2

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE 6/11/79	3. REPORT TYPE AND DATES COVERED Final

4. TITLE AND SUBTITLE NONLINEAR GUIDANCE OF AIR-TO-AIR MISSILES	5. FUNDING NUMBERS 61102F 2304/A1
6. AUTHOR(S) Jan F. Andrus	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Mathematics Department University of New Orleans New Orleans, La 70122	8. PERFORMING ORGANIZATION REPORT NUMBER AFOSR-TR- 89-1361
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AFOSR BLDG 410 BAFB DC 20332-6448	10. SPONSORING/MONITORING AGENCY REPORT NUMBER AFOSR 78-3641
11. SUPPLEMENTARY NOTES	
12a. DISTRIBUTION/AVAILABILITY STATEMENT	12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

14. SUBJECT TERMS	15. NUMBER OF PAGES
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

FINAL TECHNICAL REPORT

In response to Grant No. AFOSR 78-3641

Title: Nonlinear Guidance of Air-to-Air Missiles

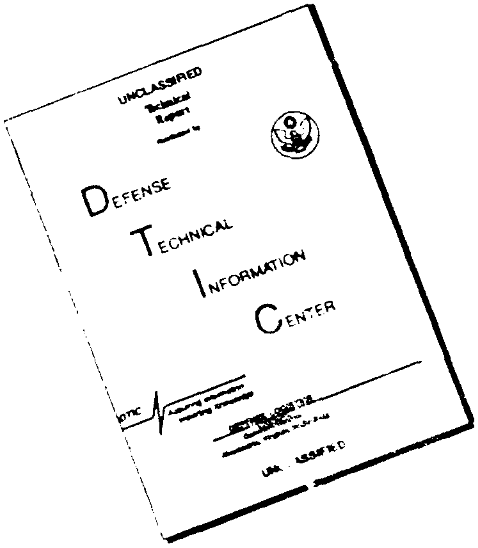
Sponsored by:

Air Force Office of Scientific Research
Bolling Air Force Base, D. C. 20332
Program Manager: Major Charles L. Nefzger
AFOSR/NM, Bldg. 410

From:

J. F. Andrus
Principal Investigator on Grant No. AFOSR 78-3641
University of New Orleans
Mathematics Department
New Orleans, Louisiana 70122

DISCLAIMER NOTICE



UNCLASSIFIED

Technical Report

DEFENSE TECHNICAL INFORMATION CENTER

DTIC

UNCLASSIFIED

THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

Abstract of Technical Progress

In earlier work the necessary conditions of optimality were derived for a problem of minimum miss-distance guidance of air-to-air missiles. The model was based upon nonlinear translational equations of motion. The solution of the necessary conditions requires a solution of a two-point boundary-condition problem. Two methods proposed for the latter solution, an elliptic integral method and a series technique, were studied and both methods were rejected in favor of a procedure based upon the quasilinearization method. The latter requires fewer assumptions and exhibits excellent convergence properties.

In order to remove the numerical integration problem and to simplify the linear two-point boundary-condition problem associated with quasilinearization, the regular method was modified, three alternative techniques being derived, and a technical report was written which discusses the convergence properties and accuracy of the three modified quasilinearization methods applied to two-point boundary-condition problems in general.

One of the latter methods was applied to the missile guidance problem and compared to a linear method of guidance. The comparison was favorable in the cases of missile encounters considered. A second technical report was prepared which describes the application and the numerical results.

Technical Report

The principal objective was the development of an optimal guidance method for air-to-air missiles. The proposed model was based upon nonlinear translational equations of motion and a quadratic performance index. Availability of a predicted time-history of the target's position was assumed.

In an earlier investigation the necessary conditions of optimality were derived and two methods were proposed for solving these conditions. They were: An elliptic integral method and a series method. These were to be placed in a form suitable for computation and applied to the solution of the necessary conditions, which reduce to a two-point boundary-condition (TPBC) problem associated with a system of nonlinear ordinary differential equations. As far as possible the following extensions (among others) to the elliptic integral and series methods were to be made:

(a) Introduction of variable missile velocity magnitude;

(b) Extension into three dimensions;

(c) Allowance for variations of more than 90° in the flight direction angles;

(d) Inclusion of the case of control angles which do not vary monotonically with time;

(e) Inclusion of an improved method for calculating time-to-go;

(f) Placing bounds upon the control variables.

After a study of the methods based upon elliptic integrals and series, it was decided that a third method based upon quasilinearization (or the generalized Newton-Raphson method) would be a more powerful computational device (as far as numerical convergence to a solution to the TPBC problem is concerned) and a more flexible tool in regard to the implementation of extensions (a) - (f) given above. On consulting with the Program Manager, it was decided that the quasilinearization method would be used rather than the methods originally proposed for study.

MODIFIED QUASILINEARIZATION METHODS

by

J. F. Andrus

Technical Report No. 69

October 1978

Revised February 1979

Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR 78-3641. The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

However, the usual quasilinearization algorithm is involved, computationally speaking, because it requires many numerical integrations of systems of differential equations and the solution of many linear TPBC problems. With this in mind the following enclosed technical report was prepared:

[1] "Modified Methods of Quasilinearization".

The latter report develops modified methods of quasilinearization in which numerical integration is unnecessary and in which the linear TPBC problems to be solved are reduced in number. The modified quasilinearization methods are applicable to any TPBC problem. A shortened version of the report has been submitted for possible publication in the SIAM Journal of Numerical Analysis.

