A Diagonal-Secant Update Technique for Sparse Unconstrained Optimization

Guangye Li

CRPC-TR92191
January, 1992

Center for Research on Parallel Computation
Rice University
P.O. Box 1892
Houston, TX 77251-1892

A DIAGONAL-SECANT UPDATE TECHNIQUE FOR SPARSE UNCONSTRAINED OPTIMIZATION

GUANGYE LI *

Abstract. This paper presents a diagonal-secant modification of the successive element correction method, a finite-difference based method, for sparse unconstrained optimization. This new method uses the gradient values more efficiently in forming the approximate Hessian than the successive element correction method. It is shown that the new method has at least the same local convergence rates as the successive element correction method for general problems and that it has better q-convergence and r-convergence rates than the successive element correction method for problems with band structures. The numerical results show that the new method may be competitive with most of the existing methods for some problems.

Keywords. Unconstrained optimization, quasi-Newton method, symmetry, sparsity, Hessian.

AMS(MOS) subject classification. 65K10, 65H10.

Abbreviated Title. Diagonal-secant update.

1. Introduction. This paper is concerned with the unconstrained minimization problem

$$\min_{x \in R^n} f(x) \tag{1.1}$$

where $f : D \subset R^n \to R$ is twice differentiable and the sparsity structure of Hessian $H(x)$ is known. To solve problem (1.1), we consider the following Newton-like method:

$$x^{k+1} = x^k - (B^k)^{-1} g(x^k),\ k = 0, 1, ..., \tag{1.2}$$

where $g(x^k) = \nabla f(x^k)$ and B^k is a symmetric $n \times n$ matrix, which is an approximation to $H(x^k)$ with the same sparsity pattern as the Hessian. The purpose of this paper is to try to find an efficient way to obtain such B^k under the assumption that $g(x)$ can be only computed as a single vector, i.e., element-by-element gradient evaluation subroutines are unavailable.

To reduce the number of gradient evaluations needed for forming the approximate Hessian in a finite-difference method, Powell and Toint [9] extended the idea of the CPR algorithm for solving systems of nonlinear equations, proposed by Curtis, Powell and Reid [2], to the symmetric case, and gave two practical methods: the direct method and the indirect lower triangular substitution method. The direct method is based on a symmetrically consistent partition of the columns of the Hessian, while the indirect method is based on a consistent partition of the columns of the lower triangular part of the Hessian. Coleman and Moré [1] connected the partition problem to a graph coloring problem and gave some partitioning algorithms which can make the number of gradient evaluations optimal or nearly optimal. The direct method can take advantage

* Center for Research on Parallel Computation, Rice University, P. O. Box 1892, Houston, Texas 77251-1892.

of the symmetry to reduce the number of gradient evaluations for some structures. Unfortunately, this is not always true. The symmetric band structure is a typical example. Coleman and Moré [1] show that $2\beta + 1$ differences are required in forming the approximate Hessian with a symmetric band structure by the direct method, where β is the half bandwidth of the Hessian. This means that the direct method requires the same number of gradient evaluations as the CPR method does for symmetric band structure.

The indirect method sometimes, especially for symmetric band structures, may exploit symmetry to a greater extent than the direct method. However, the computation of the approximate Hessian B^k requires a sequence of substitutions which makes the cost of obtaining B^k higher than in the direct method and which may magnify rounding and truncation errors.

To further reduce the number of gradient values needed for forming the approximate Hessian, in a previous work we [4] proposed a successive element correction method, the CM-successive element correction method (CMEC) for systems of nonlinear equations, which is also based on a consistent partition of the columns of the Hessian. The basic idea of the CMEC method is that some of the elements of the approximate Hessian are corrected at each iterative step by using only two gradient values. The author [5] extended the idea of the CMEC method to the symmetric case and proposed a secant modification of the CMEC method by applying Marwil [6] and Toint's [10] sparse symmetric secant (SPSB) update, and the resulting algorithm will be referred to the SCMEC method.

In this paper, we present a diagonal secant modification of the CMEC method, which is called the DSCMEC method. The idea of this modification is that instead of updating all elements of B^k, we only update the diagonal of B^k to make the updated matrix satisfy the secant equation. The computational cost of the diagonal secant update is much less than that in the SCMEC method, and we show that the new method has the same local convergence properties as the SCMEC method for general problems. The main results of this paper is that this new method has better local q-convergence and r-convergence rates for problems with band structures than the CMEC method. Our numerical results show that this new method is promising in practice.

We arrange this paper in the following way: A brief description of the CMEC method and the SCMEC method and some theoretical results for these methods are given in Section 2. The diagonal secant modification of the CMEC method and some local convergence results are given in Section 3. Some numerical results and some comparisons are given in Section 4.

