Mathematical Programming 28 (1984) 243–270
North-Holland

ESTIMATION OF SPARSE HESSIAN MATRICES AND GRAPH COLORING PROBLEMS

Thomas F. COLEMAN

Department of Computer Science, Cornell University, Ithaca, NY 14853, USA

Jorge J. MORÉ

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received 16 December 1982
Revised manuscript received 27 May 1983

Large scale optimization problems often require an approximation to the Hessian matrix. If the Hessian matrix is sparse then estimation by differences of gradients is attractive because the number of required differences is usually small compared to the dimension of the problem. The problem of estimating Hessian matrices by differences can be phrased as follows: Given the sparsity structure of a symmetric matrix A, obtain vectors $d_1, d_2, \dots, d_p$ such that $Ad_1, Ad_2, \dots, Ad_p$ determine A uniquely with p as small as possible. We approach this problem from a graph theoretic point of view and show that both direct and indirect approaches to this problem have a natural graph coloring interpretation. The complexity of the problem is analyzed and efficient practical heuristic procedures are developed. Numerical results illustrate the differences between the various approaches.

Key words: Graph Coloring, Estimation of Hessian Matrices, Sparsity, Differentiation, Numerical Differences, NP-Complete Problems, Unconstrained Minimization.

1. Introduction

Optimization algorithms which use second order information require the computation or estimation of the symmetric matrix of second derivatives $\nabla^2 f(x)$ for some problem function $f:\mathbb{R}^n \to \mathbb{R}$. In large scale problems the Hessian matrix $\nabla^2 f(x)$ is often sparse and then estimation of the Hessian matrix by differencing the gradient $\nabla f(x)$ becomes attractive because the number of differences needed is usually small relative to the dimension of the problem. For example, if $\nabla^2 f(x)$ is tridiagonal then Powell and Toint (1979) show that only 2 gradient differences are needed to estimate $\nabla^2 f(x)$. For a general sparsity structure, however, it is not easy to estimate the Hessian with a small number of gradient differences, and thus we address the following problem: Given a function $f:\mathbb{R}^n \to \mathbb{R}$ and knowledge of the sparsity structure of the Hessian matrix $\nabla^2 f(x)$, how many gradient differences are needed to estimate $\nabla^2 f(x)$?

Work supported in part by the Applied Mathematical Sciences Research Program (KC-04-02) of the Office of Energy Research of the U.S. Department of Energy under Contract W-31-109-Eng-38.

We assume that it is desirable to evaluate the gradient $\nabla f(x)$ as a single entity rather than separately evaluate the components $\partial_1 f(x), \ldots, \partial_n f(x)$ of the gradient. This would certainly be true if the components have expensive common subexpressions. The Hessian matrix can be estimated by noting that (with the appropriate differentiability assumptions) the product $\nabla^2 f(x)d$ can be estimated, for example, by forward differences,

$$\nabla^2 f(x)d = [\nabla f(x+d) - \nabla f(x)] + o(\|d\|),$$

or by central differences,

$$\nabla^2 f(x)d = \tfrac{1}{2}[\nabla f(x+d) - \nabla f(x-d)] + o(\|d\|^2).$$

The problem of estimating a sparse Hessian matrix can thus be formulated as follows: Given knowledge of the sparsity structure of a symmetric matrix A of order n, obtain vectors $d_1, d_2, \ldots, d_p$ such that $Ad_1, Ad_2, \ldots, Ad_p$ determine A uniquely. Note that since A is associated with the Hessian matrix $\nabla^2 f(x)$, the sparsity structure of A should represent the sparsity structure of $\nabla^2 f(x)$ for all x of interest. Moreover, since in a minimization problem the Hessian is usually positive definite at a minimizer, it is natural to assume that A has nonzero diagonal elements. Also note that since each evaluation of Ad is associated with the estimation of $\nabla^2 f(x)d$ by differencing the gradient, and since the evaluation of the gradient can be costly, we are interested in obtaining difference vectors $d_1, d_2, \ldots, d_p$ with p as small as possible.

If symmetry of the matrix A is ignored, then Curtis, Powell and Reid (1974) and Coleman and Moré (1983) have suggested several possible methods. Curtis, Powell and Reid observed that if the directions partition the columns into groups such that columns in a group do not have a nonzero in the same row position, then the elements of A can be determined directly. Based on this observation, Curtis, Powell and Reid proposed an algorithm designed to form a small number of such groups—each group corresponding to a direction—the CPR method. Coleman and Moré used the connection of the partition problem with a certain graph coloring problem to suggest improved partition algorithms based on graph coloring heuristics. Their numerical results show that the problem of estimating a sparse Jacobian matrix can be successfully attacked as a graph coloring problem, and that the improved algorithms are optimal or nearly optimal on practical problems.

It should be noted that direct methods based on a partition of the columns is not the only way to go in the unsymmetric case. An example of Eisenstat (1980) (see Coleman and Moré (1983)) demonstrates that allowing columns to intersect within groups and allowing columns to reside in several groups may reduce the number of directions needed by a direct method. In addition, indirect procedures are possible. For example, Newsam and Ramsdell (1983) proposed an indirect algorithm for the unsymmetric case which never requires more than $\rho_{\max}$ directions, where $\rho_{\max}$ is the maximum number of nonzeroes in any row. It is not difficult to show (see Coleman and Moré (1983)) that at least $\rho_{\max}$ directions are required to determine a general matrix A uniquely, so their algorithm is optimal. On the negative side, this method needs to solve n least squares problems in order to extract A from $Ad_1, Ad_2, \ldots, Ad_p$; in addition, the specific procedure described by Newsam and Ramsdell leads to ill-conditioned systems. A direct method, on the other hand, obtains A directly from the difference vectors $Ad_1, Ad_2, \ldots, Ad_p$. In view of the optimal or nearly optimal behavior of direct algorithms based on a partition of the columns of A, it is not clear that this indirect method is competitive.

Powell and Toint (1979) were the first to show that exploiting symmetry can result in significant gains for many sparsity structures. For example, consider an arrowhead structure. A matrix of order 5 with this structure has the form

$$A = \begin{pmatrix} \times & \times & \times & \times & \times \\ \times & \times & & & \\ \times & & \times & & \\ \times & & & \times & \\ \times & & & & \times \end{pmatrix}.$$

For this structure n directions are needed if symmetry is ignored (since one row is dense), but only 2 directions are needed if symmetry is used. That is, if

$$d_1 = (1, 0, 0, \ldots, 0)^{\mathrm{T}}, \qquad d_2 = (0, 1, 1, \ldots, 1)^{\mathrm{T}},$$

then the first direction determines the first column of A, and by symmetry, the first row. The second direction gives the remaining diagonal elements.

Powell and Toint (1979) proposed several algorithms which exploit symmetry. In particular, two methods were detailed: a direct method and an indirect lower triangular substitution procedure. The possibility of a more general indirect method was also discussed but was not pursued at length.

In this paper we analyze, from a graph theoretic point of view, direct and indirect methods for the determination of a symmetric matrix. From this vantage point the problems can be cleanly stated, their complexity analyzed, and improved algorithms can be obtained.

Section 2 is devoted to a discussion of direct methods for the symmetric problem. Direct methods based on partitions of the columns of A are considered and the corresponding partition problem is characterized as a restricted coloring problem on the adjacency graph of A. We call this restricted problem the symmetric coloring problem. Although direct methods based on partitions of the columns of A are quite natural, Section 2 ends with an example which shows that the use of more general covers of the columns of A can yield significant reductions in the number of evaluations of Ad needed to estimate A.

The symmetric coloring problem is analyzed in Section 3. We show that it is possible to relate the symmetric chromatic number of a graph G to the chromatic number of certain supergraphs of G. As a consequence of this relationship, we prove that the decision problem for the symmetric coloring problem is NP-complete.

This result is established by showing that if there is a polynomial algorithm for the symmetric coloring problem on bipartite graphs then there is also a polynomial algorithm for the general coloring problem. The graph theory needed to understand this paper is introduced as needed; for background material on NP-complete problems, see Garey and Johnson (1979).

Algorithms for the symmetric coloring problem are analyzed in Section 4. We consider a generalization of the sequential algorithm for general graph coloring, and the graph-theoretic version of the algorithm proposed by Powell and Toint (1979) for determining column partitions of symmetric matrices. We compare several algorithms for the symmetric coloring problem and conclude that on our test problems the Powell–Toint method usually requires the least number of evaluations of Ad to determine A. Our numerical results also show that algorithms which ignore the symmetry of A are not competitive.

Section 5 contains a result which strongly suggests that the direct determination of a symmetric matrix may not be the road to follow. It is shown that if A is a symmetric band matrix which is dense within the band, then symmetry cannot be used to reduce the number of evaluations of Ad needed by a direct method based on partitions of the columns of A.

