JOURNAL OF COMPUTATIONAL PHYSICS 8, 464–471 (1971)

Partial Iterative Refinements

R. H. MILLER*

Department of Astronomy and Astrophysics, Committee on Information Sciences, Institute for Computer Research, University of Chicago, Chicago, Illinois 60637

Received November 10, 1970

A few well-defined parameters may be iteratively refined by a reduced Newton–Raphson process based on the generalized inverse. The method has been used in a gravitational n-body integration to control the usual ten first integrals of motion. The discussion is based on that example to show how the details may be carried through.

I. INTRODUCTION

Multivariate Newton–Raphson processes are used in many computations to find a kind of "self-consistent" solution to a set of difference equations (see, for example, [1, 2]). Usually, the number of variables to be refined is equal to the number of conditions available; so the matrix involved is nonsingular and may be inverted without difficulty. It may happen, however, that the number of conditions is less than the number of variables in a situation for which an iterative refinement may be useful. A simple extension of the Newton–Raphson method to this case is described in this note. As frequently occurs in a computational context, the working-out of details is much more of a problem than the general theory and may, indeed, determine whether a method is practical. For that reason, this note is principally devoted to the application of the method to a practical calculation—the gravitational n-body calculation. The lines along which a general formulation can be constructed, as well as other situations to which the method apply, should be evident from this application.

The gravitational n-body calculation [3-9] is an attempt to integrate the Newtonian equations of motion for point particles with inverse-square-law forces. The forces are computed between each pair of particles, and the entire calculation is treated as precisely as computational methods allow. The process is numerically quite unstable (even though the first 10 integrals can be adequately computed) for

* Visiting Astronomer, Kitt Peak National Observatory, which is operated by the Association of Universities for Research in Astronomy, Inc., under contract with the National Science Foundation.

physical reasons that are reflected in the structure of the equations [10]. The improvement in control of the first ten integrals described here has little effect on the numerical stability (Section 3).

The evolution of an n-body system may be geometrically visualized as the motion of a representative point in the $6n$-dimensional Γ space (phase space), which is a linear vector space (it is the system space of the differential equations). The integrals of motion define a set of intersecting hypersurfaces with the trajectory contained in the intersection. The ($6n$-10)-dimensional hypersurface in which the ten first integrals are exactly conserved will be called the "integral hypersurface".

As the computed system evolves, the point representing the computed system drifts away from the point representing a real physical system (or an exactly integrated set of equations); the "difference vector" has components lying in the integral hypersurface and others orthogonal to it. The components orthogonal to the integral hypersurface are responsible for the drifts in the values of the first integrals that are observed in every real calculation. The partial refinement described in this note is a process whereby the point representing the computed system is moved toward the integral hypersurface. The components lying in the integral hypersurface cannot be determined from the integrals; accordingly, the point is moved toward the hypersurface along the normals to the hypersurface (gradients of the integrals). This merely provides an unambiguous rule for moving the point; the only justification that can be offered is that the length of the displacement vector is minimal in a least-squares sense (but not the usual least-squares problem; this problem is heavily underdetermined). Displacements having these properties may be computed by using the generalized inverse. This method differs from other gradient methods in that it deals with an underdetermined problem.

There is no guarantee that the solutions so found are any better than any that might be constructed by some other recipe. In fact, the method was developed to assist in the study of the peculiarities of numerical solutions to gravitational n-body systems, in particular, to show that the numerical instability persists even if the first integrals are (nearly) exactly conserved. The discussion gets rather philosophical and is postponed to another paper.

The notion of refinements at each integration step is not unusual; in any predictor-corrector scheme, the correction is a shift of the representative point according to some rule. The rule chosen here is that the point shall be shifted to conserve the first integrals. Although the resulting system is expected to be stable, it could turn out to be unstable.

The conditions for the method to apply are simply that the phase space be locally Euclidean, that the computed system point be close enough to the integral hypersurface that the space between is well-behaved.The notion of a normal to the surface must be well defined. In spaces of many dimensions, the topology of the integral hypersurface can become very complicated, so it is not *a priori* obvious that this kind of scheme can be made to work in practice. But it has been made to work; so the pathological special cases that might be imagined do not seem to present practical difficulties.