Through out this paper, $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\|\cdot\|$ denotes the l_2 norm of a vector, and $\|\cdot\|_\infty$ denotes the infinity norm of a vector. To specify the sparsity of a given matrix B, we use M to denote the set of index pairs (i,j), where b_{ij} is a structural nonzero element of B, i.e.,

$$M = \{(i,j) : b_{ij} \neq 0\}.$$

2. The CMEC method and the SCMEC method. The CMEC method can be formulated as follows:

ALGORITHM 2.1. *Given $x^0 \in R^n$, and a nonsingular symmetric matrix B^0, which has the same sparsity pattern as the Hessian, do the following:*
At the initial step:

1. *Using Coleman and Moré's graph coloring technique [1] compute a symmetrically consistent partition of the columns of the Hessian which divides the set $\{1, 2, ..., n\}$ into p subsets $c_1, c_2, ..., c_p$.*
2. *Set $l = 0$.*
3. *Solve $B^0 s^0 = -g(x^0)$.*
4. *Choose x^1 by $x^1 = x^0 + s^0$ or by a global strategy.*

At each iteration $k > 0$:

1. *Update B^{k-1} to B^k:*
   a. *Choose a scalar h^k.*
   b. *If $l < p$, then set $l = l + 1$, otherwise set $l = 1$.*
   c. *Set*

$$d^k = \sum_{j \in c_l} h^k e_j.$$

   where e_j is the j-th component of the unit matrix.
   d. *If $j \in c_l$ and $(i, j) \in M$, then set*

$$b_{ij}^k = \frac{1}{h^k} e_i^T (g(x^k + d^k) - g(x^k)),$$

   and set

$$b_{ji}^k = b_{ij}^k.$$

   Otherwise, set

$$b_{ij}^k = b_{ij}^{k-1}.$$

2. *Solve $B^k s^k = -g(x^k)$.*
3. *Choose x^{k+1} by $x^k + s^k$ or by a global strategy.*
4. *Check convergence.*

Note that at step 1(a), we use a uniform step length for all components of x. This is for simplicity in our theoretical discussion. In practice, one should choose different h_j^k for each component of x_j^k. (See Section 4.)

The following results for the CMEC method were given in [5]:

LEMMA 2.1. *Assume $H(x)$ satisfies the following Lipschitz condition: For $(i, j) \in M$ there exists an $\alpha_{ij} > 0$ such that*

$$|e_i^T (H(x) - H(y)) e_j| \leq \alpha_{ij} \|x - y\|, \quad x, y \in D. \tag{2.1}$$

Let x^k and B^k be generated by Algorithm 2.1. Assume $x^k \in D$ and $x^k + d^k \in D$. If b_{ij}^{k-1} is corrected at the kth iterative step, then

$$|e_i^T(B^k - H(x^k))e_j| \leq \frac{\sqrt{n}}{2}\alpha_{ij}|h^k|.$$

LEMMA 2.2. *Assume $H(x)$ satisfies Lipschitz condition (2.1). Let $\{x^j\}_{j=1}^k \subset D$ and $\{B^j\}_{j=0}^k$ be generated by Algorithm 2.1 with B^0 satisfying $\|B^0 - H(x^0)\|_F \leq \delta$. If $\{x^j + d^j\}_{j=1}^k \subset D$, then for $k \geq p$,*

$$|e_i^T(B^k - H(x^k))e_j| \leq \alpha_{ij}(\bar{e}_k + \bar{h}_k),$$

for any $(i,j) \in M$. Moreover,

$$\|B^k - H(x^k)\|_F \leq \alpha(\bar{e}_k + \bar{h}_k),$$

and for $k \geq p$,

$$\|B^k - H(x^k)\|_F \leq \alpha(2\bar{e}_k + \bar{h}_k) + \delta$$

where

$$\alpha = (\sum_{(i,j)\in M} \alpha_{ij}^2)^{\frac{1}{2}},\ \bar{e}_k = \max_{1\leq j\leq m(k)}\{\|x^k - x^{k-j}\|\},\ \bar{h}_k = \frac{\sqrt{n}}{2}\max_{0\leq j\leq m(k)} h^{k-j},$$

with $m(k) = min\{k, p-1\}$ and $h^0 = 0$.

THEOREM 2.3. *Assume that $g : D \subset R^n \to R^n$ satisfies the standard assumption for local convergence, i.e.,*

(2.2) *There exists an $x^* \in D$ such that $g(x^*) = 0$ and $H(x^*)$ is nonsingular.*

Also assume that H satisfies Lipschitz condition (2.1). Let $\{x^k\}$ be generated by Algorithm 2.1 without any global strategy. Then there exist ϵ, δ, $h > 0$ such that if $0 < |h^k| < h$, $x^0 \in D$ and $B^0 \in R^{n\times n}$ satisfy

$$\|x^0 - x^*\| \leq \epsilon,\ \|B^0 - H(x^0)\|_F \leq \delta,$$

then $\{x^k\}$ is well-defined and converges q-linearly to x^*. If $\lim_{k->\infty}|h^k| = 0$, then the convergence is q-superlinear. If there exists some constant C such that $|h^k| \leq C\|g(x^k)\|$, then the convergence is p-step q-quadratic, and the r-convergence order of $\{x^k\}$ is not less than τ_p, where τ_p is the unique positive root of the equation

(2.3) $$t^p - t^{p-1} - 1 = 0.$$

The basic idea of the SCMEC method is to make a secant modification on the matrix B^k by using the information $g(x^{k-1})$ we already have to get a better approximation to the Hessian, say $\bar{B}^k$. This method can be formulated as follows:

ALGORITHM 2.2. *Given x^0 and B^0 as in Algorithm 2.1, take the initial step as in Algorithm 2.1. At each iteration $k > 0$ do the following:*

1. *Update B^{k-1} by steps a, b, c and d of Algorithm 2.1 to get B^k.*
2. *Update B^k by Marwil and Toint's SPSB update to get $\bar{B}^k$.*
3. *Solve $\bar{B}^k s^k = -g(x^k)$.*
4. *Choose x^{k+1} by $x^k + s^k$, or by a global strategy.*
5. *Check convergence.*

THEOREM 2.4. *Algorithm 2.2 has at least the same local convergence properties as Algorithm 2.1.*

3. The diagonal secant modification of the CMEC method. Though theoretically, we could not say that the secant modification of the CMEC method using Marwil and Toint's SPSB update (SCMEC) is better than the CMEC method, according to our experiments, it usually takes fewer iterations to get the solution, and therefore, it uses fewer gradient values than the CMEC method. However, Marwil and Toint's SPSB update needs the solution of an additional banded linear system to determine the updated matrix $\bar{B}^k$. While this computatinal cost may not be too significant if the pure SPSB update is used (See [11]), according to our experience, this additional cost may be too high comparing to what we gain from the SCMEC method. In addition, the coding for the SPSB method, especially for the degenerate case, is quite complicated. To overcome these drawbacks, we consider a diagonal secant update. The basic idea is to update only the diagonal of B^k and make the updated matrix $\bar{B}^k$ satisfy the secant equation

$$\bar{B}^k s^{k-1} = y^{k-1},$$

where $s^{k-1} = x^k - x^{k-1}$ and $y^{k-1} = g(x^k) - g(x^{k-1})$. We first define

$$\alpha^+ = \begin{cases} \frac{1}{\alpha} & \text{if } x \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

for a scalar α. Now the diagonal secant update is given by

$$\bar{B}^k = B^k + \sum_{i=1}^{n} (e_i^T s^{k-1})^+ e_i^T (y^{k-1} - B^k s^{k-1}) e_i e_i^T. \tag{3.1}$$

It is easy to verify that if $e_i^T s^{k-1} \neq 0$, then the i-th row of $\bar{B}^k$ satisfies the i-th secant equation, i.e.,

$$e_i^T \bar{B}^k s^{k-1} = e_i^T y^{k-1}.$$

It can be seen from (3.1) that when $e_i^T s^{k-1} = 0$, no update is performed on the i-th element of the diagonal. In practice, we actually skip the update if $|e_i^T s^{k-1}|$ is relatively small, i.e., if

$$|e_i^T s^{k-1}| < \theta \|s^{k-1}\|_\infty, \tag{3.2}$$

then we skip the update, where $\theta > 0$ is a small scalar.

Now the update is as follows: For $i = 1, 2, ..., n$, if

$$|e_i^T s^{k-1}| \geq \theta \|s^{k-1}\|_\infty, \tag{3.3}$$

then

$$e_i^T \bar{B}^k e_i = e_i^T B^k e_i + \frac{1}{e_i^T s^{k-1}} e_i^T (y^{k-1} - B^k s^{k-1}). \tag{3.4}$$

Otherwise,

$$e_i^T \bar{B}^k e_i = e_i^T B^k e_i. \tag{3.5}$$

Note that a similar diagonal update may be found in Dennis and Schnabel [3, p. 256], where a different choice of y is used and n additional gradient component values are needed.

Now the diagonal secant modification of the CMEC method (DSCMEC) is as follows.

ALGORITHM 3.1. *Given a symmetrically consistent partition of the Hessian, x^0 and B^0 as in Algorithm 2.1, do the following:*
At the initial step :

1. *Set $l = 0$ and $\bar{B}^0 = B^0$.*
2. *Solve $\bar{B}^0 s = -g(x^0)$.*
3. *Choose x^1 by $x^0 + s$ or by a global strategy.*

At each iteration $k > 0$:

1. *Update B^{k-1} by steps a, b, c and d of Algorithm 2.1 to get B^k.*
2. *Update the diagonal of B^k by (3.4) and (3.5) to get $\bar{B}^k$.*
3. *Solve $\bar{B}^k s = -g(x^k)$.*
4. *Choose x^{k+1} by $x^k + s$ or by a global strategy.*
5. *Check convergence.*

It can be seen from Algorithm 3.1 that the number of gradient evaluations needed to form $\bar{B}^k$ at each iteration is two, the same as the CMEC method, and the additional cost of the diagonal update is only a multiplication of a sparse matrix with a vector plus n arithmetic operations, which is much less than that of the SPSB update. Furthermore, we will show that the DSCMEC method has at least the same local convergence properties as the CMEC method for general problems, and it has better local convergence rates for symmetric band structures. Our numerical experiments also show

that the DSCMEC method behaves well in practice. Now we give the local convergence properties of the DSCMEC method.

