An Algorithm for Lagrangian Subspaces

Sönke Hansen
Fb 17 Mathematik-Informatik
Universität-GHS Paderborn
Warburger Strasse 100
D-4790 Paderborn, Federal Republic of Germany

Submitted by Ludwig Elsner

ABSTRACT

Lagrangian subspaces of the symplectic vector space $\mathbb{R}^n \times \mathbb{R}^n$ can be represented by symmetric matrices. An algorithm for computing such a representation is described and analysed. It is useful for paraxial ray tracing.

1. INTRODUCTION

An n-dimensional linear subspace $\lambda \subset \mathbb{R}^n \times \mathbb{R}^n$ is called Lagrangian if the standard symplectic form vanishes on λ:

$$\sum_{j=1}^{n} (\xi_j y_j - \eta_j x_j) = 0 \qquad \text{if} \quad (x, \xi), (y, \eta) \in \lambda.$$

For every Lagrangian subspace λ there exist symmetric $n \times n$ matrices T and L such that

$$\lambda = \{(x, \xi) \in \mathbb{R}^n \times \mathbb{R}^n;\ y + L\eta = 0 \text{ with } (y, \eta) = (x, \xi + Tx)\}. \qquad (1)$$

T is regarded as defining a linear symplectic change of coordinates from (x, ξ) to (y, η). When T is held fixed, L is uniquely determined by λ. The map taking λ to (the matrix elements of) L introduces local coordinates on the Lagrangian Grassmannian

$$\Lambda(n) = \{\lambda;\ \lambda \subset \mathbb{R}^n \times \mathbb{R}^n \text{ Lagrangian subspace}\}.$$

LINEAR ALGEBRA AND ITS APPLICATIONS 129:155–176 (1990)

0024-3795/90/$3.50

Equipped with these coordinates $\Lambda(n)$ becomes a $n(n+1)/2$-dimensional real analytic manifold [1].

In this paper the following problem is studied. Given a Lagrangian subspace λ, how can matrices T and L representing λ as in (1) be computed numerically? Phrased differently, an algorithm which selects a coordinate patch on $\Lambda(n)$ valid around a given $\lambda \in \Lambda(n)$ and which then changes to these coordinates is called for. For the coordinates to be acceptable the norms of the matrices T and L may not be large. An algorithm doing this is described in Section 3 and analysed in Section 4.

The problem considered here naturally arises from dynamic and paraxial ray tracing, which is done in optics and in seismics. There one traces, along a central ray, wavefront curvatures, amplitudes, and rays infinitesimally close to the central or axial ray [2, 3, 6]. Classical methods have difficulties at caustics. Following the ideas of Maslov [7], these are overcome when working with Lagrangian manifolds. This approach leads to the consideration of curves of Lagrangian subspaces, the tangent spaces to a Lagrangian manifold along a bicharacteristic. It then becomes necessary to change coordinates on $\Lambda(n)$ because a curve in $\Lambda(n)$ may leave a coordinate patch. The algorithm presented here automatizes this change. Such an application is outlined in Section 6.

2. AUXILIARY RESULTS

$\mathbb{M}_{n,m}$ denotes the set of real $n \times n$ matrices; $\mathbb{M}_n = \mathbb{M}_{n,n}$. $I = I_n$ is the unit matrix. Uppercase letters are used for matrices, and the corresponding subscripted lowercase letters for their elements.

Recall some definitions and facts from symplectic linear algebra [1, 5]. A linear transformation on $\mathbb{R}^n \times \mathbb{R}^n$ is called symplectic if it leaves the symplectic form

$$\sigma((x,\xi),(y,\eta)) = \sum_{j=1}^{n} (\xi_j y_j - \eta_j x_j)$$

invariant. Correspondingly, a matrix $S \in \mathbb{M}_{2n}$ is called symplectic if $S^T J S = J$ holds with

$$J = \begin{pmatrix} 0 & -I_n \\ I_n & 0 \end{pmatrix}.$$

Examples of symplectic matrices are

$$\begin{pmatrix} A & 0 \\ 0 & A^{-T} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} I_n & 0 \\ B & I_n \end{pmatrix}$$

where A is nonsingular and B symmetric. J is also symplectic. The symplectic matrices form a group.

A linear subspace $\lambda \subset \mathbb{R}^n \times \mathbb{R}^n$ is called Lagrangian if it is isotropic, i.e. $\sigma((x,\xi),(y,\eta)) = 0$ for $(x,\xi),(y,\eta) \in \lambda$, and n-dimensional. Symplectic linear transformations map Lagrangian subspaces to Lagrangian subspaces. The twisted graph

$$G'(\chi) = \{(y, x, \eta, -\xi);\ (y,\eta) = \chi(x,\xi)\}$$

of a linear transformation χ on $\mathbb{R}^n \times \mathbb{R}^n$ is a Lagrangian subspace of $\mathbb{R}^{2n} \times \mathbb{R}^{2n}$ if and only if χ is symplectic.

In the following a pair $(A, B) \in \mathbb{M}_n \times \mathbb{M}_n$ is called an *L-pair* if the matrix $(A\ B)$ has rank equal to n and AB^T is symmetric. Note that AB^T is symmetric if and only if the rows of $(A\ B)$ are orthogonal with respect to the symplectic form σ,

$$\sum_{l=1}^{n} (b_{il}a_{jl} - b_{jl}a_{il}) = 0 \qquad \text{for} \quad i, j = 1, \ldots, n. \tag{2}$$

The notion of L-pairs is used here to characterize Lagrangian subspaces in terms of matrices.