In Section 6 we focus on the second approach introduced by Powell and Toint: triangular substitution methods. Again we show that there is a natural graph-theoretic interpretation of the problem. In effect, the problem reduces to another restricted coloring problem on the adjacency graph of A. We call this restricted problem the triangular coloring problem.

The triangular coloring problem is analyzed in Section 7. We show, in particular, that if there is a polynomial algorithm for the triangular coloring problem on bipartite graphs then there is also a polynomial algorithm for the general coloring problem. The proof techniques used in this section are similar, but more direct, than those used in Section 3.

Section 8 contains numerical results for both direct and triangular substitution methods for determining symmetric matrices. We conclude that triangular substitution methods require the least number of evaluations of Ad to determine a symmetric matrix A. Moreover, we show that there is an algorithm that is always nearly optimal on our problems.

We end the paper with some observations on possible directions for future research in this area.

2. Direct determination of symmetric matrices

As shown in the introduction, the problem of estimating a sparse Hessian matrix can be phrased as follows: Given the sparsity structure of a symmetric matrix A, obtain vectors $d_1, d_2, \ldots, d_p$ such that $Ad_1, Ad_2, \ldots, Ad_p$ determine A uniquely. In this section we are mainly concerned with direct methods for determining A based on partitions of the columns of A.

A *partition* of the columns of A is a division of the columns into groups $C_1, C_2, \dots, C_p$ such that each column belongs to one and only one group. A partition of the columns of A is *consistent* with the direct determination of A if whenever a_{ij} is a nonzero element of A then the group containing column j has no other column with a nonzero in row i. A partition is *symmetrically consistent* if whenever a_{ij} is a nonzero element of A then the group containing column j has no other column with a nonzero in row i, or the group with column i has no other column with a nonzero in row j.

Given a consistent partition of the columns of A, it is straightforward to determine the elements of A with p evaluations of Ad by associating each group C with a direction d with components $\delta_j = 0$ if j does not belong to C, and $\delta_j \neq 0$ otherwise. Then

$$Ad = \sum_{j \in C} \delta_j a_j$$

where $a_1, a_2, \dots, a_n$ are the columns of A, and it follows that if column j is the only column in group C with a nonzero in row i then

$$(Ad)_i = \delta_j a_{ij},$$

and thus a_{ij} is determined. In this way, every nonzero of A is directly determined.

If A is symmetric, it is possible to determine A while only requiring that the partition be symmetrically consistent. Thus, given a symmetrically consistent partition of the columns of the symmetric matrix A, if column j is the only column in its group with a nonzero in row i then a_{ij} can be determined as above, while if column i is the only column in its group with a nonzero in row j then a_{ji} can be determined. Hence, every nonzero of A is directly determined with p evaluations of Ad.

The concept of a consistent partition was introduced by Coleman and Moré (1983) in their study of direct estimation methods for general matrices. As we shall see in Section 6, this concept is also of use in methods for the indirect estimation of symmetric matrices.

Powell and Toint (1979) have considered partitions of the columns of A with the property that two columns in a group are allowed to have a nonzero in row i only if column a_i belongs to a previous group. It is clear that such partitions are symmetrically consistent. Moreover, the following example of Powell and Toint (1979) shows that the required number of groups can be reduced if we use general symmetrically consistent partitions. If

$$A = \begin{pmatrix} \times & \times & \times & \times & & \\ \times & \times & \times & & \times & \\ \times & \times & \times & & & \times \\ \times & & & \times & & \\ & \times & & & \times & \\ & & \times & & & \times \end{pmatrix}$$

then 4 groups are necessary if we consider partitions which satisfy the conditions of Powell and Toint but $\{1, 5\}$, $\{2, 6\}$, $\{3, 4\}$ is a symmetrically consistent partition of the columns of A.

Partition Problem: Obtain a symmetrically consistent partition of the columns of the symmetric matrix A with the fewest number of groups.

We are interested in partitions with the least number of groups because each group involves one evaluation of Ad, and this in turn requires the evaluation of the gradient.

How difficult is the partition problem? To approach this problem it is useful to express the partition problem in the language of graph theory.

A *graph* G is an ordered pair (V, E) where V is a finite and nonempty set of *vertices* and the *edges* E are unordered pairs of distinct vertices. The vertices u and v are *adjacent* if (u, v) is an edge with *endpoints* u and v. A *p-coloring* of a graph G is a function

$$\phi : V \to \{1, 2, \ldots, p\}$$

such that $\phi(u) \neq \phi(v)$ if u and v are adjacent. A coloring ϕ *induces* a partition of V with components

$$C_i = \{u \in V : \phi(u) = i\},$$

and such that vertices in the same component are not adjacent. The *chromatic number* $\chi(G)$ of G is the smallest p for which G has a p-coloring.

We want to associate the partition problem with a coloring of a suitable graph. In the unsymmetric case the appropriate graph is the graph $G_U(A)$ with vertex set $\{a_1, a_2, \ldots, a_n\}$ where a_j is the jth column of A and edge (a_i, a_j) if $i \neq j$ and columns a_i and a_j have a nonzero in the same row position. In graph theory terminology, $G_U(A)$ is the *intersection graph* of the columns of A.

An important observation is that ϕ is a p-coloring of $G_U(A)$ if and only if ϕ induces a consistent partition of the columns of A. Thus the chromatic number of $G_U(A)$ is the smallest number of groups in a consistent partition of the columns of the matrix A.

In the symmetric case, the appropriate graph is the graph $G_S(A)$ with vertex set $\{a_1, a_2, \ldots, a_n\}$ and edge (a_i, a_j) if and only if $i \neq j$ and $a_{ij} \neq 0$. In graph theory terminology, $G_S(A)$ is the *adjacency graph* of the symmetric matrix A.

A coloring ϕ of $G_S(A)$ does not necessarily induce a symmetrically consistent partition of the columns of a symmetric matrix A; it is necessary to restrict the class of colorings.

Definition. A mapping $\phi : V \to \{1, 2, \ldots, p\}$ is a *symmetric p-coloring* of a graph $G = (V, E)$ if ϕ is a p-coloring of G and if ϕ is not a 2-coloring for any path in G of length 3. The symmetric chromatic number $\chi_\sigma(G)$ is the smallest p for which G has a symmetric p-coloring.

A *path in* G of length l is a sequence $(v_0, v_1, \ldots, v_l)$, of distinct vertices in G such that v_{i-1} is adjacent to v_i for $1 \leq i \leq l$. Thus, if ϕ is a symmetric p-coloring of G then the situation

Red Blue Red Blue

is not allowed.

As a simple illustration of these concepts, consider the variation on the arrowhead structure of Section 1 which adds the main subdiagonal and the main superdiagonal to the structure. A matrix of order 6 with this structure has the form

$$A = \begin{pmatrix} \times & \times & \times & \times & \times & \times \\ \times & \times & \times & & & \\ \times & \times & \times & \times & & \\ \times & & \times & \times & \times & \\ \times & & & \times & \times & \times \\ \times & & & & \times & \times \end{pmatrix}.$$

For this structure is not difficult to show that $\chi(G_S(A)) = 3$, that $\chi_\sigma(G_S(A)) = 4$, and that $\chi(G_U(A)) = n$. In general, however, determining the chromatic number of $G_U(A)$ or the symmetric chromatic number of $G_S(A)$ is a hard problem. This point is discussed further in Section 3.

In most cases, $G_U(A)$ is the square of $G_S(A)$. In graph theory, the *square* G^2 of a graph $G = (V, E)$ is the graph with vertex set V and edge (u, v) if and only if there is a path in G between u and v of length $l \leq 2$. A motivation for this definition is that if A is a symmetric matrix with $a_{ij} \geq 0$ and $a_{ii} > 0$ then

$$G_S(A)^2 = G_S(A^2).$$

The following result establishes the connection between $G_U(A)$ and $G_S(A)$.

Lemma 2.1. *If A is a symmetric matrix with nonzero diagonal elements then*

$$G_U(A) = G_S(A)^2.$$

Proof. If (a_i, a_j) is an edge of $G_U(A)$ then there is an index r such that $a_{ri} \neq 0$ and $a_{rj} \neq 0$. Hence (a_i, a_j) is an edge of $G_S(A)^2$. Conversely, if (a_i, a_j) is an edge of $G_S(A)^2$ then there is a path (a_i, a_k, a_j) in $G_S(A)$ of length $l \leq 2$. If the path has length $l = 1$ then (a_i, a_j) is an edge of $G_S(A)$ and hence, $a_{ij} \neq 0$. Since $a_{ii} \neq 0$, it follows that (a_i, a_j) is an edge of $G_U(A)$. If the path has length $l = 2$ then $a_{ik} \neq 0$ and $a_{kj} \neq 0$. Since A is symmetric, (a_i, a_j) is an edge in $G_U(A)$. $\square$

We can now express the partition problem for symmetric matrices as a graph coloring problem.

