Math. Program., Ser. A (2008) 115:1–30
DOI 10.1007/s10107-007-0137-1

FULL LENGTH PAPER

Sparse quasi-Newton updates with positive definite matrix completion

Nobuo Yamashita

Received: 4 August 2005 / Accepted: 31 March 2007 / Published online: 19 June 2007

Abstract Quasi-Newton methods are powerful techniques for solving unconstrained minimization problems. Variable metric methods, which include the BFGS and DFP methods, generate dense positive definite approximations and, therefore, are not applicable to large-scale problems. To overcome this difficulty, a sparse quasi-Newton update with positive definite matrix completion that exploits the sparsity pattern $E := \{(i, j) \mid (\nabla^2 f(x))_{ij} \neq 0 \text{ for some } x \in R^n\}$ of the Hessian is proposed. The proposed method first calculates a partial approximate Hessian H_{ij}^{QN}, $(i, j) \in F$, where $F \supseteq E$, using an existing quasi-Newton update formula such as the BFGS or DFP methods. Next, a full matrix H_{k+1}, which is a maximum-determinant positive definite matrix completion of H_{ij}^{QN}, $(i, j) \in F$, is obtained. If the sparsity pattern E (or its extension F) has a property related to a chordal graph, then the matrix H_{k+1} can be expressed as products of some sparse matrices. The time and space requirements of the proposed method are lower than those of the BFGS or the DFP methods. In particular, when the Hessian matrix is tridiagonal, the complexities become $O(n)$. The proposed method is shown to have superlinear convergence under the usual assumptions.

Keywords Quasi-Newton method · Large-scale problems · Sparsity · Positive definite matrix completion

Mathematics Subject Classification (2000) 90C53 · 90C06

N. Yamashita (✉)
Department of Applied Mathematics of Physics, Graduate School of Informatics,
Kyoto University, Kyoto 606-8501, Japan
e-mail: nobuo@i.kyoto-u.ac.jp

1 Introduction

In the present paper, we consider the following unconstrained minimization problem:

$$\begin{array}{ll} \min & f(x) \\ \text{subject to} & x \in R^n. \end{array} \tag{1}$$

Throughout the present paper, it is assumed that f is twice continuously differentiable, n is huge, and $\nabla^2 f(x)$ is sparse. For solving the unconstrained minimization problem, there exist several useful methods, including steepest descent, Newton and quasi-Newton methods, and the conjugate gradient method [17]. The quasi-Newton method is easy to implement and has good convergence properties.

The quasi-Newton method generates a sequence $\{x_k\}$ by $x_{k+1} = x_k - H_k \nabla f(x_k)$ with an approximate inverse Hessian H_k. The approximate inverse Hessian usually satisfies the secant condition:

$$H_{k+1} y_k = s_k, \tag{2}$$

where

$$\begin{aligned} s_k &= x_{k+1} - x_k \\ y_k &= \nabla f(x_{k+1}) - \nabla f(x_k). \end{aligned}$$

In the present paper, the primary focus is on updates that preserve the positive definiteness of H_k. When H_k is positive definite, $d_k = -H_k \nabla f(x_k)$ becomes the descent direction of f at x_k, and so it is possible to construct a globally convergent algorithm combined with some line search techniques. Variable metric methods are quasi-Newton methods that satisfy the secant condition and positive definiteness. They include the well-known BFGS and DFP methods. The BFGS and DFP update formulae are given by

$$H_{k+1}^{\text{BFGS}} = H_k - \frac{H_k y_k s_k^T + s_k (H_k y_k)^T}{s_k^T y_k} + \left(1 + \frac{y_k^T H_k y_k}{s_k^T y_k}\right) \frac{s_k s_k^T}{s_k^T y_k} \tag{3}$$

and

$$H_{k+1}^{DFP} = H_k - \frac{H_k y_k (H_k y_k)^T}{y_k^T H_k y_k} + \frac{s_k s_k^T}{s_k^T y_k}, \tag{4}$$

respectively. It is known that both H_{k+1}^{BFGS} and H_{k+1}^{DFP} are positive definite when $s_k^T y_k > 0$ and H_k is positive definite. Moreover, the update can be calculated within $O(n^2)$ arithmetic operations, whereas the Newton method requires $O(n^3)$ arithmetic operations to solve Newton equations. The method has superlinear convergence under appropriate conditions [5, 17]. Therefore, the method is very efficient for small- and medium-scale problems. For large-scale problems, the Hessian $\nabla^2 f(x_k)$ usually becomes sparse. By exploiting the sparsity, the Newton method with the trust region technique can be implemented with little memory. Thus, the method is applicable for such problems. However, since $s_k s_k^T$ in (3) or (4) is dense, the updated matrix H_{k+1}

(or its inverse B_{k+1}) is also dense, even if the Hessian is sparse. Storing the full matrix H_{k+1} requires $O(n^2)$ memory, and thus the BFGS and DFP methods are not applicable for large-scale problems.

In order to overcome this difficulty, several methods have been proposed [7,16,22]. The limited-memory BFGS (L-BFGS) method [16] is widely used in practice. The L-BFGS method stores a few vector pairs (s^i, y^i), $i = k - m + 1, \ldots, k - 1, k$, and constructs an approximate Hessian by the BFGS method with the vector pairs. The approximate Hessian satisfies the secant condition and becomes positive definite. The time and space complexities per iteration of the L-BFGS method are O(mn), and it is shown that the L-BFGS method converges linearly [15]. However, since the L-BFGS method does not use much information of the Hessian, it converges very slowly for ill-conditioned problems.

Although the present paper focuses on updates preserving the positive definiteness of H_k, there exist several efficient and practical quasi-Newton methods that do not always preserve the positive definiteness [2,11,12,18–20]. These methods have a global convergence property when using the trust region techniques. Among these techniques, the partially separable BFGS method [11] is practically useful and has already been implemented in LANCELOT [3]. The partially separable BFGS method is applied to problems in which the objective functions f are partially separable, i.e., $f(x) = \sum_{i=1}^{l} f_i(x_{C_i})$, where $C_i \subseteq \{1, 2, \ldots, n\}$, $i = 1, \ldots, l$ and x_{C_i} denotes the $|C_i|$-dimensional vector with components x_i, $i \in C_i$. A function with a sparse Hessian is partially separable [11]. Moreover, most objective functions of practical problems in the real world are partially separable. The partially separable BFGS method generates the approximate Hessian B_k^i for each function f_i by the BFGS method, and composes full matrix B_k of B_k^i, $i = 1, \ldots, l$. Since $|C_i|$ is much smaller than n for large-scale problems, B_k^i becomes a small matrix, and so the partially separable BFGS method can be implemented with little memory. Moreover, since B_k is composed of the approximate Hessians of the functions f_i, B_k is closer to the true Hessian than that of the pure BFGS. However, since $y_{C_i}^T s_{C_i}$ is not necessarily positive, even if $y^T s > 0$, B_k^i (and hence B_k) is not always positive definite. The sufficient condition for B_k of the partially separable BFGS method to be positive definite is that all functions f_i are convex. Griewank and Toint [13] proposed a technique whereby B_k is positive definite when f is convex and $\{C_r | r = 1, \ldots, l\}$ is a maximum clique family of a chordal graph (for details regarding the chordal graph, see Sect. 2).

In the present paper, quasi-Newton updates that exploit the sparsity of the Hessian and guarantee positive definiteness, even if f is nonconvex, are proposed. Although Toint [22] and Fletcher [7] previously proposed updates that exploit sparsity, these methods involve the solution of a convex programming problem at each iteration in order to obtain approximate Hessians. Moreover, since these methods require the sparsity and secant conditions simultaneously, the approximate Hessian can be ill-posed when $(s_k)_i = 0$ for some i [21]. The method proposed herein is based on positive definite matrix completion. For a given set $F \subseteq \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$ and a partial matrix $\bar{X}_{ij}$, $(i, j) \in F$, X is said to be a positive definite matrix completion of $\bar{X}_{ij}$, $(i, j) \in F$, or $\bar{X}_{ij}$, $(i, j) \in F$ is said to have a positive definite matrix completion X, if X is an $n \times n$ positive definite matrix and $X_{ij} = \bar{X}_{ij}$,

$\forall(i,j) \in F$. The positive definite matrix completion has been investigated extensively [8,9,14]. Recently, the positive definite matrix completion has been used for the interior point method for solving the sparse semidefinite programming problem [8]. The results reported in [8,14] are as follows: (1) if F and $\bar{X}_{ij}$, $(i,j) \in F$ satisfy some properties related to a chordal graph (for the definition of "chordal", see Sect. 2), then $\bar{X}_{ij}$, $(i,j) \in F$ has a positive definite matrix completion. (2) If X is the maximum-determinant positive definite matrix completion, then $(\bar{X})^{-1}_{ij} = 0$, $(i,j) \in F$. (3) The maximum-determinant positive definite matrix completion is expressed as the products of sparse matrices. Based on these results, new sparse quasi-Newton updates are proposed. The proposed methods first calculate a partial approximate inverse Hessian H^{QN}_{ij}, $(i,j) \in F$, where F is an extension of the sparsity pattern $E = \{(i,j) \mid (\nabla^2 f(x))_{ij} \neq 0 \text{ for some } x \in R^n\}$ of the Hessian, by using the existing quasi-Newton updates, such as the BFGS method (3) and the DFP method (4). A full matrix H_{k+1}, which is the maximum-determinant positive definite matrix completion of H^{QN}_{ij}, $(i,j) \in F$, is then obtained. When the Hessian is sparse, the time and space complexities of the proposed method become much lower than those of the BFGS and DFP methods. Since the updates do not require the sparsity and secant conditions simultaneously, they do not suffer from Sorensen's example [21], i.e., the approximate Hessian does not become ill-posed, even if $(s_k)_i = 0$ for some i. Moreover, the proposed update is shown herein to have local and superlinear convergence under the usual assumptions.

The present paper is organized as follows. In Sect. 2, results regarding positive definite matrix completion are introduced. These results are based primarily on [8,14], and are adapted slightly for the present purpose. In Sect. 3, the sparse quasi-Newton updates with positive definite matrix completion are proposed, and their time and space complexities per iteration are discussed. In Sect. 4, the behavior of the proposed update for Sorensen's example, which indicates that the proposed update is better than existing sparse quasi-Newton updates, is examined. The proposed update with the DFP method, which is a special case of the proposed update, is then shown to have local and superlinear convergence under appropriate conditions in Sect. 5. A number of numerical experiments are presented in Sect. 6, and concluding remarks are presented in Sect. 7.

The following notation is used throughout the present paper. In the present paper, V is denoted by $\{1,2,\ldots,n\}$. For a given set $F \subset V \times V$, $F_i = \{j \in V \mid (i,j) \in F\}$ and $|F|$ denotes the number of elements of F. For an $n \times n$ matrix H, $\|H\|$ denotes the Frobenius norm of H, and $H \succeq 0$ indicates that H is positive definite. For a vector $z \in R^n$ and a set $S \subseteq V$, z_S denotes an $|S|$-dimensional vector with components z_i, $i \in S$. For an $n \times n$ matrix A and sets $S, U \subseteq V$, A_{SU} denotes an $|S| \times |U|$ matrix with components A_{ij}, $(i,j) \in S \times U$.

2 Positive definite matrix completion

In this section, a number of results regarding positive definite matrix completion, which will be used in subsequent sections, will be introduced. Most of these results are found in [8,14].