LEMMA 1. *Let λ be a linear subspace of $\mathbb{R}^n \times \mathbb{R}^n$. Then λ is Lagrangian if and only if there is an L-pair (A, B) such that*

$$\lambda = \{(x,\xi) \in \mathbb{R}^n \times \mathbb{R}^n;\ Ax + B\xi = 0\}. \tag{3}$$

Proof. Assume λ Lagrangian. Choose a basis $l_1, \ldots, l_n$ for λ. Consider the matrices $A, B \in \mathbb{M}_n$ for which $l_1^T J, \ldots, l_n^T J$ are the rows of $(A\ B)$. Then (A, B) is an L-pair satisfying (3). Conversely, given an L-pair (A, B) representing λ as in (3), then λ is generated by the columns of $J(A\ B)^T$. The proof is complete. ∎

The following fact is an immediate corollary to Lemma 1. For an L-pair (A, B), $E \in \mathbb{M}_n$ nonsingular, and $S \in \mathbb{M}_{2n}$ symplectic, the pair $(\tilde{A}, \tilde{B})$ defined by

$$(\tilde{A} \quad \tilde{B}) = E(A \quad B)S$$

also is an L-pair.

L-pairs are introduced merely for technical convenience. More useful representations of Lagrangian subspaces are obtained by aiming at L-pairs of the form $(A, B) = (I, L)$.

The algorithm presented in Section 3 is based on two simple results about L-pairs.

LEMMA 2. *Let (A, B) be an L-pair. Then, for all j, the jth columns of A and B cannot vanish simultaneously.*

Proof. Let $e_j \in \mathbb{R}^n$ denote the standard unit vector with 1 in the jth component and 0 in all others. The jth columns of A and B are produced by applying $(A \; B)$ to the vectors $(e_j, 0)$ and $(0, e_j)$, respectively. If these columns of A and B were both zero, the vectors $(e_j, 0)$ and $(0, e_j)$ would belong to the Lagrangian subspace defined by (A, B). This however would contradict $\sigma((e_j, 0), (0, e_j)) \neq 0$. The proof is complete. ∎

LEMMA 3. *Let (A, B) be an L-pair. Let $k \in \{1, \ldots, n\}$, and let*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

be the block partitions associated with the splitting $\mathbb{R}^n = \mathbb{R}^k \times \mathbb{R}^{n-k}$. Assume A_{11} nonsingular and $A_{21} = 0$. Then (A_{22}, B_{22}) is an L-pair.

Proof. Assume, without loss of generality, $A_{11} = I_k$. The symmetry of $A_{22}B_{22}^T$ follows from the symmetry of

$$AB^T = \begin{pmatrix} * & * \\ * & A_{22}B_{22}^T \end{pmatrix}.$$

Let $i > k$ and $j \leqslant k$. Use Equation (2) to solve for b_{ij}:

$$b_{ij} = \sum_{l=k+1}^{n} b_{jl} a_{il} - \sum_{l=k+1}^{n} a_{jl} b_{il}.$$

This means that the columns of B_{21} are linearly dependent on those of A_{22} and B_{22}. Hence $(A_{22} \;\; B_{22})$ has rank equal to $n-k$. The proof of the lemma is complete. ■

3. THE REPRESENTATION ALGORITHM

Let (A, B) be an L-pair of $n \times n$ matrices. The algorithm described here produces symmetric $n \times n$ matrices T and L such that

$$(I \quad L) = E(A \quad B)\begin{pmatrix} I & 0 \\ -T & I \end{pmatrix} \tag{4}$$

holds with a nonsingular matrix E. Equation (4) implies the desired representation (1) of the Lagrangian subspace λ given by (3).

A simplified version of the algorithm designed to arrive at Equation (4) starts as follows. The first column of A is made nonzero by adding or subtracting the first column of B. Lemma 2 is used here. Then A and B are multiplied from the left with an orthogonal matrix such that the first column of A becomes $(\alpha, 0, \ldots, 0)^T$ with $\alpha \neq 0$. Now Lemma 3 implies that the task is reduced to treating an L-pair with dimension decreased by 1.

A pivoting strategy is employed in the actual algorithm. The algorithm takes n steps to compute L-pairs $(A^{(1)}, B^{(1)}), \ldots, (A^{(n)}, B^{(n)})$ which relate to $(A^{(0)}, B^{(0)}) = (A, B)$, and are such that the first k columns of $A^{(k)}$ form an upper-triangular matrix with rank equal to k. In particular, $A^{(n)}$ is nonsingular upper triangular. With an orthogonal matrix $Q^{(k)}$, a permutation matrix $P^{(k)}$, and a diagonal matrix $D^{(k)}$ the kth step is

$$(A^{(k)} \quad B^{(k)}) = Q^{(k)}(A^{(k-1)} \quad B^{(k-1)})\begin{pmatrix} P^{(k)} & 0 \\ -P^{(k)}D^{(k)} & P^{(k)} \end{pmatrix}. \tag{5}$$

The $n \times n$ matrices $Q^{(k)}$, $P^{(k)}$, and $D^{(k)}$ are chosen in the following way. Assume that $(A^{(k-1)}, B^{(k-1)})$ is an L-pair with $a_{jj}^{(k-1)} \neq 0$ and $a_{ij}^{(k-1)} = 0$ if $j < k$ and $i > j$. The matrices $(a_{ij}^{(k-1)})_{i,j \geqslant k}$ and $(b_{ij}^{(k-1)})_{i,j \geqslant k}$ form an L-pair.

This follows from Lemma 3. Now choose $p = p_k \in \{k, \ldots, n\}$ and $s = s_k \in \{1, -1\}$ for which the sum

$$\sum_{i=k}^{n} \left(a_{ip}^{(k-1)} + sb_{ip}^{(k-1)}\right)^2$$

is maximal. It follows from Lemma 2 that this sum is positive. Set

$$Q^{(k)} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H^{(k)} \end{pmatrix},$$

where $H^{(k)}$ is an orthogonal matrix which transforms the $(n-k+1)$-dimensional vector

$$\left(a_{kp}^{k-1)} + sb_{kp}^{(k-1)}, \ldots, a_{np}^{(k-1)} + sb_{np}^{(k-1)}\right)^T,$$

$p = p_k$, $s = s_k$, into the nonzero vector

$$\left(a_{kk}^{(k)}, 0, \ldots, 0\right)^T.$$

This involves a QR factorization technique. $H^{(k)}$ may be obtained as a Householder reflection matrix; see [8]. Let $P^{(k)}$ be the matrix representing the permutation which interchanges k with p_k and leaves all other indices fixed. Let $D^{(k)}$ be the $n \times n$ matrix having $d_{kk}^{(k)} = -s_k$ as its only nonzero entry. Now define $(A^{(k)}, B^{(k)})$ by Equation (5). It is clear from the construction that $a_{jj}^{(k)} \neq 0$ and $a_{ij}^{(k)} = 0$ for $j \leqslant k$ and $i > j$. The last matrix on the right in equation (5) is symplectic. Therefore $(A^{(k)}, B^{(k)})$ also is an L-pair.