Theorem 2.2. *Let A be a symmetric matrix with nonzero diagonal elements. The mapping ϕ is a symmetric coloring of $G_S(A)$ if and only if ϕ induces a symmetrically consistent partition of the columns of A.*

Proof. Suppose that ϕ is a symmetric coloring of $G_S(A)$. Since ϕ is a coloring, ϕ induces a partition of the columns of A. If this partition is not symmetrically consistent then there is a nonzero a_{ij} and columns $a_r \neq a_j$ and $a_s \neq a_i$ such that a_j and a_r are in the same group with $a_{ir} \neq 0$, and a_i and a_s are in the same group with $a_{js} \neq 0$. Since a_j and a_r are in the same group $\phi(a_j) = \phi(a_r)$, and similarly, $\phi(a_i) = \phi(a_s)$. Hence ϕ is a 2-coloring of the path

$$P = (a_r, a_i, a_j, a_s).$$

Since ϕ is a symmetric coloring of $G_S(A)$, this implies that P is a path of length $l < 3$. Hence

$$\phi(a_r) = \phi(a_i) = \phi(a_j) = \phi(a_s).$$

This is not possible because ϕ is a coloring of $G_S(A)$. This contradiction shows that the partition must be symmetrically consistent.

Conversely, assume that ϕ induces a symmetrically consistent partition of the columns of A. To show that ϕ is a coloring of $G_S(A)$ assume that $a_{ij} \neq 0$ with $i \neq j$ but that $\phi(a_i) = \phi(a_j)$. Then columns a_i and a_j are in the same group, and since $a_{ii} \neq 0$ and $a_{jj} \neq 0$, the partition is not symmetrically consistent. Hence, ϕ is a coloring of $G_S(A)$. To show that ϕ is a symmetric coloring of $G_S(A)$ let

$$P = (a_r, a_i, a_j, a_s)$$

be a path of length $l = 3$. Then $a_{ij} \neq 0$. If a_j is the only column in its group with a nonzero in row i then since $a_{ir} \neq 0$ we must have $\phi(a_j) \neq \phi(a_r)$. Similarly, if a_i is the only column in its group with a nonzero in row j then $\phi(a_i) \neq \phi(a_s)$. Hence, ϕ is not a 2-coloring of P. □

In view of Theorem 2.2, the partition problem is equivalent to the following problem.

Symmetric Graph Coloring Problem: Obtain a minimum symmetric coloring of $G_S(A)$.

We are now faced with the question—how difficult is the symmetric graph coloring problem? It is known that the coloring problem for arbitrary graphs is NP-complete. This makes the existence of a polynomially bounded algorithm for solving the unsymmetric graph coloring problem an unlikely prospect. Is the symmetric graph coloring problem also NP-complete? We consider this question in the next section.

As a final note for this section, we remark that although we have concentrated on direct methods based on partitions of the columns of A, other direct methods are possible. In a general direct method, the groups $C_1, C_2, \ldots, C_p$ are a *covering* of the columns of A in the sense that each column of A belongs to at least one group. To show that the use of a general covering of the columns of A may lead to a decrease in the number of evaluations of Ad needed to determine A consider a matrix of the form

$$A = \begin{pmatrix} A_1 & D \\ D & A_2 \end{pmatrix} \tag{2.1}$$

where A_1 and A_2 are dense matrices of order $2n$, and D is a diagonal matrix with nonzero diagonal elements. The groups

$$\{1, 2, \ldots, 2n\}, \qquad \{j, 2n+1+j\}, \quad 1 \leq j < 2n, \qquad \{2n, 2n+1\},$$

are a covering but not a partition of the columns of A. However, for each nonzero a_{ij} there is a group containing column j such that no other column in this group has a nonzero in row i, or a group containing column i such that no other column in this group has a nonzero in row j, and thus A can be determined directly with $2n+1$ evaluations of Ad. On the other hand, we show in the next section that at least $3n$ evaluations are needed if we use partitions of the columns of A. As a matter of fact, since the partition

$$\{j, 3n+j\}, \quad \{n+j\}, \quad \{2n+j\}, \qquad 1 \leq j \leq n,$$

is symmetrically consistent, the symmetric chromatic number of $G_S(A)$ is $3n$. Examples of this type suggest that it may be worthwhile to investigate more general coverings of the columns of A.

3. The symmetric chromatic number

In this section we investigate the relationship between the symmetric chromatic number and the standard chromatic number of a graph. In particular, we show that determining the symmetric chromatic number of bipartite graphs is just as hard as determining the chromatic number of a general graph.

Let A be a symmetric matrix with nonzero elements on the diagonal. In Section 2 we proved that the chromatic number of $G_S(A)^2$ is the smallest possible number of groups in a consistent partition, and that the symmetric chromatic number of $G_S(A)$ is the smallest possible number of groups in a symmetrically consistent partition. The following result shows that the use of symmetrically consistent partitions is likely to yield a smaller number of groups.

Theorem 3.1. *Let G be a graph. Then*

$$\chi(G) \leq \chi_\sigma(G) \leq \chi(G^2).$$

Proof. Just note that if ϕ is a symmetric coloring of G then ϕ is a coloring of G, and that if ϕ is a coloring of G^2 then ϕ is a symmetric coloring of G. □

Theorem 3.1 establishes the simplest kind of bounds on the symmetric chromatic number of a graph G. Other bounds are possible. Also note that Theorem 3.1 suggests that the symmetric chromatic number of G is related to the chromatic number of certain graphs which lie between G and G^2 in the usual graph inclusion sense: If $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ then G_1 is a *subgraph* of G_2 (written $G_1 \subset G_2$) if $V_1 \subset V_2$ and $E_1 \subset E_2$.

Definition. Let $G = (V, E)$ be a graph. A graph $G_\sigma = (V_\sigma, E_\sigma)$ is a symmetric completion of G if $V_\sigma = V$ and if E_σ is obtained by requiring that E_σ contain E, and that if (v_1, v_2, v_3, v_4) is a path in G of length 3 then E_σ must contain (v_1, v_3) or (v_2, v_4).

Given a graph G there are many possible symmetric completions G_σ, but in all cases $G \subset G_\sigma \subset G^2$. Also note that if G_σ is a symmetric completion of G and (u, v) is an edge in G, then E_σ must contain all edges of the form (w_1, v) for w_1 adjacent to u, or all edges of the form (u, w_2) for w_2 adjacent to v. To prove this, note that if (w_1, u, v, w_2) is a path in G of length 3, and if E_σ does not contain (w_1, v) for some w_1, then E_σ must contain all edges of the form (u, w_2).

Theorem 3.2. *Let G be a graph. Then*

$$\chi_\sigma(G) = \min\{\chi(G_\sigma)\text{: } G_\sigma \text{ a symmetric completion of } G\}.$$

Proof. We first claim that if ϕ is a coloring of a symmetric completion G_σ then ϕ is a symmetric coloring of G. As a consequence, it follows that

$$\chi_\sigma(G) \leq \chi(G_\sigma) \tag{3.1}$$

for any symmetric completion G_σ. Suppose that ϕ is a coloring of a symmetric completion G_σ. Then ϕ is a coloring of G. To show that ϕ is a symmetric coloring of G let (v_1, v_2, v_3, v_4) be a path in G of length 3. Then E_σ contains (v_1, v_3) or (v_2, v_4). If E_σ contains (v_1, v_3) then $\phi(v_1) \neq \phi(v_3)$, and if E_σ contains (v_2, v_4) then $\phi(v_2) \neq \phi(v_4)$. Hence, ϕ is a symmetric coloring of G. This proves our claim and establishes (3.1).

We now claim that if ϕ is a symmetric coloring of G then ϕ colors some symmetric completion G_σ. A consequence of this claim is that

$$\chi(G_\sigma) \leq \chi_\sigma(G). \tag{3.2}$$

Let ϕ be a symmetric coloring of G and let (v_1, v_2, v_3, v_4) be a path in G of length 3. Then $\phi(v_1) \neq \phi(v_3)$ or $\phi(v_2) \neq \phi(v_4)$. In the first case complete G by adding (v_1, v_3) to E_σ and in the second case add (v_2, v_4). Hence, ϕ is a coloring of G_σ. This establishes our second claim and shows that (3.2) holds. □

Theorem 3.2 is very useful in the determination of the symmetric chromatic number of a graph. As an application of this result we show, as promised at the end of Section 2, that if A is of the form (2.1) where A_1 and A_2 are of order $2n$, then

$$\chi_\sigma(G_S(A)) \geq 3n, \tag{3.3}$$

and thus any symmetrically consistent partition of the columns of A needs at least $3n$ groups. The proof of (3.3) requires the notion of a clique: A subgraph $G_0 = (V_0, E_0)$ of G is a *clique* if each pair of distinct vertices in V_0 are adjacent. The clique is *induced* by V_0; the *size* of the clique is $|V_0|$.