II. Computational Details

The first ten integrals consist of the position and velocity of the centroid (3 each), the total angular momentum (3 components), and the total energy. To the first order, the error in each of these ten integrals is a linear combination of the components of a displacement in the phase space. The matrix of the linear combinations has 10 rows and $6n$ columns; its elements are the partial derivatives of each of the integrals in turn by the components of the particle positions and momenta. These latter form a basis for the phase space. The rows of the matrix are the gradients of the integrals. If the generalized inverse of this matrix is found, the displacements of particle positions and velocities computed from it will be made up of linear combinations of the gradients of the integrals—the correcting displacement must lie in the subspace spanned by the gradients. This is the feature that assures the minimal displacement in a least-squares sense.

The gradients are made up of quantities that are readily available in the calculation. It is a simple matter requiring no extra computation to construct this matrix. Let $x_i^{(\alpha)}$, $v_i^{(\alpha)}$ and $p_i^{(\alpha)}$ represent the i-component of the position, velocity, and momentum of particle number α ($\alpha = 1, 2, 3,..., n$; $i = 1, 2, 3$). The meanings of the other symbols are conventional. Then the first integrals and their gradients are given by:

Centroid in configuration space:

$$X_i = \sum_{\alpha} m^{(\alpha)} x_i^{(\alpha)}, \tag{1}$$

$$\frac{\partial X_i}{\partial x_j^{(\beta)}} = m^{(\alpha)} \delta_{ij} \delta_{\alpha\beta} = m^{(\beta)} \delta_{ij} \delta_{\alpha\beta}, \tag{2}$$

$$\frac{\partial X_i}{\partial p_j^{(\beta)}} = 0. \tag{3}$$

Total linear momentum:

$$P_i = \sum_{\alpha} p_i^{(\alpha)}, \tag{4}$$

$$\frac{\partial P_i}{\partial x_j^{(\beta)}} = 0, \tag{5}$$

$$\frac{\partial P_i}{\partial p_j^{(\beta)}} = \delta_{\alpha\beta} \delta_{ij}. \tag{6}$$

Total angular momentum:

$$L_i = \sum_{\alpha} \epsilon_{ijk} x_j^{(\alpha)} p_k^{(\alpha)}, \tag{7}$$

$$\frac{\partial L_i}{\partial x_\ell^{(\beta)}} = \epsilon_{ijk} \delta_{j\ell} \delta_{\alpha\beta} p_k^{(\alpha)} = \epsilon_{i\ell k} p_k^{(\beta)}, \tag{8}$$

$$\frac{\partial L_i}{\partial p_\ell^{(\beta)}} = \epsilon_{ijk} \delta_{k\ell} \delta_{\alpha\beta} x_j^{(\alpha)} = -\epsilon_{i\ell k} x_k^{(\beta)}. \tag{9}$$

Total energy:

$$E = -\frac{1}{2} \sum_{\alpha} \sum_{\beta \neq \alpha} \frac{G m^{(\alpha)} m^{(\beta)}}{r_{\alpha\beta}} + \frac{1}{2} \sum_{\alpha} p_i^{(\alpha)} v_i^{(\alpha)}, \tag{10}$$

$$\frac{\partial E}{\partial x_j^{(\gamma)}} = \frac{\partial \mathscr{V}}{\partial x_j^{(\gamma)}} = -F_j^{(\gamma)}, \tag{11}$$

$$\frac{\partial E}{\partial p_j^{(\beta)}} = v_j^{(\beta)} = \frac{p_j^{(\beta)}}{m^{(\beta)}}. \tag{12}$$

Here, ϵ_{ijk} is the (3-dimensional) totally antisymmetric symbol, and summation over repeated i, j, k, ℓ indices is implied, but not over repeated α, β indices. These expressions obscure the essential simplicity of the construction. In detail, the matrix of gradients (for equal masses) is

