PROCEDURES FOR SOLVING KEPLER'S EQUATION

A. W. ODELL and R. H. GOODING
Royal Aircraft Establishment, Farnborough, Hants, England

(Received 29 October, 1985; accepted 8 January, 1986)

Abstract. We review starting formulae and iteration processes for the solution of Kepler's equation, and give details of two complete procedures. The first has been in use for a number of years, but the second is entirely new. The new procedure operates with an iterative process that always gives fourth-order convergence and is taken to only two iterations. The error in the resulting solution then never exceeds 7×10^{-15} rad.

1. Introduction

Kepler's equation has aroused fresh interest during the past quarter of a century, reflecting the demands of the space age and the potential of the digital computer. In the authors' monograph on the subject (Gooding and Odell, 1985), 19 papers are cited from this period, of which seven appeared in the *Journal of the Astronautical Sciences* and seven in *Celestial Mechanics*; one of the latter is the recent paper by Shepperd (1985), which is not confined to Kepler's equation. The present paper is a greatly shortened version of the monograph, copies of the latter being available on request.

The conventional form of Kepler's equation may be written

$$E - e \sin E = M, \tag{1}$$

where M is the mean anomaly of a body in an elliptic orbit of eccentricity e, and E is the corresponding eccentric anomaly; we shall also find it convenient to write the equation as

$$f(E) = 0,$$

where

$$f(E) = E - e \sin E - M. \tag{2}$$

The elliptic interpretation of Equation (1) breaks down when the orbit becomes parabolic (as $e \to 1$ with a fixed perigee), since, at any given time, both sides of the equation are zero in the limit – it is now common practice to cover orbits of arbitrary eccentricity by working with a generalized version of Kepler's equation and universal variables (as discussed in Section 5 of this paper). The interpretation is valid for a rectilinear ellipse (for which $e = 1$), however, and there is in any case no mathematical breakdown of the equation when $e = 1$, since $f(E)$ remains strictly monotonic, with a unique E corresponding to any given value of M. To obtain a solution procedure that is accurate for all ellipses, therefore, it is legitimate to demand that the procedure be valid for $e = 1$. The main difficulty inherent in

Kepler's equation, viz the vanishing of $f'(E)$ when $e = 1$ and M is a multiple of 2π, is then automatically resolved.

Direct analytical solutions of Kepler's equation are in general very inefficient to compute. This applies, in particular, to the classical power-series expansions and to the rather complicated quadrature solution found by Siewert and Burniston (1972). Thus iterative solutions by trial and error (which is arguably no less 'direct' or 'analytical') retain their virtual monopoly. Recent papers, such as that of Danby and Burkardt (1983), have focused separate attention on these complementary components of an iterative procedure, and the present paper is no exception: Section 2 is devoted to starting formulae, i.e. to the trial component, and Section 3 to methods of ıteration, i.e. to the feedback reduction of error. Section 4 puts the components together and describes two complete solution procedures that have been incorporated in Fortran functions developed at Farnborough, listings of which are appended. The first procedure has been in use for a number of years, though not previously published. The second procedure is entirely new and should provıde a particularly accurate, rapid and generally applicable solution for Kepler's equation: it operates with a fixed number of iterations, viz. 2, and for $0 \leqslant e \leqslant 1$ and any M, the error in the solution never exceeds 7×10^{-15} rad.

2. Starting Formulae

A number of authors, in particular Smith (1979), Broucke (1980) and Bergam and Prussıng (1982), tabulate the starting formulae that they consider of most interest for intercomparıson purposes, and we do the same in Table 1. This re-lists the 12

TABLE I
Comparıson of starting formulae

S	Formula for E_0	O (error)	(a)	(b)	(c)	(d)
S_1	M	e			√	√
S_2	$M + e \sin M$	e^2			√	√
S_3	$M + e \sin M(1 + e \cos M)$	e^3			√	√
S_4	$M + e$	e		√	*	
S_5	$M + e \sin M/\{1 - \sin(M + e) + \sin M\}$	e^3			*	√
S_6	$M + e(\pi - M)/(1 + e)$	e		√		
S_7	$\mathrm{Min}\{M/(1 - e), S_4, S_6\}$	e		√		
S_8	$S_3 + \lambda e^4(\pi - S_3)$; $\lambda = 1/20\pi$	e^3		√		
S_9	$M + e \sin M/(1 - 2e \cos M + e^2)^{1/2}$	e^4		√	√	
S_{10}	See Equations (6) to (8)	1	√	√		
S_{11}	See Equations (10) and (11)	e^4	√	√	√	√
S_{12}	See Equations (20), (25) to (27)	e	√	√		√

(a) Free of 'slow convergence' phenomenon
(b) Always Newton–Raphson convergent
(c) Valıd wıthout 'range reductıon'
(d) Smoothly portable
* For S_4 and S_5, the avoıdance of range reductıon ıs trıvıal

different 'starters' from Gooding and Odell (1985), six of which were taken from other authors' papers. Supplementary information in Table 1 is as follows. Col. 3 gives the order of magnitude of the error in the starter, as a power of e. Col. 4 indicates, by a tick when appropriate, that the starter will always exploit the better-than-linear convergence of the Newton–Raphson and higher-order iteration processes; this 'freedom from slow convergence' exists so long as the starter provides good initial values when $f'(E)$ approaches zero. A tick in Col. 5 indicates that, with this starter, the Newton–Raphson process is guaranteed to converge, albeit slowly perhaps. Col. 6 is ticked when the starting formula does not demand a preliminary reduction to get M in the range $(0, \pi)$. Finally, Col. 7 ticks the starters that are 'smoothly portable', and this is a concept worth some immediate comment.

Two obvious desiderata of a starting formula for the solution of Kepler's equation, and of a computing procedure in general, are that it should be as fast and accurate as feasible. A third desideratum is that it should be as aesthetically pleasing as possible, with arbitrariness held to a minimum; this desideratum is contravened if, as applies to many procedures for solving (1), the starter is composed of a number of sub-formulae, to be used in different regions of the (e, M)-plane, patched together in a manner that is essentially *ad hoc*. The concept of 'smooth portability' may be regarded as a fourth desideratum that is closely allied to the third. What is desired amounts to the requirement that any patching should be continuous, down to the quantum level of the word-length of an arbitrary computer, and that this continuity should apply not merely to the sub-formulae themselves but also to their first derivatives; in particular we require that if the starting formula only applies to the range $(0, \pi)$ for M, then it should map this range onto (not just into) $(0, \pi)$, so that range reduction is continuous (continuity of first derivative may then be seen to be an automatic consequence of the symmetry). The significance of a starter's smooth portability in this sense is that, in conjunction with an iteration process that operates for a fixed number of iterations, the complete procedure is automatically well adapted to any computer, since even if it terminates without exploiting the full accuracy of a particular machine, at least the output E (and its derivatives with respect to e and M) will be essentially continuous.

The first three starters are defined by the first three truncations of the power-series expansion of E – there can be little point in ever going to more than three terms, in particular because the series is not convergent for all $e \leqslant 1$. The use of one of these starters, with the Newton–Raphson iterator, has been the standard method of solving Kepler's equation for many years, but it has long been recognized that convergence is not guaranteed. Full analysis of iteration from S_1, by Gooding and Odell (1985), shows that the boundary between convergence and divergence is marked by oscillation with a two-iteration cycle, for which the condition is given by the equation

$$e = \sin E_1 / \sin (2E_1 - \tan E_1),$$

where E_1 is the value of E after one iteration (and hence any odd number of iterations). The minimum value of e, such that this equation has a solution for E_1, is about 0.9733, the value of M that leads to this solution being about 0.246 rad; for any $e > 0.9733$ there is a range of values of M, roughly centred on 0.246 rad, for which the Newton–Raphson iterator diverges, the limiting range (where $e = 1$) being from zero to about 0.493 rad. With the starters S_2 and S_3, the limiting values of e for Newton–Raphson divergence are much higher, being about 0.9936 for S_2 and 0.9972 for S_3.

It was remarked by Smith (1979) that the starter $M + e$ (for $0 \leqslant M \leqslant \pi$) would always lead to convergence under Newton–Raphson iteration, and later authors have regularly referred to this fact, so we include this starter as S_4 in Table 1. Smith found S_4 superior to all the other starters he considered, with the exception of the one we denote by S_5, given by

$$S_5 = M + \frac{e \sin M}{1 - \sin(M + e) + \sin M}.$$

S_5 arises as a secant-generated combination of S_1 and S_4, this being a natural combination since $S_1 \leqslant E \leqslant S_4$ (for $0 \leqslant M \leqslant \pi$); it is not free of Newton–Raphson divergence, however (though Smith seems not to have recognized this), which occurs for values of e greater than about 0.9995. The source of the trouble is the same as for S_1, S_2 and S_3, viz. that if $0 < S < E$ (with $0 < M < E < \pi$), then E_1 (the estimate of E after one iteration) will exceed E (the true value) and sometimes be much bigger than π. If $0 < E < S \leqslant \pi$, on the other hand, then $E < E_1 < S$ and convergence is guaranteed. On this basis, the simplest possible starter for which convergence is certain (with $0 \leqslant \mathrm{M} \leqslant 2\pi$) is π itself. It would be inefficient to iterate from π, however, so for our next starter we use the associated E_1, given by one Newton–Raphson iteration, writing

$$S_6 = M + \frac{e(\pi - M)}{1 + e}. \tag{3}$$

We also have $S_6 = (M + e\pi)/(1 + e)$, which exhibits it as a weighted average of S_1 ($= M$) and π. This starter does not appear to have been given explicitly before, but the quantity concerned was introduced by Broucke (1980) as an upper bound for E over the range $\pi - (e + 1) \leqslant M \leqslant \pi$.

The development of S_6 suggests two further starters not previously presented. Broucke was very close to the first of these, since his upper-bound analysis led him to

$$\mathrm{Min}(M/(1 - e),\ M + e,\ \pi).$$

Having recognized S_6 as an upper bound, therefore, he might well.have given instead (for $0 \leqslant M < \pi$) the starter we write as

$$S_7 = \mathrm{Min}\,(M/(1 - e), M + e, (M + e\pi)/(1 + e)).$$

Again, since Equation (3) is in the form

$$S_6 = S_1 + \lambda e(\pi - S_1),$$

we are led to consider potentially superior starters of the form

$$S_8 = S_3 + \lambda e^4(\pi - S_3),$$

the rationale for taking λe^4 (rather than λe^3) as a coefficient being that, although S_3 is only correct to $O(e^2)$, it is undesirable to corrupt it with an excessive term in e^3. The choice of λ requires some care: if it is too small, then for e close to 1 and M of the order 10^{-2} to 10^{-3} rad, the Newton–Raphson process will still be in danger of diverging; if λ is too large, on the other hand, then for e close to 1 and extremely small M, S_8 will so over-estimate E that convergence becomes impracticably slow. A suitable compromise value was found to be $1/20\pi$.

