^\dagger F(x_k) \right\| \leq \frac{4\mu}{3} \|F(x_k)\|$, and (18).

From $\|F(x_k)\| \leq q^k \|F(x_0)\|$ and $q \in (0, 1)$, we have $\|F(x_k)\| \leq \|F(x_0)\|$. Hence, this computation shows $\|F(x_{k+1})\| \leq q^{k+1} \|F(x_0)\|$ with $q = 1 - \eta \left(\frac{2}{3} - \frac{8}{9}\eta\gamma\mu^2 \|F(x_0)\|\right)$.

Thus, it holds $\lim_{k\to\infty} \|F(x_k)\| = 0$, and from (20) we obtain

$$
\|x_{k+1} - x_k\| = \eta \left\| B_k^\dagger F(x_k) \right\| \leq \eta \frac{4\mu}{3} \|F(x_k)\| ,
$$

which implies that (x_k) is a Cauchy sequence. Therefore, the claim follows from the continuity of the mapping $x \mapsto \|F(x)\|$ in Ω. □

In Theorem 2, the bound (15) on the norm of the residual $\|F(x_0)\|$ at the starting point ensures that a step size $\eta > 0$ can be chosen such that all iterates stay in a neighbourhood of x_0 and that the deterioration of the Jacobian approximation B_k to the Jacobian $F'(x_k)$ is bounded.

Having this semi-local convergence theorem at hand, we consider a specific choice of a least change secant update, the first Broyden update for underdetermined systems introduced in [5]. We show that our quasi-Newton scheme with Broyden update from

(14) converges to a root if the start value $x_0 \in \mathbb{R}^n$ results in a small residual and the Jacobian of the residual function has full rank. Specifically, we show that, for a suitably chosen starting vector x_0, the bound (16) on the damping parameter satisfies $\overline{\eta} > 1$ and thus the convergence of the scheme without damping, that is, $\eta = 1$ can be concluded from Theorem 2 since $\eta = 1 \in (0, \overline{\eta})$. Hence, our quasi-Newton scheme with Broyden update (14) converges.

Corollary 1 *Assume that there exists a closed and convex $\Omega \subseteq \mathbb{R}^n$ such that F satisfies Assumption 1 in Ω.*

Let $x_0 \in \Omega$ and $\mu > 0$ such that $\left\| F'(x_0)^\top h \right\| \geq \frac{2}{\mu} \|h\|$, $h \in \mathbb{R}^m$, and $N(x_0, \varrho) \subseteq \Omega$, where $\varrho = \frac{1}{\gamma\mu}$. Assume that it holds

$$\|F(x_0)\| < \frac{3}{16\mu^2\gamma}. \tag{23}$$

Let (x_k, H_k) be generated by the quasi-Newton method with Broyden update as in (14) starting with $(x_0, F'(x_0)^\dagger)$. Then, it holds $x_k \in N(x_0, \varrho)$ and $F(\lim_{k\to\infty} x_k) = 0$.

Proof By definition of H_k, we have $H_k = B_k^\dagger$ with B_k obtained by the Broyden update (12). As shown in [24], the update function U satisfies Assumption 2 with $\alpha = \frac{\gamma}{2}$. Thus, the assumptions of Theorem 2 hold with $\delta = 0$ since it holds $\|F(x_0)\| < \frac{1}{4\gamma\mu^2} = \frac{1}{2\mu^2} \min\left\{\frac{1}{\gamma}, \frac{1}{4\alpha}\right\}$.

Thus, the damped quasi-Newton method converges for any $\eta \in (0, \overline{\eta})$ with

$$\overline{\eta} = \left(1 - 4\mu^2 \|F(x_0)\| \gamma\right) \frac{3}{4\gamma\mu^2 \|F(x_0)\|} = \frac{3}{4\gamma\mu^2 \|F(x_0)\|} - 3 > 1,$$

using (23). □

Quadratic case Consider the case where all functions F_i are quadratic as in (5). In this case, the quasi-Newton method with Broyden update starting at $x_0 = 0 \in \mathbb{R}^n$ and $B_0 = F'(x_0) = B$ converges if B has full rank with $\left\| B^\top h \right\| \geq \frac{2}{\mu} \|h\|$, $h \in \mathbb{R}^2$, and $\|c\| < \frac{3}{16\mu^2\gamma}$.

For the specific problem (9), it holds $\|F(0)\| = \frac{1}{16} < \frac{3}{2\sqrt{2}} \frac{1}{16} = \frac{3}{16\mu^2\gamma}$. Hence, the quasi-Newton method with Broyden update starting at $x_0 = 0 \in \mathbb{R}^n$ and $B_0 = F'(x_0)$ converges to a root of this problem.

5 Application to a data fitting problem

A specific class of systems of nonlinear underdetermined equations corresponds to data fitting tasks, which are known as supervised learning problems in the context of artificial neural networks. Let $X = \{(\xi^\ell, \zeta^\ell) \mid \ell = 1, \ldots, |X|\} \subseteq \mathbb{R}^{n_{in}} \times \mathbb{R}^{n_{out}}$ be a set of given data points and $g(x, \cdot) : \mathbb{R}^{n_{in}} \to \mathbb{R}^{n_{out}}$ be a parameterised model. Our goal is to find the values of the network's parameters $x \in \mathbb{R}^n$ such that $g(x, \xi^\ell) = \zeta^\ell$

for $\ell = 1, \ldots, |X|$, that is, we are interested in a solution of (1) with the residual function $F : \mathbb{R}^n \to \mathbb{R}^{|X| n_{out}}$ defined element-wise by

$$F_i(x) := \left(\zeta_j^\ell - g_j(x, \xi^\ell)\right) \qquad \text{for} \quad i = n_{out}(\ell - 1) + j. \tag{24}$$

To simplify the notation, let $\Xi \in \mathbb{R}^{|X| \times n_{in}}$ denote the matrix holding the input data vectors $\xi^1, \ldots, \xi^{|X|} \in \mathbb{R}^{n_{in}}$ as rows and $Z \in \mathbb{R}^{|X| \times n_{out}}$ denote the matrix holding the output data vectors $\zeta^1, \ldots, \zeta^{|X|} \in \mathbb{R}^{n_{out}}$ as rows.