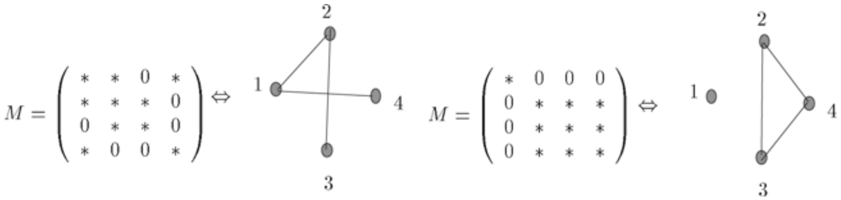
$M = \begin{pmatrix} * & * & 0 & * \\ * & * & * & 0 \\ 0 & * & * & 0 \\ * & 0 & 0 & * \end{pmatrix} \Leftrightarrow$
2
1
4
3
$M = \begin{pmatrix} * & 0 & 0 & 0 \\ 0 & * & * & * \\ 0 & * & * & * \\ 0 & * & * & * \end{pmatrix} \Leftrightarrow$
2
1
4
3

Fig. 1 $F = \{(i, j) \in V \times V \mid M_{ij} \neq 0\}$ and its related graph

Let $F \subseteq V \times V$. Throughout this section, it is assumed that $(i, i) \in F$ for $i \in V$, and $(i, j) \in F$ if $(j, i) \in F$. For a given $\bar{X}_{ij}$, $(i, j) \in F$, $X \in R^{n\times n}$ is said to be a positive matrix completion of $\bar{X}_{ij}$, $(i, j) \in F$ if X is positive definite and $X_{ij} = \bar{X}_{ij}$, $(i, j) \in F$. The problem of finding a positive definite matrix completion of $\bar{X}_{ij}$, $(i, j) \in F$ is usually formulated as a semidefinite programming problem, and thus positive definite matrix completion is not easily calculated. However, if F and $\bar{X}_{ij}$, $(i, j) \in F$ have certain properties, then the positive definite matrix completion can be calculated directly. Such properties are related to a graph $G(V, \bar{F})$ induced from F, where $G(V, \bar{F})$ is a graph having a vertex set V and an edge set $\bar{F} := F\backslash\{(i, i) \mid i = 1, \dots, n\}$ (Fig. 1).

Recall the following concepts of graph theory, which are related to positive definite matrix completion.

Definition 1
- Two vertices $u, v \in V$ are adjacent if $(u, v) \in \bar{F}$. The set of vertices adjacent to $v \in V$ is denoted by $\mathrm{Adj}(v)$.
- A graph is complete if every pair of vertices is adjacent.
- For a subset V' of V, the induced subgraph on V' is a graph $G(V', \bar{F}')$ with the edge set $\bar{F}' = \bar{F} \cap (V' \times V')$.
- A clique of a graph is an induced subgraph that is complete.
- A clique is maximal if its vertices do not constitute a proper subset of another clique.
- A vertex is simplicial if its adjacent vertices induce a clique.
- For a cycle, an edge is a cord of the cycle if it joins two nonconsecutive vertices of the cycle.
- A graph is chordal if every cycle of length greater than 3 has a chord (Fig. 2).

Example 1 Let A be given by

$$A = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

The graph related to A is described in Fig. 3. Since the graph has no cycle, it is chordal, and its simplicial vertices are 2, 3, and 4.

When $G(V, \bar{F})$ is a chordal graph, there exists a family $\{C_r \mid r = 1, \dots, l\}$ of maximal cliques of $G(V, \bar{F})$ such that $F = \cup_{i=1}^{l} C_l \times C_l$ [1]. (The family of maximal cliques of the chordal graph can be computed within $O(n + m)$ by the maximum

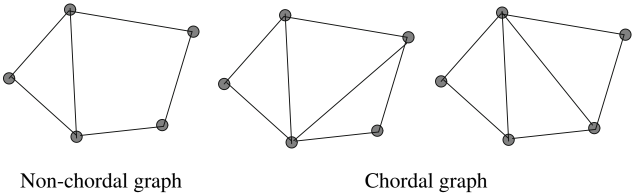

Fig. 2 Chordal graph

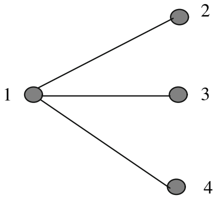

Fig. 3 Graph related to A

cardinality search [1], where m is the number of edges.) One of the necessary conditions for $\bar{X}_{ij}$, $(i, j) \in F$ to have a positive matrix completion is that $\bar{X}_{C_r C_r}$ is positive definite for all $r = 1, \ldots, l$. This condition is referred to as the clique positive definite condition. When $G(V, \bar{F})$ is chordal, it becomes a sufficient condition [14]. Moreover, [14] reported the following properties.

Theorem 1 (a) *$G(V, \bar{F})$ is a chordal graph if and only if $\bar{X}_{ij}$, $(i, j) \in F$ satisfying the clique positive definite condition has a positive definite matrix completion.*

(b) *Suppose that $G(V, \bar{F})$ is a chordal graph, and $\bar{X}_{ij}$, $(i, j) \in F$ satisfies the clique positive definite condition. A maximum-determinant positive definite matrix completion of $\bar{X}_{ij}$, $(i, j) \in F$, i.e., a solution of*

$$\begin{array}{ll} \max & \det(X) \\ \text{subject to} & X_{ij} = \bar{X}_{ij}, \ \forall (i, j) \in F \\ & X = X^T \\ & X \succeq 0 \end{array}$$

is then unique, and $X^{-1}_{ij} = 0$ for all $(i, j) \notin F$.

Example 2 The graph related to the matrix A in Example 1 has the following family $\{C_1, C_2, C_3\}$ of maximal cliques.

$$C_1 = \{1, 2\}, \quad C_2 = \{1, 3\}, \quad C_3 = \{1, 4\}$$

Since

$$A_{C_1C_1} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}, \quad A_{C_2C_2} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad A_{C_3C_3} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

$A_{ij} \in F$ satisfies the clique positive definite condition, where $F = \{(i, j) \in \{1, 2, 3, 4\} \times \{1, 2, 3, 4\} \mid A_{ij} \neq 0\}$. The maximum-determinant positive definite matrix completion $\hat{A}$ of A_{ij}, $(i, j) \in F$ and its inverse are

$$\hat{A} = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 1 & 1/2 & 1/2 \\ 1 & 1/2 & 2 & 1/2 \\ 1 & 1/2 & 1/2 & 1 \end{pmatrix}, \quad \hat{A}^{-1} = \begin{pmatrix} 5/3 & -1 & -1/3 & -1 \\ -1 & 2 & 0 & 0 \\ -1/3 & 0 & 2/3 & 0 \\ -1 & 0 & 0 & 2 \end{pmatrix}.$$

Note that A is not positive definite.

Next, the method by which to compute the maximum-determinant positive definite matrix completion of $\bar{X}_{ij}$, $(i, j) \in F$ is considered.

The family $\{C_r \mid r = 1, \ldots, l\}$ of maximal cliques of the chordal graph $G(V, \bar{F})$ can be indexed in such a way that for each $r = 1, 2, \ldots, l-1$, the following holds:

$$\exists s > r \quad \text{such that } C_r \cap (C_{r+1} \cup C_{r+2} \cdots \cup C_l) \subsetneq C_s.$$

This is called the running intersection property and is easily obtained using the clique tree [1].

Next, it is assumed that $\{C_r \mid r = 1, \ldots, l\}$ are indexed as satisfying the running intersection property. Then, the following families of subsets of $\{C_r\}$ can be defined:

$$S_r = C_r \backslash (C_{r+1} \cup C_{r+2} \cup \cdots \cup C_l), \quad r = 1, \ldots, l \tag{5}$$

$$U_r = C_r \cap (C_{r+1} \cup C_{r+2} \cup \cdots \cup C_l), \quad r = 1, \ldots, l \tag{6}$$

Example 3 The family of maximum cliques in Example 2 satisfies the running intersection property. The corresponding sets $\{S_r\}$ and $\{U_r\}$ are given by $S_1 = \{2\}$, $S_2 = \{3\}$, $S_3 = \{1, 4\}$, $U_1 = \{1\}$, $U_2 = \{1\}$ and $U_3 = \emptyset$.

Using $\{S_r\}$ and $\{U_r\}$ defined by (5) and (6), the maximum-determinant positive definite matrix completion X of $\bar{X}_{ij}$, $(i, j) \in F$ is given as follows [8, Sparse clique-factorization formula (2.16)]:

$$X = P_1^T P_2^T \cdots P_l^T Q P_l P_{l-1} \cdots P_2 P_1, \tag{7}$$

where the factors $\{P_r\}$ and Q are given by

$$[P_r]_{ij} = \begin{cases} 1 & i = j \\ (\bar{X}_{U_r U_r}^{-1} \bar{X}_{U_r S_r})_{ij} & (i, j) \in U_r \times S_r \\ 0 & \text{otherwise} \end{cases}$$

for $r = 1, \ldots, l-1$, and

$$Q_{ij} = \begin{cases} (Q_r)_{i,j} & (i, j) \in S_r \times S_r, r = 1, \ldots, l \\ 0 & \text{otherwise} \end{cases}$$

with

$$Q_r = \begin{cases} \bar{X}_{S_r S_r} - \bar{X}_{S_r U_r} \bar{X}_{U_r U_r}^{-1} \bar{X}_{U_r S_r} & r \leq l-1 \\ \bar{X}_{S_r S_r} & r = l. \end{cases}$$

Remark 1 The vertices can be indexed as $v_i > v_j$ for $v_i \in S_r$, $v_j \in S_{r'}$ with $r > r'$. For the index $\{v_1, v_2, \ldots, v_n\}$ $P_r, r = 1, \ldots, l$ are lower triangle matrices and Q is a block diagonal matrix.

Example 4 Let A, F, $\{C_r\}$, $\{S_r\}$ and $\{U_r\}$ be given in Examples 1–3. Then, we have

$$P_1 = \begin{pmatrix} 1 & 1/2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad P_2 = \begin{pmatrix} 1 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad P_3 = I, \quad Q = \begin{pmatrix} 2 & 0 & 0 & 1 \\ 0 & 1/2 & 0 & 0 \\ 0 & 0 & 3/2 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

and

$$P_1^T P_2^T P_3^T Q P_3 P_2 P_1 = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 1 & 1/2 & 1/2 \\ 1 & 1/2 & 2 & 1/2 \\ 1 & 1/2 & 1/2 & 1 \end{pmatrix}.$$

3 Sparse quasi-Newton updates with positive definite matrix completion

In this section, new sparse quasi-Newton updates are proposed.