From the equations (5) and the initial condition $(A^{(0)}, B^{(0)}) = (A, B)$ a direct relation between (A, B) and $(A^{(n)}, B^{(n)})$ follows by induction:

$$\begin{pmatrix} A^{(n)} & B^{(n)} \end{pmatrix} = Q \begin{pmatrix} A & B \end{pmatrix} \begin{pmatrix} P & 0 \\ -PT^{(n)} & P \end{pmatrix}. \tag{6}$$

Here $Q = Q^{(n)} \cdots Q^{(1)}$, $P = P^{(1)} \cdots P^{(n)}$, and $T^{(n)}$ is obtained from the recursion

$$T^{(k)} = P^{(k)T} T^{(k-1)} P^{(k)} + D^{(k)}$$

starting with $T^{(0)} = 0$. It is easily proved by induction that the diagonal

elements $t_{jj}^{(k)}$ with $j \leqslant k$ are the only matrix elements in $T^{(k)}$ which can have nonzero values. It follows that $T^{(k-1)}$ and $P^{(k)}$ commute. Hence

$$T^{(k)} = D^{(k)} + \cdots + D^{(0)}. \tag{7}$$

Finally a back substitution followed by a removal of the permutation is performed. This amounts to a multiplication of Equation (6) from the right with

$$\begin{pmatrix} P^T & 0 \\ 0 & P^T \end{pmatrix}$$

and from the left with $PA^{(n)^{-1}}$. The result of this is Equation (4) with $L = PA^{(n)^{-1}}B^{(n)}PT$ and

$$T = PT^{(n)}P^T. \tag{8}$$

The description of the algorithm is complete.

THEOREM 1. *Let A and B be real $n \times n$ matrices such that*

$$\lambda = \{(x, \xi) \in \mathbb{R}^n \times \mathbb{R}^n;\ Ax + B\xi = 0\}$$

is a Lagrangian subspace of $\mathbb{R}^n \times \mathbb{R}^n$. The algorithm in this section determines real symmetric $n \times n$ matrices T and L such that λ is given by the equations

$$\lambda: y + L\eta = 0$$

in the symplectic coordinates $(y, \eta) = (x, \xi + Tx)$. T is a diagonal matrix having only $+1$ or -1 on its diagonal.

Proof. (A, B) is an L-pair. From the description of the algorithm it is already clear that it arrives at a representation (1) for λ with $T, L \in \mathbf{M}_n$ symmetric. Equations (7) and (8) imply the assertion about T. ∎

REMARK. If (A, B) is an L-pair such that A is nonsingular, one can apply QR factorization followed by back substitution directly and arrive at (1) with $T = 0$ and $L = A^{-1}B$. The number of essential operations in floating-point arithmetic needed for this is about $5n^3/3$. The additional work

needed for computing and updating the scalar products used for pivot selection plus that for adding the columns amounts to about $7n^2$ operations. Thus the algorithm presented here, which does the selection of and the change to new coordinates in one stroke, uses essentially the same amount of time as an algorithm which only changes to predetermined coordinates—the case where T is known beforehand. In principle the desired coordinate change on $\Lambda(n)$ could also be achieved by checking through all the 2^n possible cases for T. Evidently such a procedure is not very reasonable for $n \geqslant 3$, say.

REMARK. Let λ be a Lagrangian subspace. Let $T = T_0$ and $L = L_0$ be symmetric matrices representing λ as in (1). Then (3) holds with the L-pair $(A, B) = (I + L_0T_0, L_0)$. Starting from this L-pair, the algorithm constructs matrices T_1 and L_1 such that λ is represented as in (1) with $T = T_1$ and $L = L_1$. Applied in this way, the algorithm computes a change of coordinates on $\Lambda(n)$. T_1 will differ from T_0 if $T_0 = \operatorname{diag}(\pm 1, \ldots, \pm 1)$ and if L_0 has a column with Euclidean norm greater than 1. To see this look at the value of s_1 determined in the first step of the algorithm. When used in this way the algorithm will be called the transformation algorithm.

4. ELEMENT GROWTH

The algorithm described in the previous section leads to a representation of a Lagrangian subspace in the desired form (1). It also works without pivoting, i.e. when $p_k = k$ is chosen. However, the pivoting strategy adopted in the algorithm makes it possible to derive a bound on the norm of the matrix L. The bound only depends on the dimension n.

THEOREM 2. *The symmetric matrix L computed by the algorithm in Section 3 is bounded:*

$$|l_{ij}| \leqslant n\, 4^{n-1} \qquad \text{for} \quad i, j = 1, \ldots, n.$$

Proof. The main task is to estimate the absolute values of the elements of the matrices $A^{(n)}$ and $B^{(n)}$ by the absolute value of the diagonal element in $A^{(n)}$ in the same row. To simplify the notation assume that $P^{(k)} = I$, i.e. $p_k = k$, holds for $k = 1, \ldots, n$. Essentially no generality is lost when assuming

this. Let $i, j \in \{1, \ldots, n\}$. Recall the steps in the algorithm. Observe that $a_{ij}^{(n)} = a_{ij}^{(i)}$ if $j \leqslant i$, $a_{ij}^{(n)} = a_{ij}^{(j)}$, $b_{ij}^{(n)} = b_{ij}^{(i)}$. The elements of $A^{(n)}$ satisfy

$$\begin{aligned} a_{ij}^{(n)} &= 0 \qquad \text{if} \quad j < i, \\ |a_{ij}^{(n)}| &< |a_{ii}^{(n)}| \qquad \text{if} \quad i < j. \end{aligned} \tag{9}$$