LEMMA 3.1. *Let $\{x^j\}_{j=1}^k \subset D$ and $\{B^j\}_{j=1}^k$ be generated by Algorithm 3.1. Assume that $H(x)$, B^0 and $\{x^j+d^j\}_{j=1}^k$ satisfy the assumptions of Lemma 2.2. Then there exists a constant $C_1 > 0$ such that for $k \geq p$,*

$$\|\bar{B}^k - H(x^k)\|_F \leq C_1\alpha(\bar{e}_k + \bar{h}_k), \tag{3.6}$$

and for $k < p$,

$$\|\bar{B}^k - H(x^k)\|_F \leq C_1(\alpha(2\bar{e}_k + \bar{h}_k) + \delta),$$

where

$$\bar{e}_k = \max_{1\leq j\leq m(k)}\{\|x^k - x^{k-j}\|\},\ \ \bar{h}_k = \frac{\sqrt{n}}{2}\max_{0\leq j\leq m(k)} h^{k-j},$$

$m(k) = min\{k,\ p-1\}$ and $h^0 = 0$.

Proof. Let

$$\bar{J}^{k-1} = \int_0^1 H(x^{k-1} + ts^{k-1})dt. \tag{3.7}$$

Then,

$$\bar{J}^{k-1}s^{k-1} = y^{k-1}. \tag{3.8}$$

Note that by Lipschitz condition (2.1), for $(i,j) \in M$,

$$\begin{aligned} |e_i^T(\bar{J}^{k-1} - H(x^k))e_j| &\leq |e_i^T\int_0^1 (H(x^{k-1} + t(x^k - x^{k-1})) - H(x^k))dte_j| \\ &\leq \alpha_{ij}\|x^k - x^{k-1}\|\int_0^1 (1-t)dt \leq \frac{\alpha_{ij}}{2}\|x^k - x^{k-1}\|, \end{aligned} \tag{3.9}$$

and therefore,

$$\|\bar{J}^{k-1} - H(x^k)\|_F \leq \frac{\alpha}{2}\|x^k - x^{k-1}\|.$$

Thus, from (3.1), (3.3), (3.4), (3.5), (3.7) and (3.8),

$$\begin{aligned} \|\bar{B}^k - B^k\|_F &\leq \frac{1}{\theta}\frac{\|y^{k-1} - B^k s^{k-1}\|}{\|s^{k-1}\|_\infty} \\ &\leq \frac{1}{\theta}\frac{\|(\bar{J}^{k-1} - B^k)s^{k-1}\|}{\|s^{k-1}\|_\infty} \\ &\leq \frac{1}{\theta}\frac{\|(\bar{J}^{k-1} - B^k)\|_F\|s^{k-1}\|}{\|s^{k-1}\|_\infty} \\ &\leq \frac{\sqrt{n}}{\theta}\|\bar{J}^{k-1} - B^k\|_F \end{aligned}$$

$$\leq \frac{\sqrt{n}}{\theta}(\|\bar{J}^{k-1} - H(x^k)\|_F + \|H(x^k) - B^k\|_F)$$
$$\leq \frac{\sqrt{n}}{\theta}(\frac{\alpha}{2}\|x^k - x^{k-1}\| + \|H(x^k) - B^k\|_F).$$

Therefore,

$$\begin{aligned}\|\bar{B}^k - H(x^k)\|_F &\leq \|\bar{B}^k - B^k\|_F + \|B^k - H(x^k)\|_F \\ &\leq \frac{\sqrt{n}}{\theta}(\frac{\alpha}{2}\|x^k - x^{k-1}\| + \|H(x^k) - B^k\|_F) + \|B^k - H(x^k)\|_F.\end{aligned}$$

Then, the desired results can be obtained by applying Lemma 2.2 and setting

$$C_1 = \frac{3\sqrt{n}}{2\theta} + 1. \tag{3.10}$$

Now we have the following local convergence results for Algorithm 3.1.

THEOREM 3.2. *Under the same assumptions as in Theorem 2.3, Algorithm 3.1 has at least the same local convergence results as Algorithm 2.1.*

Proof. Since $x^* \in D$ and D is an open convex set, we can choose ϵ so that $S(x^*, 2\epsilon) \equiv \{x : \|x - x^*\| < 2\epsilon\} \subset D$. Also, we can chose ϵ, δ and h so that

$$\sqrt{n}h < \epsilon, \quad 2\beta C_1(\alpha(\frac{9\epsilon}{2} + h) + \delta) < \frac{1}{2}, \tag{3.11}$$

where C_1 is defined in (3.10) , and $\beta > 0$ satisfies $\|H^{-1}(x^*)\|_F < \beta$. Without loss of generality, we assume that $C_1 \geq 1$.