Fletcher [6] showed that H_{k+1}^{DFP} is a unique solution of the following problem:

$$\begin{array}{ll} \min_H & \psi(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}}) \\ \text{subject to} & H y_k = s_k, H = H^T \\ & H \succeq 0, \end{array} \tag{8}$$

where $\psi : R^{n \times n} \to R$ is a strictly convex function defined by

$$\psi(A) = \text{trace}(A) - \ln \det(A). \tag{9}$$

When A is symmetric positive definite and its eigenvalues are $\lambda_i, i = 1, \ldots, n$, we have $\psi(A) = \sum_{i=1}^n (\lambda_i - \ln \lambda_i)$. Therefore, the minimum of ψ on $A \succeq 0$ occurs at $\lambda_i = 1, i = 1, \ldots, n$. This implies that $\psi(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}})$ denotes a kind of distance from H_k to H, and thus the solution H_{k+1} of (8) is the “nearest” positive semidefinite matrix satisfying the secant condition from H_k. On the other hand, B_{k+1}^{BFGS}, the inverse of H_{k+1}^{BFGS}, is a solution of the following problem [6]:

$$\begin{array}{ll} \min_B & \psi(H_k^{\frac{1}{2}} B H_k^{\frac{1}{2}}) \\ \text{subject to} & Bs_k = y_k,\ B = B^T \\ & B \succeq 0. \end{array} \tag{10}$$

The above problems (8) and (10) do not include the information of the sparsity of the Hessian. Taking advantage of this information, a new approximate Hessian may be constructed using less memory. Therefore, rather than (8), the following problem is considered:

$$\begin{array}{ll} \min_H & \psi(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}}) \\ \text{subject to} & Hy_k = s_k,\ H = H^T \\ & (H^{-1})_{ij} = 0,\ (i,j) \notin F \\ & H \succeq 0, \end{array} \tag{11}$$

where $F \supseteq E = \{(i,j) \mid \nabla^2 f(x)_{i,j} \neq 0 \text{ for some } x \in R^n\}$. Here, E is referred to as the sparsity pattern of the Hessian and F is referred to as an extension of E. (Of course, it is desirable to choose $F = E$, but certain properties of F are required, as will be discussed later). Throughout the present paper, it is assumed that $(i,i) \in F$ for all $i \in V$ and that $(i,j) \in F$ if $(j,i) \in F$. Fletcher [7] considered the problem (10) with the sparsity conditions $B_{ij} = 0, (i,j) \notin F$, and proposed the use of its exact solution as B_{k+1}. Since the problem is a nonlinear convex programming problem, a great deal of time is required in order to obtain the exact solution. Moreover, as shown in Sect. 4, B_{k+1} sometimes becomes unstable due to the simultaneous requirement of the sparsity and secant conditions [21]. In the present paper, the use of an approximate solution of (11) as H_{k+1} is considered rather than the exact solution. More precisely, the following new updates are proposed:

Step 1: Obtain a partial matrix H_{ij}^{QN}, $(i,j) \in F$ using existing quasi-Newton updates, such as the BFGS and DFP methods.

Step 2: Obtain a solution H_{k+1} of the following problem with H_{ij}^{QN}, $(i,j) \in F$ as given constants.

$$\begin{array}{ll} \min & \psi(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}}) \\ \text{subject to} & H_{ij} = H_{i,j}^{QN},\ (i,j) \in F \\ & H = H^T \\ & (H^{-1})_{ij} = 0,\ (i,j) \notin F \\ & H \succeq 0 \end{array} \tag{12}$$

Remark 2 If the DFP method is used in Step 1, then H^{QN} is a solution of problem (11) without the sparsity constraints, i.e., problem (8).

Remark 3 The secant condition $Hy_k = s_k$ in problem (11) is replaced with the constraints $H_{ij} = H_{i,j}^{QN}$, $(i,j) \in F$ in problem (12). Therefore, as shown in Sect. 4, H_{k+1} is stable even if $(s_k)_i = 0$ for some i.

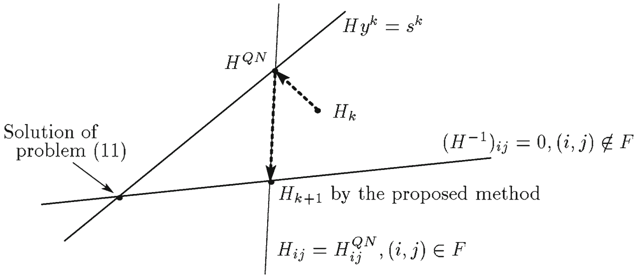
$Hy^k = s^k$
H^{QN}
H_k
Solution of problem (11)
$(H^{-1})_{ij} = 0, (i,j) \notin F$
H_{k+1} by the proposed method
$H_{ij} = H_{ij}^{QN}, (i,j) \in F$

Fig. 4 Proposed method

Remark 4 When $F = V \times V$, the proposed updates are reduced to the existing quasi-Newton updates used in Step 1.

The proposed update is illustrated in Fig. 4.

From the above remarks and Fig. 4, the updated matrix H_{k+1} is regarded as a kind of approximate solution of (11). However, problem (12) still appears to be difficult. Fortunately, as shown below, if $G(V, \bar{F})$ is chordal, then problem (12) is equivalent to finding a maximum-determinant positive definite matrix completion of H_{ij}^{QN}, $(i, j) \in F$, i.e.,

$$\begin{array}{ll} \max & \det(H) \\ \text{subject to} & H_{ij} = H_{i,j}^{QN}, (i, j) \in F \\ & H = H^T \\ & H \succeq 0. \end{array} \tag{13}$$

Theorem 2 *Suppose that $s_k^T y_k > 0$, H_k is symmetric positive definite, and $(H_k^{-1})_{ij} = 0, \forall (i, j) \notin F$. If $G(V, \bar{F})$ is a chordal graph, then problem* (12) *is equivalent to problem* (13).

Proof First, problem (12) is shown to be equivalent to

$$\begin{array}{ll} \max & \det(H) \\ \text{subject to} & H_{ij} = H_{i,j}^{QN}, (i, j) \in F \\ & H = H^T \\ & (H^{-1})_{ij} = 0, (i, j) \notin F \\ & H \succeq 0. \end{array} \tag{14}$$

Since $(H_k^{-1})_{ij} = 0, \forall (i, j) \notin F$ and $H_{ij} = H_{i,j}^{QN}$, $(i, j) \in F$ in the constraint of (12), we have

$$\begin{aligned} \operatorname{trace}(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}}) = \operatorname{trace}(H H_k^{-1}) &= \sum_{i=1}^{n} \sum_{j=1}^{n} H_{ij} (H_k^{-1})_{ji} \\ &= \sum_{i=1}^{n} \sum_{j \in F_i} H_{ij} (H_k^{-1})_{ij} = \sum_{i=1}^{n} \sum_{j \in F_i} H_{ij}^{QN} (H_k^{-1})_{ij}, \end{aligned}$$

which shows that $\mathrm{trace}(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}})$ is constant on the feasible set of (12). Moreover, we have

$$\ln\det(H_k^{-\frac{1}{2}} H H_k^{-\frac{1}{2}}) = 2\ln\det(H_k^{-\frac{1}{2}}) + \ln\det(H).$$

Therefore, problem (12) is equivalent to problem (14).

Next, problem (14) is shown to be equivalent to problem (13). Suppose that $\{C_r \mid r = 1, \ldots, r\}$ is a family of maximal cliques of $G(V, \bar{F})$. Since $s_k^T y_k > 0$ and H_k is positive definite, H^{QN} is also positive definite. Therefore, the submatrices $H^{QN}_{C_r C_r}, r = 1, \ldots, l$ are positive definite, i.e., H^{QN}_{ij}, $(i, j) \in F$ satisfies the clique positive definite condition. It then follows from Theorem 1(b) that a unique solution of (13) satisfies $(H^{-1})_{ij} = 0, (i, j) \notin F$, i.e., it is also a solution of (14). □

From this theorem, if $G(V, \bar{F})$ is a chordal graph, then the solution H_{k+1} of the problem (12) is given by the sparse clique-factorization formula (7).

Remark 5 Fletcher [7] showed that the problem (10) with the sparsity conditions $B_{ij} = 0, (i, j) \notin F$ can be efficiently solved by the Newton method if a factorization of B_k has no fill-in, which implies that $G(V, \bar{F})$ is chordal.

Next, the proposed update is described.

Matrix completion quasi-Newton (MCQN) update

Step 0 Obtain an extension F of E such that $G(V, \bar{F})$ is chordal. Calculate a family $\{C_r \mid r = 1, \ldots, l\}$ of maximum cliques of $G(V, \bar{F})$, $\{S_r \mid r = 1, \ldots, l\}$ and $\{U_r \mid r = 1, \ldots, l\}$ by (5) and (6). Choose $x_0 \in R^n$ and a positive definite matrix H_0 with $(H_0^{-1})_{ij} = 0, \forall (i, j) \notin F$. Set $k = 0$.

Step 1 If x_k satisfies the termination criterion, then stop.

Step 2 $x_{k+1} = x_k - H_k \nabla f(x_k)$.

Step 3 Obtain H^{QN}_{ij}, $(i, j) \in F$ by the existing quasi-Newton updates.

Step 4 Obtain the sparse clique-factorization formula (7) of H_{k+1} with $\bar{X}_{ij} = H^{QN}_{ij}$, $(i, j) \in F$.

Step 5 Set $k := k + 1$ and go to Step 1.

Remark 6 The proposed update is related to the partially separable BFGS method [11]. Consider the case in which $f(x) = \sum_{r=1}^{l} f_r(x_{C_r})$ and $C_i \cup C_j = \emptyset$ for all $i, j \in \{1, \ldots, l\}$ and $i \neq j$. Then, the Hessian $\nabla^2 f(x)$ forms a block diagonal. The partially separable BFGS method updates each block B_k^i of the approximate Hessian B_k as

$$B_{k+1}^i = B_k^i - \frac{B_k^i s_{C_r} s_{C_r}^T B_k^i}{s_{C_r}^T y_{C_r}} + \frac{y_{C_r} y_{C_r}^T}{s_{C_r}^T y_{C_r}},$$

whereas the MCQN update with the BFGS method updates

$$B_{k+1}^i = B_k^i - \frac{B_k^i s_{C_r} s_{C_r}^T B_k^i}{s^T y} + \frac{y_{C_r} y_{C_r}^T}{s^T y}.$$

The differences between these methods are the denominators of the second and the third terms. When $f_r, r = 1, \dots, l$ are convex, then $s_{C_r}^T y_{C_r}$ is positive for all $r = 1, \dots, l$ and $s^T y \geq s_{C_r}^T y_{C_r}$. Therefore, B_{k+1}^i updated by the partially separable BFGS method appears to be closer to the true Hessian than that of the MCQN update, and the MCQN update is regarded as a damped partially separable BFGS method. When f_r is not convex for some r, $s_{C_r}^T y_{C_r}$ is not necessarily positive, and hence B_k of the partially separable BFGS method is not always positive definite.[1] On the other hand, the MCQN update always guarantees the positive definiteness of B_k, even if f is nonconvex.

Remark 7 The MCQN update is not scale invariant in general because a linear transformation $x = Sz$ with a general nonsingular matrix S destroys the sparsity pattern E. When S is positive definite and diagonal, the MCQN update with DFP (or BFGS) method is scale invariant for $x = Sz$. Let S be a positive definite diagonal matrix and $\hat{f}(z) = f(Sz)$. Moreover, let $\hat{H}^{QN}$ and $\hat{H}_k$, respectively, be updated in Steps 3 and 4 of the MCQN update for $\hat{f}$ and z. Since the DFP and BFGS methods are scale invariant, $\hat{H}^{QN} = SH^{QN}S$. The problem (13) for $\hat{f}$ is given by

$$\begin{array}{ll} \max & \det(\hat{H}) \\ \text{s.t.} & \hat{H}_{ij} = S_{ii}S_{jj}H_{ij}^{QN}, \ \ (i, j) \in F \\ & \hat{H} = \hat{H}^T \\ & \hat{H} \succeq 0. \end{array}$$

Let $\bar{H}$ be a matrix such that $S\bar{H}S = \hat{H}$. Then, the above problem is rewritten as

$$\begin{array}{ll} \max & \det(\bar{H}) \\ \text{s.t.} & \bar{H}_{ij} = H_{ij}^{QN}, \ \ (i, j) \in F \\ & \bar{H} = \bar{H}^T \\ & \bar{H} \succeq 0. \end{array}$$

Thus, the solution of the problem is the solution of the original problem (13), and hence $\hat{H}_{k+1} = SH_{k+1}S$.