The inequality follows from the following chain of estimates valid if $i < j$:

$$\begin{aligned} |a_{ii}^{(i)}|^2 &\geqslant \sum_{l=i}^{n} \left(a_{lj}^{(i-1)} + s_j b_{lj}^{(i-1)} \right)^2 \\ &= \sum_{l=i}^{n} \left(a_{lj}^{(j-1)} + s_j b_{lj}^{(j-1)} \right)^2 \\ &= \sum_{l=i}^{n} \left(a_{lj}^{(j)} \right)^2 \\ &\geqslant |a_{jj}^{(j)}|^2 + |a_{ij}^{(j)}|^2 \\ &> |a_{ij}^{(j)}|^2. \end{aligned}$$

The first inequality is a consequence of pivoting in step i of the algorithm. The equality following it holds because the scalar products between the projections to the components i to n of the jth columns of $A^{(k)}$ and $B^{(k)}$ remain unchanged when stepping with k from $i-1$ to $j-1$. The other equality is clear from the definition of step j in the algorithm. The last inequality follows because $a_{jj}^{(j)} \neq 0$.

The pivoting in step i also implies

$$|a_{ii}^{(i)}|^2 \geqslant \sum_{l=i}^{n} |b_{lj}^{(i)}|^2 \qquad \text{if} \quad j \geqslant i.$$

So, in particular,

$$|b_{ij}^{(n)}| \leqslant |a_{ii}^{(n)}| \qquad \text{if} \quad j \geqslant i. \tag{10}$$

It remains to estimate $|b_{ij}^{(n)}|$ for $j < i$. To do this it is convenient to pass to the

scaled matrices $\tilde{A}$ and $\tilde{B}$, arising from $A^{(n)}$ and $B^{(n)}$, respectively, by dividing the ith rows by $a_{ii}^{(n)}$. $(\tilde{A}, \tilde{B})$ is an L-pair. Now (9) and (10) become

$$\begin{aligned} \tilde{a}_{ii} &= 1, \\ \tilde{a}_{ij} &= 0 \qquad \text{if} \quad i > j, \\ |\tilde{a}_{ij}| &< 1 \qquad \text{if} \quad i < j, \\ |\tilde{b}_{ij}| &\leqslant 1 \qquad \text{if} \quad i \leqslant j. \end{aligned} \tag{11}$$

Assume $j < i$. The rows of $(\tilde{A}\ \tilde{B})$ are orthogonal with respect to the symplectic form σ:

$$\sum_{k=1}^{n} \left(\tilde{b}_{ik}\tilde{a}_{jk} - \tilde{b}_{jk}\tilde{a}_{ik} \right) = 0.$$

Use the equations in (11) to solve this equation for $\tilde{b}_{ij}$. Estimate this solution and obtain, using the inequalities in (11),

$$\begin{aligned} |\tilde{b}_{ij}| &\leqslant \sum_{k>j} |\tilde{b}_{ik}\tilde{a}_{jk}| + \sum_{k \geqslant i} |\tilde{b}_{jk}\tilde{a}_{ik}| \\ &\leqslant \sum_{j<k<i} |\tilde{b}_{ik}| + 2(n-i+1) \\ &= \beta_{ij}. \end{aligned}$$

The last equation is a definition. Add β_{ij} to the inequality just shown, and get $\beta_{ij-1} \leqslant 2\beta_{ij}$. Hence

$$|\tilde{b}_{ij}| \leqslant n\, 2^{n-1}. \tag{12}$$

Similar arguments apply to the back substitution, i.e. to the passage from $(\tilde{A}, \tilde{B})$ to (I, L). Again use (11) and obtain

$$\max_{i,j} |l_{ij}| \leqslant 2^{n-1} \max_{i,j} |\tilde{b}_{ij}|.$$

Now combine this with (12). The proof of the theorem is complete. ■

REMARK. It is not clear how large the computed matrices L can actually become. Certainly the bound derived in the theorem is overly pessimistic. Numerical experiments were carried out to see how the algorithm performed in practice. For the task described in the remark at the end of the last section tests with collections of randomly generated matrices $T_0 = \operatorname{diag}(\pm 1, \cdots, \pm 1)$ and L_0 symmetric were made. it was found that the resulting matrices L_1 had row sum norm greater than 1 in less than 5% and greater than 2 in less than 1% of all cases. The dimensions n were taken in the range from 3 to 8. This range of dimensions is of interest in applications to ray tracing; e.g., $n = 8$ occurs when dealing with symplectic transformations on space-time.

5. ERROR ANALYSIS

When the computations in the algorithm are done in floating-point arithmetic the computed matrices T and L will, because of rounding errors, not represent the given Lagrangian subspace exactly. Here estimates on these errors are given. For simplicity, all permutations $P^{(k)}$ are assumed to be equal to the identity. This assumption will not restrict the generality of the error analysis.

To handle the accumulation of errors the following elementary estimate will be useful.

LEMMA 4. *Let* $\alpha \geq 0$, $b \geq 0$, *and let* $a_k \geq 0$, $a_{k+1} \leq (1+\alpha)a_k + b$ *for* $k = 0, 1, 2, \ldots$. *Then* $a_k \leq (1+\alpha)^k a_0 + k(1+\alpha)^{k-1} b$ *for* $k = 1, 2, \ldots$.