We consider a shallow neural network model of the form $g(x, \xi^\ell) = \sum_{i=1}^{n_h} v_i \phi(w_i^\top \xi^\ell) \in \mathbb{R}$ with parameters $x = (w_1^\top, w_2^\top, \ldots, w_{n_h}^\top)^\top \in \mathbb{R}^{n_h n_{in}}$, fixed hidden-to-output weights $v \in \mathbb{R}^{n_h}$ and element-wise applied activation function $\phi : \mathbb{R} \to \mathbb{R}$. Now, working in the framework of artificial networks, we adopt the terms from this context and refer to n_h as the width of the shallow neural network g.

The setting where the activation function ϕ satisfies $0 < m \leq \left|\phi'(z)\right|$ and $\left|\phi''(z)\right| \leq M$ for $z \in \mathbb{R}$ is discussed in [16, 21] for specific choices of the weights v. From [16] we have the following result.

Lemma 1 *Let $F : \mathbb{R}^n \to \mathbb{R}^{|X|}$ be defined element-wise by (24) with $g(x, \xi^\ell) = \sum_{i=1}^{n_h} v_i \phi(w_i^\top \xi^\ell)$ with parameters $x = (w_1^\top, w_2^\top, \ldots, w_{n_h}^\top)^\top \in \mathbb{R}^{n_h n_{in}}$. Assume that there exist constants $m, M > 0$ such that $m \leq \left|\phi'(z)\right|$ and $M \geq \left|\phi''(z)\right|$ for any $z \in \mathbb{R}$.*

Then, for any closed and convex set $\Omega \subseteq \mathbb{R}^n$ the Jacobian F' is Lipschitz continuous with $\gamma = M \|v\| \|\Xi\|_{2,\infty} \|\Xi\|$ and $\left\|F'(x_0)^\top h\right\| \geq \frac{2}{\mu} \|h\|$, $h \in \mathbb{R}^{|X|}$, holds for any $x_0 \in \Omega$ with $\mu = \frac{2}{m \|v\| \sigma_{min}(\Xi)}$, where $\sigma_{min}(\Xi)$ is the smallest singular value of the matrix $\Xi \in \mathbb{R}^{|X| \times n_{in}}$ and $\|\Xi\|_{2,\infty}$ denotes the largest Euclidean norm of the rows of the matrix Ξ.

As a direct consequence of this property and Corollary 1, we have

Corollary 2 *Let $F : \mathbb{R}^n \to \mathbb{R}^{|X|}$ be defined as in Lemma 1.*

Let $x_0 \in \mathbb{R}^{n_h n_{in}}$ satisfy

$$\|F(x_0)\| < \frac{3m^2 \|v\| \sigma_{min}(\Xi)^2}{64M \|\Xi\|_{2,\infty} \|\Xi\|}, \tag{25}$$

and let (x_k, H_k) be the iterates obtained by the quasi-Newton method with Broyden update and $H_0 := F'(x_0)^\dagger$. Then the sequence (x_k) converges to a root of F.

For activation functions such as the logistic sigmoid function $\phi(z) = \frac{1}{1+\exp(-z)}$ or the softplus function $\phi(z) = \log(1 + \exp(z))$, which are typically used in artificial neural network models, there is no lower bound $m > 0$ of the derivative. In [16], it is suggested to add a small linear term, that is, to consider $\phi_t(z) := (1 - t)\phi(z) + tz$ for some activation function ϕ. A different approach is discussed in [17], where this assumption is relaxed. Instead, in order to show that $\left\|F'(x_0)^\top h\right\| \geq \frac{2}{\mu} \|h\|$, $h \in \mathbb{R}^{|X|}$, holds for some $\mu > 0$, results from randomised numerical linear algebra are employed to prove that this bound is satisfied up to a given probability if the start value is obtained randomly and the network is sufficiently wide. Specifically, we have the following result [17].

Lemma 2 *Assume that there exists an* $M > 0$ *such that* $\sup\{\sup_{z\in\mathbb{R}} |\phi'(z)|, \sup_{z\in\mathbb{R}} |\phi''(z)|\} \leq M$, *and the input data* ξ^ℓ *is normalised, that is, it holds* $\|\xi^\ell\| = 1$ *for* $\ell = 1, \ldots, |X|$. *Define* $g : \mathbb{R}^{n_h \times n_{in}} \times \mathbb{R}^{n_{in}}$ *by* $g(x, \xi^\ell) = v^\top \phi(x\xi^\ell)$ *for the vector* $v \in \mathbb{R}^{n_h}$ *defined by*

$$v_i := \begin{cases} -\frac{\|Z\|}{\sqrt{n_h |X|}} & i = 1, \ldots, \lceil \frac{n_h}{2} \rceil, \\ \frac{\|Z\|}{\sqrt{n_h |X|}} & i = \lceil \frac{n_h}{2} \rceil + 1, \ldots, 2\lceil \frac{n_h}{2} \rceil, \\ 0 & else. \end{cases}$$

Let $F : \mathbb{R}^{n_h \times n_{in}} \to \mathbb{R}^{|X|}$ *be the residual function defined in* (24). *Then the Jacobian* F' *is Lipschitz continuous with constant* $\gamma > 0$.

