Abbreviated Lecture Notes on

Ellipsoidal Error Bounds for Trajectory Calculations

Prof. W. Kahan
Mathematics Dept., and
Elect. Eng. & Computer Science Dept.
University of California at Berkeley
For Presentation Oct. 12, 1993

Abstract:
A practical way is outlined to bound error accrued during numerical calculations of trajectories. The algorithm accommodates uncertainties in the governing differential equation as well as error due to the numerical process. The accrued error is constrained to lie within an ellipsoid that can be proved to grow, as time τ increases to $+\infty$, bigger than the worst possible accrual by a factor no worse than $1 + ß\sqrt{\tau}$ for some constant $ß$, rather than exponentially bigger, until nonlinearity in the differential equation forces a singularity to manifest itself.

Introduction:
A *trajectory* is the graph of $y(\tau)$, the solution of an *Autonomous Initial Value Problem*

$$(\ dy/d\tau =)\quad y'(\tau) = f(y(\tau)) \quad \text{for} \quad \tau \ge 0\ , \quad y(0) = y_0\ . \qquad \text{(AIVP)}$$

Here y_0 is a given vector and f a given vector-valued function of its vector argument. Perturbing y_0 to $y_0 + \delta y_0$ and f to $f + \delta f$ perturbs the trajectory to the solution $y + \delta y$ of a perturbed AIVP

$$(y + \delta y)'(\tau) = (f + \delta f)(\ (y + \delta y)(\tau)\ ) \quad \text{for} \quad \tau \ge 0\ , \quad (y + \delta y)(0) = y_0 + \delta y_0\ .$$

Given only some kind of bound for δy_0 and δf, how can we infer a bound upon δy ?

Actually, in practice we know $y + \delta y$ but not y ; this will be overlooked now to simplify the exposition. To the same end, we shall not discuss how the truncation and rounding errors incurred by the numerical process that solves the AIVP can be incorporated into δf along with errors due to idealizations that model a complex physical situation by a simplified expression f.

Were all perturbations $\delta\ldots$ infinitesimal, the accrued error δy would satisfy the *Adjoint* or *Variational Initial Value Problem* associated with the given AIVP :

$$\delta y'(\tau) = J(\tau)\cdot\delta y(\tau) + \delta f\ , \quad \delta y(\tau) = \delta y_0\ . \qquad \text{(VIVP)}$$

Here *Jacobian* matrix $J(\tau) := f'(y) = \partial f(y)/\partial y$ at $y = y(\tau)$; this $J(\tau)$ is the matrix of first partial derivatives of f evaluated at $y(\tau)$, the presumed-to-be-known trajectory. Perturbation $\delta f = \delta f(y(\tau))$; however, whereas $J(\tau)$ can be computed from f' and $y(\tau)$, all we know about δf is an upper bound. Likewise for δy_0. Presumably computable, perhaps functions of $y(\tau)$, these bounds upon δf and δy_0 are the data from which we wish to infer a bound upon $\delta y(\tau)$. Computing this bound for accrued error along with $y(\tau)$ is the problem addressed in these notes.

In practice the perturbations $\delta\ldots$ are not infinitesimal. The effect in practice of their finiteness is to increase the bound upon δf by an amount roughly proportional to the product of the square of the computed bound upon $\delta y(\tau)$ and a bound upon the second derivative of f. This increase turns the linear VIVP into something nonlinear like a Riccati equation, whose solution may

become infinite at a finite time τ even in some cases when $\delta y(\tau)$ is known to stay bounded for all finite τ. Despite its importance, this nonlinear contribution is not discussed in these notes. Instead, to simplify the exposition of a subject that is already complicated enough, δy is assumed to stay so small that its second-order contributions may safely be neglected.

We assume $J(\tau)$ and δf to be continuous for all τ; in practice δf may be only piecewise continuous. This is a technicality we could dispatch by converting differential equations into equivalent Volterra integral equations. For simplicity's sake we won't do that either.