We first show by induction on k that

$$\|x^{k+1} - x^*\| \leq \frac{1}{2}\|x^k - x^*\|, \; k = 0, 1, 2, ... \tag{3.12}$$

Note that by Lipschitz condition (2.1) and (3.11),

$$\begin{aligned}\|H^{-1}(x^*)(\bar{B}^0 - H(x^*))\|_F &\leq \|H^{-1}(x^*)\|_F(\|\bar{B}^0 - H(x^0)\|_F + \|H(x^0) - H(x^*)\|_F) \\ &\leq \beta(\delta + \alpha\epsilon) < \frac{1}{2}.\end{aligned} \tag{3.13}$$

Therefore, by Dennis and Schnabel's Theorem 3.1.4 [3],

$$\|(\bar{B}^0)^{-1}\|_F \leq 2\beta,$$

which shows that x^1 is well-defined. By Lipschitz condition (2.1),

$$\begin{aligned}\|x^1 - x^*\| &\leq \|(\bar{B}^0)^{-1}\|_F(\|g(x^*) - g(x^0) - H(x^0)(x^* - x^0)\| \\ &\quad + \|\bar{B}^0 - H(x^0)\|_F\|x^* - x^0\|) \\ &\leq 2\beta(\frac{\alpha}{2}\|x^0 - x^*\| + \|\bar{B}^0 - H(x^0)\|_F)\|x^* - x^0\|\end{aligned}$$

$$\leq \quad 2\beta(\frac{3\alpha}{2}+\delta)\|x^*-x^0\| \leq \frac{1}{2}\|x^0-x^*\|. \tag{3.14}$$

This means that (3.12) holds for $k = 0$. Now suppose (3.12) holds for $k = 1, 2, ..., m-1$. We show that it also holds for $k = m$. By (3.12),

$$\|x^m + d^m - x^*\| \leq \|x^m - x^*\| + \|d^m\| \leq \|x^0 - x^*\| + \sqrt{n}h < 2\epsilon.$$

Thus, $\{x^k + d^k\}_{k=1}^m \subset S(x^*, 2\epsilon) \subset D$. By Lemma 3.1, there exists an integer $1 \leq j_0 \leq min\{m, p-1\}$ such that

$$\begin{aligned} \|\bar{B}^m - H(x^m)\|_F &\leq C_1(\alpha(2\|x^m - x^{m-j_0}\| + \bar{h}_m) + \delta) \\ &\leq C_1(\alpha(2(\|x^m - x^*\| + \|x^* - x^{m-j_0}\|) + \bar{h}_m) + \delta) \\ &\leq C_1(\alpha(4\|x^* - x^{m-j_0}\| + \bar{h}_m) + \delta). \end{aligned} \tag{3.15}$$

Thus, using a similar argument to (3.13), we obtain

$$\|H^{-1}(x^*)(\bar{B}^m - H(x^*))\|_F \leq \beta C_1(\alpha(5\epsilon + h) + \delta) < \frac{1}{2}$$

and

$$\|(\bar{B}^m)^{-1}\|_F \leq 2\beta,$$

which shows that x^{m+1} is well defined. Using (3.15) and a similar argument as (3.14), we have

$$\begin{aligned} \|x^{m+1} - x^*\| &\leq 2\beta(\frac{\alpha}{2}\|x^m - x^*\| + \|\bar{B}^m - H(x^m)\|_F)\|x^* - x^m\| \\ &\leq 2\beta(\frac{\alpha\epsilon}{2} + C_1(\alpha(4\epsilon + \bar{h}_m) + \delta))\|x^* - x^m\| \\ &\leq 2\beta C_1(\alpha(\frac{9}{2}\epsilon + h) + \delta)\|x^* - x^m\| \leq \frac{1}{2}\|x^m - x^*\|, \end{aligned} \tag{3.16}$$

which completes the induction step. It follows from (3.12) that $\{x^k\}$ converges to x^* at least q-linearly.

Note that for $k \geq p$, by (3.6), inequality (3.15) is changed to

$$\begin{aligned} \|\bar{B}^k - H(x^k)\|_F &\leq C_1\alpha(\|x^k - x^*\| + \|x^* - x^{k-p+1}\| + \bar{h}_k) \\ &\leq C_1\alpha(2\|x^* - x^{k-p+1}\| + \bar{h}_k), \end{aligned}$$

and therefore, (3.16) is changed to

$$\|x^{k+1} - x^*\| \leq 2C_1\alpha\beta(\frac{5}{2}\|x^* - x^{k-p+1}\| + \bar{h}_k)\|x^k - x^*\|. \tag{3.17}$$

Since $\{\bar{h}_k\}$ is a sub-sequence of $\{h^k\}$, $h^k \to 0$ implies $\bar{h}_k \to 0$. Therefore, by (3.17), $\{x^k\}$ converges to x^* q-superlinearly if $h^k \to 0$. By Dennis and Schnabel's Lemma 4.1.16 [3],

$$|h^k| \leq C\|g(x^k)\|$$

is equivalent to

$$|h^k| \leq C_2 \|x^k - x^*\|,$$

where $C_2 > 0$ is a constant. Therefore, if $|h^k| \leq C\|g(x^k)\|$, inequality (3.17) can be rewritten as

$$\|x^{k+1} - x^*\| \leq C_3 \|x^* - x^{k-p+1}\| \|x^k - x^*\| \leq C_3 \|x^* - x^{k-p+1}\|^2, \tag{3.18}$$

where $C_3 > 0$ is a constant, which implies that $\{x^k\}$ converges to x^* at least p-step q-quadratically.