Proof. Observe that, in the case of equality, $a_{k+1} = (1+\alpha)a_k + b$ for $k = 0, 1, 2, \ldots$, the following solution formula holds:

$$a_k = (1+\alpha)^k a_0 + b \sum_{j=0}^{k-1} (1+\alpha)^j.$$

The estimate follows from this. ∎

Let $\tilde{A}^{(0)} = A$ and $\tilde{B}^{(0)} = B$. A floating-point implementation of the kth step (5) of the algorithm with $P^{(k)} = I$ reads

$$\tilde{C}^{(k-1)} = \tilde{A}^{(k-1)} - \tilde{B}^{(k-1)}\tilde{D}^{(k)} + \delta C^{(k-1)}, \tag{13}$$

$$\begin{pmatrix} \tilde{A}^{(k)} & \tilde{B}^{(k)} \end{pmatrix} = \tilde{Q}^{(k)} \begin{pmatrix} \tilde{C}^{(k-1)} & \tilde{B}^{(k-1)} \end{pmatrix} + \begin{pmatrix} \delta A^{(k)} & \delta B^{(k)} \end{pmatrix}. \tag{14}$$

Here $\tilde{A}^{(k)}$ is upper triangular in the first k columns, $\tilde{Q}^{(k)}$ is unitary, and $\tilde{D}^{(k)}$ is zero except for the kth diagonal element, which is $+1$ or -1 with the sign chosen according to the strategy described in Section 3.

For the errors caused by roundoff one has, with μ denoting the machine unit,

$$\|\delta C^{(k-1)}\| \leqslant \mu\big(\|\tilde{A}^{(k-1)}\| + \|\tilde{B}^{(k-1)}\|\big) + O(\mu^2) \tag{15}$$

and (cf. [9, pp. 152–160]), with a moderately growing function ϕ,

$$\|\delta A^{(k)}\| \leqslant \phi(n)\mu\|\tilde{C}^{(k-1)}\| + O(\mu^2), \tag{16}$$

$$\|\delta B^{(k)}\| \leqslant \phi(n)\mu\|\tilde{B}^{(k-1)}\| + O(\mu^2). \tag{17}$$

In general, because of errors, $(\tilde{A}^{(k)}, \tilde{B}^{(k)})$ will not be an L-pair.

LEMMA 5. *Assume* $[1+\phi(n)]n\mu < 0.1$. *Then*

$$\|\tilde{A}^{(k)}\| \leqslant 1.24\big[\|A\| + n\|B\|\big] + O(n^2\mu^2), \tag{18}$$

$$\|\tilde{B}^{(k)}\| \leqslant 1.11\|B\| + O(n\mu^2), \tag{19}$$

$$\|\tilde{C}^{(k-1)}\| \leqslant 1.4\big[\|A\| + (n+1)\|B\|\big] + O(n^2\mu^2), \tag{20}$$

for $k = 1, 2, \ldots, n$.

Proof. The assumption implies $[1+\mu+\phi(n)\mu]^n \leqslant 1.11$. Equations (14) and (17) and the unitarity of $\tilde{Q}^{(k)}$ imply

$$\|\tilde{B}^{(k)}\| \leqslant \big[1+\phi(n)\mu\big]\|\tilde{B}^{(k-1)}\| + O(\mu^2).$$

Lemma 4 leads to

$$\|\tilde{B}^{(k)}\| \leqslant \big[1+\phi(n)\mu\big]^n\|B\| + O(n\mu^2)$$

for $k \leqslant n$. This proves (19).

Equations (13) and (15) imply

$$\|\tilde{C}^{(k-1)}\| \leqslant (1+\mu)\big(\|\tilde{A}^{(k-1)}\| + \|\tilde{B}^{(k-1)}\|\big) + O(\mu^2). \tag{21}$$

From (14) and (16) one obtains

$$\|\tilde{A}^{(k)}\| \leqslant [1+\phi(n)\mu]\|\tilde{C}^{(k-1)}\| + O(\mu^2).$$

Inserting first (21) and then (19) one arrives at

$$\|\tilde{A}^{(k)}\| \leqslant [1+\phi(n)\mu+\mu](\|\tilde{A}^{(k-1)}\| + 1.11\|B\|) + O(n\mu^2).$$

Again, applying Lemma 4 and using the assumption, one deduces (18). The remaining inequality (20) follows upon insertion of (18) and (19) into (21). The proof of the lemma is complete. ■

The important goal is to estimate the error

$$(\delta A \quad \delta B) = (\tilde{A}^{(n)} \quad \tilde{B}^{(n)}) - \tilde{Q}(A \quad B)\begin{pmatrix} I & 0 \\ -T & I \end{pmatrix}, \tag{22}$$

where $\tilde{Q} = \tilde{Q}^{(n)} \cdots \tilde{Q}^{(1)}$ and $T = \tilde{D}^{(n)} + \cdots + \tilde{D}^{(1)}$.

LEMMA 6. *Assume* $[1+\phi(n)]n\mu < 0.1$. *Then*

$$\|\delta A\| \leqslant \mu n\, 1.4[1+\phi(n)][\|A\| + (n+2)\|B\|] + O(n^3\mu^2),$$

$$\|\delta B\| \leqslant \mu n\, 1.11\phi(n)\|B\| + O(n^2\mu^2).$$

Proof. Consider the matrices

$$E^{(k)} = \tilde{A}^{(k)} - \tilde{Q}^{(k)}(\tilde{A}^{(k-1)} - \tilde{B}^{(k-1)}\tilde{D}^{(k)}),$$

$$F^{(k)} = \tilde{B}^{(k)} - \tilde{Q}^{(k)}\tilde{B}^{(k-1)}$$

and

$$U^{(k)} = \tilde{Q}^{(n)} \cdots \tilde{Q}^{(k+1)},$$

$$T^{(k)} = \tilde{D}^{(n)} + \cdots + \tilde{D}^{(k+1)}$$

for $k < n$, $U^{(n)} = I$, and $T^{(n)} = 0$. Equation (22) can be written in the

following way:

$$(\delta A \quad \delta B) = \sum_{k=1}^{n} U^{(k)} \begin{pmatrix} E^{(k)} & F^{(k)} \end{pmatrix} \begin{pmatrix} I & 0 \\ -T^{(k)} & I \end{pmatrix}.$$