Despite all our simplifications, the computation of a bound upon δy remains so challenging that every previously published scheme I know about is prone to producing bounds too big by a factor that grows like an exponential function of τ when $J(\tau)$ behaves in a way that the scheme dislikes. The algorithm described here is far less pessimistic; its bound cannot grow too big by a factor bigger than $1 + ß\sqrt{\tau}$ for some constant ß that depends upon moduli of continuity of $J(\tau)$ and of the bound for δf. The proof of this claim is too long to fit here. Space and time barely suffice for an outline of my algorithm.

My algorithm adjoins to the given AIVP another differential equation whose solution, intended to be computed simultaneously with $y(\tau)$, is a symmetric matrix $A(\tau)$ that describes an ellipsoid centered at $y(\tau)$ and surely big enough to enclose $\delta y(\tau)$, but not too much bigger. The adjoined differential equation's dimension is about half the square of y's dimension, so $A(\tau)$ may well cost enormously more to compute than $y(\tau)$. Since previously published schemes typically cost twice as much as mine, I need not apologize for it.

The Reachable Set:

Let us now write z in place of δy and u in place of δf in the VIVP above; it becomes

$$z'(\tau) = J(\tau)\cdot z(\tau) + u(\tau) \quad \text{for } \tau \geq 0, \quad z(0) = z_0 . \qquad \text{(VIVP)}$$

The matrix $J(\tau)$ is assumed computable; but for $u(\tau)$ and z_0 only bounds are available. We construe these bounds as constraints that restrict $u(\tau)$ and z_0 to certain small regions; say

$$u(\tau) \in \hat{\mathbb{U}}(\tau) \quad \text{and} \quad z_0 \in \mathbb{A}$$

for given centrally symmetric convex bodies $\hat{\mathbb{U}}(\tau)$ and $\mathbb{A}$ characterized by parameters to be discussed later. For instance these bodies could be spheres characterized by their radii, which are then upper bounds for the lengths of $u(\tau)$ and z_0. Parallelepipeds have been used too, characterized by the matrices that map a unit hypercube onto them. But we use ellipsoids for reasons to be discussed later.

The *Reachable Set* $\hat{\mathbb{O}}(\tau)$ consists of all values that $z(\tau)$ can take compatible with the given constraints imposed by $\hat{\mathbb{U}}(\tau)$ and $\mathbb{A}$ upon u and z_0 in the VIVP. ("Reachable Set" is a term coined by Control theorists.) From the given hypotheses about $\hat{\mathbb{U}}(\tau)$ and $\mathbb{A}$, it follows that $\hat{\mathbb{O}}(\tau)$ must be a centrally symmetric convex body too. But generally the shape of $\hat{\mathbb{O}}(\tau)$ is not so simple as the shapes of $\hat{\mathbb{U}}(\tau)$ and $\mathbb{A}$. Regardless of whether the latter are ellipsoids or parallelepipeds, $\hat{\mathbb{O}}(\tau)$ need not be any of those. The best we can expect to do computationally is to approximate $\hat{\mathbb{O}}(\tau)$ by one of those simpler figures. Thus, our task is to compute whatever parameters characterize a simpler centrally symmetric convex body, ellipsoid or parallelepiped, that circumscribes reachable set $\hat{\mathbb{O}}(\tau)$ as tightly as is possible at a tolerable cost.

Ellipsoidal Bounds:
Any (open) ellipsoid $\mathbb{A}$ centered at the origin o is characterized by an appropriate symmetric positive definite matrix A as follows:

$$\mathbb{A} \text{ consists of all vectors } x \text{ that satisfy } x^T A^{-1} x < 1 .$$

The eigenvectors of A point in the directions of principal semi-axes of $\mathbb{A}$, and the eigenvalues of A are the squared lengths of those semi-axes. Flattened ellipsoids, those with some semi-axes of zero length, are represented by singular positive semi-definite matrices A, for which we must understand the expression $x^T A^{-1} x$ to require that x be confined to the range of A.