The estimate of the r-convergence order of Algorithm 3.1 can be easily obtained by using (3.18) and applying Ortega and Rheinboldt's Theorem 9.2.9 [8].

Theorem 3.2 did not say that Algorithm 3.1 has a better local convergence property than Algorithm 2.1 in general. However, we will show that it may actually have a better q-convergence rate and a better r-convergence order than Algorithm 2.1 for problems with band structures.

LEMMA 3.3. *Let $\bar{B}^k$ be generated by Algorithm 3.1. If (3.3) is satisfied, then the distance between the i-th diagonal element of $\bar{B}^k$ and the i-th diagonal element of $\bar{J}^{k-1}$ defined by (3.7) is independent of the i-th diagonal element of B^k, and it depends only on the off-diagonal elements on the i-th row of B^k.*

Proof. Let Ω_i denote the set of the column indices of the nonzero elements on the i-th row of $H(x)$, i.e.

$$\Omega_i = \{m \in 1, 2, ..., n : \ e_i^T H(x) e_m \neq 0\},$$

and let

$$\bar{\Omega}_i = \{m \in \Omega_i : \ m \neq i\}.$$

From (3.1) and (3.8), for all $i \in \{1, 2, ..., n\}$ such that $|e_i^T s^{k-1}| \geq \theta \|s^{k-1}\|_\infty$, we have

$$\begin{aligned}
|e_i^T(\bar{B}^k - \bar{J}^{k-1})e_i| &= |e_i^T B^k e_i + \frac{1}{e_i^T s^{k-1}} e_i^T(\bar{J}^{k-1} - B^k)s^{k-1} - e_i^T \bar{J}^{k-1} e_i| \\
&= |e_i^T B^k e_i + \frac{1}{e_i^T s^{k-1}} e_i^T(\bar{J}^{k-1} - B^k) \sum_{m \in \Omega_i} e_m^T s^{k-1} e_m - e_i^T \bar{J}^{k-1} e_i| \\
&= |\sum_{m \in \bar{\Omega}_i} e_i^T(\bar{J}^{k-1} - B^k) e_m \frac{e_m^T s^{k-1}}{e_i^T s^{k-1}}| \\
&\leq \frac{1}{\theta} \sum_{m \in \bar{\Omega}_i} |e_i^T(\bar{J}^{k-1} - B^k) e_m|. \qquad (3.19)
\end{aligned}$$

Now we give the main results of this new method, i. e., it has better local convergence rates than the CMEC method for problems with band structures. Before we prove those results, we give some insight by an example with tridiagonal structure explaining why this diagonal secant update technique is specially good for the band structures.

The following picture shows the elements corrected by the CMEC method in first three iterative steps. The elements at 'a' positions are corrected at the first iteration, the ones at 'b' positions are corrected at the second iteration and the ones at 'c' positions are corrected at the third iteration.

$$\begin{bmatrix} a & ab & 0 & 0 & 0 & 0 \\ ab & b & bc & 0 & 0 & 0 \\ 0 & bc & c & ac & 0 & 0 \\ 0 & 0 & ac & a & ab & 0 \\ 0 & 0 & 0 & ab & b & bc \\ 0 & 0 & 0 & 0 & bc & c \end{bmatrix}.$$

Note that after three iterations all off-diagonal elements are corrected twice while all diagonal elements are corrected only once, i.e., the diagonal elements are not as good approximations to the corresponding elements of the Hessian as the off-diagonal ones. By updating diagonal elements using a diagonal secant update we may refresh the diagonal at every iteration and make all elements be better approximations to the elements of the Hessian.

We state the above fact as the following lemma without proof.

LEMMA 3.4. *Assume that $H(x)$ has a band structure with a bandwidth at least three. Let B^k for $k \geq p$, be generated by Algorithm 2.1. Then the off-diagonal elements of B^k are corrected at least twice in every p iterations.*

Now we have the following better estimate for $\bar{B}^k$ than that in Lemma 3.1.

LEMMA 3.5. *Let $\{x^j\}_{j=1}^k \subset D$ and $\{B^j\}_{j=1}^k$ be generated by Algorithm 3.1. Assume that $H(x)$, B^0 and $\{x^j + d^j\}_{j=1}^k$ satisfy the assumptions of Lemma 2.2. Also assume that $H(x)$ has a band structure with a bandwidth at least three. If (3.3) is satisfied for all $i \in \{1, 2, ..., n\}$, then there exists a constant $C_4 > 0$ such that for $k \geq p$,*

$$\|\bar{B}^k - H(x^k)\|_F \leq C_4 \alpha (\hat{e}_k + \hat{h}_k) \tag{3.20}$$

where

$$\hat{e}_k = \max_{1 \leq j \leq p-2} \{\|x^k - x^{k-j}\|\} \text{ and } \hat{h}_k = \frac{\sqrt{n}}{2} \max_{0 \leq j \leq p-2} \{h^{k-j}\}.$$