The origin of 'slow convergence' with S_4, S_6 and S_8 is that, for these starters, zero M does not generate zero S when $e \neq 0$. An obvious way to remedy this is to consider only starters of the form $M + (e \sin M)G(e, M)$ for suitable functions G. The starters S_1, S_2, S_3 and S_5 are of this form, but all these can lead to Newton–Raphson divergence, the avoidance of which is in conflict with the avoidance of slow convergence. One approach to the dilemma is to take the starter in this form, but with G having a convenient denominator such that $1/G(1,0)$ vanishes but $(e \sin M)G(e, M)$ has a finite limit at $(1, 0)$. The first candidate to suggest itself for $G(e, M)$ is $(1 - e \cos M)^{-1}$, but this gives $\cot \frac{1}{2}M$ for $G(1, M) \sin M$, and so violates the requirement for a finite limit. This starter would not even have novelty, since it is identical with the outcome (E_1) of one Newton–Raphson iteration from the starter S_1 – the divergence associated with S_1 can in fact be attributed directly to the infinite limit.

The starter just rejected may be written as $M + \alpha$, where $\alpha = (e \sin M)/(1 - e \cos M)$, and from this may be derived the more complicated starter, correct to $O(e^3)$, $M + \alpha(1 - \frac{1}{2}\alpha^2)$, which is considered by Smith (1979), Ng (1979), Broucke (1980) and Danby and Burkardt (1983). This also has an infinite limit, however, and again lacks novelty since, as pointed out by Danby and Burkardt, it is just the E_1 resulting from an iteration of the process we shall consider in Section 3 and attribute to Chebyshev.

The next idea is to reduce the power of the denominator in $G(e, M)$ by introducing a square root. The obvious candidate is $(1 - e \cos M)^{1/2}$, which gives $\sqrt{2}$ (rad) for the limit of $G(1, M) \sin M$. The starter given by this $G(e, M)$ is free of Newton–Raphson divergence, though not of slow convergence, and we exclude it from Table 1 for only one reason, viz. that there is a very similar starter which is superior. This is given by

$$S_9 = M + \frac{e \sin M}{(1 - 2e \cos M + e^2)^{1/2}}. \tag{4}$$

The function $G(e, M)$ associated with (4) may be seen to have, as its power-series expansion in e, the Legendre series of general term $e^j P_j(\cos M)$. This matches the series that formally solves Kepler's equation as far as the e^3 term, so that the error in S_9 is $O(e^4)$. The limit of $G(1, M) \sin M$ is now 1 rad. As S_9 is the starter-component of the first of the two Kepler-solution procedures that we describe in detail in Section 4, we give it no further attention here.

The only defect of S_9 is the one suffered by all the starters so far considered, viz. the propensity for slow convergence when (e, M) is close to $(1, 0)$. Because $f'(E)$ is zero at this point, fast convergence is only possible in its vicinity if a starter is employed that reflects the true behaviour of E. It was noted by Ng (1979) that E behaves like the cube root of $6M$ when (e, M) is close to $(1, 0)$, and the remaining starters we consider all reflect this fact.

The simplest cube-root-reflecting starter is just $\sqrt[3]{6M}$, which is obviously appropriate when $e = 1$ and $M \approx 0$, so that $f(E) = E^3/6 + O(E^5)$. In fact it works very well for any $e(\leqslant 1)$ and over the whole basic range of $M(0 \leqslant M \leqslant \pi)$, though it is obviously not very efficient in general. To get a more efficient starter, we follow Ng in neglecting E^5 and higher powers of E, in which case Equation (1) may be written

$$E^3 + 3qE - 2r = 0, \qquad (5)$$

where

$$q = 2(1 - e)/e, \qquad r = 3M/e \qquad (6)$$

in the notation of Ng (1979). Equation (5) is the classical cubic equation in reduced form (no E^2 term) and Ng gave its solution as

$$[r + (q^3 + r^2)^{1/2}]^{1/3} + [r - (q^3 + r^2)^{1/2}]^{1/3}.$$

This may be better expressed, however, both to avoid a pair of cube roots where only one is necessary and also to avoid (when $q^3 \ll r^2$) the subtraction of almost equal quantities. We define

$$s = [(r^2 + q^3)^{1/2} + r]^{1/3}, \qquad (7)$$

therefore, and then write

$$S_{10} = s - \frac{q}{s}. \qquad (8)$$

Given the different merits of S_9 and S_{10}, we now looked for a formula, intricate if necessary, that would combine these merits. Returning to our consideration of starters of the form $M + (e \sin M)G(e, M)$, in which a particular candidate for $G(e, M)$ was $(1 - e \cos M)^{-1/2}$, we observe that if the square root is replaced by a cube root, then near $(1, 0)$ the starter behaves like $\sqrt[3]{2M}$. Thus behaviour of the form $\sqrt[3]{6M}$ is given by taking $G(e, M)$ as $\sqrt[3]{3}\ (1 - e \cos M)^{-1/3}$. The resulting starter is like S_1 (and unlike S_{10}) in being valid, in a single formula, over the complete range of M and for any $e(\leqslant 1)$, but some refinements are possible. First, to make it (like S_9) agree with the power series for E to terms in e^3, we replace $\sqrt[3]{3}$ by a more complicated numerator. A natural approach is to write

$$G(e, M) = \frac{1 + \frac{2}{3}e \cos M + \frac{1}{36}e^2(1 + 19 \cos 2M) + e^3(\alpha + \beta \cos M + \gamma \cos 2M)}{(1 - e \cos M)^{1/3}},$$

where α, β and γ are disposable constants (multiplying the fourth power of e in the starter) that are assigned with the overriding object of restoring the right numerator for (e, M) near $(1, 0)$. This condition leads to

$$\alpha + \beta + \gamma = \sqrt[3]{3} - \tfrac{20}{9},$$

and two more conditions on the constants can be obtained by making the starter approximately correct for two additional values of M (with $e = 1$), the obvious values being $\frac{1}{3}\pi$ and $\frac{2}{3}\pi$.

The other refinement we make, before formally labelling our starter as S_{11}, is more important. It arises because $1 - e \cos M$ is not quite the right quantity to be under the cube root in the denominator of $G(e, M)$. By splitting $1 - e \cos M$ into the sum of $e(1 - \cos M)$ and $1 - e$, we get

$$1 - e \cos M \approx \tfrac{1}{18} r^2 + \tfrac{1}{2} q \tag{9}$$

when q and r, given by (6), are both small. But if $q^3 \ll r^2 \ll q (\ll 1)$, we may expand (7) to give

$$s = (2r)^{1/3}\left(1 + \frac{q^3}{12r^2} + \cdots\right)$$

and on substitution in (8) it becomes clear that the dominating term $\frac{1}{2}q$ in (9) may be excessive: it was found, in practice, to be a source of slow convergence in cases such as $(e, M) = (0.9999, 0.0001)$. However, $q^3/8$ (in place of $\frac{1}{2}q$) should give no difficulty, the significance of this being that

$$1 - g \cos M \approx \tfrac{1}{18} r^2 + \tfrac{1}{8} q^3$$

if

$$g = 1 - (1 - e)^3.$$

The only problem with $1 - g \cos M$ is that it behaves like $1 - 3e \cos M$ when e is small. What we require, for the coefficient of $\cos M$, is a linear combination of e and g, and the combination $\sigma e + (1 - \sigma) g$ is suitable if $\sigma = (1 - e)^2$, since if $e \approx 0$ this approximates to e, with a matching first derivative, and if $e \approx 1$ it approximates to g, with matching first and second derivatives. The use of σ leads to $1 - e[1 + ee_1(1 + e_1)^2] \cos M$ for the quantity under the cube root of the desired starter, where $e_1 = 1 - e$, but there is an induced effect on the numerator of $G(e, M)$, if the starter is to remain correct to $O(e^3)$. We finally get

$$S_{11} = M + \frac{e \sin M [1 + \frac{2}{3} e \cos M + \frac{1}{36} e^2 (1 - 48 \cos M + 19 \cos 2M) + e^3(\alpha + \beta \cos M + \gamma \cos 2M)]}{\{1 - e[1 + ee_1(1 + e_1)^2] \cos M\}^{1/3}} \tag{10}$$

where

$$e_1 = 1 - e, \qquad (\alpha, \beta, \gamma) = -\tfrac{1}{6}(\sqrt[3]{3} - \tfrac{8}{9})(1, -9, 2); \tag{11}$$

α, β and γ have convenient values, only their sum having a precisely determined value.

When S_{11} is substituted for E in (1), the (numerically) maximum residual is about 0.26 rad and occurs at $(e, M) = (1, 2.6$ rad). This is greater than the maximum residual with S_9 (0.16 rad, occurring at (1,0)), but it occurs at a point at which there are no convergence problems. The starter qualifies as the best of all those in Table 1, in receiving a full quota of ticks, and we would be recommending it as a standard all-purpose starter, for use in particular whenever S_9 is inadequate, if it were not for the discovery of a final starter (S_{12}) that is much faster to compute and has fewer arbitrary features. The description of this is deferred to Section 4.

3. Iterative Processes

Newton's method for locating the root of an equation, referred to here as the Newton–Raphson process, is so much better known than any other method that it is worth pointing out that, in addition to the more advanced processes that stem from the Newton–Raphson, there are also more elementary methods that can be used in the solving of Kepler's equation. The secant method and *regula falsi* (the method of false position) are examples, and so is the bisection method. The latter iterates by bisecting the interval in which E must lie, starting (if $0 \leqslant M \leqslant \pi$) with the interval $(0, \pi)$ and testing the mid-point of the current interval during each iteration. After n iterations the maximum error in the best estimate of E is $\pi/2^{n+1}$, so 34 iterations would give an accuracy of 10^{-10} rad; this is not as inefficient as it may seem, since the computation of each $\sin E$ (to test-substitute in (1)) will be very fast if auxiliary tables of $\sin E$ and $\cos E$ are stored in the computer for $E = \pi/2^i$ with $i = 2, 3, \ldots, n$. A method that is rather less primitive, and particularly well adapted to use with hand calculators of the Hewlett-Packard type, iterates from E_i to E_{i+1} via the formula

$$E_{i+1} = M + e \sin E_i;$$

the obvious starter to use is S_1, so that $E_0 = M$. The process converges for all e ($\leqslant 1$), but convergence becomes infinitely slow if $e = 1$ and M approaches zero (or any multiple of 2π).

The last-mentioned process gives linear convergence, i.e. of first order only; thus, if ε_i is the error in E_i, ε_i is of order $e^i \varepsilon_0$, so that (for example) if $e < 0.1$, each successive iteration gives at least one more significant figure in E. The advantage of the Newton–Raphson process and the more advanced processes that stem from it is that their convergence is of second order at least, except when $f'(E) \approx 0$ and an inadequate starter is in use. Convergence of order k means that ε_{i+1} is of order ε_i^k so that, roughly speaking, the number of significant figures in E is multiplied by k during each iteration. The concept will not be analysed further here, as Danby and Burkardt (1983) give a good account, but it must be emphasized that even high-order convergence reverts to linear if $f'(E) \approx 0$ and the starter is poor.