A theorem of Fritz John (1948) asserts that any centrally symmetric convex body $\mathbb{U}$ in an N-dimensional space can be circumscribed by an ellipsoid $\mathbb{Y}$ tightly enough that the boundary of $\mathbb{U}$ lies inside $\mathbb{Y}$ but outside $\mathbb{Y}/\sqrt{N}$. A short proof of his theorem is posted in lecture notes at <http://www.cs.berkeley.edu/~wkahan/MathH110/NORMlite.pdf>. His chosen candidate $\mathbb{Y}$ is the circumscribing ellipsoid of minimum volume; its characterizing matrix Y minimizes $\det(Y)$. Thus, little can be lost in spaces of modest dimension if the bounding bodies $\mathbb{U}(\tau)$ and $\mathbb{A}$ containing respectively $u(\tau)$ and z_0 above are taken to be ellipsoids; otherwise they can be circumscribed by ellipsoids at the cost of worsening our bound $\mathbb{A}(\tau)$ upon the reachable set $\mathbb{O}(\tau)$ by at worst a factor $\sqrt{N}$. As error bounds go, this degree of exacerbated pessimism will go unremarked.

Thus we may assume that, along with the Jacobian matrix $J(\tau)$ of partial derivatives, we are supplied with symmetric positive (semi-)definite matrices $U(\tau)$ and A_0 whose ellipsoids $\mathbb{U}(\tau)$ and $\mathbb{A}$, we are told, contain $u(\tau)$ and z_0. Often $U(\tau)$ and A_0 will be diagonal. Our task is to compute a symmetric positive definite matrix $A(\tau)$ whose ellipsoid $\mathbb{A}(\tau)$ circumscribes the reachable set $\mathbb{O}(\tau)$ as tightly as possible at a tolerable cost.

The Auxiliary Differential Equation:
First let's summarize the situation as it stands now. The given AIVP

$$y'(\tau) = f(y(\tau)) \quad \text{for} \quad \tau \ge 0 , \quad \text{and} \quad y(0) = y_0 \qquad \text{(AIVP)}$$

is being solved numerically for y, but unknown perturbations $\delta y(0) = z_0$ and $\delta f = u(\tau)$ induce in y an accrued perturbation $\delta y = z$ that satisfies

$$z'(\tau) = J(\tau)\cdot z(\tau) + u(\tau) \quad \text{for} \quad \tau \ge 0 , \quad \text{and} \quad z(0) = z_0 . \qquad \text{(VIVP)}$$

Here $J(\tau) = f'(y(\tau))$ is known, and so are symmetric positive definite matrices $U(\tau)$ and $A(0) = A_0$ for which

$$u^T U^{-1} u < 1 \text{ for } \tau \ge 0 , \text{ and } z_0^T A_0^{-1} z_0 < 1 \text{ at } \tau = 0 .$$

By constraining u and z_0, these inequalities compel the accrual z to lie in a Reachable Set $\mathbb{O}(\tau)$ about which we wish to circumscribe an ellipsoid $\mathbb{A}(\tau)$ by computing its symmetric positive definite matrix $A(\tau)$ such that

$$z^T A^{-1} z < 1 \text{ for every } z \text{ in } \mathbb{O} , \text{ for every } \tau \ge 0 .$$

(Here all matrices except A_0 are functions of τ.) Infinitely many matrices A fulfill these requirements; we seek one of the smaller ones.

Here is how to construct one. First let α be any strictly positive (piecewise-) continuous scalar function of τ; later we shall exhibit a good choice for α. Then compute $A(\tau)$ as the solution of the following *Auxiliary Differential Equation*:

$$A' = J\cdot A + A\cdot J^T + \alpha\cdot U + A/\alpha \quad \text{for } \tau \ge 0, \quad \text{and} \quad A(0) = A_0 . \qquad \text{(ADE)}$$

The solution $A(\tau)$ of this ADE is intended to be computed numerically and simultaneously with the solution $y(\tau)$ of the AIVP .

Theorem: $z^T A^{-1} z < 1$ for every z in the reachable set $\hat{\mathbb{O}}$.