Proof. Without loss of generality we assume that $\theta \leq 1$. By Lemma 3.4, for $k \geq p$ and $(i, j) \in M$, $i \neq j$ there exists at least one integer $0 \leq q \leq p-2$ such that b_{ij}^{k-q} is corrected at the $(k-q)$th step. Let m be the smallest one among all such q's. Then,

$$e_i^T B^k e_j = e_i^T B^{k-m} e_j.$$

Thus, from Lemma 2.1,

$$|e_i^T (B^k - H(x^k)) e_j| = |e_i^T (B^{k-m} - H(x^k)) e_j$$

$$
\begin{aligned}
&= |e_i^T(B^{k-m} - H(x^{k-m}))e_j| + |e_i^T(H(x^{k-m}) - H(x^k))e_j| \\
&= \leq \alpha_{ij}(\frac{\sqrt{n}}{2}|h^{k-m}| + \|x^k - x^{k-m}\|) \\
&= \leq \alpha_{ij}(\hat{e}_k + \hat{h}_k). \qquad (3.21)
\end{aligned}
$$

Therefore, from (3.9), (3.19) and Lipschitz condition (2.1),

$$
\begin{aligned}
|e_i^T(\bar{B}^k - H(x^k))e_i| &\leq |e_i^T(\bar{B}^k - \bar{J}^{k-1})e_i| + |e_i^T(\bar{J}^{k-1} - H(x^k))e_i| \\
&\leq \frac{1}{\theta}\sum_{m\in\bar{\Omega}_i}|e_i^T(\bar{J}^{k-1} - B^k)e_m| + \frac{\alpha_{ii}}{2}\|x^k - x^{k-1}\| \\
&\leq \frac{1}{\theta}(\sum_{m\in\bar{\Omega}_i}|e_i^T(\bar{J}^{k-1} - H(x^k))e_m| + |e_i^T(B^k - H(x^k))e_m|) \\
&\quad + \frac{\alpha_{ii}}{2}\|x^k - x^{k-1}\| \\
&\leq \frac{1}{\theta}\sum_{m\in\bar{\Omega}_i}(\frac{\alpha_{im}}{2}\|x^k - x^{k-1}\| + \alpha_{im}(\hat{e}_k + \hat{h}_k)) + \frac{\alpha_{ii}}{2}\|x^k - x^{k-1}\| \\
&\leq \frac{1}{\theta}\sum_{m\in\Omega_i}\alpha_{im}(\frac{1}{2}\|x^k - x^{k-1}\| + \hat{e}_k + \hat{h}_k) \\
&\leq \frac{1}{\theta}\sum_{m\in\Omega_i}\alpha_{im}(\frac{3}{2}\hat{e}_k + \hat{h}_k) \\
&\leq \frac{3}{2\theta}(\hat{e}_k + \hat{h}_k)\sum_{m\in\Omega_i}\alpha_{im} \\
&\leq \frac{3}{2\theta}(\hat{e}_k + \hat{h}_k)\sqrt{l}\,\alpha_i, \qquad (3.22)
\end{aligned}
$$

where $l = max\{l_i : i = 1, 2, ...n\}$, l_i is the number of nonzero elements on the i-th row, and

$$\alpha_i = (\sum_{m\in\Omega_i}\alpha_{im}^2)^{\frac{1}{2}}.$$

Let $C_5 = \frac{3\sqrt{l}}{2\theta}$. Then, by (3.21) and (3.22),

$$
\begin{aligned}
\|\bar{B}^k - H(x^k)\|_F^2 &= \sum_{(i,j)\in M, i\neq j}|e_i^T(B^k - H(x^k))e_j|^2 + \sum_{i=1}^n|e_i^T(\bar{B}^k - H(x^k))e_i|^2 \\
&\leq (\hat{e}_k + \hat{h}_k)^2\sum_{(i,j)\in M}\alpha_{ij}^2 + C_5^2(\hat{e}_k + \hat{h}_k)^2\sum_{i=1}^n\alpha_i^2 \\
&= (\hat{e}_k + \hat{h}_k)^2\alpha^2 + C_5^2(\hat{e}_k + \hat{h}_k)^2\alpha^2 \\
&\leq (C_5^2 + 1)(\hat{e}_k + \hat{h}_k)^2\alpha^2. \qquad (3.23)
\end{aligned}
$$

Let

$$C_4 = \sqrt{C_5^2 + 1}.$$

Then (3.20) follows from (3.23).

By using Lemma 3.5 and an argument similar to the proof of Theorem 3.2, we have the following convergence results for Algorithm 3.1.

THEOREM 3.6. *Assume that $H(x)$ has a band structure with a bandwidth at least three and that $g(x)$ and $H(x)$ satisfy the hypotheses in Theorem 2.3. Also assume that $|h^k| \leq C\|g(s^k)\|$ and that (3.3) is satisfied for all $i \in \{1, 2, ..., n\}$ and all k sufficiently large. Then Algorithm 3.1 is locally at least $p-1$ step q-quadratically convergent, and the r-convergence order is not less than τ where τ is the unique positive root of the equation*

$$t^{p-1} - t^{p-2} - 1 = 0.$$

Comparing Theorem 3.6 with Theorem 2.3 and Theorem 2.4, we see that the DSCMEC method has a better q-convergence rate and a better r-convergence order than the CMEC method for problems with band structures. This theoretical result explains our numerical results in Section 4.