Next, the time and space complexities per iteration of the MCQN update are estimated. In order to obtain H_{ij}^{QN} in Step 3, the BFGS or DFP update formula may be employed. Let us assume the use of the BFGS method. Step 3 is then calculated as follows:

$$H_{i,j}^{QN} = (H_k)_{i,j} + \rho s_i s_j - \frac{(H_k y_k)_i (s_k)_j + (s_k)_j (H_k y_k)_j}{s_k^T y_k} \quad \forall (i, j) \in F, \tag{15}$$

where

$$\rho = \frac{1}{s_k^T y_k} + \frac{(y_k)^T H_k y_k}{(s_k^T y_k)^2}.$$

[1] Griewank and Toint [13] presented a technique for B_k to be positive definite when f is convex and $\{C_r | r = 1, \dots, l\}$ is a family of maximum cliques of a chordal graph.

First, the time complexity per iteration is estimated. To compute $(H^{QN}_{U_rU_r})^{-1}$ for each r, $O(|C_r|^3)$ arithmetic operations are required. Therefore, the calculation of $H_k v$ for given $v \in R^n$ requires $O(\sum_{i=1}^{l} |C_r|^3)$ arithmetic operations, and thus the time complexity of Step 2 is $O(\sum_{i=1}^{l} |C_r|^3)$. In Step 3, first $H_k y_k$ is calculated, followed by the computation of H^{QN}_{ij}, $(i, j) \in F$. The calculation of $H_k y_k$ is $O(\sum_{i=1}^{l} |C_r|^3)$. Moreover, since $|F| \leq \sum_{r=1}^{l} |C_r|^2$, $O(\sum_{r=1}^{l} |C_r|^2)$ arithmetic operations are required for (15). Consequently, the time complexity of Step 3 is $O(\sum_{r=1}^{l} |C_r|^3)$. Step 4 is a dummy step because the factorization (7) of H^{k+1} is computed whenever $H_k v$ is computed for given v. Consequently, when $\nabla f(x_k)$ is given, the time complexity per iteration of the MCQN update is $O(\sum_{r=1}^{l} |C_r|^3)$. If $((H_k)_{U_rU_r})^{-1}$ is stored for all $r = 1, \ldots, l$, the time complexity can be reduced to $O(\sum_{r=1}^{l} |C_r|^2)$. For clarification, note that

$$(H_{k+1})_{U_rU_r} = H^{QN}_{U_rU_r} = (H_k)_{U_rU_r} + \rho s_{U_r} s_{U_r}^T - \frac{(H_k y_k)_{U_r} s_{U_r}^T + s_{U_r} (H_k y)_{U_r}^T}{s_k^T y_k}.$$

Thus, using the Sherman–Morrison formula, $((H_{k+1})_{U_rU_r})^{-1}$ can be computed from $((H_k)_{U_rU_r})^{-1}$ within $O(|C_r|^2)$ arithmetic operations. By using the stored $((H_k)_{U_rU_r})^{-1}$, the time complexity of the computations of $H_k v$ becomes $O(\sum_{r=1}^{l} |C_r|^2)$.

Next, the space complexity is estimated. When $((H_{k+1})_{U_rU_r})^{-1}$ is not stored for all r, only $(H_k)_{ij}, (ij) \in F$ need be stored. Therefore, the space complexity is $O(|F|)$. When $((H_{k+1})_{U_rU_r})^{-1}$ is stored for each r, the space complexity becomes $O(\sum_{i=1}^{l} |C_r|^2)$.

When the Hessian is sparse, in general, C_r becomes much less than n. Since $l \leq n$, $\sum_{r=1}^{l} |C_r|^2$ is much smaller than n^2. For example, as shown below, when the Hessian is tridiagonal, $l = n$ and $|C_r| = 2$ for all $r = 1, \ldots, n$. Then, the time and space complexities become $O(n)$.

In Step 0 of the proposed update, the chordal extension $G(V, \bar{F})$ of $G(V, \bar{E})$ must be obtained. The problem of finding a minimum chordal extension of a general graph is NP complete. The minimum chordal extension is obtained via the minimum fill-in Cholesky factorization of a positive definite matrix with sparsity pattern E. Therefore, various existing heuristic methods, such as minimum degree ordering and nested dissection ordering, may be employed for the minimum fill-in Cholesky factorization. On the other hand, when the sparsity pattern E has a special structure, the minimum chordal extension $G(V, \bar{F})$ can easily be obtained. The following are practical examples in which F becomes E [8].

Band matrix Suppose that a sparsity pattern E is given by $E = \{(i, j) \in V \times V \mid |i - j| \leq \beta\}$ with a positive integer β. Let

$$C_r = \{i \in V \mid (r-1)\kappa < i \leq \beta + r\kappa\}, \quad r = 1, \ldots, l$$

with a positive integer κ and the smallest positive integer l satisfying $\beta + l\kappa \geq n$ and $F = \cup_{r=1}^{l} C_r \times C_r$. Then, $G(V, \bar{F})$ is chordal and $\{C_r \mid r = 1, \ldots, l\}$ is a family of

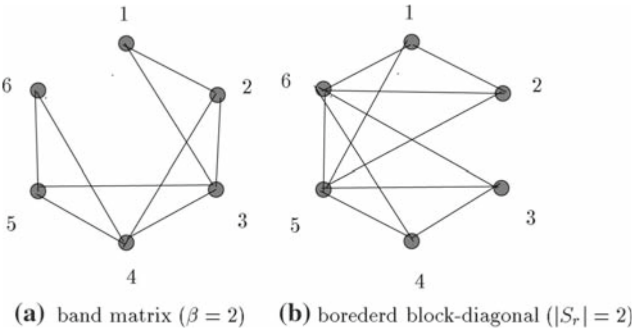
1
2
3
4
5
6
(a) band matrix ($\beta = 2$)
1
2
3
4
5
6
(b) borederd block-diagonal ($|S_r| = 2$)

Fig. 5 Special cases of $G(V, \bar{F})$

its maximum cliques. Figure 5a shows the case for $n = 6$ and $\beta = 2$, and the graph is verified to be chordal.

Note that the integer κ corresponds to $|C_r|$. Moreover, as κ becomes large, l becomes small and $|F|$ becomes large. If $\kappa = 1$, then $l = n - \beta$, $|C_r| = \beta$ and $F = E$.

Now let us consider (7) the case in which the Hessian is tridiagonal, i.e., $\beta = 1$ and $\kappa = 1$. In this case, we have $S_r = \{r\}, r = 1, \ldots, l-1$, $S_l = \{n-1, n\}$, $U_r = \{r+1\}, r = 1, \ldots, l-1$ and $U_l = \emptyset$. Therefore, $P_l = I$ and $P_r, r = 1, \ldots, l-1$ are given by

$$[P_r]_{ij} = \begin{cases} 1 & i = j \\ H^{QN}_{r+1,r}/H^{QN}_{(r+1),(r+1)} & (i,j) = (r+1, r) \\ 0 & \text{otherwise} \end{cases}$$

and Q_r are given by

$$Q_r = \begin{cases} H^{QN}_{r,r} - (H^{QN}_{r,r+1})^2/H^{QN}_{r+1,r+1} & r \leq l-1 \\ H^{QN}_{S_r S_r} & r = l. \end{cases}$$

Therefore, P_r and Q_r can be computed with $O(1)$ arithmetic operations, and the space complexity is $O(1)$. For given v, Hv can be computed with $O(n)$ arithmetic operations.

Bordered block-diagonal Consider the case in which the Hessian has the following form:

$$\begin{pmatrix} B_{S_1S_1} & 0 & \cdots & 0 & B_{S_1S_0} \\ 0 & B_{S_2S_2} & \cdots & 0 & B_{S_2S_0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & B_{S_lS_l} & B_{S_lS_0} \\ B_{S_0S_1} & B_{S_0S_2} & \cdots & B_{S_0S_l} & B_{S_0S_0}, \end{pmatrix}.$$

Let $C_r = S_0 \cup S_r$. Then, $E = F = \cup_{r=1}^{l} C_r \times C_r$ and $G(\bar{F}, V)$ is a chordal graph. Next, suppose that n is even and $|S_r| = 2, r = 0, \ldots, l$. Then, we have $l = n/2 - 1$, $S_0 = \{n-1, n\}$, $S_r = \{2r-1, 2r\}$ and $U_r = S_0, r = 1, \ldots, l$ (the case in which $n = 6$ is illustrated in Fig. 5b). Therefore, $H^{QN}_{U_r U_r}$ becomes a 2×2 matrix for each r, and thus P_r and Q_r can be calculated within $O(1)$ arithmetic operations. Consequently, the time complexity per iteration becomes $O(n)$.

4 Behavior of the MCQN update on Sorensen's example

In this section, the behavior of the proposed update on the following Sorensen's example [21] is shown.

$$f(x) = \frac{1}{8}(x_1 - 1)^2(x_1 + 1)^2 x_3^2 + x_2^2 + (x_2 - x_3)^2 \tag{16}$$

with $x_0 = (0, 0, \sqrt{432/55} - \varepsilon)^T$, $x_1 = (-5/6, 1, \sqrt{432/55})^T$, $\varepsilon = 10^{-6}$. As shown in [21, p. 149], if the secant condition $B_1 s_0 = y_0$ is imposed, then

$$(B_1)_{13} = \frac{1 + 5(B_1)_{11}/6}{\varepsilon},$$

which leads to numerical difficulty. This observation applies to most existing sparse quasi-Newton updates.

The Hessian of f has the following form:

$$\begin{pmatrix} * & 0 & * \\ 0 & * & * \\ * & * & * \end{pmatrix}$$

Therefore, its sparsity pattern E is bordered diagonal, and thus $G(V, \bar{E})$ is chordal and its maximum cliques are $C_1 = \{1, 3\}$ and $C_2 = \{2, 3\}$. When B_0 is the identity matrix, the new matrix B_1 updated by the MCQN update with the BFGS method becomes

$$B_1 = \begin{pmatrix} 0.3421 & 0 & 0.2373 \\ 0 & 2.0629 & -1.7167 \\ 0.2373 & -1.7167 & 2.5931 \end{pmatrix}$$

This shows that the proposed method does not suffer from Sorensen's problem.

Next, the behavior of the MCQN update with the BFGS method on the problem is shown in Table 1.

After nine iterations, the method obtain an approximate stationary point of f. Moreover, even if the true Hessians are singular (see $k = 7, 8, 9$), the approximate Hessians B_k are still positive definite and stable.