To simplify the notation in the following analysis of generalized Newton processes, we confine ourselves to the consideration of a single iteration. We suppose that $f(x)$ is a given monotonic (and continuous) function and that an approximation to a root of $f(x)=0$ is given by $x=\xi$. We denote $f(\xi)$ by η and suppose that the derivatives η', $\eta'', \ldots, \eta^{(k)}$ (for some k in due course to be associated with convergence of order $k+1$) are also available; for the particular f of (2), of course, all the higher derivatives are known at once from η and η'. The problem is simply to use η, η' etc to derive a value of δ such that $\xi+\delta$ is much closer than ξ is to the root of $f(x)=0$. We note that, from the assumptions about f, the inverse function (f^{-1}) is defined, and we denote its derivatives, at $f^{-1}(\eta)=\xi$, by ξ', ξ'' etc; thus,

$$\xi' = 1/\eta', \qquad \xi'' = -\eta''/(\eta')^3 \text{ etc.}$$

From the Taylor expansion of $f(\xi+\delta)$, we see that we should naturally like δ to be a root of

$$\eta + \eta'\delta + \tfrac{1}{2}\eta''\delta^2 + \cdots + \frac{1}{k!}\eta^{(k)}\delta^k = 0. \tag{12}$$

For $k=1$, we immediately have the Newton–Raphson formula, with its quadratic convergence, but for higher values of k we have to locate the appropriate root of a polynomial. Before taking this further, however, we consider the much simpler analysis arising from the Taylor expansion of $f^{-1}(0)=f^{-1}(\eta-\eta)$ which, if it were exact, would give the root of $f(x)=0$ at once. If we take this expansion to be our $\xi+\delta$ in fact, then (on subtracting ξ from both sides of the resulting equation) we have

$$\delta = -\xi'\eta + \tfrac{1}{2}\xi''\eta^2 - \cdots + \frac{1}{k!}\xi^{(k)}(-\eta)^k. \tag{13}$$

For $k=1$, we again have the Newton–Raphson formula, since $\xi'=1/\eta'$.

For $k=2$, (13) gives (on rewriting η as f, to produce a more familiar expression)

$$\delta = -\frac{f}{f'}(1+\tfrac{1}{2}ff''/f'^2), \tag{14}$$

which is equation 20 of Danby and Burkardt (1983). Taking $k=3$ gives, similarly, their equation 22. The iterative process defined by (14) is often quoted in the literature: Bergam and Prussing (1982) describe it simply as 'second-order Newton' whilst Ng (1979) attributes it to Schröder, and Broucke (1980) to Chebyshev; Traub (1961) has some historical notes on the general approach, which he traces back to Euler. In spite of this pedigree, the cubic convergence given by (14) is generally inferior to that given by Halley's formula, which we shortly introduce as equation (17), since polynomials usually give better representations of naturally occurring functions than they do of the inverses of these functions. This is obviously true for the Kepler function, and herein lies a complete explanation of the 'strange attractors' of Broucke (1980). The most striking example of 'strange behaviour' is when

(14) gives $\delta = 0$, from $ff'' + 2f'^2 = 0$, and hence a false solution of Kepler's equation. A full analysis of this is given by Gooding and Odell (1985), who show that the simplest example of the phenomenon occurs with the starter S_1 ($E_0 = M$), for (e, M) such that $\cos M = e = \sqrt{2/3} \approx 0.8165$; then $-ff'' = 2f'^2 = 2/9$ and (14) falsely implies that the starting value is already a solution! Neither the Halley process nor even the bare Newton–Raphson process has any difficulty with such examples. We cannot too strongly stress the advantages of the Halley process over the Chebyshev process, since the difficulty in getting the message across can be seen from the recent paper by Peters (1984).

We now return from (13) to the harder-to-solve, but more rewarding, (12). For $k = 2$, we have

$$\eta + \eta'\delta + \tfrac{1}{2}\eta''\delta^2 = 0, \tag{15}$$

for which the schoolboy's solution is available if we are willing to extract a square root and select a sign. Another, more general, approach is available, however: we denote the Newton–Raphson solution of (12) by δ_1, so $\delta_1 = -\eta/\eta'$, and linearize (15) by rewriting it as

$$\eta + \delta(\eta' + \tfrac{1}{2}\eta''\delta_1) = 0;$$

this is an approximation, but nothing has been lost since (15) was already only an approximation of (12). We denote the solution of (16) as δ_{12}, for a reason that will become apparent, writing (with f for η)

$$\delta_{12} = -\frac{f}{f' + \tfrac{1}{2}\delta_1 f''}; \tag{16}$$

this is Equation (17) of Danby and Burkardt (1983), and their Equation (18) follows by the natural extension to $k = 3$, using δ_{12} to generate a linear equation in what we call δ_{123}. Substituting for δ_1 in (16) we get a self-contained formula for an iteration process with cubic convergence, viz.

$$\delta_{12} = -\frac{f}{f' - \tfrac{1}{2}ff''/f'}\cdot \tag{17}$$

This is the formula of Halley (1694) for iterating the root of an equation; it has often been rediscovered, and given other names, as remarked by Traub (1961). It is interesting to note that it can also be obtained, perhaps more naturally, from the root of the unique bilinear function (if $\eta' \neq 0$) that passes through (ξ,η) and matches the derivatives η' and η''.

The basis for the notation δ_1,δ_{12},δ_{123} etc is that any existing δ,δ_{ex} say, can be used to linearize (12) so that a new δ is obtained from

$$\eta + \delta\left(\eta' + \tfrac{1}{2}\eta''\delta_{ex} + \cdots + \frac{1}{n!}\eta^{(n)}\delta_{ex}^{n-1}\right) = 0,$$

where $n \leqslant k$. Thus a chain of suffices, of the form $n_1 n_2 \ldots n_N$, can be developed such that δ, with this suffix, is obtained from δ_{ex} with $ex = n_1 n_2 \ldots n_{N-1}$. Clearly n_1 must be 1 and $n_j > 1$ if $j > 1$; also $n_j \geqslant n_i$ when $j > i$, if the development is to be useful. To judge the possible merit of suffix chains such as 13, 122 etc, we must consider the order of convergence developed: if O_N is the order associated with the suffix chain $1n_2 \ldots n_N$, then O_{N-1} can be evaluated recursively, and it is not hard to see that (with $O_1 = 2$)

$$O_N = \min(O_{N-1} + 1,\ n_N + 1);$$

thus only the progression δ_1, δ_{12}, $\delta_{123}, \ldots$, can be efficient.

The rationale for proceeding to δ_{12} and beyond is, of course, to obtain the maximum benefit from one iteration before advancing (if necessary) to the next. This is particularly appropriate if, as with (2), the computation of higher derivatives is trivial. The development described does not exhaust the possibilities, however, since at each stage, say when the suffix chain is about to reach n_N, another option is available: instead of computing the new δ, with suffix $1n_2 \ldots n_N$, we use the existing δ_{ex} to compute a new η, associated with $\xi + \delta_{ex}$ (as new ξ) and given by

$$\eta + \eta'\delta_{ex} + \tfrac{1}{2}\eta''\delta_{ex}^2 + \cdots + \frac{1}{n_N!}\eta^{(n_N)}\delta^{n_N};$$

we also compute the corresponding new η', η'' etc as required. In Gooding and Odell (1985), where the description is amplified with a sketch, this step is described as 'rectification'. What we are doing is to abandon the original point (ξ, η), that certainly lies on the function $y = f(x)$, in favour of a new point that in general does not, but instead lies on a polynomial curve, of degree n_N, that is hyperosculatory to the function at the original point – the object is to postpone the start of the next iteration and its computation of entirely fresh derivatives.

After rectification the development of the suffix chain can continue, necessarily starting with a new 1. Thus the notation extends automatically and we have unambiguous chains such as 12312123 (two rectifications). To clarify, we observe that the overall formula for δ_{121}, which has the simplest new suffix chain now available, is given by the combination of an initial Newton–Raphson component ($\delta_1 = -f/f'$) and then, after second-degree rectification, a second Newton–Raphson component. The rectified f and f' are given by $f + f'\delta_1 + \frac{1}{2}f''\delta_1^2$ and $f' + f''\delta_1$, respectively, and the overall formula is

$$\delta_{121} = -\frac{f(f'^2 - \frac{1}{2}ff'')}{f'(f'^2 - ff'')}. \tag{18}$$

Clearly (18) is another cubic-convergence formula, being a third special case (with $\lambda = -\frac{1}{2}$) of the general cubic formula

$$\delta = -\frac{f[f'^2 + \lambda ff'']}{f'[f'^2 + (\lambda - \frac{1}{2})ff'']}$$

that covers (17) and (14), with $\lambda = 0$ (for Halley) and $\lambda = \frac{1}{2}$ (for Chebyshev) respectively.

Though δ_{121} does not give better convergence than δ_{12}, both being cubic, δ_{131} does do better than δ_{13}, since the latter is still only cubic whilst the former is quartic. The overall formula is

$$\delta_{131} = -\frac{f(f'^3 - \frac{1}{2}ff'f'' + \frac{1}{3}f^2f''')}{f'(f'^3 - ff'f'' + \frac{1}{2}f^2f''')} \tag{19}$$

which is a special case of the general quartic formula given by Gooding and Odell (1985). The components of (19) are still both Newton–Raphson, and in looking for an iterator to associate with our final starter (S_{12}), we also looked at δ_{1312}, δ_{1231} and δ_{12312}, in all of which there is at least one Halley component; as we shall see in Section 4, we selected the δ_{1231} process.

Though δ_{131} does better than δ_{121} (quartic convergence as opposed to cubic), δ_{141}, δ_{151} etc still only give quartic convergence, not quintic. This is obvious, since they cannot do better than $\delta_{1\infty 1}$, which is defined as the limit of δ_{1n1} and hence represents two distinct iterations of the Newton–Raphson process; but the composition of two quadratic iterations must be quartic, if viewed as a single iteration.

4. Two Kepler Procedures Used at RAE

4.1. The Older Procedure

The earliest procedures used at the Royal Aircraft Establishment, in solving Kepler's equation, were based on the starters, S_1, S_2 and S_3, with iteration by the Newton–Raphson process, and some computer programs still contain Fortran versions of these procedures. Since about 1977, however, the most efficient procedure in general use has been the one listed in Appendix A as the Fortran function EKEPL1 (this procedure, like the earlier ones, has also been used with the name EAFKEP); it is based on the starter S_9, defined by Equation (4), with iteration by the Halley process.

As remarked in Section 2, the starter is formally correct to order e^3, and it leads to a maximum residual error of about 0.16 rad, when substituted in (1); this occurs when $e = 1$ and M is a multiple of 2π, since the limiting value of (4) is then 1 rad. For low values of M (taking $0 \leqslant M \leqslant \pi$) the starter, which is a monotonic function of M, always over-estimates E, assuming $e > 0$, but there is a cross-over value of M, above which the starter underestimates E. For the cross-over M, (4) generates an immediate solution of (1), and this value of M is plotted against e in Figure 1, together with the corresponding value of E.