$U^{(k)}$ is unitary. $T^{(k)}$ is diagonal with diagonal elements equal to -1, 0, or -1. Hence

$$\|\delta A\| \leqslant \sum_{k=1}^{n} \left(\|E^{(k)}\| + \|F^{(k)}\|\right),$$

$$\|\delta B\| \leqslant \sum_{k=1}^{n} \|F^{(k)}\|.$$

A comparison of the definitions of $E^{(k)}$ and $F^{(k)}$ with (13) and (14) shows

$$E^{(k)} = \tilde{Q}^{(k)} \delta C^{(k-1)} + \delta A^{(k)},$$

$$F^{(k)} = \delta B^{(k)}.$$

With (15), (16), (17), and the estimates proven in Lemma 5 one obtains

$$\begin{aligned} \|E^{(k)}\| &\leqslant \|\delta C^{(k-1)}\| + \|\delta A^{(k)}\| \\ &\leqslant \mu \|\tilde{A}^{(k-1)}\| + \mu \|\tilde{B}^{(k-1)}\| + \phi(n)\mu \|\tilde{C}^{(k-1)}\| + O(\mu^2) \\ &\leqslant \mu\, 1.4\left[1 + \phi(n)\right]\left[\|A\| + (n+1)\|B\|\right] + O(n^2\mu^2) \end{aligned}$$

and

$$\begin{aligned} \|F^{(k)}\| &\leqslant \mu\phi(n)\|\tilde{B}^{(k-1)}\| + O(\mu^2) \\ &\leqslant 1.11\,\mu\phi(n)\|B\| + O(n\mu^2). \end{aligned}$$

Adding these inequalities leads to estimates for $\|\delta A\|$ and $\|\delta B\|$ proving the lemma. ■

For simplicity, back substitution is viewed here as the computation of an approximate left inverse Z with residue R to the upper triangular matrix $\tilde{A}^{(n)}$,

$$I = Z\tilde{A}^{(n)} + R, \tag{23}$$

and the subsequent multiplication

$$\tilde{L} = Z\tilde{B}^{(n)} + S, \tag{24}$$

with error

$$\|S\| \leqslant \mu\|Z\|\,\|\tilde{B}^{(n)}\| + O(\mu^2). \tag{25}$$

THEOREM 3. *Let (A, B) be an L-pair defining a Lagrangian subspace λ as in* (3). *Let T and $\tilde{L}$ be the matrices computed with a floating-point implementation of the algorithm in Section* 3, (13), (14), (23), (24), *with errors as in* (15), (16), (17), (22), (25). *Assume $\|Z\|\,\|\delta A\| < 0.5$ and $\|R\| < 0.5$. Assume $[1 + \phi(n)]n\mu < 0.1$, μ denoting the machine unit. Then*

$$\tilde{L} = L + \delta L,$$

where L is the symmetric matrix representing λ with respect to the coordinates defined by T as in (1). *The error δL satisfies the estimate*

$$\begin{aligned}\|\delta L\| \leqslant\ & \mu\, 1.4n[1 + \phi(n)]\,\|A\|\|Z\| \\ & + \mu\, 1.4(n+1)(n+2)[1 + \phi(n)]\,\|B\|\|Z\| \\ & + \|R\|\|L\| \\ & + O(\|Z\|n^3\mu^2).\end{aligned}$$

Proof. (22), (23), and (24) imply

$$\begin{aligned}(I \quad \tilde{L}) &= Z(\tilde{A}^{(n)} \quad \tilde{B}^{(n)}) + (R \quad S) \\ &= Z\tilde{Q}(A \quad B)\begin{pmatrix} I & O \\ -T & I \end{pmatrix} + Z(\delta A \quad \delta B) + (R \quad S).\end{aligned}$$

The assumption of the theorem implies that $I - Z\,\delta A - R$ is invertible.

Define L by

$$(I - Z\delta A - R)L = \tilde{L} - Z\delta B - S.$$

Then (4) holds with an invertible matrix E. Hence T and L represent λ exactly. The estimate on $\delta L = \tilde{L} - L$ follows from

$$\|\delta L\| \leqslant (\|Z\|\|A\| + \|R\|)\|L\| + \|Z\|\|\delta B\| + \|S\|$$

and Lemma 6 and (25). The proof is complete. ∎

REMARK. In the theorem conditions are formulated which imply that the computed solution is close to an exact solution. Note that this exact solution is, in general, different from the solution obtained with the algorithm in exact arithmetic. This is related to the fact that T is not everywhere continuous as a function of the L-pair (A, B).

REMARK. Theorem 3 provides *a posteriori* error estimates. For a full forward error analysis a bound on the condition of the back substitution is necessary. Note that a bound on this condition is implicit—for exact arithmetic—in the proof of Theorem 2. No attempt is made here to extend Theorem 2 to floating-point arithmetic.

6. APPLICATION TO WAVEFRONT TRACING

Wavefronts in a medium with smoothly varying refractive index $\nu(x) > 0$ can be obtained from solutions of the eikonal equation $|\nabla\phi(x)| = \nu(x)$, $x \in \mathbb{R}^3$. (Here $|\cdot|$ is the Euclidean norm.) The wavefront at time t is given by the equation $t = \phi(x)$. The Hessian of ϕ,

$$W(x) = \nabla^2\phi(x),$$

contains information about the wavefront curvature. More precisely, the wavefront curvature form is the quadratic form defined by $\nu(x)^{-1}W(x)$ restricted to the orthogonal complement of $\nabla\phi(x)$. Rays $x(t)$ hit the wavefronts orthogonally: $dx/dt = \nu^{-2}\nabla\phi(x(t))$. The propagation of wavefront

curvatures is governed by Ricatti-type ordinary differential equations. In the case considered here,

$$\frac{dW}{dt} + WW + C = 0$$

for $W(t) = W(x(t))$. Here $C + (\nabla \nu)(\nabla \nu)^T + \nabla^2 \nu = 0$. This follows from differentiating the eikonal equation twice. A solution to this equation may blow up in a pole. This actually does happen at caustics, e.g. at foci. For more details see [5] and [6].

EXAMPLE. Consider the case $\nu \equiv 1$. Then $\phi(x) = 1 - |x|$ solves the eikonal equation for $x \neq 0$. Here $W(x) = -|x|^{-1}(I - \xi\xi^T)$ with $\xi = \nabla\phi(x)$. Along a ray $x(t) = (t-1)\xi$, $|\xi| = 1$, $W(t)$ will cease to exist at $t = 1$.