Table 1 Behavior of MCQN for Sorensen's example

Iteration	B_k	$\nabla^2 f(x_k)$	$\|\nabla f(x_k)\|$
$k=1$	$\begin{pmatrix} 0.3421 & 0 & 0.2373 \\ 0 & 2.0629 & -1.7167 \\ 0.2373 & -1.7167 & 2.5931 \end{pmatrix}$	$\begin{pmatrix} 8.5091 & 0 & 0.7136 \\ 0 & 4 & -2 \\ 0.7136 & -2 & 2.0233 \end{pmatrix}$	4.13
$k=2$	$\begin{pmatrix} 3.6201 & 0 & -3.4288 \\ 0 & 1.5396 & -0.8149 \\ -3.4288 & -0.8149 & 3.5658 \end{pmatrix}$	$\begin{pmatrix} 26.3805 & 0 & -17.7974 \\ 0 & 4 & -2 \\ -17.7974 & -2 & 10.5672 \end{pmatrix}$	8.41
$k=3$	$\begin{pmatrix} 4.2031 & 0 & -1.3889 \\ 0 & 2.3174 & 0.6037 \\ -1.3889 & 0.6037 & 10.4431 \end{pmatrix}$	$\begin{pmatrix} 9.9069 & 0 & -16.6557 \\ 0 & 4 & -2 \\ -16.6557 & -2 & 22.3021 \end{pmatrix}$	6.07
$k=4$	$\begin{pmatrix} 2.0461 & 0 & 0.6472 \\ 0 & 2.3485 & 0.1128 \\ 0.6472 & 0.1128 & 10.4969 \end{pmatrix}$	$\begin{pmatrix} 3.3856 & 0 & -2.4441 \\ 0 & 4 & -2 \\ -2.4441 & -2 & 3.1845 \end{pmatrix}$	4.21
$k=5$	$\begin{pmatrix} 2.0584 & 0 & 1.3355 \\ 0 & 2.3740 & 0.3049 \\ 1.3355 & 0.3049 & 9.4283 \end{pmatrix}$	$\begin{pmatrix} -0.0516 & 0 & 0.0531 \\ 0 & 4 & -2 \\ 0.0532 & -2 & 2.2249 \end{pmatrix}$	1.51
$k=6$	$\begin{pmatrix} 2.1656 & 0 & 1.1738 \\ 0 & 2.1526 & 1.0876 \\ 1.1738 & 1.0876 & 8.2456 \end{pmatrix}$	$\begin{pmatrix} -0.003 & 0 & 0.0001 \\ 0 & 4 & -2 \\ 0.0001 & -2 & 2.25 \end{pmatrix}$	3.70E−1
$k=7$	$\begin{pmatrix} 2.1462 & 0 & 1.0924 \\ 0 & 2.0614 & 1.1157 \\ 1.0924 & 1.1157 & 8.2908 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 2.25 \end{pmatrix}$	1.58E−2
$k=8$	$\begin{pmatrix} 2.0837 & 0 & 1.0747 \\ 0 & 1.5396 & 1.0563 \\ 1.0747 & 1.0563 & 8.2706 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 2.25 \end{pmatrix}$	7.23E−4
$k=9$	$\begin{pmatrix} 2.0231 & 0 & 1.0629 \\ 0 & 2.0294 & 1.0483 \\ 1.0629 & 1.0483 & 8.2687 \end{pmatrix}$	$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 2.25 \end{pmatrix}$	2.34E−5

5 Local and superlinear convergence of the MCQN update with the DFP method

In this section, the MCQN update with the DFP method in Step 3 is shown to have local and superlinear convergence.

This is proven in a manner similar to [17, 8.4 Convergence Analysis], where the superlinear convergence of the BFGS method is demonstrated using the following property of the function ψ defined by (9):

$$0 < \psi(B_{k+1}^{\mathrm{BFGS}}) \leq \psi(B_k) + \frac{y_k^T y_k}{y_k^T s_k} - \frac{\|B_k s_k\|^2}{s_k^T B_k s_k} - \ln \frac{y_k^T s_k}{\|s_k\|^2} + \ln \frac{s_k B_k s_k}{\|s_k\|^2}. \qquad (17)$$

Here, $B_k = H_k^{-1}$ and $B_{k+1}^{\mathrm{BFGS}} = (H_{k+1}^{\mathrm{BFGS}})^{-1}$. Since the MCQN update generates H_k and (17) is the inequality for B_k, the proof technique cannot be applied directly to show the superlinear convergence of the MCQN update. Moreover, since H_{k+1} is the maximum-determinant positive definite matrix completion of H_{ij}^{QN}, $(i, j) \in F$, we have $\det(H_{k+1}) \geq \det(H^{QN})$, and thus $\det(B_{k+1}) \leq \det(B^{QN})$, where $B^{QN} = (H^{QN})^{-1}$. Therefore, when the MCQN update with the BFGS method is considered in Step 3, i.e., $B^{QN} = B_{k+1}^{\mathrm{BFGS}}$, it is difficult to derive inequalities like (17) due to the definition of ψ. Taking these difficulties into account, the MCQN update with the DFP method is considered because the update formula (4) of the DFP method has a form similar to that of B_{k+1}^{BFGS}. An inequality similar to (17) will be derived for H_{k+1} updated by the MCQN update with the DFP method.

For the present purposes, the following assumptions are necessary:

Assumption 1 Let x_* be a solution of (1), and let $\mathcal{C} = \{x \in R^n \mid \|x - x_*\| \leq b\}$ with a positive constant b.

(i) The objective function f is twice continuously differentiable on $\mathcal{C}$.
(ii) There exist positive constants m and M such that

$$m\|z\|^2 \leq z^T(\nabla^2 f(x))^{-1} z \leq M\|z\|^2 \quad \forall z \in R^n$$

for all $x \in \mathcal{C}$.

If the second-order sufficient optimality condition holds at the solution x_* and b is sufficiently small, then Assumption 1(ii) holds. From Assumption 1(i), $\nabla^2 f(x)$ is Lipschitz continuous on $\mathcal{C}$. Then, from Lemmas 4.1.12 and 4.1.15 in [5], there exist L_1 and L_2 such that for all $x_k, x_{k+1} \in \mathcal{C}$

$$\|y_k - \nabla^2 f(x_*) s_k\| \leq L_1 \|s_k\|^2 \tag{18}$$

and

$$\|y_k - \nabla^2 f(x_*) s_k\| \leq L_2 \varepsilon_k \|s_k\|, \tag{19}$$

where ε_k is defined by

$$\varepsilon_k = \max\{\|x_{k+1} - x_*\|, \|x_k - x_*\|\}. \tag{20}$$

Moreover, there exists a positive constant L_3 such that for all $z_1, z_2 \in \mathcal{C}$

$$\|\nabla f(z_1) - \nabla f(z_2)\| \leq L_3 \|z_1 - z_2\|. \tag{21}$$

Therefore, we have

$$\|y_k\| = \|\nabla f(x_{k+1}) - \nabla f(x_k)\| \leq L_3 \|s_k\| \quad \text{for all } x_k, x_{k+1} \in \mathcal{C}. \tag{22}$$

From Eq. (8.12) of [17] we have

$$y_k = \bar{G}_k s_k, \tag{23}$$

where $\bar{G}_k$ is the average Hessian defined by $\bar{G}_k = \int_0^1 \nabla^2 f(x_k + t s_k) dt$.

For convenience in the present analysis, the following notations are used. Similar notations are used in [17]:

$$G_* = \nabla^2 f(x_*), \quad H_* = \nabla^2 f(x_*)^{-1},$$

$$\tilde{s}_k = H_*^{-1/2} s_k, \quad \tilde{y}_k = H_*^{1/2} y_k, \quad \tilde{H}_k = H_*^{-1/2} H_k H_*^{-1/2}, \quad \tilde{H}^{QN} = H_*^{-1/2} H^{QN} H_*^{-1/2},$$

$$\cos \tilde{\theta}_k = \frac{\tilde{y}_k^T \tilde{H}_k \tilde{y}_k}{\|\tilde{y}_k\| \|\tilde{H}_k \tilde{y}_k\|}, \quad \tilde{q}_k = \frac{\tilde{y}_k^T \tilde{H}_k \tilde{y}_k}{\|\tilde{y}_k\|^2},$$

$$\tilde{M}_k = \frac{\|\tilde{s}_k\|^2}{\tilde{y}_k^T \tilde{s}_k}, \quad \tilde{m}_k = \frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{y}_k^T \tilde{y}_k}.$$

Here, $\tilde{\theta}_k$ is the angle between $\tilde{y}_k$ and $\tilde{H}_k \tilde{y}_k$.

Frequent use will be made of the following inequality in the present analysis:

$$h(t) := t - \ln t - 1 \geq 0 \quad \forall t > 0. \tag{24}$$

The inequality can be shown by the fact that h is strictly convex on $t > 0$, and its minimum is attained at $t = 1$.

First, the following two basic lemmas are given:

Lemma 1 *Suppose that Assumption 1 holds. Then, there exists $c \in (0, \infty)$ and $\gamma \in (0, b)$ such that*

$$\ln \tilde{m}_k \geq -2c\varepsilon_k$$

$$\tilde{M}_k \leq 1 + c\varepsilon_k$$

whenever $\varepsilon_k < \gamma$.

Proof Note that $x_k, x_{k+1} \in C$ when $\varepsilon_k < \gamma$. Since

$$y_k - G_* s_k = (\bar{G}_k - G_*) s_k$$

from (23), we have

$$\begin{aligned} \tilde{y}_k - \tilde{s}_k &= G_*^{-1/2}(y_k - G_* s_k) \\ &= G_*^{-1/2}(\bar{G}_k - G_*) s_k \\ &= G_*^{-1/2}(\bar{G}_k - G_*) G_*^{-1/2} \tilde{s}_k. \end{aligned}$$

Thus, there exists a positive constant $\bar{c}$ such that

$$\|\tilde{y}_k - \tilde{s}_k\| \le \|G_*^{-1/2}\|^2 \|\tilde{s}_k\| \|\bar{G}_k - G_*\| \le \bar{c}\|\tilde{s}_k\|\varepsilon_k, \tag{25}$$

where the first inequality follows from the Cauchy–Schwartz inequality and the second inequality follows from the Lipschitz continuity of $\nabla^2 f$. It follows from the triangle inequality that

$$\|\tilde{y}_k\| - \|\tilde{s}_k\| \le \|\tilde{y}_k - \tilde{s}_k\| \le c\|\tilde{s}_k\|\varepsilon_k$$

and

$$-\|\tilde{y}_k\| + \|\tilde{s}_k\| \le \|\tilde{y}_k - \tilde{s}_k\| \le c\|\tilde{s}_k\|\varepsilon_k,$$

which can be rewritten as

$$(1 - \bar{c}\varepsilon_k)\|\tilde{s}_k\| \le \|\tilde{y}_k\| \le (1 + \bar{c}\varepsilon_k)\|\tilde{s}_k\|. \tag{26}$$

Suppose that γ is sufficiently small. Then, it may be assumed that $1 - \bar{c}\varepsilon_k > 0$. Moreover, squaring both sides of (25), we have

$$\begin{aligned} 2\tilde{y}_k^T \tilde{s}_k &\ge \|\tilde{y}_k\|^2 + (1 - \bar{c}^2\varepsilon_k^2)\|s_k\|^2 \\ &\ge (1 - \bar{c}^2\varepsilon_k^2)\|s_k\|^2 + (1 - \bar{c}\varepsilon_k)^2\|s_k\|^2 \\ &= 2(1 - \bar{c}\varepsilon_k)\|s_k\|^2, \end{aligned} \tag{27}$$

where the second inequality follows from the first inequality of (26). It then follows from the second inequality of (26) that

$$\tilde{m}_k = \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{y}_k\|^2} \ge \frac{(1 - \bar{c}\varepsilon_k)\|s_k\|^2}{(1 + \bar{c}\varepsilon_k)^2\|s_k\|^2} = \frac{(1 + \bar{c}\varepsilon_k)^2 - 3\bar{c}\varepsilon_k - \bar{c}^2\varepsilon_k^2}{(1 + \bar{c}\varepsilon_k)^2} = 1 - \frac{3\bar{c}\varepsilon_k + \bar{c}^2\varepsilon_k^2}{(1 + \bar{c}\varepsilon_k)^2}.$$