Two features of the Halley iteration process in EKEPL1 are worth discussing. The first concerns the convergence criterion, which is specified by the satisfying of (1) to a residual whose magnitude does not exceed 10^{-4} rad. This gives very much better accuracy than at first sight appears, since the criterion is effectively applied to an E_i

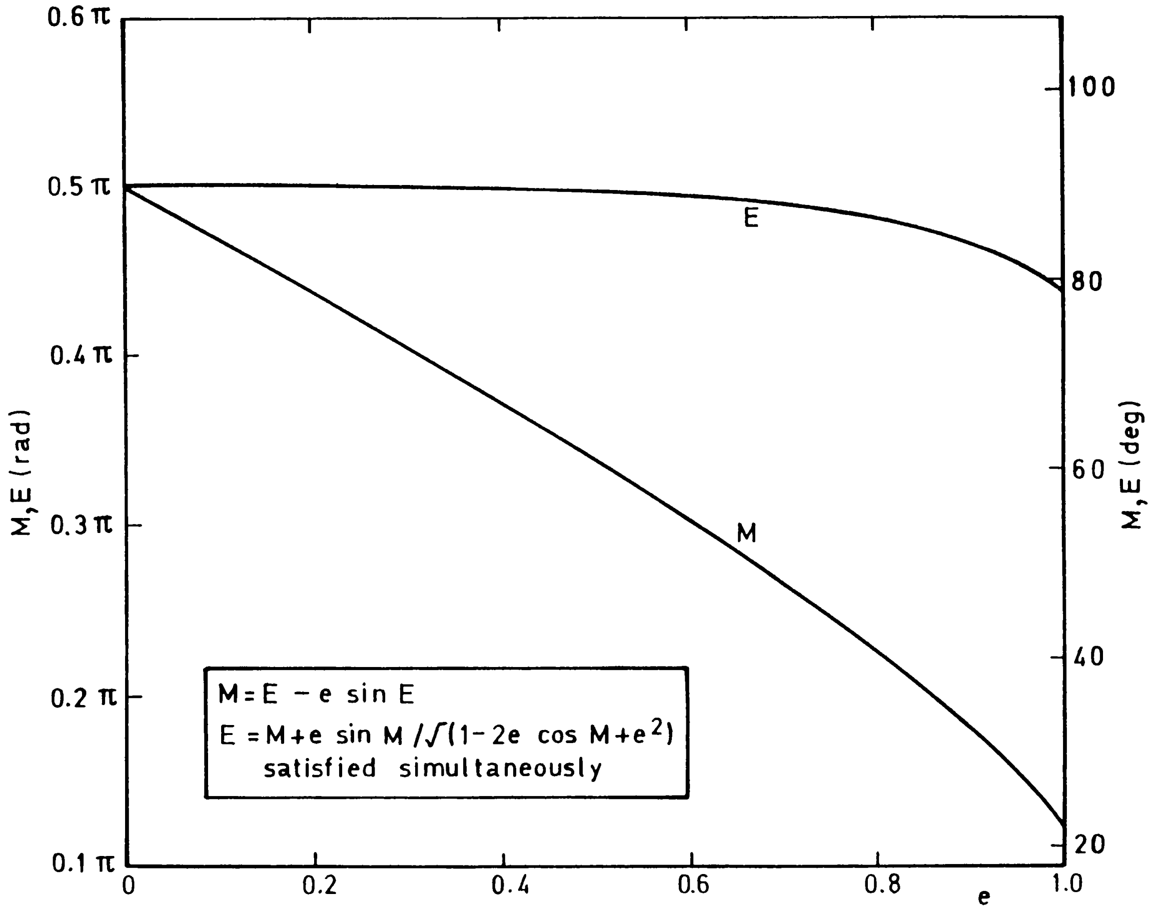


Fig. 1. Cross-over values of M and E for the S_q starter.

that is an iteration behind the current best estimation (E_{i+1}) of the process. Thus it follows from the cubic convergence–cf. the analysis of Danby and Burkardt (1983) and the remarks of Smith (1961) on convergence termination – that f_{i+1} is $O(f_i^3)$, so long as f'_i is not too small. Hence the 10^{-4} criterion actually implies an accuracy of the order of 10^{-12} rad, except when (e, M) is in the vicinity of $(1, 0)$ where the degeneration to linear convergence is bound to make EKEPL1 behave badly.

When convergence degenerates, the solution of (1) is known to behave like the solution of $E^3 = 6M$ and, as shown by Gooding and Odell (1985), the residual in (1) then reduces by a factor of 8 on each iteration; thus the initial residual of about 0.16 rad reduces to about 4×10^{-5} rad after the fourth iteration, which makes the fifth iteration the terminal one with an (implicit) residual of about 5×10^{-6} rad. With e limited to 0.9, however, there will never be more than three iterations and the maximum residual in (1) is about 10^{-10} rad.

The other feature of EKEPL1 worth referring to concerns its automatic applicability to arbitrary values of M, as Table 1 promises for the starter. The procedure operates by the refinement of ψ, equal to $E - M$, rather than E itself, which avoids the theoretical possibility, for very large values of M and computers of low precision, that the convergence criterion might fail to operate and the process continue indefinitely. This would be a consequence of rounding error, of course, to which otherwise little attention is given in EKEPL1. For the other procedure (EKEPL2), to which the rest of Section 4 is devoted, rounding error has been allowed for very carefully, since this procedure was designed for maximum accuracy under all circumstances.

4.2. The New Procedure – Starter

The difficulty in solving Kepler's equation increases as e rises from zero to unity, but in one respect (as we have seen in discussing the starter S_{10}) there is an immediate simplification with $e = 1$, since the starter $\sqrt[3]{6M}$ works so well. It was decided to base the starter for a new procedure on two principles, therefore: first, that the starter for arbitrary e should be obtained, via linear interpolation, from the starter for $e = 1$; second, that this special starter, which would now always be used, should be made as efficient as possible. We consider these points in turn.

If E_{01} denotes a starting value (E_0) for $e = 1$, then the linearly interpolated starter for general e is given by

$$E_0 = eE_{01} + e_1 M, \tag{20}$$

where $e_1 = 1 - e$ as in (11). To assess the interpolation principle, before the E_{01} starter had been selected, it was decided to pretend that an exact E_{01} would be available and to look at the value of $f(E_0)$ given by (2), i.e. to look at $M_0 - M$, where

$$M_0 = E_0 - e \sin E_0. \tag{21}$$

We also have

$$M = E_{01} - \sin E_{01} \tag{22}$$

from the assumption about E_{01}, so if we regard E_{01} as the independent variable, with the resulting M, E_0 and M_0 given by (22), (20) and (21) in turn, we can do the analysis without actually having to solve either of the two versions of Kepler's equation. We get

$$\begin{aligned} M_0 - M = ee_1 E_{01}\{1 - \tfrac{1}{6}E_{01}^2(2 + e + e^2) + \\ + \tfrac{1}{120}E_{01}^4(2 + e + 11e^2 + e^3 + e^4) + O(E_{01}^6)\}. \end{aligned} \tag{23}$$

Equation (23) can be differentiated with respect to E_{01} to locate the approximate value of E_{01} (and hence M) that, for a given e, makes $M_0 - M$ maximum; in particular, we find (from the smaller root of the quadratic equation in E_{01}^2) that E_{01} is about 1.05 rad and 0.79 rad for $e = 0$ and 1 respectively, the corresponding values of M being 0.18 rad and 0.08 rad.

These results were confirmed by direct numerical analysis of Equations (20) to (22). Figure 2 plots results for the full range of e, showing (for each e) both the maximum value of $|M_0 - M|$ and the value of M to which it applies. The overall maximum value of $|M_0 - M|$ is about 0.149 rad, occurring for an e of about 0.46.

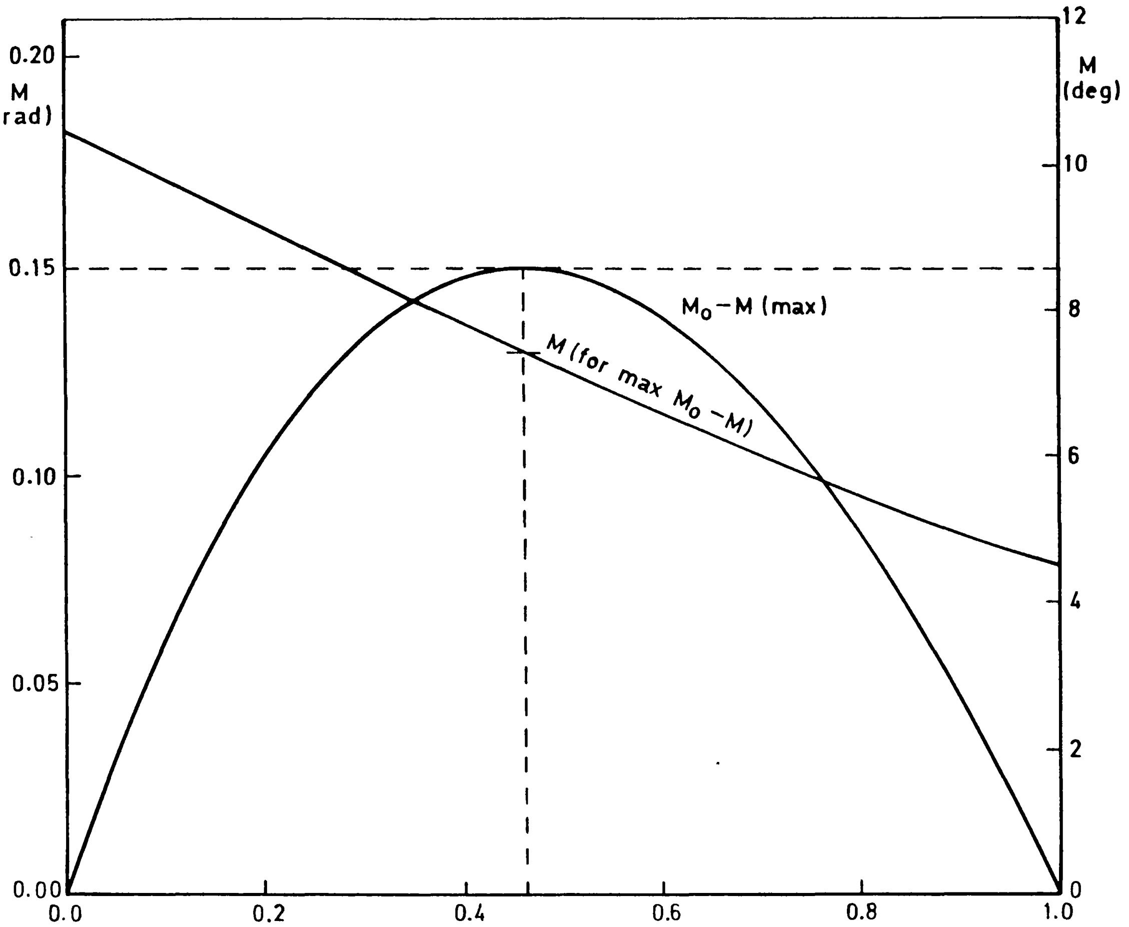


Fig. 2. Maximum residual ($M_0 - M$) for hypothetical starter linearly interpolated from the exact solution for $e = 1$.

(For each e there is an M for which the error is zero, and for larger M the value of $M_0 - M$ is negative, but the maximum value of $|M_0 - M|$ is always for $M_0 > M$).