With $W = W(t)$ is associated the Lagrangian subspace $\lambda = \lambda(t)$,

$$\lambda = \{(\hat{x}, \hat{\xi});\ W\hat{x} - \hat{\xi} = 0\}.$$

Application of the algorithm results in a representation

$$\lambda : \hat{y} + L\hat{\eta} = 0 \qquad \text{for} \quad (\hat{y}, \hat{\eta}) = (\hat{x}, \hat{\xi} + T\hat{x})$$

with $L = -(W + T)^{-1}$. Here caustics correspond to points where $L(t)$ becomes singular. The differential equation for W transforms to a differential equation for L,

$$\frac{dL}{dt} + (LT + I)(I + TL) + LCL = 0. \tag{26}$$

Equation (26) is an ordinary differential equation on $\Lambda(n)$ written in the coordinates induced by T. Its solution is a curve of Lagrangian subspaces, $\lambda(t)$, which are the tangent spaces, along a bicharacteristic curve, of a Lagrangian manifold solving the (generalized) eikonal equation [1, 5, 6]. A bicharacteristic curve is a solution to Hamilton's canonical equations,

$$\dot{x} = \frac{\partial H}{\partial \xi}, \qquad \dot{\xi} = -\frac{\partial H}{\partial x}, \tag{27}$$

where $H = |\xi| / \nu(x)$ is the Hamilton function.

Let $\|\cdot\|$ be a matrix norm. Assume $\|T\| \leqslant 1$ and $\|C\| \leqslant \gamma$, γ a constant. Then, for a solution to (26), $\|dL/dt\| \leqslant (1+\gamma)(1+\|L\|)^2$. Comparison with the solution to the initial-value problem

$$\dot{z} = (1+\gamma)(1+z)^2, \qquad z(t_0) = \|L(t_0)\|,$$

leads to an estimate

$$\|L(t)\| \leqslant \frac{1}{(1+\gamma)(t_1 - t)} \qquad \text{for} \quad t_0 < t < t_1,$$

where

$$t_1 - t_0 = \frac{1}{(1+\gamma)(1+\|L(t_0)\|)}$$

is a lower bound for the life span of the solution $L(t)$.

To trace $\lambda(t)$ along a bicharacteristic curve, proceed in the following way: Start with $L(t)$ and T representing $\lambda(t)$. Solve (26) as long as $\|L(t)\| < l$ for some fixed (large) threshold l. In case the threshold l is reached, change, with the transformation algorithm (Section 3, last remark) to a new representation of $\lambda(t)$ with different matrices $L(t)$ and T. Continue in this way. If l is chosen large enough, then it is guaranteed that every integration of (26) proceeds at least for some positive time span which depends only on l and γ. This is evident from Theorem 2 and the estimates on solutions to (26) shown above. Hence, with the foregoing procedure, a curve of Lagrangian tangent spaces in $\Lambda(n)$ can be continued indefinitely along a bicharacteristic curve.

This procedure was implemented and tested. The results obtained for two examples are reproduced below. The first example is the example above for a particular ray. Here the computed solution can be compared with the analytic solution, which may be obtained from analytic continuation. The second example models a ray system issuing from a point source at the origin and passing through a strongly refracting circular lens embedded in a homogeneous medium. The Runge-Kutta-Fehlberg method of order 4(5) with automatic stepsize control [4] was used to compute both the bicharacteristic curves and the solutions to the paraxial-ray-tracing equations (26) numerically. In the integration of (27) and (26) the relative and absolute error tolerances were 10^{-5} and 10^{-3}, respectively. To avoid stepping over a pole in $L(t)$ inadvertantly, the stepsizes were taken to be less than the estimated

lifespan $t_1 - t_0$. Blowup was tested for with the threshold value $l = 5$ and $\|L\|$ equal to the sum of the absolute values of the elements of the upper triangular part of L. The integrations started with the parameter value $t = 0$ and stopped when $t = 10$ was reached.

NUMERICAL EXAMPLE 1. Here $\nu \equiv 1$. The initial values at $t = 0$ are $x(0) = (-1,0,0)$, $\xi(0) = (1,0,0)$, and a Lagrangian subspace $\lambda(0)$ given by $W(0) = \text{diag}(0, -1, -1)$. The transformation algorithm is used to obtain the following representation for $\lambda(0)$:

$$L_T(0) = \text{diag}(1.0, 0.5, 0.5), \qquad T = \text{diag}(-1, -1, -1).$$

[Here and in the following the dependence of L on t and T is emphasized by use of the notation $L_T(t)$.] Starting from this initial value, the numerical integration proceeds from 0 to 1.705 using 5 steps. The integration stops at 1.705 because L exceeds the threshold:

$$L_T(1.705) = \text{diag}(1.0, -2.394, -2.394), \qquad T = \text{diag}(-1, -1, -1).$$

The transformation algorithm changes to

$$L_T(1.705) = \text{diag}(1.0, -0.414, -0.414), \qquad T = \text{diag}(-1, 1, 1).$$

Starting from this initial value, the numerical integration proceeds from 1.705 to 10.135 using 30 steps. The integration terminates at 10.135. At the output time $t = 10$ the computed result equals the exact result

$$L_T(t) = \text{diag}\left(1, \frac{1-t}{t}, \frac{1-t}{t}\right), \qquad T = \text{diag}(-1, 1, 1).$$

up to a relative error 2.174E − 06.