To establish (3.3) we first need to note that the size of a clique is a lower bound on the chromatic number of the graph. Now consider a symmetric completion G_σ of $G_S(A)$ and note that a_j for $1 \leq j \leq 2n$, and a_j for $2n < j \leq 4n$ are cliques of size $2n$ in $G_S(A)$ and hence in G_σ. We claim that G_σ has a clique of size $3n$. To establish this claim, note that for each edge (a_i, a_{2n+i}) of $G_S(a)$ we must have that a_i is adjacent to each a_j for $2n < j \leq 4n$, or that a_{2n+i} is adjacent to each a_j for $1 \leq j \leq 2n$. Hence, at least half of the vertices $a_1, \ldots, a_{2n}$ are adjacent in G_σ to all of the vertices a_j for $2n < j \leq 4n$, or half of the vertices $a_{2n+1}, \ldots, a_{4n}$ are adjacent in G_σ to all of the vertices a_j for $1 \leq j \leq 2n$. In either case, G_σ has a clique of size $3n$ and thus $\chi(G_\sigma) \geq 3n$. This establishes our claim, and then Theorem 3.2 yields (3.3).

As another application of Theorem 3.2 we show that the determination of the symmetric chromatic number is a difficult problem even if the graph is bipartite: A graph G is *bipartite* if and only if G is 2-colorable. Equivalently, a graph $G = (V, E)$ is bipartite if and only if V is the union of two disjoint sets V_1 and V_2 such that any edge in G has one endpoint in V_1 and the other in V_2.

Theorem 3.3. *Let $G = (V, E)$ be a graph. If $\chi(G) \geq 3$ then there is a bipartite graph B with $|V|(1 + |E|)$ vertices such that*

$$\chi_\sigma(B) = \chi(G). \tag{3.4}$$

Proof. Let $v_1, \ldots, v_n$ be the vertices of G, and let $e_1, \ldots, e_m$ be the edges of G. For each edge $e_l = (v_i, v_j)$ define a bipartite graph B_l with vertices

$$\{v_i, v_j, w_1^{(l)}, \ldots, w_n^{(l)}\}$$

and edges

$$(v_i, w_k^{(l)}), \quad (v_j, w_k^{(l)}), \qquad k = 1, \ldots, n.$$

Now define a bipartite graph B by setting

$$V(B) = V(G) \cup \{w_k^{(l)}: 1 \leq k \leq n, 1 \leq l \leq m\}$$

and

$$E(B) = \bigcup_{l=1}^{m} E(B_l).$$

We now show that (3.4) holds for this bipartite graph B. To prove that

$$\chi_\sigma(B) \le \chi(G), \tag{3.5}$$

define a symmetric completion B_σ by setting $E(B_\sigma) = E(B) \cup E(G)$. Since $\chi(G) \ge 3$, it is clear that $\chi(B_\sigma) = \chi(G)$, and hence Theorem 3.2 shows that (3.5) holds. Next we establish that any symmetric completion B_σ satisfies

$$\chi(G) \le \chi(B_\sigma). \tag{3.6}$$

To show this first note that if G is a subgraph of B_σ then (3.6) trivially holds. On the other hand, if (u, v) is an edge of G but not an edge of B_σ then since $(u, w_i^{(l)}, v, w_j^{(l)})$ is a path in B of length 3, we must have that $w_i^{(l)}$ and $w_j^{(l)}$ are adjacent in B_σ. This implies that the n vertices

$$\{w_1^{(l)}, w_2^{(l)}, \ldots, w_n^{(l)}\}$$

form a clique in B_σ and hence, $n \le \chi(B_\sigma)$. Since $\chi(G) \le n$, we have shown that (3.6) holds. Theorem 3.2 and inequality (3.6) now imply that $\chi(G) \le \chi_\sigma(B)$. In view of (3.5), this yields (3.4). □

The proof of Theorem 3.3 provides a polynomial algorithm for obtaining the bipartite graph B. Thus the techniques of Theorem 3.3 can be used to show that if there is a polynomial algorithm for determining the symmetric chromatic number of a bipartite graph then there is also a polynomial algorithm for determining the chromatic number of a general graph. These techniques also show that the decision problem for the symmetric chromatic number problem on bipartite graphs is NP-complete.

4. Algorithms for symmetric graph coloring

The literature on graph coloring algorithms is extensive, but there are no algorithms for the symmetric graph coloring problem. In this section we introduce some possible algorithms and investigate their behavior.

Let us first consider coloring algorithms for the standard coloring problem. These algorithms can be described best with the help of some additional graph theory terminology: Given a graph $G = (V, E)$ and a nonempty subset W of V, the *subgraph* $G[W]$ *induced* by W has vertex set W and all edges (u, v) such that $(u, v) \in E$ with u and v in W.

Algorithm: Let $G = (V, E)$ be a graph with vertices ordered $v_1, v_2, \ldots, v_n$ and set

$$V_k = \{v_1, v_2, \ldots, v_k\}.$$

For $k = 1, 2, \ldots, n$ the sequential coloring algorithm sets $\phi(v_k)$ to the smallest positive integer such that ϕ is a coloring of $G[V_k]$.

Additional information and references for sequential coloring algorithms are provided by Coleman and Moré (1983). At this point we just note that the numerical results of Coleman and Moré show that there are orderings of the vertices so that the sequential coloring algorithm yields optimal or near optimal results on graphs of the form $G_U(A)$ for m by n matrices A with a wide variety of sparsity patterns.

A symmetric coloring of a graph $G=(V,E)$ can be obtained by applying a sequential coloring algorithm to G^2. If $G=G_S(A)$ for a symmetric matrix A then Lemma 2.1 shows that this is equivalent to applying a sequential coloring algorithm to $G_U(A)$. This approach has been studied by McCormick (1983), but as already noted, this approach is not usually appropriate because in many cases $\chi(G^2)$ is considerably larger than $\chi_\sigma(G)$.

A reasonable approach to the symmetric coloring problem is to extend the idea behind the sequential coloring algorithm.

Algorithm: Let $G=(V,E)$ be a graph with vertices ordered $v_1, v_2, \ldots, v_n$. For $k=1,2,\ldots,n$ the symmetric sequential coloring algorithm sets $\phi(v_k)$ to the smallest positive integer such that ϕ is a symmetric coloring of $G[V_k]$.

The behavior of the symmetric sequential coloring algorithm is even more dependent on the ordering of the vertices than the standard sequential coloring algorithm. To illustrate this point, consider the bipartite graph with vertices

$$V=\{u_1, u_2, v_1, v_2, \ldots, v_n\} \tag{4.1}$$

and edges

$$E=\{(u_i, v_j)\colon 1 \leqslant i \leqslant 2,\ 1 \leqslant j \leqslant n\}. \tag{4.2}$$

The symmetric sequential coloring algorithm with the ordering

$$\{v_1, v_2, \ldots, v_n, u_1, u_2\}$$

requires 3 colors, but requires $n+1$ colors with the ordering

$$\{u_1, u_2, v_1, v_2, \ldots, v_n\}.$$

This example shows that the symmetric sequential coloring algorithm may require $\alpha_1 n \chi_\sigma(G)$ colors for some positive constant α_1. This result is in contrast to the results obtained by Coleman and Moré (1983) for the unsymmetric coloring algorithm. They showed that if $G=G_U(A)$ for some m by n matrix A, then any reasonable sequential coloring algorithm requires, at worst, $\alpha_2 m^{1/2}\chi(G)$ colors for some positive constant α_2.

Our numerical results for various types of symmetric sequential coloring algorithms show that at the kth stage these algorithms tend to produce a large number of 2-colored paths of length 2. Thus the number of forbidden colors for v_k increases, and then we may obtain poor results. This is illustrated by the above example. If $\phi(u_1)=\phi(u_2)$ and we assign a color to v_i, then (u_1, v_i, u_2) is a 2-colored path of length 2 and hence we cannot have $\phi(v_j)=\phi(v_i)$ for $j \neq i$.

We have obtained better results with the algorithm of Powell and Toint (1979). As originally proposed, the Powell and Toint algorithm determines a symmetrically consistent partition. Due to the equivalence established in Theorem 2.2, this method implicitly determines a symmetric coloring. Translation of their partitioning procedure into a symmetric coloring algorithm requires the concept of the degree of a vertex: Given a graph $G=(V, E)$ the *degree* of a vertex v is the number of edges with v as an endpoint.

Algorithm: Let $G=(V, E)$ be a graph.

For $k=1, 2, \ldots,$

(a) Let U_k be the uncolored vertices. If U_k is empty then terminate the algorithm.