After this analysis it seemed probable that linear interpolation, based on a sound starter for $e = 1$, would give good results. It remained to choose the special starter with efficiency in mind, the assessment being in parallel with the choice of the iterator, to be described in Section 4.3. All the desiderata of Section 2 were to be satisfied if possible, and attention was directed to the use of the bilinear formula,

$$E_{01} = \frac{\alpha + \beta M}{\gamma + M}, \tag{24}$$

with suitable constants α, β and γ. This is very fast to compute, following a normalization of M to the range, $(0, \pi)$, and it was soon found that good results could be obtained if the lower limit of the range was changed from zero to a small positive quantity. A second bilinear formula could then be patched in to cover the bottom of the range, but it was found that three such formulae would be necessary to guarantee an accuracy (in satisfying Equation (1)) of 10^{-13} rad with the two-iteration convergence process that was emerging.

The patching of bilinears was not very appealing, however, neither of the last two desiderata being satisfied. Even the accuracy remained unsatisfactory, for very small M, being poor in relative (as opposed to absolute) terms. The reason for this lies in the impossibility of representing a cube root of a bilinear formula, when the range covered extends to zero. An obvious solution to the difficulty would be to use the simplest cube-root starter, $\sqrt[3]{6M}$, over the entire range. One objection to this is the resulting discontinuity at $M = \pi$, where the starter would give about 2.66 rad, for E_{01}, instead of π; this could be dealt with in various ways, a surprisingly efficacious one being just to replace $\sqrt[3]{6M}$ by $\sqrt[3]{\pi^2 M}$. The real objection to a universal cube-root starter, however, is the time wasted in computing it.

The natural compromise was to patch the $\sqrt[3]{6M}$ starter into a single bilinear formula, if a way of doing this could be found that left all the desiderata satisfied. The transition point would inevitably be, to some extent, arbitrary, but the arbitrariness ends here since the constants in (24) can be chosen to satisfy three obvious conditions: that E_{01} is continuous at the transition; that its derivative with respect to M is likewise; and that $E_{01} = \pi$ when $M = \pi$. (Continuity of the derivative is then automatic, at $M = \pi$, as remarked in Section 2.)

To make our starter 'smoothly portable' at $M = \pi$, we re-express (24) as

$$U = \frac{aW}{b - W}, \tag{25}$$

where

$$W = \pi - M, \qquad U = \pi - E_{01}. \tag{26}$$

It is assumed, of course, that π is held to the maximum possible accuracy on the computer used (Fortran variable *PI* in the listing of Appendix B), but no other constants have to be approximated. It turned out that a very suitable transition value was given by $M = \frac{1}{6}$ rad, since (25) can then be patched to both E_{01} (= 1 rad) and its derivative by taking

$$a = (\pi - 1)^2/(\pi + 2/3) \tag{27a}$$

and

$$b = 2(\pi - 1/6)^2/(\pi + 2/3). \tag{27b}$$

Figure 3 plots the cube-root and bilinear curves that, between them, yield the adopted E_{01}, the unused segment of each curve being included for completeness. The true plot of E against M (for $e = 1$) is also provided. The unused segment of the cube-root curve yields about 2.66 rad at $M = \pi$, whilst the unused segment of the bilinear curve yields about 0.63 rad at $M = 0$. The slope of the bilinear curve, at $M = \pi$, is given by a/b, which is about 0.26; the true slope is 0.5 exactly, but (as has been remarked) this is of no consequence.

Equations (20) and (25) to (27) determine the starter S^{12}. The maximum value of

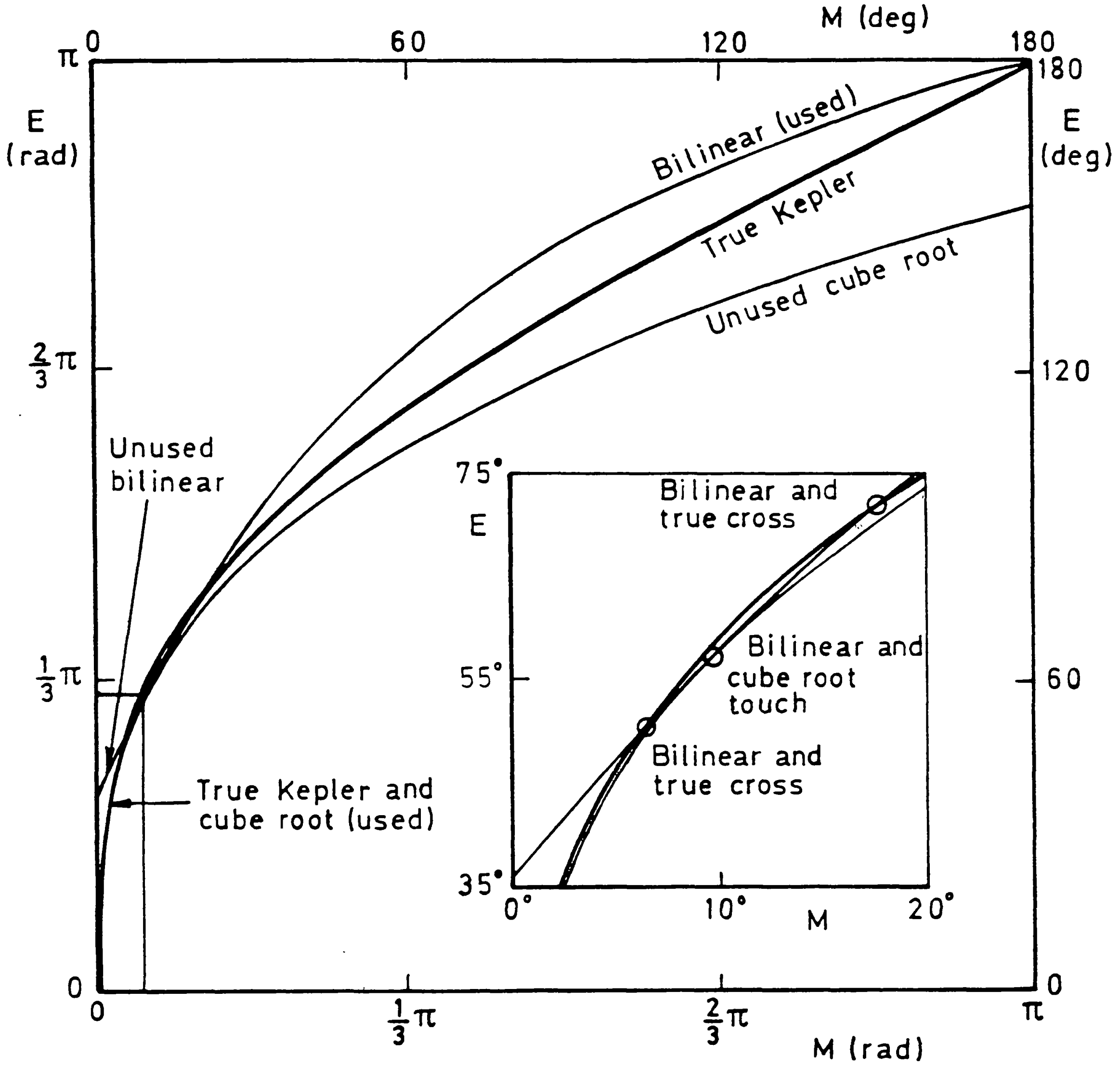


Fig. 3. The S_{12} starter ($e = 1$) – cube root and bilinear functions.

$|M_0 - M|$ that results, with M_0 given by (21), is plotted against e in Figure 4, which effectively supersedes Figure 2 now that a real starter (rather than a hypothetical one) is available. For values of e up to about 0.689, the maximum $|M_0 - M|$ occurs for values of M for which the solution of (1) is always around 0.2 rad, but for higher values of e this peak is exceeded by another, associated with values of M for which E is about ten times greater. Thus Figure 4 is the upper envelope of two curves, and extensions of these are shown at the intersection point. The overall maximum value of $|M_0 - M|$, which occurs for $e = 1$ and an M of about 1.7 rad, is about 0.385 rad. For every e it happens that the maximum $|M_0 - M|$ occurs with $M_0 > M$, and if e lies between about 0.398 and 0.977, M_0 actually exceeds M for all M in the range $(0, \pi)$.

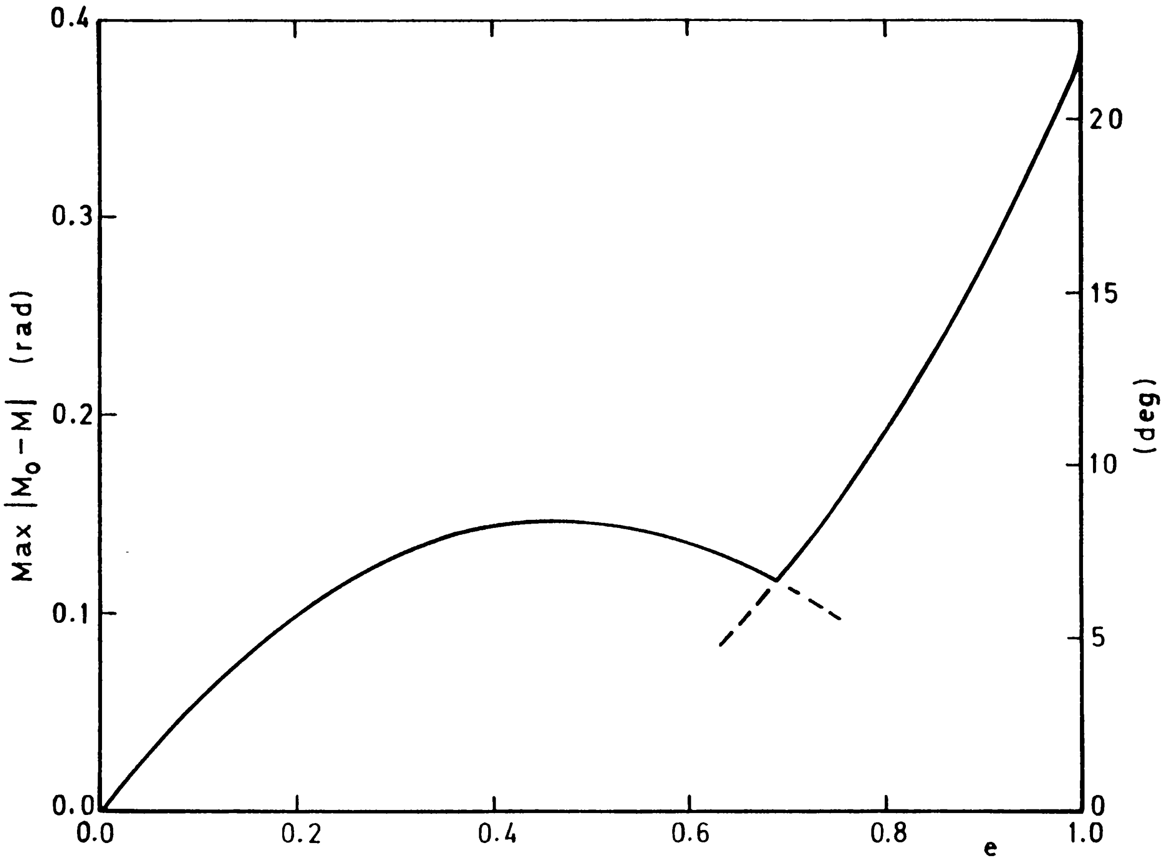


Fig. 4. Maximum residual ($M_0 - M$) for the S_{12} starter.