Since $\varepsilon_k \le \gamma$, there exists a positive constant c_1 such that

$$\tilde{m}_k \ge 1 - c_1\varepsilon_k. \tag{28}$$

From (24), we have

$$\frac{-c_1\varepsilon_k}{1 - c_1\varepsilon_k} - \ln(1 - c_1\varepsilon_k) = 1 - \frac{1}{1 - c_1\varepsilon_k} + \ln\left(\frac{1}{1 - c_1\varepsilon_k}\right) = -h\left(\frac{1}{1 - c_1\varepsilon_k}\right) \le 0,$$

and thus

$$\frac{-c_1\varepsilon_k}{1 - c_1\varepsilon_k} \le \ln(1 - c_1\varepsilon_k). \tag{29}$$

Since γ is chosen to be sufficiently small, it may be assumed that $c_1\varepsilon_k < \frac{1}{2}$. Thus, from (29), we have

$$\ln(1 - c_1\varepsilon_k) \geq \frac{-c_1\varepsilon_k}{1 - c_1\varepsilon_k} \geq -2c_1\varepsilon_k.$$

It then follows from (28) that

$$\ln \tilde{m}_k \geq \ln(1 - c_1\varepsilon_k) \geq -2c_1\varepsilon_k. \tag{30}$$

From (27), we have

$$\tilde{M}_k = \frac{\|\tilde{s}_k\|^2}{\tilde{y}_k^T \tilde{s}_k} \leq \frac{1}{1 - \bar{c}\varepsilon_k} = \frac{1 - \bar{c}\varepsilon_k + \bar{c}\varepsilon_k}{1 - \bar{c}\varepsilon_k} = 1 + \frac{\bar{c}\varepsilon_k}{1 - \bar{c}\varepsilon_k}.$$

Since $\varepsilon_k \leq \gamma$, there exists a positive constant c_2 such that

$$\tilde{M}_k \leq 1 + c_2\varepsilon_k. \tag{31}$$

Letting $c = \max\{c_1, c_2\}$, we have the desired inequalities from (30) and (31). □

Lemma 2 *Assuming that Assumption 1 holds and $H^{QN} = H_{k+1}^{DFP}$. Then we have*

$$\psi(\tilde{H}_{k+1}) \leq \psi(\tilde{H}^{QN}),$$

where ψ is defined by (9).

Proof The determinant term and the trace term of ψ are investigated separately. Since H^{QN} is feasible for problem (13) and H_{k+1} is the unique maximizer of (13), we have $\det(H^{QN}) \leq \det(H_{k+1})$. Moreover, since $H_*^{-1/2}$ is positive definite by Assumption 1, we have

$$\begin{aligned} \det(\tilde{H}^{QN}) &= \det(H_*^{-1/2})\det(H^{QN})\det(H_*^{-1/2}) \\ &\leq \det(H_*^{-1/2})\det(H_{k+1})\det(H_*^{-1/2}) \\ &= \det(\tilde{H}_{k+1}). \end{aligned} \tag{32}$$

Next, $\operatorname{trace}(\tilde{H}^{QN}) = \operatorname{trace}(\tilde{H}_{k+1})$ is shown. Since $H_{ij}^{QN} = (H_{k+1})_{ij}, \forall (i,j) \in F$ and $(G_*)_{ij} = 0, \forall (i,j) \notin F$, we have

$$\begin{aligned} \operatorname{trace}(\tilde{H}^{QN}) &= \operatorname{trace}(H_*^{-1/2} H^{QN} H_*^{-1/2}) = \operatorname{trace}(H^{QN} G_*) \\ &= \sum_{i=1}^n \sum_{j=1}^n H_{ij}^{QN} (G_*)_{ji} = \sum_{i=1}^n \sum_{j=1}^n H_{ij}^{QN} (G_*)_{ij} \\ &= \sum_{i=1}^n \sum_{j \in F_i} H_{ij}^{QN} (G_*)_{ij} = \sum_{i=1}^n \sum_{j \in F_i} (H_{k+1})_{ij} (G_*)_{ij} \\ &= \operatorname{trace}(H_{k+1} G_*) = \operatorname{trace}(\tilde{H}_{k+1}). \end{aligned} \tag{33}$$

Combining (32) and (33), we have the desired inequality. □

By using the above lemmas, the following key inequality, which corresponds to (17), is shown.

Lemma 3 *Suppose that Assumption 1 holds and $H^{QN} = H_{k+1}^{DFP}$. Suppose also that γ is the constant specified in Lemma 1. If $\varepsilon_k \leq \gamma$, then we have*

$$\psi(\tilde{H}_{k+1}) + \ln \frac{1}{\cos^2 \tilde{\theta}_k} - \left[1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln \left(\frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right)\right] \leq \psi(\tilde{H}_k) + 3c\varepsilon_k. \quad (34)$$

Proof By Assumption 1(ii) and (23), we have

$$\frac{y_k^T s_k}{y_k^T y_k} = \frac{y_k^T \bar{H}_k y_k}{y_k^T y_k} \geq m$$

and

$$\frac{y_k^T y_k}{y_k^T s_k} = \frac{z_k^T \bar{H}_k z_k}{z_k^T z_k} \leq M,$$

where $z_k = \bar{H}_k^{1/2} y_k$ and $\bar{H}_k = \bar{G}_k^{-1}$.

Since H^{QN} is obtained from H_k by the DFP formula (4), we have

$$\begin{aligned} \tilde{H}^{QN} &= H_*^{-1/2} H^{QN} H_*^{-1/2} \\ &= H_*^{-1/2} H_k H_*^{-1/2} + H_*^{-1/2} \left(-\frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k} \right) H_*^{-1/2} \\ &= \tilde{H}_k - \frac{\tilde{H}_k H_*^{1/2} y_k y_k^T H_*^{1/2} \tilde{H}_k}{y_k^T H_*^{1/2} H_*^{-1/2} H_k H_*^{-1/2} H_*^{1/2} y_k} + \frac{H_*^{-1/2} s_k s_k^T H_*^{-1/2}}{y_k^T H_*^{1/2} H_*^{-1/2} s_k} \\ &= \tilde{H}_k - \frac{\tilde{H}_k \tilde{y}_k \tilde{y}_k^T \tilde{H}_k}{\tilde{y}_k^T \tilde{H}_k \tilde{y}_k} + \frac{\tilde{s}_k \tilde{s}_k^T}{\tilde{y}_k^T \tilde{s}_k}. \end{aligned} \quad (35)$$

Since $\text{trace}(zz^T) = \|z|^2$ for $z \in R^n$, it then follows from (35) that

$$\text{trace}(\tilde{H}^{QN}) = \text{trace}(\tilde{H}_k) - \frac{\|\tilde{H}_k y_k\|^2}{\tilde{y}_k \tilde{H}_k \tilde{y}_k} + \frac{\|\tilde{s}_k\|^2}{\tilde{y}_k^T \tilde{s}_k}. \quad (36)$$

In a manner similar to Exercise 8.9 in [17], from (35), we have

$$\det(\tilde{H}^{QN}) = \det(\tilde{H}_k) \frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{y}_k \tilde{H}_k \tilde{y}_k}. \quad (37)$$

Moreover, by simple calculations, we have

$$\frac{\tilde{y}_k^T \tilde{s}_k}{\tilde{y}_k \tilde{H}_k \tilde{y}_k} = \frac{\tilde{y}_k^T \tilde{s}_k}{\|\tilde{y}_k\|^2} \frac{\|\tilde{y}_k\|^2}{\tilde{y}_k \tilde{H}_k \tilde{y}_k} = \frac{\tilde{m}_k}{\tilde{q}_k} \tag{38}$$

and

$$\frac{\|\tilde{H}_k y_k\|^2}{\tilde{y}_k \tilde{H}_k \tilde{y}_k} = \frac{\tilde{y}_k \tilde{H}_k \tilde{y}_k}{\|\tilde{y}_k\|^2} \frac{\|\tilde{H}_k y_k\|^2 \|\tilde{y}_k\|^2}{(\tilde{y}_k \tilde{H}_k \tilde{y}_k)^2} = \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}. \tag{39}$$

It then follows from (36) to (39) that

$$\begin{aligned}\psi(\tilde{H}^{QN}) &= \operatorname{trace}(\tilde{H}^{QN}) - \ln\det(\tilde{H}^{QN}) \\ &= \operatorname{trace}(\tilde{H}_k) + \tilde{M}_k - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} - \ln\det(\tilde{H}_k) - \ln \tilde{m}_k + \ln \tilde{q}_k \\ &= \psi(\tilde{H}_k) + \tilde{M}_k - \ln(\tilde{m}_k) - 1 + 1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln\left(\frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right) + \ln\cos^2 \tilde{\theta}_k.\end{aligned}$$

Then, from Lemmas 1 and 2, we have

$$\psi(\tilde{H}_{k+1}) \le \psi(\tilde{H}^{QN}) \le \psi(\tilde{H}_k) + 3c\varepsilon_k + \ln\cos^2 \tilde{\theta}_k + 1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln\left(\frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right),$$

and thus

$$\psi(\tilde{H}_{k+1}) + \ln\frac{1}{\cos^2 \tilde{\theta}_k} - \left[1 - \frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k} + \ln\left(\frac{\tilde{q}_k}{\cos^2 \tilde{\theta}_k}\right)\right] \le \psi(\tilde{H}_k) + 3c\varepsilon_k,$$

which is the desired inequality. □

Using the inequality (34), the local and superlinear convergence will be shown. First, the local convergence is shown. To this end, the following relationship between $\psi(\tilde{H}_k)$ and the distance $\|H_k - H_*\|$ is needed.

Lemma 4 *Suppose that Assumption* 1 *holds. Suppose also that* $H \in R^{n\times n}$ *is symmetric positive definite and* $\tilde{H} = H_*^{-\frac{1}{2}} H H_*^{-\frac{1}{2}}$.

(a) *Let* $\mu_i, i = 1, \dots, n$ *be the eigenvalues of* H. *Then,* $\psi(H) = \sum_{i=1}^n (\mu_i - \ln \mu_i)$ *and* $\psi(H) - n \ge 0$.
(b) *For any* $\rho > 0$, *there exists* δ *such that* $\psi(\tilde{H}) - n < \delta$ *implies* $\|H - H_*\| < \rho$.
(c) *For any* $\delta > 0$, *there exists* ρ *such that* $\|H - H_*\| < \rho$ *implies* $\psi(\tilde{H}) - n < \delta$.

Proof To show (a), note that $\det(H) = \Pi_{i=1}^n \mu_i$ and $\operatorname{trace}(H) = \sum_{i=1}^n \mu_i$. Thus, we have $\psi(H) = \sum_{i=1}^n (\mu_i - \ln \mu_i)$. It then follows from (24) that $\psi(H) - n = \sum_{i=1}^n (\mu_i - \ln \mu_i - 1) \ge 0$.