(b) Sort the vertices of $G[U_k]$ in decreasing order of degree in $G[U_k]$.

(c) Build a vertex set W_k by examining the vertices in U_k in the order determined in (b), and adding a vertex v to W_k if there is not a path in $G[U_k]$ between v and some vertex in W_k of length $l \leq 2$.

(d) For each $v \in W_k$ let $\phi(v)=k$.

This is the graph-theory version of the direct Powell–Toint method. Note that the coloring ϕ produced by this algorithm is such that if a vertex w is adjacent to vertices v_1 and v_2 with $\phi(v_1)=\phi(v_2)$ then

$$\phi(w)<\phi(v_1)=\phi(v_2).$$

As a consequence, ϕ is a symmetric coloring of G.

It is certainly possible to envision modifications to the Powell–Toint algorithm. For example, we can modify step c by allowing the assignment of a vertex v to W_k if this does not lead to the creation of a 2-colored path of length 3 at step d. Thapa (1982) has proposed a modification along these lines. We have not considered any such modifications because there is no guarantee that they will perform better than the original algorithm.

Numerical results for some of the coloring algorithms mentioned above can be found in Table 1. The graphs G used in the numerical results are of the form $G_S(A)$ where A is a sparse symmetric matrix of order n. The sparsity patterns used are those in the Everstine (1979) collection where the dimensions n range from 59 to 2680. In addition to the dimension n of the problem, we have included the density matd of the matrix A, the maximum number maxr of nonzeroes in any row of A, and the number of colors required by the algorithm. The totals for Table 1 appear in Table 2.

The sl algorithm of Table 1 is a sequential coloring algorithm on G^2. The ordering used is known in the graph theory literature (Matula, Marble and Isaacson (1972)) as the *smallest–last* ordering. To define this ordering for a graph $G=(V, E)$, assume that the vertices $v_{k+1}, \ldots, v_n$ have been selected, and choose v_k so that the degree of v_k in the subgraph induced by

$$V-\{v_{k+1}, \ldots, v_n\}$$

is minimal. Thus the sl algorithm is a sequential coloring algorithm on G^2 with the

Table 1
Direct methods

n	matd(%)	maxr	sl	ssl	dpt
59	7.67	6	6	5	6
66	7.35	6	6	5	6
72	4.28	5	5	3	3
87	7.15	13	13	9	10
162	4.50	9	10	9	10
193	9.38	30	32	27	27
198	3.55	12	12	10	9
209	3.99	17	17	13	13
221	3.34	12	12	9	10
234	1.52	10	10	5	5
245	2.43	13	13	10	10
307	2.68	9	11	11	10
310	2.55	11	11	10	9
346	2.69	19	20	16	15
361	2.27	9	11	11	10
419	2.03	13	15	12	12
492	1.30	11	11	10	9
503	2.38	25	25	20	20
512	1.34	15	16	14	13
592	1.46	15	15	12	11
607	1.39	14	17	13	13
758	1.04	11	12	10	10
869	0.96	14	15	13	11
878	0.97	10	11	11	11
918	0.88	13	14	11	11
992	1.70	18	22	20	18
1005	0.85	27	27	19	20
1007	0.85	10	11	11	11
1242	0.68	12	14	12	12
2680	0.35	19	19	14	14

smallest–last ordering for G^2. Software for the sl algorithm is described by Coleman and Moré (1982).

The sequential coloring algorithm with the smallest–last ordering can be guaranteed to work well for many graphs. Given a graph $G=(V,E)$ and a nonempty $W \subset V$, let $\delta(G[W])$ be the smallest degree of any vertex in $G[W]$. The sequential coloring algorithm on a graph G with the smallest–last ordering for G requires no

Table 2
Totals for Table 1

maxr	sl	ssl	dpt
408	433	355	349

more than

$$\max\{1+\delta(G[W]) \colon W \subset V\} \tag{4.3}$$

colors. This is not difficult to show; just note that the color assigned to v_k does not exceed $1+d_k$ where d_k is the degree of v_k in $G[\{v_1, v_2, \ldots, v_k\}]$. By the definition of the smallest–last ordering, it is clear that

$$d_k \leq \max\{\delta(G[W]) \colon W \subset V\}$$

and the result follows immediately. For more information on the smallest–last ordering, see Coleman and Moré (1982, 1983), and Matula and Beck (1983).

The ssl algorithm for Table 1 is a symmetric sequential coloring algorithm on G with the smallest–last ordering for G^2. The dpt algorithm is the Powell–Toint (1979) algorithm described above.

The above discussion on the smallest-last ordering shows that the number of colors required by the sl and ssl algorithms is bounded by

$$\max\{1+\delta(G^2[W]) \colon W \subset V\}. \tag{4.4}$$

Unfortunately, (4.4) can be a poor upper bound on $\chi_\sigma(G)$. For example, if G is the bipartite graph with vertices (4.1) and edges (4.2) then $\chi_\sigma(G)=3$ but (4.4) is $n+2$.

The results of Tables 1 and 2 show that the sl algorithm produces optimal or nearly optimal results as a coloring algorithm for G^2. This can be verified by noting that maxr is a lower bound on the chromatic number of G^2. The sl algorithm, however, does not produce nearly optimal results as a symmetric coloring algorithm for G.

On these problems the dpt and ssl algorithms never require more colors than the sl algorithm. On the other hand, the dpt and ssl algorithms only represent a 20% improvement over sl; it would have been reasonable to expect a 50% improvement over an algorithm which disregards symmetry. Also note that on these problems the dpt algorithm is, in general, superior to the ssl algorithm.

One final point. It is not possible to determine if the results produced by ssl and dpt are nearly optimal because we do not have a good computable lower bound on the symmetric chromatic number of G.

5. Direct methods and band matrices

Let A be a symmetric band matrix with bandwidth β. Assume furthermore that A is dense within the band so that

$$a_{ij} \neq 0 \Leftrightarrow |i-j| \leq \beta. \tag{5.1}$$

The purpose of this section is to present one result—unfortunately, a negative one. We prove that

$$\chi_\sigma(G_S(A)) = \chi(G_S(A)^2) = 2\beta + 1. \tag{5.2}$$

Since Coleman and Moré (1983) have shown that if A satisfies (5.1) then there are algorithms for coloring $G_U(A) = G_S(A)^2$ which are optimal, this result shows that symmetry is not important to direct methods based on a partition of the columns of A.

It is best to phrase our results in terms of band graphs: A graph $G = (V, E)$ is a *band graph with bandwidth* β if there is an ordering of the vertices $v_1, v_2, \ldots, v_n$ such that

$$(v_i, v_j) \in E \Leftrightarrow 0 < |i - j| \leq \beta.$$

The ordering $v_1, v_2, \ldots, v_n$ is a natural ordering of the band graph.

The notion of a band graph was introduced by Coleman and Moré (1983) in their study of coloring algorithms for graphs of the form $G_U(A)$. In this connection, note that if A is a symmetric matrix and (5.1) holds then $G_U(A)$ is a band graph with a bandwidth of 2β. Also note that if A is a symmetric matrix then $G_S(A)$ is a band graph with bandwidth β if and only if there is a permutation of the rows and columns of A such that the permuted matrix satisfies (5.1).

We have already noted that the size of a clique is a lower bound on the chromatic number of a graph. In particular, if G is a band graph then G^2 has a clique of size $2\beta + 1$ whenever $|V| \geq 2\beta + 1$, and thus $2\beta + 1 \leq \chi(G^2)$. We now extend this result.

Lemma 5.1. *Let $G = (V, E)$ be a band graph with bandwidth β, and assume that $|V| \geq 3\beta + 1$. Then every symmetric completion G_σ of G has a clique of size $2\beta + 1$.*

Proof. The proof is by induction on the size of the clique. Clearly, any set of $\beta + 1$ vertices induces a clique in G and hence in G_σ. For the induction step we assume that for some indices l and m,

$$\{v_l, \ldots, v_{\beta+l}, v_{\beta+m+1}, \ldots, v_{2\beta+l}\}, \quad 1 \leq l \leq m \leq \beta + 1, \tag{5.3}$$

or

$$\{v_m, \ldots, v_{\beta+l-1}, v_{\beta+m}, \ldots, v_{2\beta+m}\}, \quad 1 \leq l \leq m \leq \beta + 1, \tag{5.4}$$

induces a clique in G_σ of size $\sigma = 2\beta + l - m + 1$. We now show that there is a clique in G_σ of size $\sigma + 1$ and of the required form.