Let $x_0 \in \mathbb{R}^{n_h \times n_{in}}$ *with entries i.i.d. from* $\mathcal{N}(0, 1)$ *and* $c > 0$. *Assume that the problem is large enough, that is for* $\delta > 0$ *it holds*

$$\begin{aligned} \sqrt{n_h n_{in}} &\geq \frac{8}{c} M(1 + (1+\delta)M)\,|X|\,\tilde{\kappa}(\Xi) \quad \text{with} \quad \tilde{\kappa}(\Xi) \\ &:= \frac{\sqrt{\frac{n_{in}}{|X|}}\,\|\Xi\|}{(\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g\phi'(g)])^2 \sigma_{min}^2(\Xi * \Xi)}, \end{aligned} \tag{26}$$

where $*$ *denotes the Khatri-Rao product of matrices as defined in [17] and* σ_{min} *refers to the smallest singular value of the matrix. Then with probability at least* $1 - \frac{1}{|X|} - \exp\left(-\delta^2 \frac{|X|}{2\|\Xi\|^2}\right)$ *we have* $\left\|F'(x_0)^\top h\right\| \geq \frac{2}{\mu}\|h\|$, $h \in \mathbb{R}^{|X|}$, *and*

$$\|F(x_0)\| < c\,\frac{1}{\mu^2 \gamma}. \tag{27}$$

Corollary 3 *Let* $X \subseteq \mathbb{R}^{n_{in}} \times \mathbb{R}$ *be a given data set and let* F *be defined as in Lemma 2. Assume* n_h *is large enough such that* (26) *is satisfied with* $c = \frac{3}{16}$. *Let* x_0 *be generated randomly with entries i.i.d. from* $\mathcal{N}(0, 1)$. *Then the quasi-Newton method* (14) *with Broyden update starting with* $(x_0, F'(x_0)^\dagger)$ *converges to a root of* F *with probability at least* $1 - \frac{1}{|X|} - \exp\left(-\delta^2 \frac{|X|}{2\|\Xi\|^2}\right)$.

We report results of experiments for training shallow neural networks with varying number of parameters $n_{in} n_h$ in the next section.

6 Numerical experiments

In this section, we show results of experiments for solving the quadratic problem (9) with varying $\|c\|$ and root-finding problems arising in the context of training shallow neural networks with varying size with our quasi-Newton method with Broyden update from (14). To showcase the potential of our method, we also include results concerning the computation of eigenpairs of an eigenvalue problem and the supervised training of a multilayer neural network for the classification of the Iris data set.

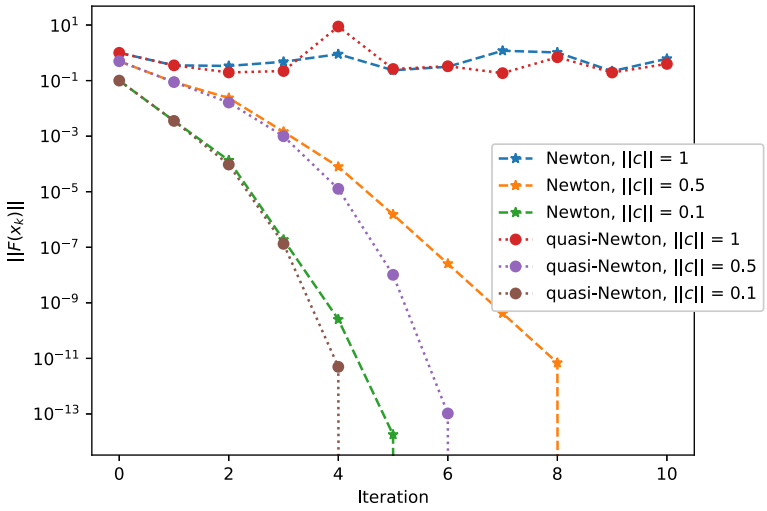

Fig. 1 Norm of residual $\|F(x_k)\|$ per iteration for Newton and quasi-Newton method with Broyden update for solving the quadratic root-finding problem (9) with different values of $\|c\|$

First, we apply the quasi-Newton method with Broyden update starting with $(0, F'(0)^{\dagger})$ to solve the quadratic problem (9) with varying values of $\|c\|$. As specific choices, we consider $c \in \{(1,0)^{\top}, (0.5,0)^{\top}, (0.1,0)^{\top}\}$; see the discussion at the end of Sect. 4. While the quasi-Newton method does not converge for $c = (1,0)^{\top}$, it converges to a root of the system in less than 10 iterations for the choices of c with smaller norm, see Fig. 1. For a comparison, we have included the results of applying the Newton method with update $x_{k+1} = x_k - F'(x_k)^{\dagger}F(x_k)$ which shows a similar convergence behaviour.

As a second test problem, we consider the computation of eigenpairs of an eigenvalue problem as an underdetermined root-finding problem. Given a symmetric matrix $A \in \mathbb{R}^{n\times n}$, we define the residual function $F : \mathbb{R}^{n+1} \to \mathbb{R}^n$ by

$$F(x) = F(\lambda, u) = (A - \lambda I_n)u, \tag{28}$$

where $\lambda \in \mathbb{R}$ is the first element of $x \in \mathbb{R}^{n+1}$ and $u \in \mathbb{R}^n$ consists of the remaining elements of the parameter vector x, that is, $x = (\lambda, u)$. We assume that A has distinct eigenvalues. In this case, the Jacobian $F'(\lambda, u) = \left(-u, A - \lambda I\right)$ has full rank. We investigate the convergence properties of Newton's method and of our quasi-Newton method with Broyden update for the following two matrices with dimension $n = 10$:

1. The system matrix arising from the finite difference discretisation of the 1d Poisson problem with zero boundary conditions, which is a tridiagonal matrix with elements $A_{ii} = 2$, $i = 1, \ldots, n$, and $A_{i,i-1} = A_{i-1,i} = -1$, $i = 2, \ldots, n$. The eigenvalues of this matrix are given by $\mu_j = 4\sin^2\left(\frac{j\pi}{2(n+1)}\right)$, $j = 1, \ldots, n$ [6].
2. A matrix with eigenvalues $\mu_j = j$, $j = 1, \ldots, 10$, and randomly generated eigenvectors obtained as rows from an orthogonal matrix $U \in \mathbb{R}^{10\times 10}$, that is