$$\begin{array}{c|cccccccccccccccccc}
 & x^{(1)} & x^{(2)} & \cdots & y^{(1)} & y^{(2)} & \cdots & z^{(1)} & z^{(2)} & \cdots & p_x^{(1)} & p_x^{(2)} & \cdots & p_y^{(1)} & p_y^{(2)} & \cdots & p_z^{(1)} & p_z^{(2)} & \cdots \\
\hline
X_x & 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\
X_y & 0 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\
X_z & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\
P_x & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\
P_y & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots \\
P_z & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 1 & \cdots \\
L_x & 0 & 0 & \cdots & p_z^{(1)} & p_z^{(2)} & \cdots & -p_y^{(1)} & -p_y^{(2)} & \cdots & 0 & 0 & \cdots & -z^{(1)} & -z^{(2)} & \cdots & y^{(1)} & y^{(2)} & \cdots \\
L_y & -p_z^{(1)} & -p_z^{(2)} & \cdots & 0 & 0 & \cdots & p_x^{(1)} & p_x^{(2)} & \cdots & z^{(1)} & z^{(2)} & \cdots & 0 & 0 & \cdots & -x^{(1)} & -x^{(2)} & \cdots \\
L_z & p_y^{(1)} & p_y^{(2)} & \cdots & -p_x^{(1)} & -p_x^{(2)} & \cdots & 0 & 0 & \cdots & -y^{(1)} & -y^{(2)} & \cdots & x^{(1)} & x^{(2)} & \cdots & 0 & 0 & \cdots \\
E & -F_x^{(1)} & -F_x^{(2)} & \cdots & -F_y^{(1)} & -F_y^{(2)} & \cdots & -F_z^{(1)} & -F_z^{(2)} & \cdots & v_x^{(1)} & v_x^{(2)} & \cdots & v_y^{(1)} & v_y^{(2)} & \cdots & v_z^{(1)} & v_z^{(2)} & \cdots
\end{array} \tag{13}$$

The rows of this matrix are linearly independent. This can be shown directly by seeking a linear combination of the rows that is identically zero for coefficients not all zero. When this is done, the six coefficients that multiply (grad X_i) and (grad P_j) naturally split off and must each be zero if X_i and P_i are zero. Since the dynamics always permits this shift of origin (in the absence of external forces), these six coefficients are all zero. The system of equations remains:

$$\epsilon_{ijk} J_j p_k^{(\alpha)} + dF_i^{(\alpha)} = 0, \tag{14}$$

$$\epsilon_{ijk} J_j x_k^{(\alpha)} + dv_i^{(\alpha)} = 0. \tag{15}$$

Here d is the coefficient multiplying (grad E) and J_j are three coefficients that multiply the terms of (grad L_j) in the linear combination. If the first of these equations is multiplied by $x_i^{(\alpha)}$ and the second by $p_i^{(\alpha)}$ (summation on i implied, but not on α), the indices may be relabeled to give

$$0 = -d[x_i^{(\alpha)} F_i^{(\alpha)} + p_i^{(\alpha)} v_i^{(\alpha)}], \tag{16}$$

which cannot be satisfied for all $x_i^{(\alpha)}$, $p_i^{(\alpha)}$, $F_i^{(\alpha)}$ unless $d = 0$. It is interesting to notice that the square bracket of Eq. (16), if summed over particles (α), is the expression that appears in the Lagrange–Jacobi identities [11]. The second term is twice the total kinetic energy and the first can be transformed to give the total potential energy. The Lagrange–Jacobi identities give the sum over particles as $\frac{1}{2}\, d^2I/dt^2$ (I is the "total moment of inertia" or, more properly, the trace of the inertia tensor), a quantity that cannot always be zero. The "virial theorem" merely asserts that the time-average is zero, not that the quantity is instantaneously zero at all times. The requirement of Eq. (16), with no sum over particles, is even more stringent.

With $d = 0$, Eqs. (14) and (15) contain only J_j. Finding three nonzero coefficients J_j is the same as finding a single 3-vector that is parallel to each $p_k^{(\alpha)}$ and $x_k^{(\alpha)}$. This cannot be done even in a two-dimensional system.

Thus, the rows of the matrix of gradients are linearly independent. The importance of this result, computationally, is twofold: (1) the amount of computation cannot be reduced by solving a smaller system, and (2) each of the first integrals can be forced to zero, rather than a restricted set determined by some linear combination. Experimentally, this second statement is confirmed: the refinement process does indeed bring each of the components of angular momentum and the total energy separately to zero. One's prejudices, carried over from quantum mechanics and from the more thorough treatments of classical mechanics, might have led him to expect that L_x, L_y, and L_z might not be linearly independent while L_z, L^2 might be. The demonstration of linear independence settles that issue as well.

As a practical matter, the six integrals of the centroid position and velocity present no problem numerically, particularly if both centroids are on the origin. They need not be included in the iterative refinement. This leaves only the three total angular momentum components and the total energy. For a 32-particle system, the resulting matrix is 4 × 192. Computation of generalized inverses for such strongly nonsquare matrices is quite fast.

The refinement entails computation of a "vector" whose components are the errors in the integrals, followed by the computation of the generalized inverse to the matrix of gradients, computation of the correction in phase space, and finally, the correction of the phase point. Since the argument is based on the assumption of first-order deviations, iteration requires recomputation of the forces and other elements of the matrix of gradients to allow nonlinear effects to enter. It is evident that the n-body calculation must be designed to mesh with the partial iterative refinement–many possible strategies for handling n-body calculations would fit awkwardly if at all with an iterative refinement procedure.