4.3. The New Procedure – Iterator

To achieve smooth portability, a basic requirement of the new procedure (EKEPL2) was that it should operate with a fixed number of iterations, independent of e as well as M. It was inconceivable that a single iteration would suffice, so the hope was that a sufficiently accurate two-iteration process could be found. The criterion for 'sufficient accuracy' was, for convenience, based on the precision of the computer employed for the investigation; this was the PR1ME 750, computing to double precision, so an objective of 10^{-13} rad was set.

With an overall maximum residual error of about 0.385 rad for the starter finally selected, it was certain that two iterations of neither a quadratic nor a cubic iterator would be adequate. Two iterations of an iterator with quartic convergence might well be good enough, however, since $[(0.385)^4/4!]^4/4! \approx 3 \times 10^{-14}$. We considered five quartic iterators, of which the first is the most obvious one, which in each iteration computes the quantity denoted by δ_{123} in Section 3 (and given also by Equation 18 of Danby and Burkardt (1983)). The other four iterators compute the quantities that are denoted by δ_{131}, δ_{1312}, δ_{1231} and δ_{12312} in Section 3. With these iterators there is an effective separation of an iteration into two components;

each component is either Newton–Raphson or Halley, so the four iterations can also be labelled NN, NH, HN and HH, respectively. All five iterators approximate the root of the hyper-osculating cubic curve, and an exact root of the cubic equation could of course be found, but only at the expense of a cube-root extraction.

The first iterator was rejected quite quickly. It can easily be seen to be inferior to the last two: these three iterators all start by making the same approximation (given by δ_{12}) to the root of the osculating quadratic curve and then switching attention to the more accurate cubic; the first iterator just applies the secant method to this curve, but the other two apply Newton–Raphson and Halley, respectively, via a preliminary 'rectification'. (In terms of the iterator finally selected, it was found that the first iterator produced $(M - M_2)$ residuals about two orders of magnitude larger.)

Of the remaining four iterators, a general investigation had shown that NN and HN are preferable to NH and HH in situations (not applying to our starter) where the current estimate of the root is poor. This is because, after the first part of the iteration, the cubic curve may have diverged so much from the true curve that it is safer to conclude the iteration by using Newton–Raphson rather than Halley. (With the starter $E_0 = \pi$, for example, which ought always to be safe, a single NH or HH iteration, for $e = 1$ and $M \approx 0$, leads to $E_1 \approx -4.6$ rad, which is hopeless, whereas NN and HN lead to $E_1 = 0.7$ rad.) Further, both NN and HN are faster than either NH or HH. It was decided to use NN or HN, therefore, and as NN was found to be not quite good enough our final choice of iterator is HN, giving δ_{1231}.

4.4. The New Procedure – Implementation and Results

The Fortran-77 function (EKEPL2) implementing our new procedure is listed in Appendix B. It will be seen that there are essentially five components: reduction of the range of M to the interval $(0, \pi)$, cube-root/bilinear starter for the reduced M and unit e; linear interpolation to a starter for the actual e; two applications of the HN (δ_{1231}) iterator; and shift of E back to the original interval.

The procedure was designed to avoid the build-up of truncation error in the vicinity of $(e, M) = (1, 0)$, but some remarks are called for on the avoidance of rounding error in this region (see also Section 5, on the use of universal variables). The relative error, as the solution to (1) tends to zero (behaving like $\sqrt[3]{6M}$), will be out of control unless the quantities f and f' (computed as F and FD during the iterative process) are computed in a special way whenever a suitable test criterion is satisfied.

The problem is the usual one of avoiding the subtraction of almost equal quantities. There is no difficulty with $f'(= 1 - e\cos E)$, since this is naturally computed as $e_1 + 2e\sin^2(\frac{1}{2}E)$ when the criterion is satisfied (the fact that e_1 is computed as the difference of two almost equal quantities causes no difficulty). It is not so simple for f. We have, from (2),

$$f(E) = e_1 \sin E + (E - \sin E) \tag{28}$$

and the problem is to compute an accurate value of the second term, of the form

$$E - \sin E = \tfrac{1}{6}E^3 - \tfrac{1}{120}E^5 + \cdots, \tag{29}$$

when E is small. We have not proposed an efficient universal algorithm for this computation, since the algorithm should be tailored, like the sine and cosine algorithms, to the particular computer, but we do give in Appendix C an inefficient universal algorithm (EMKEPL). This simply computes terms of (29) until there is no change in value.

This leaves the matter of the test criterion, which is largely arbitrary. It is natural to use the same criterion for both f and f', and since we can write, from (28), $f(E) = (\alpha + \beta)E$, where $\alpha \approx e_1$ and $\beta \approx \frac{1}{6}E^2$ when E is small, it seemed natural to use $e_1 + \frac{1}{6}E^2$ as a single (composite) test quantity. The essentially arbitrary value of 0.1 was chosen as the criterion value, but it is the only arbitrary value in the EKEPL2 procedure (not counting the cube-root/bilinear transition in the starter).

Implementation and testing on the PR1ME 750 computer confirmed that the looked for limiting accuracy (of around 10^{-13} rad) would always be met. Thus in no case would a more accurate result be obtained by taking the HN process to a third iteration. It was obviously of interest to know the true (ultimate) accuracy of the procedure, however, i.e. the behaviour of its truncation error, so it was decided to investigate this on computers of greater precision. The first choice was the Honeywell 870M, recently installed by the RAE's central computing service, and the second choice was the Cray 1S that is connected to the Honeywell. The accuracy of the Honeywell is such that (in double precision) it could detect any truncation error in excess of about 10^{-18} rad, and it was quickly found that the maximum error in the solution of Kepler's equation by EKEPL2 is less than 10^{-14} rad, occurring for an eccentricity of about 0.85 (with M about 0.033 rad and hence E about 0.21 rad). To get an accurate picture, however, it was necessary to go to the Cray, for which the accuracy is such that an error of about 10^{-28} rad would be detectable.

The Cray results confirmed that the maximum error in E is just under 7×10^{-15} rad, occurring for $e \approx 0.853$. However, we have usually been concerned with the residual in (1), when a solution is substituted back, rather than with the error in the solution itself, and the maximum residual $(M - M_2)$ is actually less than the maximum error $(E - E_2)$, being about 1.2×10^{-15} rad and occurring for $e \approx 0.835$. (It is worth remarking here that the linear interpolation of (20) could be replaced by a more elaborate form of weighting if desired; with e and e_1 replaced by e^2 and $1 - e^2$, for example, the maximum residual drops from 1.2×10^{-15} rad to 2.5×10^{-17} rad.) Figure 5 shows the variation, with e, of both the maximum error (for a given e) and the maximum residual; the tenth roots of both quantities have been plotted, on a linear scale, to produce curves of reasonable shape (points plotted for e less than about 0.2 are essentially estimates, since here even the Cray output is contaminated by rounding error, but it is obvious that the estimation was easy).

Figure 6 shows the variation in (tenth-root of) the M residual for the fixed value of e, viz 0.835, that gives the overall maximum residual; it is plotted against E, rather

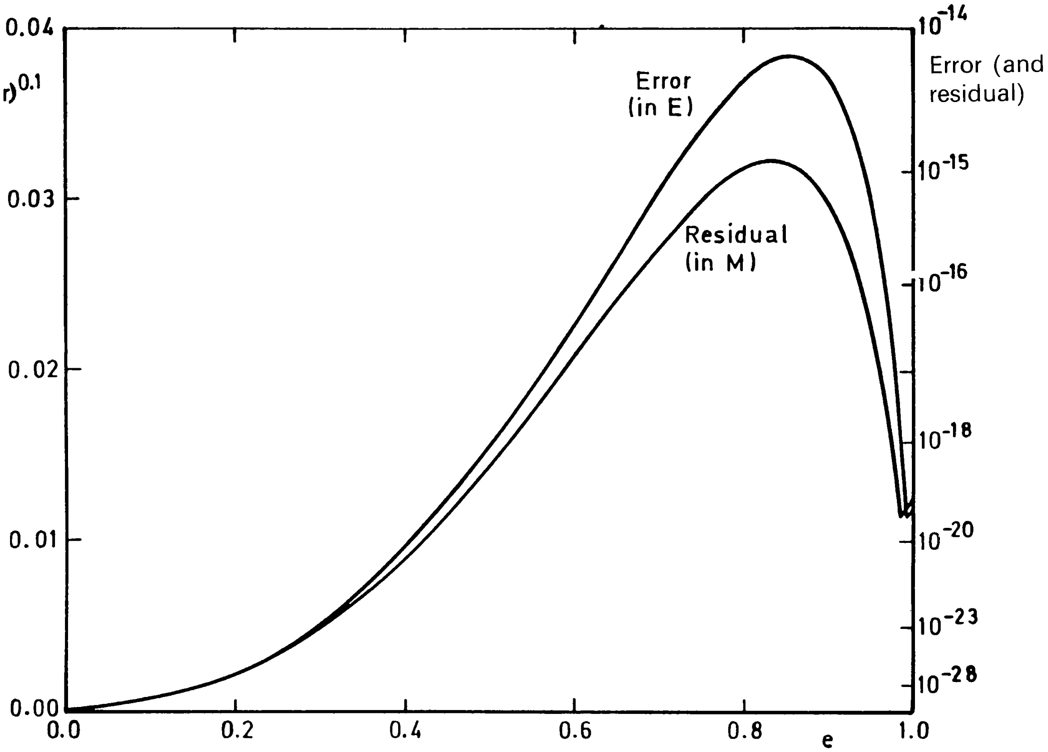


Fig. 5. Maximum error $(E - E_2)$ and residual $(M - M_2)$, in radians, for procedure EKEPL2.

than M, to produce a more instructive curve – a separate plot of M against E is provided. A plot of the relative residual is also provided; this crosses the curve for the absolute residual where $M = 1$, of course, and peaks only slightly before the absolute peak.

The 'error' plotted in Figure 5 must be understood to be the negative of the error in the normal sense, since E_2, the solution of E, is always (if $0 < M < \pi$) an underestimate rather than an over-estimate. This may be shown, for the particular iterator, to be a corollary of the fact that, throughout the open interval $(0, \pi)$, $f^{iv}(E) < 0$. Further, if e lies between 0.398 and 0.977, as specified at the end of Section 4.2, then the error does not vanish between 0 and π, and Figure 6 is an example of this.

An evident feature of Figure 6 is the existence of a second (local) maximum, at a value of M of about 1.3 rad. As e increases between 0.835 and 1, the residual at the second maximum increases, whilst at the first maximum it falls. Eventually, for e about 0.984, the second maximum becomes dominant, and this explains the little up-turns in Figure 5, right at the end of both curves (cf. the same effect in Figure 4). Each 'second maximum' can in fact be tracked beyond $e = 1$; only when e reaches 1.35 does its value for the M-residual curve equal the peak value of the first maximum.