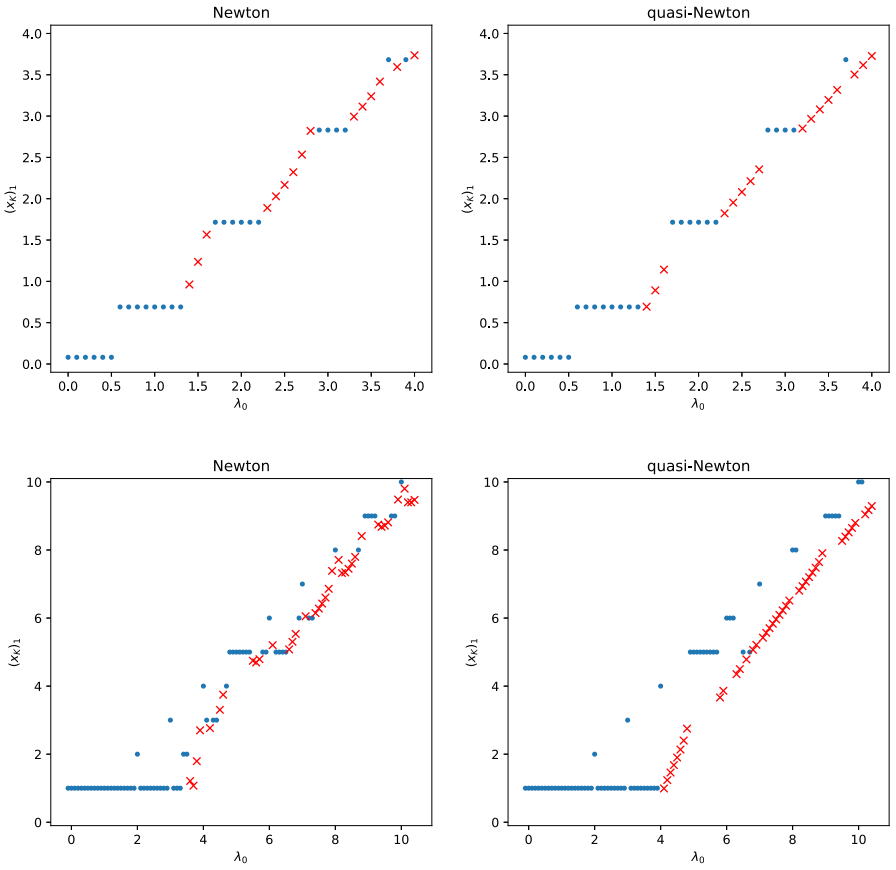

Fig. 2 First element $(x_K)_1$ of the final iterates versus different start values with $\lambda_0 = (x_0)_1$ as the first element obtained by the Newton method (left) and our quasi-Newton method with Broyden update (right) for the Poisson matrix (top) and the randomly generated matrix (bottom). Red cross (x) marks results, where $(x_K)_i = 0, i = 2, \ldots, n+1$ (Color figure online)

$A = UMU^\top$, with $M \in \mathbb{R}^{10\times 10}$ being the diagonal matrix with values $M_{jj} = \mu_j$, $j = 1, \ldots, 10$.

We test our methods for starting vectors that differ only in the first element $\lambda_0 = (x_0)_1$, corresponding to the eigenvalue, and stop the iteration when the iterate x_k satisfies the following condition $\|F(x_k)\| < 10^{-15}$. We denote the final iterate by x_K. We initialise the part of the starting vector corresponding to the eigenvector as follows:

1. as $(x_0)_i = \frac{1}{\sqrt{n}}$, $i = 2, \ldots, n+1$, in the case of the Poisson matrix, and
2. as $(x_0)_i = 1$, $i = 2, \ldots, n+1$, in the case of the randomly generated matrix.

We observe that for some start values λ_0 the methods converge to solutions with $(x_K)_i = 0$, $i = 2, \ldots, n+1$, which do not represent eigenvectors. These results are marked by a red cross (x) in Fig. 2 that shows the eigenvalue approximations $(x_K)_1$ for different choices of the first entry of the starting vector $\lambda_0 = (x_0)_1$. These

results suggest that our quasi-Newton method could be extended in order to solve eigenvalue problems if a normalisation condition on the vector $(x_k)_i, i = 2, \ldots, n+1$, is appropriately included in the algorithm. While we do not consider this extension in this work, we expect that this normalisation can be implemented by 1) augmenting the residual function with an additional entry $F(\lambda, u)_{n+1} = \|u\| - 1$, or 2) normalising the quasi-Newton update at every step in order to enforce that the part of the solution corresponding to the eigenvector has norm 1, that is, in a way similar to the power iteration.

Both iterative methods, the Newton method and our quasi-Newton method, find several distinct eigenpairs of the problems. With proper initialisation, all eigenvalues of the randomly generated matrix can be obtained together with the corresponding eigenvectors. For the Poisson matrix the methods are able to find five of the ten distinct eigenvalues and corresponding eigenvectors.

Next, we consider the supervised training of shallow neural networks. To investigate the relation between the convergence of the quasi-Newton method and the width of the shallow neural network in a data fitting problem, we consider a setting similar to the experiments in [17]. Specifically, we consider a randomly generated data set with $|X| = 100$ samples and vary the number of hidden nodes n_h and input dimension n_{in}. To this end, each input sample is drawn element-wise from the standard normal distribution and normalised, that is $(\hat{\xi}^\ell)_i \sim \mathcal{N}(0, 1), i = 1, \ldots, n_{in}$, and $\xi^\ell = \frac{\hat{\xi}^\ell}{\|\hat{\xi}^\ell\|}$, $\ell = 1, \ldots, |X|$. Similarly, the output values $\zeta_\ell \in \mathbb{R}$ are generated from the standard normal distribution, that is $\zeta^\ell \sim \mathcal{N}(0, 1), \ell = 1, \ldots, |X|$. As a shallow neural network model, we consider the setting from Lemma 2 with activation function the softplus function defined by $\phi(z) := \log(1+\exp(z))$. We apply the quasi-Newton method with Broyden update from (14) starting from $(x_0, F'(x_0)^\dagger)$ with x_0 randomly generated as in Lemma 2. We stop after 30 iterations or when the Euclidean norm of the residual function falls below 10^{-6}. This choice provides a reasonable bound since we run the experiments in `Python` using the `pytorch` framework with single precision floating point values according to IEEE 754 standard, hence the values have a precision of about 7 decimal digits. Thus, we cannot expect to obtain a smaller error. To account for the random generation of the data set and start value, we perform 10 independent runs per combination of number of hidden nodes and input dimension.