III. Experience With The Procedure

It is convenient to set an acceptance tolerance and a failure exit if the required tolerance cannot be attained in some given number of tries. In the case of the 32-body system, a sum of squares of the errors in the four integral components of 10^{-15} and a failure exit at six tries was a workable combination. In about 30 different calculations of several thousand refinements each, one calculation reached the failure exit during an extremely close encounter. The failure happened so infrequently that there seemed to be no point in taking special precautions for this case.

Convergence to this tolerance (for a system with $E \approx -250$) required 2 iterations normally, occasionally needing 3 or 4. The energy was usually the most difficult to control—the error in the energy was usually significantly larger than the errors in angular momentum components. Relative errors are not meaningful since most calculations were run with all total angular momentum components being zero.

The calculation was designed with the integrations for all particles being done at the same time, unlike the calculations of Aarseth and Wielen [6-9]. Variable time steps were used, and the iterative refinement procedure was called each 16 normal integration steps. No special pains were taken to produce a fast-running program, but tolerance requirements as stringent as this are usable only with multiple precision or with a reasonably long word. It is not meaningful to quote a figure for the comparative computing speeds of calculations run with and without the partial refinements because, as mentioned earlier, the iterative refinement procedure can effectively mesh only with a program designed with that goal in mind. Partial refinements would normally be used only where the expected computational advantages are great enough to justify the extra cost.

The iterative refinement procedure made little change in the gravitational n-body calculation, by the measure of the rate at which representative points of two computed systems separate in phase space. The details of studies using the partial iterative refinements as a diagnostic tool with the gravitational n-body problem are given in a companion paper [10]. That paper goes into the reasons why partial refinements are not sufficient to stabilize the gravitational n-body calculation.

One gets the *impression* that calculations run with the control on first integrals have fewer close encounters than those without it. It is only an impression, and would be very difficult to establish convincingly. The principal difficulty is that two calculations started from identical initial conditions, one with and one without the iterative refinement, very soon become distinct calculations—their trajectories are quite different. One trajectory may quite properly contain more close encounters than the other.

The partial refinement procedure should improve other calculations with which it might be used. Its failure to stabilize the gravitational n-body calculation follows from the peculiarities of that calculation already alluded to, and from the severity of the test criterion. Clearly, once the first integrals are controlled, they are no longer useful as indicators of the quality of a calculation.

Quite independently of the work described here, P. Nacozy [12] has applied a similar method of partial refinements to the first integrals in a gravitational n-body calculation. Nacozy's interpretation of the utility of partial refinements differs from that presented here, evidently because of very different measures of their effectiveness.

Acknowledgments

Most of this work was carried out at the Kitt Peak National Observatory, while the writer was there as Consulting Astronomer. The use of the computational facilities at Kitt Peak is gratefully acknowledged. This work was also partially supported by the U. S. Atomic Energy Commission under Contract No. AT(11-1)-2094.

References

1. R. Kippenhahn, A. Weigert, and E. Hofmeister, *Methods in Computational Phys.* **7** (1967), 129–90.
2. L. G. Henyey and R. D. Levee, *Methods in Computational Phys.* **4** (1965), 333–48.
3. J. Pasta and S. Ulam, Los Alamos Report LA-1557 (unpublished).
4. S. von Hoerner, *Z. Astrophysik* **50** (1960), 184.
5. S. von Hoerner, *Z. Astrophysik* **57** (1963), 47.
6. S. Aarseth, *Mon. Not. Roy. Astron. Soc.* **126** (1963), 223.
7. S. Aarseth, *Mon. Not. Roy. Astron. Soc.* **132** (1966), 35.
8. R. Wielen, *Ver. Astron. Rechen-Instituts Heidelberg*, Nr. 19 (1967).

9. R. Wielen, *Bull. Astron.* (3), 3 (1968), 127.
10. R. H. Miller, *J. Computational Phys.*, companion paper.
11. S. Chandrasekhar, "Principles of Stellar Dynamics," Dover, New York, 1960. Section 5.1.iii.
12. P. Nacozy, *in* "Proceedings of IAU Colloquium 10 on the Gravitational *n*-body Problem" (M. Lecar, Ed.), Reidel, Dordrecht, Holland, to appear.