Some remarks on the computing time for EKEPL2 may be of interest. The basic time, when $M > \frac{1}{6}$ rad (after range reduction) so that no cube root is required by the starter, is about 1 ms, 0.8 ms and 0.5 ms on the PRIME, Honeywell and Cray

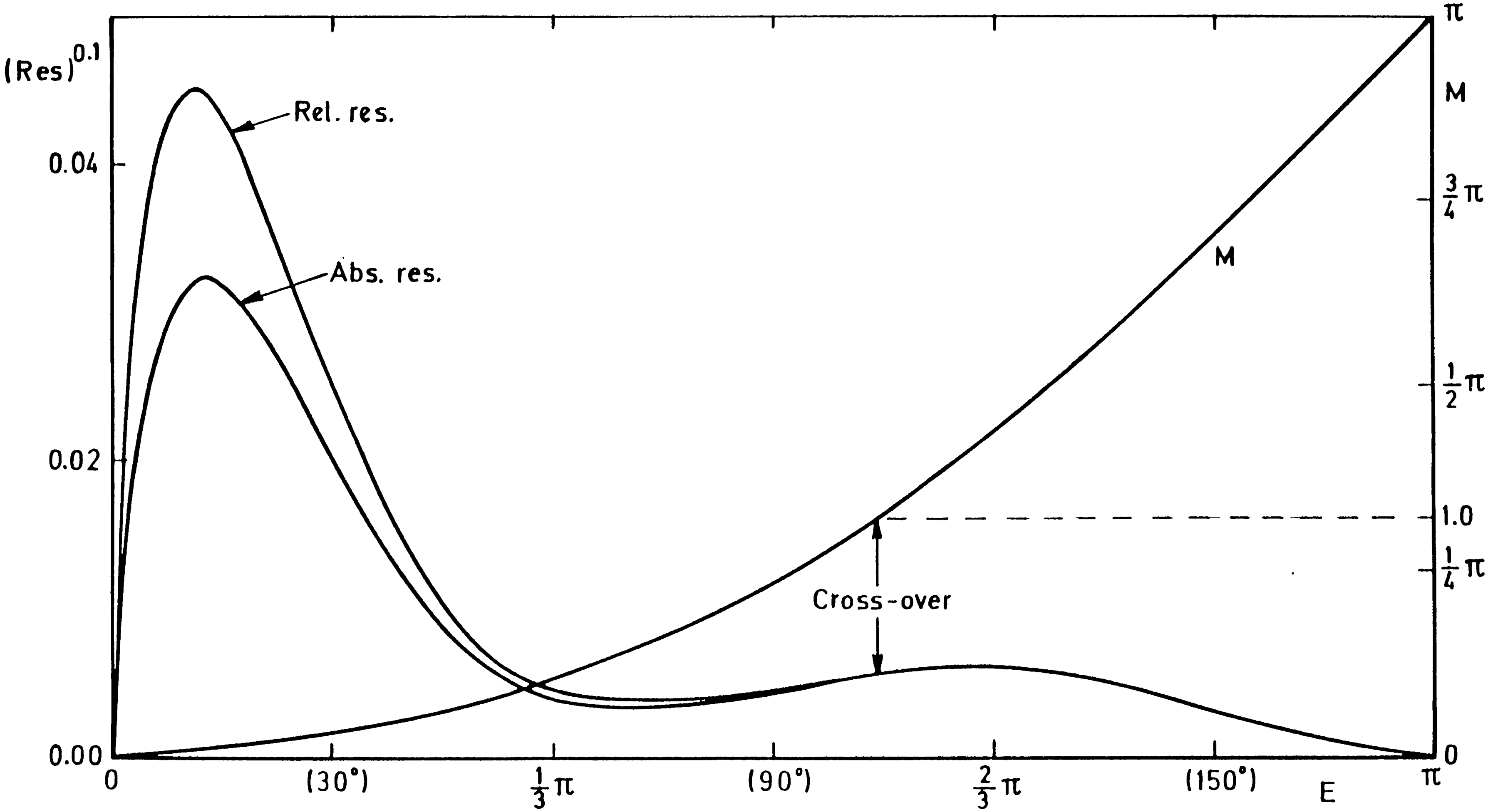


Fig. 6. Absolute and relative residuals (rad) for EKEPL2 with $e = 0.835$; plus corresponding M.

computers, respectively (about the same as for three iterations of EKEPL1). Taking the cube root on the PR1ME costs about 0.4 ms, but there is only 1 chance in 6π that this is necessary, assuming a uniform distribution of time and hence M, so the average computing time rises by only about 2%; the percentage rise is similar on the other two computers. There is also a penalty for the avoidance of rounding error, but even with the inefficient EMKEPL listed in Appendix C, and assuming the worst case ($e = 1$), the average computing time does not rise by more than 4% on any of the three computers.

In our new procedure the iterative process is quartic and gives a maximum error less than 10^{-14} rad after only two iterations, so a third iteration would presumably reduce this to not much more than 10^{-60} rad. No attempt has been made to verify this!

5. Unified and Universal Formulations

A number of papers on Kepler's equation, in particular those of Bergam and Prussing (1982), Burkardt and Danby (1983) and Shepperd (1985), have emphasized the advantages of a universal formulation such that a single equation covers all the conic sections that can define a two-body orbit, including the three degenerate 'rectilinear orbits'. Shepperd attributes the universal approach to Stumpff (1947), whilst other papers usually refer to one of the text-books that popularized it, in particular Battin (1964) and Herrick (1971); the latter gives extensive material on the subject, with a careful distinction between unified variables and universal variables – unified variables break down for a circular orbit. The universal formulation does not provide a magic avoidance of the solution of (1), however, as Herrick was at pains to observe (page xix, *loc cit*), and we believe that this remark merits some amplification.

First, we show how a 'unified equation' can be developed from the elliptic equation, (1), which (in the usual notation, with t measured from a unique perifocus if $e \neq 0$), can be rewritten as

$$a^{3/2}(E - e \sin E) = \sqrt{\mu} t. \tag{30}$$

We start by defining the two transcendental functions of Battin (1964), closely related to the Stumpff functions and given by

$$S(x) = \frac{1}{3!} - \frac{x}{5!} + \frac{x^2}{7!} - \cdots \tag{31}$$

and

$$C(x) = \frac{1}{2!} - \frac{x}{4!} + \frac{x^2}{6!} - \cdots. \tag{32}$$

The relation to the sine and cosine functions is evident and we can at once rewrite

(30) as

$$a^{3/2}\{(1-e)E+eE^3S(E^2)\}=\sqrt{\mu}t. \tag{33}$$

We now replace a, e and E by α, q and X, defined by

$$\alpha=1/a,$$
$$q=a(1-e)$$

and

$$X=\sqrt{a}E,$$

these being unified variables (α and q are universal, but X still depends on the existence of a perifocus). Then (33) gives

$$qX+(1-\alpha q)X^3S(\alpha X^2)=\sqrt{\mu}t, \tag{34}$$

which is the unified equation. (It has been tacitly assumed that $a<0$ for a hyperbola, so that q is universally the perifocal distance.)

We wish to comment on (34), but before doing so, we show how the universal equation, free of the perifocal reference, can be developed in the same way, from the elliptic equation, (1). We start from the version of Equation (30) that relates to t_0, an arbitrary but well-defined epoch, viz

$$a^{3/2}(E_0-e\sin E_0)=\sqrt{\mu}t_0.$$

Subtraction of this from (30), with some elaboration, gives

$$a^{3/2}\{(1-e\cos E_0)(E-E_0)+e\sin E_0\;[1-\cos(E-E_0)]$$
$$+e\cos E_0\;[E-E_0-\sin(E-E_0)]\}=\sqrt{\mu}(t-t_0). \tag{35}$$

We write

$$\hat{X}=\sqrt{a}(E-E_0),$$
$$r_0=a(1-e\cos E_0)$$

and

$$\mathbf{r}_0\cdot\mathbf{V}_0=\sqrt{\mu a}\;e\sin E_0,$$

using the position and velocity vectors ($\mathbf{r}_0$ and $\mathbf{V}_0$) at t_0, and measure t now from t_0. Then the three left-hand-side terms of (35) translate directly to the terms of the universal equation, which is

$$r_0\hat{X}+(\mathbf{r}_0\cdot\mathbf{V}_0/\sqrt{\mu})\hat{X}^2C(\alpha\hat{X}^2)+(1-\alpha r_0)\hat{X}^3S(\alpha\hat{X}^2)=\sqrt{\mu}t. \tag{36}$$

This gist of our comments about Equation (34) is that, for an orbit known to be elliptic, there is no advantage in solving this equation rather than (1), and there may be a distinct disadvantage. The only basis for an advantage comes from the avoidance of rounding error when E is (numerically) small, since $S(\alpha X^2)$ is then computed without the loss of accuracy that can arise from $E - e \sin E$. But $S(\alpha X^2)$ is just $(E - \sin E)/E^3$ and we have seen in Section 4 how, by a version of the procedure EMKEPL, the rounding error in $E - \sin E$ can be avoided.

The disadvantage of solving (34), rather than (1), arises with values of t that are greater than an orbital period. In defining $S(x)$, the periodic nature of the sine function has been lost, and this can lead to serious error in $S(\alpha X^2)$ where it would be negligible in $\sin E$ (unless, of course, t is range-reduced by a multiple of the orbital period first). It is not possible to overcome the difficulty by the (inefficient) computation of a great many terms, since the terms initially grow in magnitude and the rounding error can be enormous (as indicated by a comment in the listing of EMKEPL in Appendix C).

A final comment concerns the solution of Kepler's equation as e approaches zero, since it is sometimes feared that the disappearance of perifocus causes real difficulty in orbit computations – cf. the solution of (36) rather than (34). This is a misconception, however, and the solution of (1) remains valid all the way to the trivial case when, in the limit, an arbitrary point of the circle can be regarded as 'perifocus'. Other parts of an orbit computation may need, of course, to be formulated with the necessary care to avoid breakdown due to division by zero, etc.

6. Conclusion

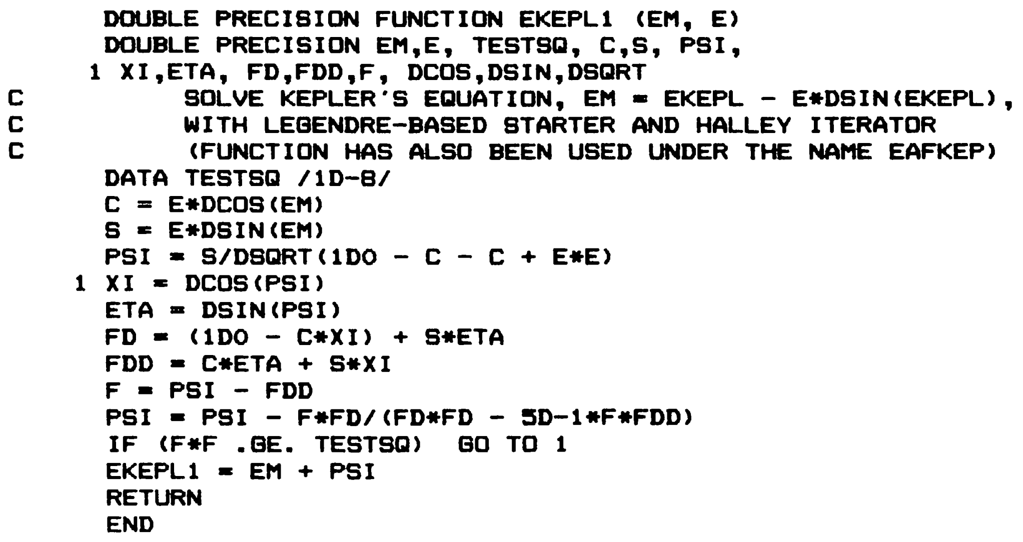
We have looked at a number of starting values for an iterative solution of Kepler's equation, listing twelve of them in Table 1. In parallel, we have considered convergence processes that are superior to the standard method of Newton (and Raphson).