If $l = m$ then the cliques induced by (5.3) or (5.4) are of size $2\beta + 1$, so assume that $l < m$. Also assume that the clique is of the form (5.3); the proof for the case when the clique is of the form (5.4) is similar. Finally, assume that there is an index k such that

$$(v_k, v_{\beta+m}) \notin E_\sigma, \quad l \leq k \leq m - 1. \tag{5.5}$$

If there is no such index k then

$$\{v_l, \ldots, v_{\beta+l}, v_{\beta+m}, \ldots, v_{2\beta+l}\}$$

induces a clique of size $\sigma+1$ in G_σ and we are done. Now consider vertices

$$v_r, \quad v_s, \qquad m \leq r \leq \beta+l, \quad \beta+m+1 \leq s \leq 2\beta+m.$$

It follows that

$$(v_k, v_r, v_{\beta+m}, v_s)$$

is a path in G of length 3, and in view of (5.5), we must have that

$$(v_r, v_s) \in E_\sigma, \qquad m \leq r \leq \beta+l, \quad \beta+m+1 \leq s \leq 2\beta+m. \tag{5.6}$$

Since (5.6) trivially holds when $s = \beta+m$, it follows that

$$\{v_m, \ldots, v_{\beta+l}, v_{\beta+m}, \ldots, v_{2\beta+m}\}$$

induces a clique in G_σ of size $\sigma+1$ and of the form (5.4). Thus, in all cases there is a clique of size $\sigma+1$. □

Theorem 5.2. *Let $G=(V,E)$ be a band graph with bandwidth β. If $|V| \geq 3\beta+1$ then*

$$\chi_\sigma(G) = \chi(G^2) = 2\beta+1.$$

Proof. Since the size of a clique is a lower bound on the chromatic number of a graph, Lemma 5.1 shows that $2\beta+1 \leq \chi(G_\sigma)$ for every symmetric completion G_σ of G. Hence, Theorems 3.1 and 3.2 yield that

$$2\beta+1 \leq \chi_\sigma(G) \leq \chi(G^2). \tag{5.7}$$

To complete the proof, just note that the mapping ϕ defined on V by

$$\phi(v_i) = i \bmod (2\beta+1)$$

is a coloring of G^2, and hence, $\chi(G^2) \leq 2\beta+1$. This bound and (5.7) establish our result. □

6. Triangular substitution methods

The result of Section 5 shows that direct methods may not be able to take advantage of the symmetry of the matrix A. In this section we explore a type of indirect method which is able, in particular, to produce the desired results for banded matrices. We consider the lower triangular substitution methods of Powell and Toint (1979); upper triangular substitution methods are entirely analogous, but following Powell and Toint, we only consider the lower triangular methods.

Let A be a symmetric matrix and let L be the lower triangular part of A; that is, L is a lower triangular matrix such that $A-L$ is strictly upper triangular. A lower triangular substitution method is based on the result of Powell and Toint (1979) that if $C_1, C_2, \ldots, C_p$ is a consistent partition of the columns of L then A can be determined indirectly with p evaluations of Ad. It is not difficult to establish this result. With each group C associate a direction d with components $\delta_j \neq 0$ if j belongs to C and $\delta_j = 0$ otherwise. Then

$$Ad = \sum_{j \in C} \delta_j a_j$$

where $a_1, a_2, \ldots, a_n$ are the columns of A. To determine a_{ij} with $i \geq j$ note that if column j is the only column in group C with a nonzero in row $i \geq j$ then

$$(Ad)_i = \delta_j a_{ij} + \sum_{l>i, l \in C} \delta_l a_{li}. \tag{6.1}$$

This expression shows that a_{ij} depends on $(Ad)_i$ and on elements of L in rows $l > i$. Thus L can be determined indirectly by first determining the nth row of L and then solving for the remaining rows of L in the order $n-1, n-2, \ldots, 1$. Another consequence of (6.1) is that computing a_{ij} requires, at most, ρ_i operations where ρ_i is the number of nonzeroes in the i-th row of A. Thus computing all of A requires less than

$$\sum_{i=1}^{n} \rho_i^2$$

arithmetic operations, and this makes a triangular substitution method attractive in terms of the overhead. On the other hand, computing all of A with a direct method requires about τ arithmetic operations where τ is the number of nonzeroes in A. Another difference between direct methods and triangular substitution methods is that in a triangular substitution method the computation of a_{ij} requires a sequence of substitutions which may magnify errors considerably, while in a direct method there is no magnification of errors. Note, however, that Powell and Toint (1979) show that magnification of errors can only occur when the ratio of the largest to the smallest component of d is large.

Powell and Toint (1979) also noted that the number of groups in a consistent partition of the columns of L depends on the ordering of the rows and columns of A. Thus, if π is a permutation matrix and L_π is the lower triangular part of $\pi^{\mathrm{T}} A \pi$ then we may have

$$\chi(G_{\mathrm{U}}(L_\pi)) < \chi(G_{\mathrm{U}}(L)).$$

For example, if A has an arrowhead structure, then it is possible to choose the permutation π so that the chromatic number of $G_{\mathrm{U}}(L_\pi)$ is any integer in the interval $[2, n]$. Since Powell and Toint were unaware of the existence of the smallest–last ordering in the graph theory literature, it is interesting to note that the algorithm proposed by Powell and Toint (1979) for choosing the permutation matrix π is the smallest–last ordering of $G_{\mathrm{S}}(A)$. We have already seen in Section 4 that the smallest–last ordering is a useful ordering in connection with sequential coloring algorithms. There are good reasons for choosing this ordering to define the permutation matrix π; we shall return to this point later on in this section.

In graph theory terminology, the lower triangular substitution method of Powell and Toint consists of choosing the smallest–last ordering for the vertices of $G_S(A)$ and then coloring $G_U(L_\pi)$ with a sequential coloring algorithm. If column j is in position $\pi(j)$ of the smallest–last ordering of $G_S(A)$ then the permutation matrix π can be identified with the smallest–last ordering by setting the jth column of π to the $\pi(j)$th column of the identity matrix. Thus the vertices of $G_U(L_\pi)$ have the induced ordering

$$a_{\pi(1)}, a_{\pi(2)}, \ldots, a_{\pi(n)}. \tag{6.2}$$

Powell and Toint used the sequential coloring algorithm with this ordering to color $G_U(L_\pi)$. There is no compelling reason for using the sequential coloring algorithm with this ordering and, in fact, our numerical results show that the use of other orderings in the sequential coloring algorithm tends to reduce the number of evaluations of Ad needed by the triangular substitution method.

To further explore the properties of triangular substitution methods, we characterize a coloring of $G_U(L_\pi)$ as a restricted coloring of $G_S(A)$ in the following sense.

Definition. A mapping $\phi: V \to \{1, 2, \ldots, p\}$ is a *triangular p-coloring* of a graph $G = (V, E)$ if ϕ is a p-coloring of G and if there is an ordering $v_1, v_2, \ldots, v_n$ of the vertices of G such that ϕ is not a 2-coloring for any path (v_i, v_k, v_j) with $k > \max(i, j)$. The *triangular chromatic number* $\chi_\tau(G)$ of G is the smallest p for which G has a triangular p-coloring.

For some graphs it is not difficult to determine $\chi_\tau(G)$. For example, if G is a band graph with bandwidth β then $\chi_\tau(G) = 1 + \beta$. Band graphs thus show that the triangular chromatic number of a graph can be considerably smaller than the symmetric chromatic number of a graph. Our next result shows that the triangular chromatic number of $G_S(A)$ is the smallest number of evaluations of Ad needed to determine a symmetric matrix A with a triangular substitution method.

Theorem 6.1. *Let A be a symmetric matrix with nonzero diagonal elements. The mapping ϕ is a triangular coloring of $G_S(A)$ if and only if ϕ is a coloring of $G_U(L_\pi)$ for some permutation matrix π.*

Proof. It is sufficient to show that ϕ is a triangular coloring of $G_S(A)$ with the ordering $a_1, a_2, \ldots, a_n$ of the vertices of $G_S(A)$ if and only if ϕ is a coloring of $G_U(L)$.

First assume that ϕ is a triangular coloring of $G_S(A)$ and let (a_i, a_j) be an edge of $G_U(L)$. Then (a_i, a_j) is an edge of $G_S(A)$ or there is an index $k > \max(i, j)$ such that (a_i, a_k) and (a_k, a_j) are edges of $G_S(A)$. Since ϕ is a triangular coloring of $G_S(A)$, we must have that $\phi(a_i) \neq \phi(a_j)$. Hence ϕ is a coloring of $G_U(L)$.

Now assume that ϕ is a coloring of $G_U(L)$. Then ϕ is a coloring of $G_S(A)$ because $G_S(A)$ is a subgraph of $G_U(L)$ whenever A has nonzero diagonal elements. If (a_i, a_k, a_j) is a path in $G_S(A)$ with $k > \max(i, j)$ then (a_i, a_j) is an edge of $G_U(L)$ and hence $\phi(a_i) \neq \phi(a_j)$. Thus ϕ is not a 2-coloring of (a_i, a_k, a_j). $\square$

An important consequence of Theorem 6.1 is that it shows that triangular substitution methods are implicitly trying to solve a restricted graph coloring problem.