4. Numerical Results. To see the numerical behavior of the CMEC and DSCMEC method, we solved seven example problems by Powell and Toint's direct method (PTD), Powell and Toint's indirect method (PTID), Marwil and Toint's sparse PSB method (SPSB), the CMEC method, and the DSCMEC method. In this section we compare the numerical results from these six methods.

The global strategy used to force convergence from far away points was a line search backtracking strategy as described by Dennis and Schnabel [3]. We choose the step length in finite differences for each element as

$$h_j^k = \sqrt{macheps}\, x_j^k,$$

where *macheps* is the machine precision. The stopping tests we used are the ones given by Dennis and Schnabel [3] and all tests were run with the same accuracy requirement ($\epsilon = 10^{-5}$). For the DSCMEC method we choose θ in (3.2) to be 10^{-8}. For the SPSB method, the CMEC method, the SCMEC method and the DSCMEC method, the initial approximations to the Hessian were computed by the PTD method. All tests were run on the Jilin University Honeywell DPS-8 in double precision.

One of the test problems is the Chained Rosenbrock function given by Toint [11]. Two others can be found in Moré, Garbow and Hillstrom [7]. They are the Extended Rosenbrock function and the Discrete boundary value function. The fourth one is an extension of Example 9.2.2 in [3], where the dimension was only two. The remaining three examples are variations of the Broyden banded function (see [7]). Here, we only made changes on the lower and upper half bandwidths to have five diagonal, seven diagonal and nine diagonal structures. We recorded the average number of iterations (NITER) and the average number of gradient evaluations (NGRAD) needed to solve the seven problems using each of the method mentioned above in Table 1. To see the effect of increasing the number of groups in the partition of the columns of the Hessian, in Table

TABLE 1

Algorithms	PTD	PTID	SPSB	CMEC	SCMEC	DSCMEC
NITER	11.0	11.0	31.6	20.9	16.1	14.9
NGRAD	53.6	40.0	35.4	44.6	35.1	32.6

TABLE 2

Algorithms	Five-diagonal		Seven-diagonal		Nine-diagonal	
	NITER	NGRAD	NITER	NGRAD	NITER	NGRAD
PTD	7	43	7	57	7	71
PTID	7	29	7	36	7	43
SPSB	31	35	39	44	34	40
CMEC	14	31	17	38	19	43
SCMEC	11	25	13	30	14	33
DSCMEC	10	23	12	28	14	33

2, we list the number of iterations (NITER) and the number of gradient evaluations (NGRAD) for solving the five-diagonal, seven-diagonal and nine-diagonal variations of the Broyden banded function by using the six methods. From the numerical results we see that the DSCMEC method uses the least number of gradient values for the test problems. It needs more iterations than the PTD and the PTID methods. However, the difference is less significant than those between the PTD method and other three methods.

5. Concluding Remarks. We have shown that the DSCMEC method has at least the same local convergence properties as the CMEC method for general sparse problems and that it has better local convergence rates than the CMEC method for problems with band structures. Our numerical results show that the DSCMEC method may be competitive with most existing methods for problems with band structures.

Achnowledgement. The author is grateful to Professor John E. Dennis and Dr. Karen A. Williamson for their helpful suggestions and corrections on the preliminary draft of this paper.

REFERENCES

[1] T. COLEMAN AND J. MORÉ, *Estimation of sparse hessian matrices and graph coloring problems*, Mathematical Programming, 28 (1984), pp. 243–270.

[2] A. CURTIS, M. POWELL, AND J. REID, *On the estimation of sparse jacobian matrices*, Journal of Applied Mathematics, 13 (1974), pp. 117–119.

[3] J. DENNIS AND R. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983.

[4] G. LI, *Successive column correction algorithms for solving sparse nonlinear systems of equations*, Mathematical Programming, 43 (1989), pp. 187–207.

[5] ——, *Successive element correction algorithms for sparse unconstrained optimization*, TR91-34, Department of Mathematical Sciences, Rice University, (1991).

[6] E. S. MARWIL, *Exploiting sparsity in newton-like methods*, Ph.D. thesis, Cornell University, Ithaca, NY, (1978).

[7] J. MORÉ, B. GARBOW, AND K. HILLSTROM, *Testing unconstrained optimization software*, ACM Transactions on Mathematical Software, 7 (1981), pp. 17–41.

[8] J. ORTEGA AND W. RHEINBOLDT, *Iterative Solution of Nonlinear Equations*, Academic Press, New York and London, 1970.

[9] M. POWELL AND P. TOINT, *On the estimation of sparse hessian matrices*, SIAM Journal on Numerical Analysis, 16 (1979), pp. 1060–1074.

[10] P. TOINT, *On sparse and symmetric matrix updating subject to a linear equation*, Mathematics of Computation, 31 (1977), pp. 954–961.

[11] ——, *Some numerical results using a sparse matrix updating formula in unconstrained optimization*, Mathematics of Computation, (1978), pp. 839–851.