Figure 3 shows the number of successful runs (red) out of 10, that is, the number of runs for which the quasi-Newton method reduced the norm of the residual $\|F(x_k)\|$ below 10^{-6}. Similar to the results in [17] for the gradient method, we can observe a phase transition between successful and non-successful settings depending on the number of parameters $n_h n_{in}$. However, the phase transition for the quasi-Newton method is less strict than those for the gradient method shown in [17].

Next, we compare the decrease of the norm of the residual $\|F(x_k)\|$ for iterates obtained by the quasi-Newton method with Broyden update, the gradient method $x_{k+1} = x_k - \eta F'(x_k)^\top F(x_k)$ applied to minimise the least squares objective function $\frac{1}{2}\|F(x)\|^2$ with $\eta = 0.15$ similar to [17], and the Newton method $x_{k+1} = x_k - F'(x_k)^\dagger F(x_k)$. To this end, we consider the training of a shallow network with softplus activation with 50 hidden nodes and input dimension $n_{in} = 50$. The residual norm per iteration and per time is shown in Fig. 4. As expected, the Newton method converges

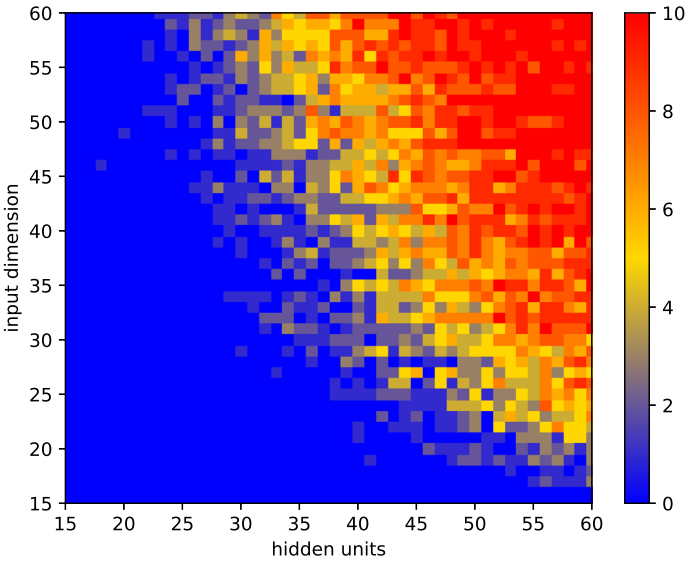

Fig. 3 Number of successful runs (encoded by colour) of quasi-Newton method with first Broyden update from (14) from a random initialisation for varying problem size. Red corresponds to 10 successful runs out of 10 (Color figure online)

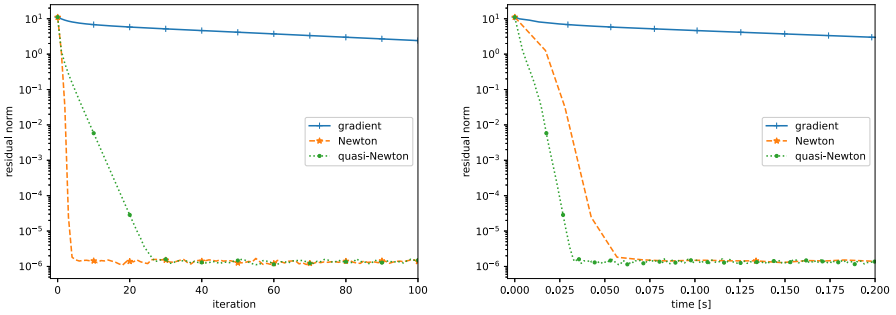

Fig. 4 Training of a shallow neural network with softplus activation and 50 hidden nodes to fit $|X| = 100$ randomly generated data samples with input dimension $n_{in} = 50$. Comparison of gradient method with step size $\eta = 0.15$, quasi-Newton method and Newton method. The pictures depict the values of the norm of the residuals per iteration and with respect to CPU time. Markers at every 10 iterations

in the least number of iterations and the quasi-Newton method is almost as good while the gradient method shows a rather slow decrease of the norm of the residual. On the other hand, in terms of CPU time, the quasi-Newton method is considerably faster than the Newton method.

As a final test problem, we consider the training of a multilayer neural network to perform classification of a real data set to demonstrate the potential of our quasi-Newton algorithm for networks with more than one hidden layer, which are not covered by the theoretical results in Sect. 5. We define the computation of a multilayer neural network with weights $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, $\ell = 1, \ldots, L+1$, and bias values $b^{(\ell)} \in \mathbb{R}^{n_\ell}$,

$\ell = 1, \ldots, L$, recursively by

$$\begin{aligned} \gamma^{(0)} &= \xi, \\ \gamma^{(\ell)} &= \phi\left(W^{(\ell)}\gamma^{(\ell-1)} + b^{(\ell)}\right), \quad \ell = 1, \ldots, L, \\ \gamma^{(L+1)} &= W^{(L+1)}\gamma^{(L)}, \end{aligned} \tag{29}$$

where $n_\ell \in \mathbb{N}$ is the width of the ℓ-th layer, $\gamma^{(\ell)} \in \mathbb{R}^{n_\ell}$ is the activation of the nodes in the ℓ-th layer. We obtain the corresponding model function $g : \mathbb{R}^n \times \mathbb{R}^{n_{in}} \to \mathbb{R}^{n_{out}}$ with parameters $x = (vec(W^{(1)}), b^{(1)}, \ldots, vec(W^{(L+1)})) \in \mathbb{R}^n$ collecting all weights and bias terms, by setting $g(x, \xi) = \gamma^{(L+1)}$, the output of the multilayer neural network with parameters x for input ξ. The total number of parameters is given by $n = \sum_{\ell=1}^{L} n_\ell(n_{\ell-1} + 1) + n_{L+1}n_L$. In our experiment, we choose $L = 4$, $n_i = 30$, $i = 1, \ldots, 4$, and use the hyperbolic tangent as element-wise activation function, that is, let $\phi(z) = \tanh(z)$. Our goal is to find a set of parameters of the network that correctly classifies the species in the Iris data set with input-label pairs $(\xi^\ell, \zeta^\ell) \in \mathbb{R}^4 \times \{0, 1, 2\}$, where the integer labels represent the type of the iris plant, that is, either Iris Setosa, Iris Versiolour, or Iris Virginica. To this end, we aim to solve the equations $g(x, \xi^\ell) - \zeta^\ell = 0$, where $\zeta^\ell \in \{0, 1, 2\}$ is the label indicating the true species. To illustrate the generalisation properties of the trained network, we select 10 data points of each species randomly as test data and use only the remaining 120 data points for training. That is, our problem has $m = 120$ equations and $n = 3900$ unknowns.