NUMERICAL EXAMPLE 2. Here $\nu(x_1, x_2, x_3) = 1 + 0.4b[(x_1 - 3)^2 + x_2^2 + x_3^2]$, with a C^2 function b, $b(s) = 0$ if $s > 1$,

$$b(s) = -(3s+1)(s-1)^3 \qquad \text{if} \quad 0 \leqslant s \leqslant 1.$$

The initial values at $t = 0$ are $x(0) = (0,0,0)$, $\xi(0) = (0.99, \sqrt{1 - 0.99^2}, 0)$, and $L_T(0) = 0$, $T = \text{diag}(-1, -1, -1)$. The Lagrangian subspace $\lambda(0)$ corresponds to a point source. The ray starts in a direction slightly off the axis of symmetry. The following sequence of numerical approximations to La-

grangian subspaces results on applying, in alternation, the integration algorithm to (26) and the transformation algorithm to the latest values for $L_T(t)$ and T:

$$L_T(0.628) = \begin{pmatrix} -1.689 & 0 & 0 \\ 0 & -1.689 & 0 \\ 0 & 0 & -1.689 \end{pmatrix}, \qquad T = \mathrm{diag}(-1, -1, -1),$$

$$L_T(0.628) = \begin{pmatrix} -0.386 & 0 & 0 \\ 0 & -0.386 & 0 \\ 0 & 0 & -0.386 \end{pmatrix}, \qquad T = \mathrm{diag}(1, 1, 1),$$

$$L_T(2.947) = \begin{pmatrix} -0.921 & -0.436 & 0 \\ -0.436 & -2.002 & 0 \\ 0 & 0 & -1.976 \end{pmatrix}, \qquad T = \mathrm{diag}(1, 1, 1),$$

$$L_T(2.947) = \begin{pmatrix} -0.795 & 0.145 & 0 \\ 0.145 & 0.666 & 0 \\ 0 & 0 & 0.669 \end{pmatrix}, \qquad T = \mathrm{diag}(1, -1, -1),$$

$$L_T(3.749) = \begin{pmatrix} -4.295 & 0.898 & 0 \\ 0.898 & -0.098 & 0 \\ 0 & 0 & 0.274 \end{pmatrix}, \qquad T = \mathrm{diag}(1, -1, -1),$$

$$L_T(3.749) = \begin{pmatrix} 0.566 & -0.118 & 0 \\ -0.118 & 0.114 & 0 \\ 0 & 0 & 0.274 \end{pmatrix}, \qquad T = \mathrm{diag}(-1, -1, -1),$$

$$L_T(4.656) = \begin{pmatrix} -0.330 & -1.537 & 0 \\ -1.537 & -3.127 & 0 \\ 0 & 0 & -1.249 \end{pmatrix}, \qquad T = \mathrm{diag}(-1, -1, -1),$$

$$L_T(4.656) = \begin{pmatrix} 0.321 & -0.212 & 0 \\ -0.212 & -0.431 & 0 \\ 0 & 0 & -0.357 \end{pmatrix}, \qquad T = \mathrm{diag}(-1, 1, 1),$$

$$L_T(5.811) = \begin{pmatrix} -2.964 & -0.713 & 0 \\ -0.713 & -0.762 & 0 \\ 0 & 0 & -0.631 \end{pmatrix}, \qquad T = \mathrm{diag}(-1, 1, 1),$$

$$L_T(5.811) = \begin{pmatrix} -0.428 & -0.103 & 0 \\ -0.103 & -0.615 & 0 \\ 0 & 0 & -0.631 \end{pmatrix}, \qquad T = \mathrm{diag}(1, 1, 1),$$

$$L_T(10.067) = \begin{pmatrix} -0.835 & -0.012 & 0 \\ -0.012 & -0.856 & 0 \\ 0 & 0 & -0.856 \end{pmatrix}, \qquad T = \mathrm{diag}(1, 1, 1).$$

There occur six solution intervals during each of which T remains unchanged, the intervals of integration. These intervals are separated by the times 0.628, 2.947, 3.749, 4.656, and 5.811, where the blowup test stops an integration. (These blowup points do not bear an immediate relation to caustics.) The integration finally stops at 10.067. The numbers of steps taken by the Runge-Kutta-Fehlberg method in the intervals are 4, 22, 9, 7, 10, and 42, respectively.

REMARK. In view of Theorem 2, the choice of the threshold value in the foregoing examples seems to be rather optimistic. However, as remarked at the end of Section 4, the matrices L obtained with the transformation algorithm are almost always small in practice. The results obtained with the examples suggest that, for the application considered here, answers to the questions of error accumulation and computational effort are likely to depend more heavily on the numerical properties of the method for integrating the ordinary differential equations than on the numerical properties of the transformation algorithm.

REMARK. The application given in this section only serves as an indication for the possible uses of the algorithm presented here. Full dynamic ray tracing [2, 6] requires the solution of ordinary differential equations for the amplitudes, called transport equations, in addition to equations like (27) and (26). The main motivation for the algorithm presented here is its use in an extension of the application to full dynamic ray tracing. This extension will be presented elsewhere.

REFERENCES

1 V. I. Arnold, On a characteristic class entering into conditions of quantization. *Funktsional. Anal. i Prilozhen.* 1:1–14 (1967); English transl., *Functional Anal. Appl.* 1:1–13 (1967).

2 V. Červený, I. A. Molotkov, and I. Pšenčik, *Ray Method in Seismology*, Univ. Karlova, Prague, 1977.

3 G. A. Deschamps, Ray techniques in electromagnetics, *Proc. IEEE* 60:1022–1035 (1972).

4 E. Fehlberg, Klassische Runge-Kutta-Formeln vierter und niedrigerer Ordnung mit Schrittweiten-Kontrolle und ihre Anwendung auf Wärmeleitungsprobleme, *Computing* 6:61–71 (1970).

5 V. Guillemin and S. Sternberg, *Geometric Asymptotics*, Amer. Math. Soc. Surveys 14, Providence, 1977.

6 A. Hanyga, Dynamic ray tracing on Lagrangian manifolds, *Geophys. J. Roy. Astron. Soc.* 79:51–63 (1984).

7 V. P. Maslov, *Theory of perturbations and asymptotic methods* (in Russian), Moskov. Gos. Univ., Moscow, 1965; French transl., *Théorie des Perturbations et Méthodes Asymptotiques*, Dunod, Paris, 1972.
8 G. W. Stewart, *Introduction to Matrix Computations*, Academic, New York, 1973.
9 J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford U. P., London, 1965.

Received April 1988; final manuscript accepted 26 January 1989