Next, (b) and (c) are shown. Let $\lambda_i, i = 1, \ldots, n$ be the eigenvalues of $\tilde{H}$. We then have

$$\|\tilde{H} - I\| = \sqrt{\sum_{i=1}^{n} (\lambda_i - 1)^2}. \tag{40}$$

Moreover, since $\|H - H_*\| = \|H_*^{\frac{1}{2}}(\tilde{H} - I)H_*^{\frac{1}{2}}\|$ and H_* is positive definite, there exist positive constants a_1 and a_2 such that

$$a_1\|\tilde{H} - I\| \le \|H - H_*\| \le a_2\|\tilde{H} - I\|.$$

It then follows from (40) that

$$a_1\sqrt{\sum_{i=1}^{n} (\lambda_i - 1)^2} \le \|H - H_*\| \le a_2\sqrt{\sum_{i=1}^{n} (\lambda_i - 1)^2}. \tag{41}$$

On the other hand, from (a), we have

$$0 \le \psi(\tilde{H}) - n = \sum_{i=1}^{n} (\lambda_i - \ln \lambda_i - 1) = \sum_{i=1}^{n} h(\lambda_i). \tag{42}$$

Since h is continuous and strictly convex on $(0, \infty)$ and its minimum attains at 1, we have (c). In order to show (b), let $\mathcal{L}(\alpha) = \{(\lambda_1, \lambda_2, \ldots, \lambda_n) \mid \sum_{i=1}^{n} h(\lambda_i) \le \alpha, \lambda_i > 0, i = 1, \ldots, n\}$. Then, it follows that $\mathcal{L}(\alpha)$ is compact, $\mathcal{L}$ is continuous for all $\alpha > 0$ and $\mathcal{L}(0) = \{(1, 1, \ldots, 1)\}$. Therefore, for any ρ there exist δ such that $\sum_{i=1}^{n} (\lambda_i - 1)^2 \le \rho$ for all $(\lambda_1, \lambda_2 \ldots, \lambda_n) \in \mathcal{L}(\delta)$. From (41), (42) and the definition of $\mathcal{L}$, we then have (b). □

Theorem 3 *Suppose that Assumption 1 holds and $H^{QN} = H_{k+1}^{DFP}$. Then, for any $\alpha \in (0, 1)$, there exists τ_x and τ_H such that $\|x_0 - x_*\| \le \tau_x$ and $\|H_0 - \nabla^2 f(x_*)^{-1}\| \le \tau_H$ imply*

$$\|x_{k+1} - x_*\| \le \alpha\|x_k - x_*\|$$

for all k.

Proof Suppose that $\alpha \in (0, 1)$. The following inequalities will be shown to hold for all k.

$$\|x_{k+1} - x_*\| \le \alpha\|x_k - x_*\| \tag{43}$$

$$\|H_k - \nabla^2 f(x_*)^{-1}\| \le \frac{\alpha}{2L_3}, \tag{44}$$

where L_3 is the Lipschitz constant of ∇f in (22).

First, note that by choosing τ_x to be sufficiently small, we have

$$L_1 M \tau_x < \frac{\alpha}{2}, \quad \tau_x \leq \gamma \tag{45}$$

where L_1, M and γ are the constants specified in (18), Assumption 1(ii) and Lemma 1, respectively. Moreover, by choosing τ_x and τ_H to be sufficiently small, if necessary, from Lemma 4(b) and (c), there exists δ such that

$$\psi(\tilde{H}_0) - n < \frac{\delta}{2}, \tag{46}$$

$$\psi(\tilde{H}) - n < \delta \implies \|H - \nabla^2 f(x_*)^{-1}\| \leq \frac{\alpha}{2L_3}. \tag{47}$$

and

$$\frac{3c\tau_x}{1-\alpha} \leq \frac{\delta}{2}, \tag{48}$$

where H is a symmetric positive definite matrix, $\tilde{H} = H_*^{-\frac{1}{2}} H H_*^{-\frac{1}{2}}$, and c is the constant specified in Lemma 3.

The inequalities (43) and (44) are shown by induction. When $k = 0$, the inequality (44) holds from (46) and (47). Moreover, we have

$$\begin{aligned}
\|x_1 - x_*\| &= \|x_0 - H_0 \nabla f(x_0) - x_*\| \\
&\leq \|x_0 - x_* - \nabla^2 f(x_*)^{-1} \nabla f(x_0)\| \\
&\quad + \|(H_0 - \nabla^2 f(x_*)^{-1})(\nabla f(x_0) - \nabla f(x_*))\| \\
&\leq \|\nabla^2 f(x_*)^{-1}(\nabla f(x_*) - \nabla f(x_0) + \nabla^2 f(x_*)(x_0 - x_*))\| \\
&\quad + L_3 \|H_0 - \nabla^2 f(x_*)^{-1}\| \|x_0 - x_*\| \\
&\leq L_1 \|\nabla^2 f(x_*)^{-1}\| \|x_0 - x_*\|^2 + \frac{\alpha}{2} \|x_0 - x_*\| \\
&\leq \left(L_1 M \tau_x + \frac{\alpha}{2} \right) \|x_0 - x_*\| \\
&\leq \alpha \|x_0 - x_*\|,
\end{aligned}$$

where the second inequality follows from (21), the third inequality follows from (18) and (47), the forth inequality follows from Assumption 1(ii) and $\|x_0 - x_*\| \leq \tau_x$, and the final inequality follows from (45).

Next, (43) and (44) are assumed to hold for $k = 0, 1, \ldots, l$, and the inequalities for $k = l + 1$ are given. Similar to the case in which $k = 0$, we have

$$\begin{aligned}\|x_{l+1}-x_*\| &= \|x_l - H_l\nabla f(x_l) - x_*\| \\ &\le \|x_l - x_* - \nabla^2 f(x_*)^{-1}\nabla f(x_l)\| \\ &\quad + \|(H_l - \nabla^2 f(x_*)^{-1})(\nabla f(x_l) - \nabla f(x_*))\| \\ &\le \|\nabla^2 f(x_*)^{-1}(\nabla f(x_*) - \nabla f(x_l) + \nabla^2 f(x_*)(x_l - x_*))\| \\ &\quad + L_3\|H_l - \nabla^2 f(x_*)^{-1}\|\|x_l - x_*\| \\ &\le L_1\|\nabla^2 f(x_*)^{-1}\|\|x_l - x_*\|^2 + \frac{\alpha}{2}\|x_l - x_*\| \\ &\le \left(L_1 M\|x_l - x_*\| + \frac{\alpha}{2}\right)\|x_l - x_*\| \\ &\le \left(L_1 M(\alpha)^l\tau_x + \frac{\alpha}{2}\right)\|x_l - x_*\| \\ &\le \alpha\|x_l - x_*\|,\end{aligned}$$

where the fifth inequality follows from the fact that $\|x_l - x_*\| \le (\alpha)^l\|x_0 - x_*\|$. This shows (43) for $k = l + 1$. Next, (44) is shown using (34) in Lemma 3. Summing up the inequalities (34) with $k = 0, 1, \ldots, l$, we have

$$\psi(\tilde{H}_{l+1}) + \sum_{k=0}^{l}\left(\ln\frac{1}{\cos^2\tilde{\theta}_k} - \left[1 - \frac{\tilde{q}_k}{\cos^2\tilde{\theta}_k} + \ln\left(\frac{\tilde{q}_k}{\cos^2\tilde{\theta}_k}\right)\right]\right) \le \psi(\tilde{H}_0) + 3c\sum_{k=0}^{l}\varepsilon_k.$$

Since $0 < \cos\tilde{\theta}_k \le 1$ and the term in the square brackets is nonpositive by (24), we have

$$\psi(\tilde{H}_{l+1}) - n \le \psi(\tilde{H}_0) - n + 3c\sum_{k=0}^{l}\varepsilon_k. \tag{49}$$

From (43), we have

$$\varepsilon_k \ = \|x_{k+1} - x_*\| \le (\alpha)^{k+1}\tau_x$$

for $k = 0, \ldots, l$, and thus

$$\sum_{k=0}^{l}\varepsilon_k \le \frac{1-(\alpha)^{l+1}}{1-\alpha}\tau_x \le \frac{\tau_x}{1-\alpha}.$$

It then follows from (49), (46) and (48) that

$$\begin{aligned}\psi(\tilde{H}_{l+1}) - n &\le \psi(\tilde{H}_0) - n + \frac{3c\tau_x}{1-\alpha} \\ &\le \frac{\delta}{2} + \frac{\delta}{2} = \delta.\end{aligned}$$

From (47), we have $\|H_{l+1} - \nabla^2 f(x_*)^{-1}\| \le \frac{\alpha}{2L_3}$, which is (44) for $k = l + 1$. □

Next, the superlinear convergence is shown. The following are the sufficient conditions for the superlinear convergence of quasi-Newton methods [4].

$$\lim_{k\to\infty} \frac{\|(B_k - G_*)s_k\|}{\|s_k\|} = 0. \tag{50}$$

Using (34) and Theorem 3, we will show that

$$\lim_{k\to\infty} \frac{\|(H_k - H_*)y_k\|}{\|y_k\|} = 0. \tag{51}$$

In order to show the superlinear convergence, the following relation between (51) and the superlinear convergence condition (50) is necessary.

Lemma 5 *Suppose that Assumption 1 holds and that $H^{QN} = H_{k+1}^{\mathrm{DFP}}$. Suppose also that $\|x_0 - x_*\| \leq \tau_x$ and $\|H_0 - \nabla^2 f(x_*)\| \leq \tau_H$ with the constants τ_x and τ_H specified in Theorem 3 for sufficiently small $\alpha \in (0, 1)$. Then, (51) implies that (50) holds.*

Proof Let $\lambda_i^k, i = 1, \ldots, n$ be the eigenvalues of H_k. Since the inequality (44) holds for sufficiently small α, it may be assumed that there exists $\lambda_{\min} > 0$ such that $\lambda_i^k \geq \lambda_{\min}$ for all i and k. Moreover, since $y_k = G_* s_k + (\bar{G}_k - G_*)s_k$ from (23), we have

$$\begin{aligned}\|(H_k - H_*)y_k\| &= \|(H_k - H_*)G_* s_k + (H_k - H_*)(\bar{G}_k - G_*)s_k\| \\ &\geq \|H_k(G_* - B_k)s_k\| - \|H_k - H_*\|\|\bar{G}_k - G_*\|\|s_k\| \\ &\geq \lambda_{\min}\|(B_k - G_*)s_k\| - \|H_k - H_*\|\|\bar{G}_k - G_*\|\|s_k\|.\end{aligned}$$

It then follows from (22) that

$$\frac{\|(H_k - H_*)y_k\|}{\|y_k\|} \geq \frac{\lambda_{\min}\|(B_k - G_*)s_k\|}{L_3\|s_k\|} - \frac{\|H_k - H_*\|\|\bar{G}_k - G_*\|}{L_3}.$$

Since $\bar{G}_k = \int_0^1 \nabla^2 f(x_k + ts_k)dt$ and $x_k \to x_*$ by Theorem 3, the second term of the right-hand side of the inequality converges to 0 as $k \to \infty$. Then, it follows from (51) that

$$\lim_{k\to\infty} \frac{\|(B_k - G_*)s_k\|}{\|s_k\|} = 0,$$

which is the desired inequality. □

The main result of this section can now be shown.