To obtain a suitable start value x_0, we apply 100 (or 200) iterations of an inexact damped Newton method, where the iterates are computed as $x_{k+1} = x_k - \eta d_k$ with $\eta = 10^{-1.5}$ and the direction d_k being an approximate solution of $F'(x_k)d_k = F(x_k)$ computed by the `torch.linalg.lstsq` algorithm with parameters `rcond=1e-5, driver='gelsd'`. Notice that the model with parameters obtained from 200 iterations of this method reaches already an accuracy of 100% on the training data. We compare a damped version of our quasi-Newton method with Broyden update, that is, the iterates are defined by $x_{k+1} = x_k - \eta H_k F(x_k)$ with $\eta = 0.1$, with the damped Newton method with the same step size, and the gradient method applied to minimise $\frac{1}{2}\|F(x)\|^2$ with step size $\eta = 10^{-6}$.

Both the damped quasi-Newton method and the damped Newton method outperform the gradient method with respect to the norm of the residual. Compared to the damped Newton method our damped quasi-Newton method shows a similar performance for the first 50 iterations and reaches a slightly higher objective value in later iterations.

We determine the predicted label for a given input as the element from $\{0, 1, 2\}$, which is closest to the output of the network, that is, $p(x, \xi) = \operatorname{argmin}_{t\in\{0,1,2\}} |t - g(x, \xi)|$ gives the label obtained by the network with parameters x for the input ξ. Then, the accuracy of the network with parameters x is given by $\frac{1}{|X|}|\{\ell \in \{1, \ldots, |X|\} \mid h(x, \xi^\ell) = \zeta^\ell\}|$, where X is either the training data set consisting of 120 samples or the test data set consisting of 30 samples. All methods find parameters such that the corresponding model classifies all data from the training

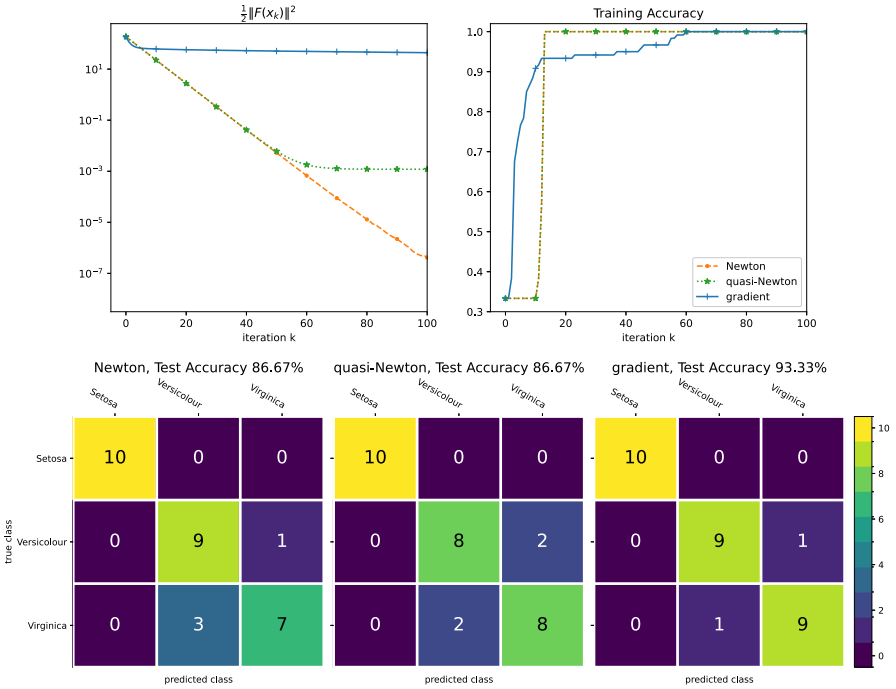

Fig. 5 Results from training of a multilayer neural network with $L = 4$ hidden layers with 30 nodes each and tanh activation to fit the Iris data set. Comparison of the damped Newton method, the damped quasi-Newton method with Broyden update with step size $\eta = 0.1$ each, and the gradient method with step size $\eta = 10^{-6}$. Least squares objective function and accuracy versus iteration (top) with markers at every 10 iterations. Confusion matrix for data from test set with 30 samples (bottom) showing the number of instances of a true class t which are predicted as class p by the model with parameters obtained by each of the algorithms for each combination of true and predicted classes (t, p)

data set correctly. On the test set, the model with parameters obtained by the gradient method gives a slightly higher accuracy (93.33%) compared to those obtained from our quasi-Newton method and the Newton scheme (86.67% each), see bottom of Fig. 5. The confusion tables for the models obtained by the three methods break down the number of instances of a specific class versus the number of predicted instances, that is, in the field true class t and predicted class p, the number of instances with label t which are classified as class p is given. Notably, all samples of type Setosa in the test set are correctly classified by each of the models obtained by the three tested algorithms, while some of the instances of type Versicolour and Virginica are mistaken.

7 Conclusion

The convergence of a quasi-Newton method for the solution of systems of nonlinear underdetermined equations was investigated. In particular, a new approach for the computation of the Moore–Penrose inverse of the approximate Jacobian coming from

the Broyden update was presented and a semi-local convergence result for a damped quasi-Newton method with least change secant update was shown. These theoretical results were illustrated for the case of systems of quadratic equations, and validated in the context of eigenvalue problems. Further, the potential of the proposed quasi-Newton method was showcased for the supervised training of overparameterised neural networks on synthetic data and for the classification of the Iris data set by a multilayer neural network.