Triangular Graph Coloring Problem: Obtain a minimum triangular coloring of $G_S(A)$.

From the graph coloring point of view, it is clear that we may think of algorithms which determine a triangular coloring of $G_S(A)$ directly instead of first determining an ordering π and then coloring $G_U(L_\pi)$. We shall not pursue this type of algorithm in this paper; we restrict ourselves to triangular substitution methods.

The following result can be used to justify the choice of the smallest–last ordering to define the permutation matrix π.

Theorem 6.2. *Let $G = (V, E)$ be a graph with the vertices ordered $v_1, v_2, \ldots, v_n$. For any $W \subset V$ let $\delta(G[W])$ be the smallest degree in the subgraph induced by W, and let $d(w; W)$ be the degree of w in $G[W]$. If*

$$V_k = \{v_1, v_2, \ldots, v_k\},$$

then

$$\max\{\delta(G[W]) \colon W \subset V\} \le \max\{d(v_k; V_k) \colon 1 \le k \le n\}. \tag{6.3}$$

Equality holds in (6.3) *if $v_1, v_2, \ldots, v_n$ is a smallest–last ordering.*

Proof. Given $W \subset V$, let k be the smallest index such that $G[W] \subset G[V_k]$. Then $v_k \in W$ and hence

$$\delta(G[W]) \le d(v_k; W) \le d(v_k; V_k).$$

Thus (6.3) holds. Moreover, if $v_1, v_2, \ldots, v_n$ is a smallest–last ordering then

$$d(v_k; V_k) = \delta(G[V_k]),$$

so that equality holds in (6.3) for a smallest–last ordering.

It is interesting to interpret this result of Matula (1968) in terms of matrices. In this case $1 + d(v_k; V_k)$ is the number of nonzeroes in the kth row of the lower triangular part of the adjacency matrix. Thus, for any ordering π of the columns of a symmetric matrix A, the smallest–last ordering minimizes the maximum number of nonzeroes in any row of L_π. This result was established independently by Powell and Toint (1979).

Another interesting consequence of Theorem (6.2) can be obtained by noting that if ϕ is a triangular coloring of G, and if $v_1, v_2, \ldots, v_n$ is the associated ordering of the vertices of G, then ϕ requires at least

$$\max\{1+d(v_k;\, V_k)\colon 1 \leqslant k \leqslant n\}. \tag{6.4}$$

colors. This is not difficult to prove. If $l = d(v_k;\, V_k)$ then there are l vertices in V_k adjacent to v_k. Without loss of generality, assume that $v_1, \ldots, v_l$ are adjacent to v_k. Hence, (v_i, v_k, v_j) is a path in G for $1 \leqslant i < j \leqslant l$. Since ϕ is a coloring of G we must have that $\phi(v_k) \neq \phi(v_i)$ for $1 \leqslant i \leqslant l$, and since ϕ is a triangular coloring of G we must also have that

$$\phi(v_i) \neq \phi(v_j), \quad i \neq j,\ i, j = 1, 2, \ldots, l.$$

Thus ϕ needs at least $1 + l$ colors, as we wanted to show.

We have shown that any triangular coloring of G requires at least (6.4) colors. Since equality holds in (6.3) for a smallest–last ordering, we also have that

$$\max\{1+d(v_k;\, V_k)\colon 1 \leqslant k \leqslant n\} \leqslant \chi_\tau(G) \tag{6.5}$$

holds for a smallest–last ordering. Clearly, we do not want an ordering which violates this inequality, so from this point of view the smallest–last ordering is quite satisfactory.

Theorem 6.3. *Let $G = (V, E)$ be a graph. Then*

$$\chi(G) \leqslant \max\{1 + \delta(G[W])\colon W \subset V\} \leqslant \chi_\tau(G) \leqslant \chi(G^2).$$

Proof. The first inequality is a standard upper bound on the chromatic number of a graph due to Szekeres and Wilf (1968). This inequality is also a consequence of the result that the number of colors required by a sequential coloring algorithm with the smallest–last ordering is bounded by (4.3). The second inequality is an immediate consequence of Theorem 6.2 and the fact that (6.5) holds for a smallest–last ordering. The third inequality follows since if ϕ is a coloring of G^2 then ϕ is a triangular coloring of G. $\square$

7. The triangular chromatic number

We have shown that the triangular chromatic number of $G_S(A)$ is the chromatic number of $G_U(L_\pi)$ for some permutation matrix π. In this section we consider the problem of determining the chromatic number of $G_U(L)$ for an arbitrary lower triangular matrix L and the problem of determining the triangular chromatic number of a general graph G. We show that both of these problems are just as hard as the general graph coloring problem.

We first prove that given a p-colorable graph G with $p \geqslant 3$, we can construct a lower triangular matrix L so that $G_U(L)$ has the same chromatic number as G.

Theorem 7.1. *Let $G=(V,E)$ be a graph. If $\chi(G)\geq 3$ then there is a lower triangular matrix L of order $|V|+|E|$ such that*

$$\chi(G_U(L))=\chi(G). \tag{7.1}$$

Proof. Let $v_1,\ldots,v_n$ be the vertices of G, and let $e_1,\ldots,e_m$ be the edges of G. Now define an m by n matrix B by setting

$$b_{li}=b_{lj}=1,\quad b_{lk}=0,\qquad k\neq i,j,$$

for each edge $e_l=(v_i,v_j)$ of G, and let

$$L=\begin{pmatrix} I_n & \\ B & I_m\end{pmatrix} \tag{7.2}$$

where I_n and I_m are the identity matrices of order n and m, respectively. It should now be clear that G is a subgraph of $G_U(L)$ and that (7.1) holds. □

Theorem 7.1 extends a result of Coleman and Moré (1983) in which it is shown that (7.1) holds for a general matrix. It is interesting to note that the lower triangular matrix (7.2) is quite sparse; it has at most 3 nonzero elements per row. Theorem 7.1 shows that even if we were able to determine the correct permutation matrix π, determining the chromatic number of $G_U(L_\pi)$ is still an intractable problem. In particular, any polynomial algorithm for determining the chromatic number of a graph is bound to fail on graphs $G_U(L)$ where L is of the form (7.2).

We now prove that the determination of the triangular chromatic number is a difficult problem even if the graph is bipartite.

Theorem 7.2. *Let $G=(V,E)$ be a graph. If $\chi(G)\geq 3$ then there is a bipartite graph B with $|V|(1+|E|)$ vertices such that*

$$\chi_\tau(B)=\chi(G). \tag{7.3}$$

Proof. The proof is very similar to that of Theorem 3.3. Let $v_1,\ldots,v_n$ be the vertices of G, and let $e_1,\ldots,e_m$ be the edges of G. For each edge $e_l=(v_i,v_j)$ define a bipartite graph B_l with vertices

$$\{v_i, v_j, w_1^{(l)},\ldots,w_n^{(l)}\}$$

and edges

$$(v_i,w_k^{(l)}),\quad (v_j,w_k^{(l)}),\qquad k=1,\ldots,n.$$

Now define a bipartite graph B by setting

$$V(B)=V(G)\cup\{w_k^{(l)}: 1\leq k\leq n, 1\leq l\leq m\}$$

and

$$E(B)=\bigcup_{l=1}^{m} E(B_l).$$

We now show that (7.3) holds for this bipartite graph B. To prove that

$$\chi_\tau(B) \leq \chi(G), \tag{7.4}$$

let ϕ be a coloring of G, and extend ϕ to a coloring of B by setting $\phi(w_k^{(l)})$ to any color that does not agree with $\phi(v_i)$ or $\phi(v_j)$. Since $\chi(G) \geq 3$ this is possible. We now show that the extended ϕ is a triangular coloring of B for any ordering of the vertices of B which orders the vertices of G first. To establish this claim note that the only paths in B of length 2 are of the form

$$(v_i, w_k^{(l)}, v_j), \quad e_l = (v_i, v_j) \in E, \tag{7.5}$$

or of the form

$$(w_i^{(r)}, v_k, w_j^{(s)}). \tag{7.6}$$

If the path is of the form (7.5) then $\phi(v_i) \neq \phi(v_j)$ because ϕ is a coloring of G and (v_i, v_j) is an edge of G. Thus ϕ is not a 2-coloring of paths of the form (7.5). On the other hand, ϕ can be a 2-coloring for paths of the form (7.6) because the vertices of G are ordered first. Hence ϕ is a triangular coloring of B and thus (7.4) follows. To complete the proof we now show that

$$\chi(G) \leq \chi_\tau(B). \tag{7.7}$$

If $\chi_\tau(B) \geq n$ then (7.7) holds trivially, so assume that $\chi_\tau(B) < n$. Let ϕ be an optimal triangular coloring of B and let π be an ordering of the vertices of B associated with the triangular coloring ϕ. We now show that ϕ is a coloring of G. Assume that $\phi(v_i) = \phi(v_j)$ for some edge $e_l = (v_i, v_j)$ of G. Since $(v_i, w_k^{(l)}, v_j)$ is a path in B and ϕ is a triangular coloring of G we must have that

$$\pi(w_k^{(l)}) < \max\{\pi(v_i), \pi(v_j)\}, \quad 1 \leq k \leq n. \tag{7.8}$$