Two solution procedures are particularly recommended, with listings appended. The first, EKEPL1, has been used at RAE for a number of years, but not previously published. With iteration by the Halley process, which normally gives cubic convergence, and a criterion of 10^{-4} rad for the satisfying of the equation after the penultimate iteration, an accuracy better than 10^{-12} rad (residual in M) is normally achieved. The procedure never requires more than five iterations. When $e \cos M$ is close to unity, the convergence degenerates to linear, and it is only then that the accuracy is poor.

The other procedure, EKEPL2, is entirely new. The number of iterations is fixed at two, and quartic convergence gives an accuracy better than 10^{-14} rad (both in E and in residual M) for all values of e and M. Convergence never degenerates, this being at the expense of a cube-root extraction whenever M is within 1/6 rad of a multiple of 2π. The procedure is very efficient, however, and requires an average time of only about 1 ms on current computers. The avoidance of rounding error (when $e \cos M \approx 1$) requires the use of a subordinate procedure (EMKEPL), the function of which is akin to the use of unified variables.

Appendix A

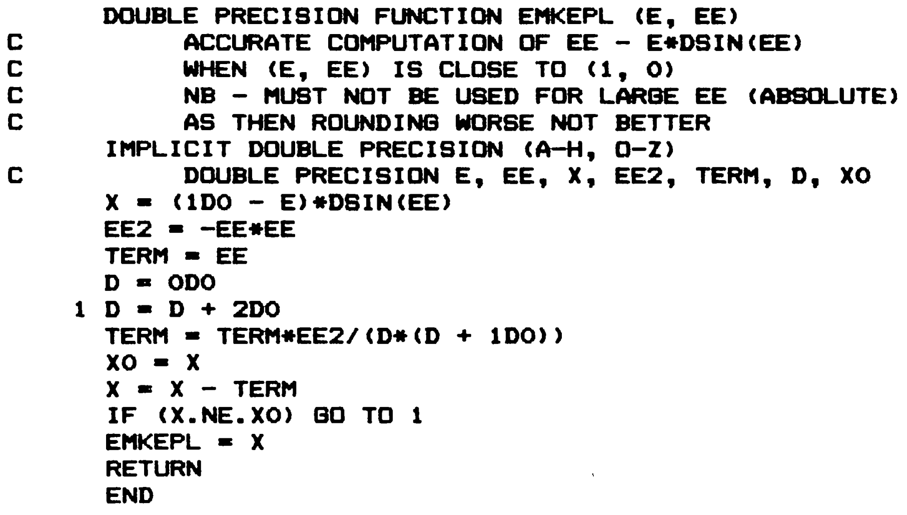
THE EKEPL1 PROCEDURE



```
      DOUBLE PRECISION FUNCTION EKEPL1 (EM, E)
      DOUBLE PRECISION EM,E, TESTSQ, C,S, PSI,
     1 XI,ETA, FD,FDD,F, DCOS,DSIN,DSQRT
C           SOLVE KEPLER'S EQUATION, EM = EKEPL - E*DSIN(EKEPL),
C           WITH LEGENDRE-BASED STARTER AND HALLEY ITERATOR
C           (FUNCTION HAS ALSO BEEN USED UNDER THE NAME EAFKEP)
      DATA TESTSQ /1D-8/
      C = E*DCOS(EM)
      S = E*DSIN(EM)
      PSI = S/DSQRT(1D0 - C - C + E*E)
    1 XI = DCOS(PSI)
      ETA = DSIN(PSI)
      FD = (1D0 - C*XI) + S*ETA
      FDD = C*ETA + S*XI
      F = PSI - FDD
      PSI = PSI - F*FD/(FD*FD - 5D-1*F*FDD)
      IF (F*F .GE. TESTSQ)  GO TO 1
      EKEPL1 = EM + PSI
      RETURN
      END
```

Appendix C

AN UNSOPHISTICATED EMKEPL PROCEDURE



```
      DOUBLE PRECISION FUNCTION EMKEPL (E, EE)
C           ACCURATE COMPUTATION OF EE - E*DSIN(EE)
C           WHEN (E, EE) IS CLOSE TO (1, 0)
C           NB - MUST NOT BE USED FOR LARGE EE (ABSOLUTE)
C           AS THEN ROUNDING WORSE NOT BETTER
      IMPLICIT DOUBLE PRECISION (A-H, O-Z)
C           DOUBLE PRECISION E, EE, X, EE2, TERM, D, XO
      X = (1D0 - E)*DSIN(EE)
      EE2 = -EE*EE
      TERM = EE
      D = 0D0
    1 D = D + 2D0
      TERM = TERM*EE2/(D*(D + 1D0))
      XO = X
      X = X - TERM
      IF (X.NE.XO) GO TO 1
      EMKEPL = X
      RETURN
      END
```

Appendix B

THE EKEPL2 PROCEDURE

```
      DOUBLE PRECISION FUNCTION EKEPL2(EM, E)
C          KEPLER'S EQUATION, EM = EKEPL - E*DSIN(EKEPL) WITH
C          E IN RANGE 0 TO 1 INCLUSIVE, SOLVED ACCURATELY
C          (IMPLICIT DOUBLE PRECISION (A-H,O-Z) COULD REPLACE
C           THE NEXT THREE LINES)
      DOUBLE PRECISION EM, E, PI, TWOPI, PINEG, SW, AHALF, ASIXTH,
     1 ATHIRD, A, B, EMR, EE, W, E1, FDD, FDDD, F, FD, DEE,
     2 DMOD, DSIN, DCOS, EMKEPL
      LOGICAL L
      PARAMETER (PI=3.141592653589793238462643383280O,TWOPI=2D0*PI,PINE
     1 G=-PI, SW=1D-1, AHALF=0.5D0,ASIXTH=AHALF/3D0,ATHIRD=ASIXTH*2D0,
     2 A=(PI-1D0)**2/(PI+2D0/3D0),B=2D0*(PI-ASIXTH)**2/(PI+2D0/3D0) )
C1         RANGE-REDUCE EM TO LIE IN RANGE -PI TO PI
      EMR = DMOD(EM,TWOPI)
      IF(EMR.LT.PINEG) EMR = EMR + TWOPI
      IF(EMR.GT.PI) EMR = EMR - TWOPI
      EE = EMR
      IF (EE) 1,4,2
    1 EE = -EE
C          (EMR IS RANGE-REDUCED EM & EE IS ABSOLUTE VALUE OF EMR)
C2         STARTER FOR E = 1 BY CUBE ROOT OR BILINEAR FUNCTION
    2 IF (EE.LT.ASIXTH) THEN
        EE = (6D0*EE)**ATHIRD
       ELSE
        W = PI - EE
        EE = PI - A*W/(B - W)
      END IF
      IF(EMR.LT.0D0) EE = -EE
C3         INTERPOLATE FOR E
      EE = EMR + (EE - EMR)*E
C4         DO TWO ITERATIONS OF HALLEY, EACH FOLLOWED BY NEWTON
      E1 = 1D0 - E
      L = (E1 + EE*EE/6D0) .GE. SW
      DO 3 ITER=1,2
      FDD = E*DSIN(EE)
      FDDD = E*DCOS(EE)
      IF (L)  THEN
        F = (EE - FDD) - EMR
        FD = 1D0 - FDDD
       ELSE
        F = EMKEPL(E,EE) - EMR
        FD = E1 + 2D0*E*DSIN(AHALF*EE)**2
      END IF
      DEE = F*FD/(AHALF*F*FDD - FD*FD)
      F = F + DEE*(FD + AHALF*DEE*(FDD + ATHIRD*DEE*FDDD))
C*         TO REDUCE THE DANGER OF UNDERFLOW REPLACE THE LAST LINE BY
C*    W = FD + AHALF*DEE*(FDD + ATHIRD*DEE*FDDD)
      FD = FD + DEE*(FDD + AHALF*DEE*FDDD)
    3 EE = EE + DEE - F/FD
C*         IF REPLACING AS ABOVE, THEN ALSO REPLACE THE LAST LINE BY
C*  3 EE = EE - (F - DEE*(FD - W))/FD
C5         RANGE-EXPAND
    4 EKEPL2 = EE + (EM - EMR)
      RETURN
      END
```

References

Battin, R. H.: 1964, *Astronautical Guidance*, McGraw-Hill, N.Y. etc.

Bergam, M. J. and Prussing, J. E.: 1982, 'Comparison of Starting Values for Iterative Solutions to a Universal Kepler's Equation', *J. Astr. Sci.* **30**, 75–84.

Broucke, R.: 1980, 'On Kepler's Equation and Strange Attractors', *J. Astr. Sci.* **28**, 255–265.

Burkardt, T. M. and Danby, J. M. A.: 1983, 'The Solution of Kepler's Equation, II', *Celes. Mech.* **31,** 317–328.

Danby, J. M. A. and Burkardt, T. M.: 1983, 'The Solution of Kepler's Equation, I', *Celes. Mech.* **31**, 95–107.

Gooding, R. H. and Odell, A. W.: 1985, 'A Monograph on Kepler's Equation', RAE Technical Report 85080.

Halley, E.: 1694, 'Methodus nova accurata e facilis inveniendi radices aequationum quarum cumque generaliter, sine praevia reductione', *Phil. Trans. Roy. Soc.* **18**, 136–148.

Herrick, S.: 1971, *Astrodynamics, Volume 1*, Van Nostrand Reinhold, London, etc.

Ng, E. W.: 1979, 'A General Algorithm for the Solution of Kepler's Equation for Elliptic Orbits', *Celes. Mech.* **20**, 243–249.

Peters, R. D.: 1984, 'Rapidly Converging Series Approximation to Kepler's Equation', *Adv. Astr. Sci.* **54** (Part 2), 1039–1047.

Shepperd, S. W.: 1985, 'Universal Keplerian State Transition Matrix', *Celes. Mech.* **35**, 129–144.

Siewert, C. E. and Burniston, E. E.: 1972, 'An Exact Analytical Solution of Kepler's Equation', *Celes. Mech.* **6**, 294–304.

Smith, G. R.: 1979, 'A Simple, Efficient, Starting Value for the Solution of Kepler's Equation for Elliptic Orbits', *Celes. Mech.* **19**, 163–166.

Smith, O. K.: 1961, 'Terminating the Iterative Solution of Kepler's Equation', *J. Amer. Rocket Soc.* **31**, 1598.

Stumpff, K.: 1947, 'Neue Formeln und Hilfstafeln zur Ephemeridenrechnung', *Astron. Nachrichten* **275**, 108–128.

Traub, J. F.: 1961, 'On a Class of Iteration Formulas and some Historical Notes', *Comm. A.C.M.* **4**, 276–278.