A Least-change secant update

Given an approximation B_k to $F'(x_k)$, secant information in form of vectors $y \in \mathbb{R}^m$ and $s \in \mathbb{R}^n$, such as $y = F(x_{k+1}) - F(x_k)$ and $s = x_{k+1} - x_k$, and an affine subspace $\mathcal{B} \subseteq \mathbb{R}^{m\times n}$, the least change secant update of B_k in $\mathcal{B}$ with respect to s, y and a norm $|||\cdot|||$ is the unique solution B_{k+1} of

$$\min_{B\in M(\mathcal{B},Q(y,s))} |||B - B_k|||,$$

where $Q(y, s) = \{B \in \mathbb{R}^{m\times n} \mid Bs = y\}$ and

$$M(\mathcal{B}, Q(y, s)) = \begin{cases} \mathcal{B} & Q(y, s) = \emptyset \\ \operatorname{argmin}_{B\in\mathcal{B}} \min_{\overline{B}\in Q(y,s)} \left|\left|\left| B - \overline{B} \right|\right|\right| & Q(y, s) \neq \emptyset, \end{cases}$$

see, e.g., [24]. Hence in this setting, the update function is specified by $U(x_k, x_{k+1}, B) = \min_{B\in M(\mathcal{B},Q(y,s))} |||B - B_k|||$. Several specific choices of $\mathcal{B}$, s, y and the norm $\|\cdot\|$ for rectangular matrices are discussed in [5]. In particular, we consider the first Broyden update in (12) for which $\mathcal{B} = \mathbb{R}^{m\times n}$, $y = F(x_{k+1}) - F(x_k)$, $s = x_{k+1} - x_k$ and $|||\cdot||| = \|\cdot\|_F$, the Frobenius norm.

B Moore–Penrose inverse of Broyden update

To show the update formula for the Moore–Penrose inverse of the first Broyden update, we show a slightly more general result assuming $s \in \mathcal{R}(B^\top)$ and allowing for an arbitrary $y \in \mathbb{R}^n$. For the specific choice $s = x_{k+1} - x_k = -\eta B_k^\dagger F(x_k)$, we immediately have $s \in \mathcal{R}(B_k^\top)$. Hence, we can use the update formula in (13) for the update of the Moore–Penrose inverse of the first Broyden update.

Lemma 3 *Let $B_k \in \mathbb{R}^{m\times n}$, $m \leq n$ have full row rank, and $s \in \mathcal{R}(B_k^\top)$ and $y \in \mathbb{R}^n$. Let $B_{k+1} = B_k + \frac{(y-B_k s)s^\top}{s^\top s} \in \mathbb{R}^{m\times n}$. Then*

$$H_{k+1} = B_k^\dagger + \frac{(s - B_k^\dagger y)s^\top B_k^\dagger}{s^\top B_k^\dagger y}, \tag{30}$$

is the Moore–Penrose inverse of B_{k+1}.

Proof We show that H_{k+1} satisfies the properties of a Moore–Penrose inverse, that is, it holds

$$\begin{aligned} B_{k+1}H_{k+1}B_{k+1} = B_{k+1}, \quad H_{k+1}B_{k+1}H_{k+1} = H_{k+1}, \\ (B_{k+1}H_{k+1})^\top = B_{k+1}H_{k+1}, \quad (H_{k+1}B_{k+1})^\top = H_{k+1}B_{k+1}. \end{aligned} \tag{*}$$

To this end, we first note that it holds $B_k B_k^\dagger = I_m$ since B_k has full row rank. Thus, we have

$$\begin{aligned} B_{k+1}H_{k+1} &= \left(B_k + \frac{(y - B_k s)s^\top}{s^\top s}\right)\left(B_k^\dagger + \frac{(s - B_k^\dagger y)s^\top B_k^\dagger}{s^\top B_k^\dagger y}\right) \\ &= B_k B_k^\dagger + \frac{(y - B_k s)s^\top B_k^\dagger}{s^\top s} + \frac{B_k(s - B_k^\dagger y)s^\top B_k^\dagger}{s^\top B_k^\dagger y} \\ &\quad + \frac{(y - B_k s)s^\top}{s^\top s}\frac{(s - B_k^\dagger y)s^\top B_k^\dagger}{s^\top B_k^\dagger y} \\ &= I_m + \frac{(y - B_k s)s^\top B_k^\dagger}{s^\top s} + \frac{(B_k s - B_k B_k^\dagger y)s^\top B_k^\dagger}{s^\top B_k^\dagger y} \\ &\quad + \frac{(y - B_k s)s^\top B_k^\dagger}{s^\top B_k^\dagger y} - \frac{(y - B_k s)s^\top B_k^\dagger}{s^\top s} \\ &= I_m - \frac{(y - B_k s)s^\top B_k^\dagger}{s^\top B_k^\dagger y} + \frac{(y - B_k s)s^\top B_k^\dagger}{s^\top B_k^\dagger y} \\ &= I_m, \end{aligned}$$

by expanding the product to obtain the second equality, expanding the last fraction and using $B_k B_k^\dagger = I_m$ to obtain the third equality, and exploiting the cancellation of terms in the last equalities.