Moreover, since $(w_r^{(l)}, v_i, w_s^{(l)})$ and $(w_r^{(l)}, v_j, w_s^{(l)})$ are paths in B and (7.8) holds, then

$$\phi(w_r^{(l)}) \neq \phi(w_s^{(l)}), \quad r \neq s, \; r, s = 1, 2, \ldots, n.$$

Thus ϕ uses at least n colors. This contradicts the assumption that $\chi_\tau(B) < n$, so we must have that $\phi(v_i) \neq \phi(v_j)$ whenever (v_i, v_j) is an edge of G. Thus ϕ is a coloring of G and as a consequence (7.7) holds. □

The similarity between the proofs of Theorem 7.2 and Theorem 3.3 is apparent. In particular, we have used the same bipartite graph B in both proofs, and we have shown that

$$\chi_\tau(B) = \chi_\sigma(B) = \chi(G).$$

The main difference between the two proofs is that in Theorem 3.3 we argue in terms of completions while in Theorem 7.2 we use colorings. It should be clear, however, that we can define a triangular completion of a graph and prove a result analogous to Theorem 3.2.

8. Numerical results

We have examined several algorithms for determining symmetric matrices; the direct methods of Section 4 and the triangular substitution methods of Section 6. We have already determined that the Powell–Toint direct method had the best overall performance of the direct methods. In this section we present numerical results for the triangular substitution methods and compare their performance with the Powell–Toint method.

The numerical results for the algorithms under consideration appear in Table 3. The totals for Table 3 are in Table 4. The graphs used in these results are the same as those in Section 4. They are graphs G of the form $G_S(A)$ where A is a sparse symmetric matrix of order n with a sparsity pattern from the Everstine (1979) collection.

Table 3
Direct and triangular substitution methods

n	matd(%)	maxr	dpt	slpt	slsl
59	4.68	4	6	4	4
66	4.43	3	6	3	3
72	2.84	3	3	3	3
87	4.15	5	10	6	7
162	2.56	5	10	6	6
193	4.95	12	27	17	17
198	2.03	5	9	6	6
209	2.23	7	13	9	9
221	1.89	5	10	7	6
234	0.98	3	5	6	4
245	1.42	6	10	7	7
307	1.50	6	10	8	7
310	1.43	5	9	7	7
346	1.49	7	15	12	11
361	1.27	5	10	7	7
419	1.13	7	12	8	8
492	0.75	5	9	6	6
503	1.29	9	20	14	13
512	0.77	7	13	10	10
592	0.81	6	11	9	8
607	0.78	6	13	9	8
758	0.59	5	10	7	7
869	0.54	6	11	8	7
878	0.54	5	11	7	7
918	0.49	6	11	8	7
992	0.90	10	18	14	14
1005	0.48	10	20	14	13
1007	0.47	5	11	7	7
1242	0.38	7	12	8	8
2680	0.19	7	14	11	9

All of the triangular substitution methods that we consider use the smallest–last ordering to define the permutation matrix π and then use a sequential coloring algorithm to color $G_U(L_\pi)$. They only differ in the ordering used by the sequential coloring algorithm.

For each problem Table 3 presents the dimension n of the problem, the density matd of the matrix L_π, the maximum number of nonzeroes maxr in any row of the matrix L_π, and the number of colors required by the coloring algorithms.

The dpt algorithm of Table 3 is an implementation of the Powell–Toint direct method described in Section 4. The slpt algorithm is the triangular substitution method of Powell and Toint (1979). As noted in Section 6, this algorithm uses the sequential coloring algorithm with the induced ordering (6.2) to color $G(L_\pi)$. The slsl algorithm uses the sequential coloring algorithm with the smallest–last ordering to color $G(L_\pi)$.

Tables 3 and 4 show that triangular substitution methods are able to determine the symmetric matrix A with fewer evaluations of Ad. Also note that triangular substitution methods represent an improvement of 45% over the slo algorithm of Section 4. Thus triangular substitution methods fulfill the expected improvement over a method that disregards symmetry.

Table 4
Totals for Table 3

maxr	dpt	slpt	slsl
182	349	248	236

Since Theorems 6.2 and 6.3 yield that maxr is a lower bound on $\chi_\tau(G)$, these numerical results show that the triangular substitution methods slpt and slsl are nearly optimal on these problems. On the average, the slsl algorithm is less than two colors away from $\chi_\tau(G)$. In the same vein, note that the slo algorithm is, on the average, less than one color away from $\chi(G^2)$.

Finally note that with the exception of one problem, the slsl algorithm never performs worse than the slpt algorithm.

9. Concluding remarks

We have analyzed direct and indirect methods for determining symmetric matrices. The emphasis is on methods which can be efficiently and reliably implemented in a computing environment. We have found that the triangular substitution method slsl requires the least number of evaluations of Ad to determine A, and that slsl is always nearly optimal on our test problems.

Although triangular substitution methods can determine A in nearly optimal fashion, recall that we mentioned at the beginning of Section 6 that the cost of obtaining A and the errors involved in determining A are higher with triangular substitution methods than with direct methods. Thus, it seems that it would be useful to study direct methods further. In this vein, note that the example at the end of Section 2 shows that general direct methods can provide vast improvements on direct methods based on partitions of the columns of A.

It may also be worthwhile to study triangular substitution methods further. A topic of interest is the existence of other reasonable choices for the ordering that defines the permutation matrix π. We have obtained good results if π is chosen via the incidence degree ordering of Coleman and Moré (1983). If the sequential coloring algorithm with the smallest-last ordering is then used to color $G[L_\pi]$, then this algorithm needs a total of 233 colors for the problems of Section 8. We have not presented detailed numerical results for this algorithm because the theoretical justification for the incidence degree ordering is not strong enough.

There are other open questions which deserve further study and which could lead to useful improvements on current algorithms. For instance, it would be interesting to investigate the ratio

$$\rho \equiv \frac{\chi_\sigma(G)}{\chi_\tau(G)}.$$

Our numerical results suggest that $\rho \geq 1$, but we have been unable to establish this conjecture. Similarly, we don't know if $\rho < 2$ always. Note that band graphs show that $2-\rho$ may be arbitrarily small.

References

T.F. Coleman and J.J. Moré, "Estimation of sparse Jacobian matrices and graph coloring problems", *SIAM Journal on Numerical Analysis* 20 (1983) 187–209.

T.F. Coleman and J.J. Moré, "Software for estimating sparse Jacobian matrices", Technical Report ANL-82-37, Argonne National Laboratory (Argonne, Illinois, 1982).

A.R. Curtis, M.J.D. Powell and J.K. Reid, "On the estimation of sparse Jacobian matrices", *Journal of the Institute of Mathematics and its Applications* 13 (1974) 117–119.

S. Eisenstat, Personal communication (1980).

G.C. Everstine, "A comparison of three resequencing algorithms for the reduction of matrix profile and wavefront", *International Journal on Numerical Methods in Engineering* 14 (1979) 837–853.

M.R. Garey and D.S. Johnson, Computers and intractability (W.H. Freeman, San Francisco, CA, 1979).

D.W. Matula, G. Marble and J.D. Isaacson, "Graph coloring algorithms", in: R. Read, ed., *Graph theory and computing* (Academic Press, New York, 1972), pp. 104–122.

D.W. Matula and L.L. Beck, "Smallest-last ordering and clustering and graph coloring algorithms", *Journal of the Association for Computing Machinery* 30 (1983) 417–427.

S.T. McCormick, "Optimal approximation of sparse Hessians and its equivalence to a graph coloring problem", *Mathematical Programming* 26 (1983), 153–171.

G.N. Newsam and J.D. Ramsdell, "Estimation of sparse Jacobian matrices", *SIAM Journal on Algebraic and Discrete Methods* (1983), to appear.

M.J.D. Powell and Ph.L. Toint, "On the estimation of sparse Hessian matrices", *SIAM Journal on Numerical Analysis* 16 (1979) 1060–1074.

G. Szekeres and H.S. Wilf, "An inequality for the chromatic number of a graph", *Journal on Combinatorial Theory* 4 (1968) 1–3.

M.N. Thapa, "Optimization of unconstrained functions with sparse Hessian matrices: Newton-type methods", Technical Report SOL 82-8, Systems Optimization Laboratory, Stanford University (Stanford, CA, 1982).