Proof of the Theorem:
The theorem's inequality starts true at $\tau = 0$, so if increasing τ ever makes it false it must do so for the first time at some $\tau > 0$ at which both $z^T A^{-1} z = 1$ and $(z^T A^{-1} z)' \ge 0$. To show that the last inequality is incompatible with the previous equality we need only expand $(z^T A^{-1} z)'$:

$$\begin{aligned}(z^T A^{-1} z)' &= (Jz + u)^T A^{-1} z - z^T A^{-1}(JA + AJ^T + \alpha U + A/\alpha) A^{-1} z + z^T A^{-1}(Jz + u)\\ &= u^T A^{-1} z + z^T A^{-1} u - \alpha z^T A^{-1} U A^{-1} z - z^T A^{-1} z/\alpha\\ &= -(\alpha A^{-1} z - U^{-1} u)^T U (\alpha A^{-1} z - U^{-1} u)/\alpha - (1 - u^T U^{-1} u)/\alpha\\ &< 0 , \text{ as claimed, so the theorem is true for all } \tau > 0 .\end{aligned}$$

How to Choose α :
Having just proved that ellipsoid $\mathbb{A}$ encloses reachable set $\hat{\mathbb{O}}$, we are dismayed to observe from the ADE that $\mathbb{A}$ may well enclose a great deal more than $\hat{\mathbb{O}}$ if α is either too big or too small. Can α be so chosen that $\mathbb{A}$ is not much bigger than $\hat{\mathbb{O}}$? In fact there is a way, albeit impractical, so to choose α that $\mathbb{A}$ and $\hat{\mathbb{O}}$ will share a common support plane (tangent) for all $\tau \ge 0$ provided the normal to that plane is fixed in advance. Even so, $\mathbb{A}$ may still exceed the size of $\hat{\mathbb{O}}$ enormously in directions parallel to that plane. Apparently, choosing α well is a subtle problem. All the more surprising, then, is the existence of a computationally simple choice that always turns out to be adequate:

$$\text{Choose} \quad \alpha := \sqrt{\big(\, \text{trace}(A) \,/\, \text{trace}(U) \,\big)} .$$

For this choice and some others, I have proved that the diameter of $\mathbb{A}$ cannot exceed that of $\hat{\mathbb{O}}$ by a factor bigger than $1 + ß\sqrt{\tau}$, where ß is a constant that depends upon various attributes of J and U. My proof is still so long that I am too embarrassed to publish it.

Computational Experience:
My earliest experiments with ellipsoidal bounds are still the most satisfying. The AIVP was the equations of motion of one of the moons of Jupiter, and the ADE included crude bounds for the gravitational influences of the rest of the solar system plus the contributions of all numerical errors. The computation ran for hundreds of orbits during which the bounding ellipsoid $\mathbb{A}$ became ever more needle-shaped, growing roughly like $(1 + \tau)^2$. Computation was halted only because I wished not to hog the University of Toronto's IBM 7094 in 1968 . The results confirmed that, in a situation where coffin-shaped or more general parallepiped-shaped bounds became infinite after a few orbits, ellipsoidal bounds did not.

Acknowledgements:
I am indebted to Gerry Gabel, who programmed the differential equation solver that I most preferred on the 7094. I have yet to see a better program; it was a predictor-corrector with automatically chosen continually varying order and stepsize. The late Prof. Tom Hull endured my many attempts to explain why ellipsoidal bounds could succeed where all shapes promoted previously were doomed to exponentially excessive growth, and I am grateful for his patience. And I am grateful to the late Prof. Fred C. Schweppe of MIT for his encouragement and for including some of the foregoing material in his book *Uncertain Dynamic Systems* (1973, Prentice-Hall, NJ), where it has languished through no fault of his. These notes were extracted from voluminous classroom handouts for a rarely offered graduate course, Math. 273 at the Univ. of Calif. at Berkeley, at the instigation of Prof. Jerrold Marsden for presentation at his workshop on *Integration Algorithms for Classical Mechanics*, 14-17 Oct. 1993, held at the Fields Inst. for Research in Math. Sciences, then at the Univ. of Waterloo, Ontario, Canada.