This property immediately shows the first three conditions in (*). It remains to check $(H_{k+1}B_{k+1})^\top = H_{k+1}B_{k+1}$. As above, we start with expanding the product $H_{k+1}B_{k+1}$, that is, we have

$$\begin{aligned} H_{k+1}B_{k+1} &= \left(B_k^\dagger + \frac{(s - B_k^\dagger y)s^\top B_k^\dagger}{s^\top B_k^\dagger y}\right)\left(B_k + \frac{(y - B_k s)s^\top}{s^\top s}\right) \\ &= B_k^\dagger B_k + \frac{B_k^\dagger(y - B_k s)s^\top}{s^\top s} + \frac{(s - B_k^\dagger y)s^\top B_k^\dagger B_k}{s^\top B_k^\dagger y} \\ &\quad + \frac{(s - B_k^\dagger y)s^\top B_k^\dagger(y - B_k s)s^\top}{s^\top B_k^\dagger y s^\top s}. \end{aligned}$$

Since $B_k^\dagger$ is the Moore–Penrose inverse of B_k, it holds $(B_k^\dagger B_k)^\top = B_k^\dagger B_k$. From $s \in \mathcal{R}(B_k^\top)$ we have $B_k^\dagger B_k s = s$ and thus $s^\top B_k^\dagger B_k = ((B_k^\dagger B_k)^\top s)^\top = s^\top$. By using these properties, we obtain

$$\begin{aligned} H_{k+1}B_{k+1} &= B_k^\dagger B_k + \frac{(B_k^\dagger y - B_k^\dagger B_k s)s^\top}{s^\top s} + \frac{(s - B_k^\dagger y)s^\top}{s^\top B_k^\dagger y} + \frac{(s - B_k^\dagger y)s^\top}{s^\top s} \\ &\quad - \frac{(s - B_k^\dagger y)s^\top}{s^\top B_k^\dagger y} \\ &= B_k^\dagger B_k + \frac{(B_k^\dagger y - s)s^\top}{s^\top s} + \frac{(s - B_k^\dagger y)s^\top}{s^\top s} \\ &= B_k^\dagger B_k. \end{aligned}$$

Hence, we have $(H_{k+1}B_{k+1})^\top = (B_k^\dagger B_k)^\top = B_k^\dagger B_k = H_{k+1}B_{k+1}$, which shows the remaining property. Thus, H_{k+1} is the Moore–Penrose inverse of B_{k+1}. □

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability statement The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest This study does not have any conflicts to disclose.

References

1. Allen-Zhu, Z., Li, Y., Song, Z.: A convergence theory for deep learning via over-parameterization. In: International Conference on Machine Learning, pp. 242–252. PMLR (2019)
2. Bao, J.-F., Li, C., Shen, W.-P., Yao, J.-C., Guu, S.-M.: Approximate Gauss–Newton methods for solving underdetermined nonlinear least squares problems. Appl. Numer. Math. **111**, 92–110 (2017)
3. Bergamaschi, L., De Simone, V., di Serafino, D., Martínez, A.: BFGS-like updates of constraint preconditioners for sequences of KKT linear systems in quadratic programming. Numerical Linear Algebra Appl. **25**(5), e2144 (2018)
4. Björck, Å.: Numerical Methods for Least Squares Problems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1996)
5. Bourji, S.K., Walker, H.F.: Least-change secant updates of nonsquare matrices. SIAM J. Numer. Anal. **27**(5), 1263–1294 (1990)
6. Briggs, W.L., Henson, V.E., McCormick, S.F.: A Multigrid Tutorial, 2nd edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2000)

7. Broyden, C.G.: A class of methods for solving nonlinear simultaneous equations. Math. Comput. **19**(92), 577–593 (1965)
8. Dembo, R.S., Eisenstat, S.C., Steihaug, T.: Inexact Newton methods. SIAM J. Numer. Anal. **19**(2), 400–408 (1982)
9. Dennis Jr, J.E., Schnabel, R.B.: Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Volume 16 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1996)
10. di Serafino, D., Toraldo, G., Viola, M.: Using gradient directions to get global convergence of Newton-type methods. Appl. Math. Comput. **409**(125612), 14 (2021)
11. Kantorovich, L.V.: On Newton's method. In: Trudy Matematicheskogo Instituta im. V. A. Steklova, vol. 28, pp. 104–144 (1949) (**in Russian**)
12. Kotsireas, I.S., Pardalos, P.M., Semenov, A., Trevena, W.T., Vrahatis, M.N.: Survey of methods for solving systems of nonlinear equations, part I: root-finding approaches. arXiv:2208.08530 (2022)
13. Kotsireas, I.S., Pardalos, P.M., Semenov, A., Trevena, W.T., Vrahatis, M.N.: Survey of methods for solving systems of nonlinear equations, part II: optimization based approaches. arXiv:2208.08532 (2022)
14. Martínez, J.M.: Quasi-Newton methods for solving underdetermined nonlinear simultaneous equations. J. Comput. Appl. Math. **34**(2), 171–190 (1991)
15. Ortega, J.M., Rheinboldt W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2000). Reprint of the 1970 original
16. Oymak, S., Soltanolkotabi, M.: Overparameterized nonlinear learning: gradient descent takes the shortest path? In: International Conference on Machine Learning, pp. 4951–4960. PMLR (2019)
17. Oymak, S., Soltanolkotabi, M.: Toward moderate overparameterization: global convergence guarantees for training shallow neural networks. IEEE J. Sel. Areas Inf. Theory **1**(1), 84–105 (2020)
18. Polyak, B.: Convexity of nonlinear image of a small ball with applications to optimization. Set-Valued Anal. **9**(1–2), 159–168 (2001)
19. Polyak, B.T.: Gradient methods for solving equations and inequalities. USSR Comput. Math. Math. Phys. **4**(6), 17–32 (1964)
20. Polyak, B.T., Tremba, A.: Solving underdetermined nonlinear equations by Newton-like method. arXiv:1703.07810 (2017)
21. Polyak, B.T., Tremba, A.: New versions of Newton method: step-size choice, convergence domain and under-determined equations. Optim. Methods Softw. **35**(6), 1272–1303 (2020)
22. Simonis, J.P.: Inexact Newton Methods Applied to Under-Determined Systems. Worcester Polytechnic Institute (2006)
23. Song, C., Ramezani-Kebrya, A., Pethick, T., Eftekhari, A., Cevher, V.: Subquadratic overparameterization for shallow neural networks. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds) Advances in Neural Information Processing Systems (2021)
24. Walker, H.F., Watson, L.T.: Least-change secant update methods for underdetermined systems. SIAM J. Numer. Anal. **27**(5), 1227–1262 (1990)