Theorem 4 *Suppose that Assumption 1 holds. Suppose also that $\|x_0 - x_*\| \leq \tau_x$ and $\|H_0 - \nabla^2 f(x_*)^{-1}\| \leq \tau_H$ hold for sufficiently small $\tau_x, \tau_H > 0$. The sequence $\{x_k\}$ generated by the MCQN update with the DFP method then converges to x_* superlinearly.*

Proof From Lemma 5 it is sufficient to show (51). Summing the inequalities (34) in Lemma 3, we have

$$\sum_{k=0}^{\infty}\left(\ln\frac{1}{\cos^2\tilde{\theta}_k}-\left[1-\frac{\tilde{q}_k}{\cos^2\tilde{\theta}_k}+\ln\left(\frac{\tilde{q}_k}{\cos^2\tilde{\theta}_k}\right)\right]\right)\leq\psi(\tilde{H}_0)+3c\sum_{k=0}^{\infty}\varepsilon_k<\infty,$$

where the first inequality follows from the fact that $\psi(\tilde{H}_k)>0$ for all k, and the last inequality follows from the local linear convergence of $\{x_k\}$ (Theorem 3). Since $0<\cos\tilde{\theta}_k\leq 1$, $\ln(1/\cos^2\tilde{\theta}_k)$ must be nonnegative. Moreover, the term in square brackets is nonpositive from (24). Therefore, we have

$$\lim_{k\to\infty}\cos\tilde{\theta}_k=1,\quad \lim_{k\to\infty}\tilde{q}_k=1. \tag{52}$$

Furthermore, we have

$$\begin{aligned}\frac{\|H_*^{-1/2}(H_k-H_*)y_k\|^2}{\|H_*^{1/2}y_k\|^2}&=\frac{\|(\tilde{H}_k-I)\tilde{y}_k\|^2}{\|\tilde{y}_k\|^2}\\&=\frac{\|\tilde{H}_k\tilde{y}_k\|^2-2\tilde{y}_k^T\tilde{H}_k\tilde{y}_k+\|\tilde{y}_k\|^2}{\|\tilde{y}_k\|^2}\\&=\frac{\tilde{q}_k^2}{\cos^2\tilde{\theta}_k}-2\tilde{q}_k+1,\end{aligned}$$

where the final equality follows from the fact that

$$\frac{\tilde{q}_k^2}{\cos^2\tilde{\theta}_k}=\frac{\left(\tilde{y}_k^T\tilde{H}_k\tilde{y}_k\right)^2}{\|\tilde{y}_k\|^4}\frac{\|\tilde{y}_k\|^2\|\tilde{H}_k\tilde{y}_k\|^2}{\left(\tilde{y}_k^T\tilde{H}_k\tilde{y}_k\right)^2}=\frac{\|\tilde{H}_k\tilde{y}_k\|^2}{\|\tilde{y}_k\|^2}.$$

It then follows from (52) and the positive definiteness of H_* that we have the desired inequality (51). □

As in the proofs, the superlinear convergence under Assumption 1 and the assumptions that (a) $\{H_k\}$ is uniformly positive definite and (b) $\sum_{k=0}^{\infty}\varepsilon_k<\infty$ can be shown. The assumptions on the initial data, i.e., the assumptions that $\|x_0-x_*\|\leq\tau_x$ and $\|H_0-\nabla^2 f(x_*)^{-1}\|\leq\tau_H$ hold for sufficiently small τ_x and τ_H, are sufficient conditions for (a) and (b).

6 Numerical experiments

In this section, numerical results are reported for the proposed MCQN update, as well as for the BFGS and the L-BFGS methods.

The following problems were solved with the initial points indicated in [7,10]:

Problem 1 (TRIDIA [10]) $f(x) = (x_1 - 1)^2 + \sum_{i=2}^{n} i(x_{i-1} - 2x_i)^2$, $x^0 = (1, \ldots, 1)^T$

Problem 2 (the chained Rosenbrock problem [7]) $f(x) = \sum_{i=1}^{n-1} 100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2$, $x^0 = (-1.2, 1, -1.2, 1, \ldots, -1.2, 1)^T$

Problem 3 (the boundary value problem [7]) $f(x) = \frac{1}{2}x^T Tx - e_n^T x - \frac{1}{(n+1)^2} \sum_{i=1}^{n}(\cos x_i + 2x_i)$, where $e_n = (1, 1, \ldots, 1)^T$ and

$$T = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix}.$$

$x^0 = (1/(n+1), 2/(n+1), \ldots, n/(n+1))^T$.

Problem 1 is a convex quadratic minimization problem, and Problems 2 and 3 are nonconvex and nonlinear. The sparsity patterns of all problems are tridiagonal. Therefore, the chordal extensions of their sparsity pattern can easily be obtained.

The following termination criterion is employed:

$$\|\nabla f(x_k)\| \leq n10^{-5} \quad \text{or} \quad k \geq 50000.$$

The second criterion implies that the method fails to obtain a solution. In order to obtain the step size t_k and set $x_{k+1} = x_k - t_k H_k \nabla f(x_k)$, Wolfe's rule was employed:

$$f(x_k + t_k d_k) - f(x_k) \leq 10^{-4} t_k \nabla f(x_k)^T d_k,$$

$$|\nabla f(x_k + t_k d_k)^T d_k| \leq -0.9 \nabla f(x_k)^T d_k$$

For the L-BFGS method, $m = 5$, which is the number of stored vectors of L-BFGS, was set, and the scaling factor $s_{k-1}^T y_{k-1}/\|y_{k-1}\|^2$ was employed. All of the algorithms were implemented in Matlab 6.1.

Problems of various dimensions, i.e., $n = 10$,100, and 1,000, were solved by the MCQN update with the DFP method and the MCQN update with the BFGS method, the BFGS method, and the L-BFGS method, and a problem with $n = 10{,}000$ was solved by these MCQN updates and the L-BFGS method. (The BFGS method could not be implemented for $n = 10{,}000$.) The results are listed in Table 2. The table lists the total number of iterations, and the symbol "F" denotes that the number is over 50,000.

Table 2 shows that the number of iterations of the MCQN update with the BFGS method is less than those of the other methods. In particular, the MCQN update was superior to the BFGS and the L-BFGS methods for Problem 3. (It should, however, be noted that the partially separable BFGS method converges in a few iterations for Problem 1 for any large n.) On the other hand, for Problems 1 and 2, the L-BFGS method was competitive with the other methods, even if the problem is ill-conditioned.

Table 2 Number of iterations by MCQN, BFGS, and L-BFGS

Problem	n	BFGS	L-BFGS@	MCQN with DFP	MCQN with BFGS
Problem 1	10	15	31	20	29
	100	108	126	167	72
	1000	662	415	1498	192
	10000	–	1191	11626	528
Problem 2	10	78	68	76	60
	100	487	527	665	341
	1000	4525	4979	6574	3207
	10000	–	49580	F	31737
Problem 3	10	15	24	15	15
	100	107	299	49	50
	1000	571	3117	86	54
	10000	–	F	2600	402

Note that although the MCQN update with the DFP method has a nice theoretical convergence property, its numerical performance is not very good.

7 Concluding remarks

In the present paper, a sparse quasi-Newton update was proposed using a positive definite matrix completion. Using the DFP method, the proposed update was shown to have local and superlinear convergence under the usual assumptions. The proposed update requires lower space and time complexities than those for existing variable metric methods that have superlinear convergence. In particular, when the Hessian has a special structure, such as band matrix or bordered block-diagonal, as discussed in Sect. 4, the complexities are drastically decreased. The simple numerical results suggest that the proposed method is very promising.

Only three test problems were solved in the numerical experiments of Sect. 6. Thus, the behaviors of the MCQN update must be investigated for many more problems arisen in practical situation. Moreover, the MCQN update should be compared with not only the L-BFGS and the BFGS methods but also other practical efficient algorithms, such as the partially separable BFGS method.

Acknowledgements I would like to thank Prof. M. Fukushima of Kyoto University, Prof. P. Toint of University of Namur, Prof. K. Murota of University of Tokyo, Mr. Dinh Hoang Tien of Kyoto University and the anonymous referee for their constructive comments and suggestions, which helped to improve the readability of the paper. This study was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan.

References

1. Blair, J.R.S., Peyton, B.: An introduction to chordal graphs and clique trees. In: George, A., Gilbert, J.R., Liu, J.W.H. (eds.) Graph Theory and Sparse Matrix Computation, pp. 1–29. Springer, New York (1993)

2. Coleman, T.F., Garbow, B., Moré, J.J.: Software for estimating sparse Hessian matrices. ACM Trans. Math. Softw. **11**, 363–378 (1985)
3. Conn, A.R., Gould, N.I.M., Toint, Ph.L.: LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A). In: Springer Series in Computational Mathematics, vol. 17. Springer, New York (1992)
4. Dennis, J.E. Jr., Moré, J.J.: A characterization of superlinear convergence and its application to quasi-Newton methods. Math. Comput. **28**(126), 549–560 (1974)
5. Dennis, J.E. Jr., Schnabel, R.B.: Numerical methods for unconstrained optimization and nonlinear equations. Prentice-Hall Inc., New Jersey (1983)
6. Fletcher, R.: A new result for quasi-Newton formulae. SIAM J. Optim. **1**, 18–21 (1991)
7. Fletcher, R.: An optimal positive definite update for sparse Hessian matrices. SIAM J. Optim. **5**, 192–218 (1995)
8. Fukuda, M., Kojima, M., Murota, K., Nakata, K.: Exploiting sparsity in semidefinite programming via matrix completion I: general framework. SIAM J. Optim. **11**, 647–674 (2000)
9. George, A., Liu, J.W.H.: Computer Solution of Large Sparse Positive Definite Systems. Prentice-Hall, Englewood Cliff (1981)
10. Gould, N.I.M., Orban, D., Toint, Ph.L.: CUTEr, a constrained and unconstrained testing environment: revisited. ACM Trans. Math. Softw. **29**, 373–394 (2003)
11. Griewank, A., Toint, Ph.L.: On the unconstrained optimization of partially separable functions. In: Powell, M.J.D. (ed.) Nonlinear Optimization 1981, pp. 301–312. Academic, London (1982)
12. Griewank, A., Toint, Ph.L.: Partitioned variable metric updates for large structured optimization problems. Numer. Math. **39**, 119–137 (1982)
13. Griewank, A., Toint, Ph.L.: On the existence of convex decomposition of partially separable functions. Math. Program. **28**, 25–29 (1984)
14. Grone, R., Johnson, C.R., Sá, E.M., Wolkowicz, H.: Positive definite completions of partial Hermitian matrices. Linear Algebra Appl. **58**, 109–124 (1984)
15. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Math. Program. **45**, 503–528 (1989)
16. Nocedal, J.: Updating quasi-Newton matrices with limited storage. Math. Comput. **35**, 773–782 (1980)
17. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (1999)
18. Powell, M.J.D., Toint, Ph.L.: On the estimation of sparse Hessian matrices. SIAM J. Numer. Anal. **16**, 1060–1074 (1979)
19. Powell, M.J.D., Toint, Ph.L.: The Shanno-Toint procedure for updating sparse symmetric matrices. IMA J. Numer. Anal. **1**, 403–413 (1981)
20. Schnabel, R.B., Toint, Ph.L.: Forcing sparsity by projecting with respect to a non-diagonally weighted Frobenius norm. Math. Program. **25**(1), 125–129 (1983)
21. Sorensen, D.C.: Collinear scaling and sequential estimation in sparse optimization algorithm. Math. Program. Stud. **18**, 135–159 (1982)
22. Toint, P.L.: On sparse and symmetric matrix updating subject to a linear equation. Math. Comput. **31**, 954–961 (1977)