One of the modified methods of quasilinearization was applied to the air-to-air missile problem, the details being given in the second enclosed technical report:

[2] "A Guidance Method for Air-to-Air
Missiles Using Quasilinearization"

The latter report contains a development of all pertinent equations in the case of encounters in three-dimensional space. It shows how the control variables may be bounded and gives an argument (in the case of encounters in two-dimensional space) for monotonic variation of the control variable. The report then describes numerical simulations of several two-dimensional missile-target encounters. The modified method of quasilinearization is compared to a linear guidance method, the former method providing more accurate commands than the linear method. The new method allows for variable missle velocity magnitude, determines time-to-go, and permits large changes in the flight direction angles. However, the largest change which has been simulated is 90°. The method converged to the optimal solution to the nonlinear guidance problem in all encounters considered.

The only technical investigator involved in this study has been Professor J. F. Andrus, Ph.D. in Mathematics, University of Florida, June 1958. His thesis was entitled "Partially Ordered, Ideal Preserving Groups". He is also author of a number of papers in numerical analysis, optimal control, and other areas of applied mathematics.

Modified Quasilinearization Methods

by

J. F. Andrus
Department of Mathematics
University of New Orleans

I. Introduction

The quasilinearization (or generalized Newton - Raphson) method [2] has proven to be a highly effective iterative method for the numerical solution of many two-point boundary-condition (TPBC) problems. Frequently only a rough approximation is required to start the convergence, and - when convergence takes place - it is quadratic in character.

There are also several difficulties associated with the method. They are:

(a) The differential equations must be linearized. This calls for differentiation of the righthand sides of the differential equations.

(b) After each iteration the approximate solution must be saved for use in the next iteration. This presents an awkward interpolation problem. Alternatively, each solution may be regenerated by means of integration during all succeeding iterations.

(c) During each iteration the linear differential equations must be integrated (usually numerically) several times in order to obtain the solution to a linear two-point boundary-value (TPBV) problem. Sometimes finite difference techniques are used instead to solve the TPBV problem. The solution is obtained by solving a large system of linear algebraic equations.

The present paper contains the development of modified quasilinearization methods in which the problems in items (b)and (c) have been considerably reduced. This is achieved by dividing the solutions into subarcs and, after each iteration, using the new approximate solution to obtain constant or linear approximations over each subinterval. The equations are linearized about the constant or linear approximations, thereby alleviating the problems of storage and interpolation.

Moreover, one can frequently solve the linear differential equations in terms of definite integrals or even in closed form over each subinterval. In such cases the problem of numerical integration is considerably reduced.

Since the final values of the dependent variables on each subinterval can be expressed linearly in terms of the unknown initial values, the solutions to the complete linear TPBV problem can be easily obtained because the final values of the dependent variables can be expressed linearly in terms of the unknown initial values, which can then be determined by solving a system (usually small) of linear algebraic equations. (See [1] for an example.)

It is argued that-under some reasonable assumptions- the modified methods of quasilinearization will converge. Moreover, convergence becomes quadratic in nature as the maximum subinterval size is decreased. In order to prove convergence it is necessary that the maximum interval size be chosen sufficiently small. It is also proven that the converged approximate solutions to the original differential equations are of nearly third-order accuracy (from the standpoint of numerical integration) in the case of the constant representations of the solutions over subarcs and of fourth-order accuracy in the cases of the linear methods.

One would ask how the constant solutions, for example, could lead to nearly third-order accuracy. The answer is that the solution over a typical subinterval obtained from the k-th iteration is not constant: it is merely approximated by a certain constant for use in the (k+1)-th iteration. Therefore, the quasilinearization iteration itself contributes to the accuracy of the numerical integration.

II. The Quasilinearization Method

Consider the TPBC problem consisting of the differential equations

$$\vec{y}' = \vec{f}(t,\vec{y}) \qquad (1)$$

and some boundary conditions applying at the initial and final points, $(t_0,\vec{y}_0)$ and $(t_F,\vec{y}_F)$. Here $\vec{y}$ and $\vec{f}$ signify column vectors with n components, and $\vec{y}' = d\vec{y}/dt$.

Let $\vec{y}_k(t)$ be the k-th approximation to the solution to the TPBC problem. For the (k+1)-th iteration, the standard quasilinearization algorithm determines the solution $\vec{y}_{k+1}(t)$ of the linearized problem consisting of the linearized boundary conditions and the linear differential equations

$$\vec{y}\,' = \vec{f}(t,\vec{y}_R) + \vec{f}_y(t,\vec{y}_R)(\vec{y}-\vec{y}_R)$$

where $\vec{f}_y$ is an n x n matrix of first partial derivatives.

III. Modified Quasilinearization Methods

Three modified methods of quasilinearization will be considered. For simplicity it will be assumed that t_o and t_F are fixed. This situation can always be accomplished by effecting a change in the independent variable. The interval $[t_o,t_F]$ will be divided into N subintervals $[t_{i-1},t_i]$, $i = 1,2,\ldots,N$, where $t_N = t_F$. The usual method of quasilinearization is to be modified as follows: After the solution $\vec{y}(t)$ to the linear TPBV problem has been obtained from the k-th iteration, the O.D.E.'s (1) are linearized (for the next iteration) on the i-th subinterval about

Method 1 $\quad \vec{y}_R^{(i)*} \equiv \frac{1}{2}\left[\vec{y}(t_{i-1}) + \vec{y}(t_i)\right]$

Method 2 $\quad \vec{y}_R^{(i)*} = \vec{y}(t_{i-1}) + \dfrac{t-t_{i-1}}{t_i-t_{i-1}}\left[\vec{y}(t_i) - \vec{y}(t_{i-1})\right]$

Method 3 $\quad \vec{y}_R^{(i)*} = \vec{y}(t_{i-1}) + (t-t_{i-1})\,\vec{f}\left[t_{i-1},\vec{y}(t_{i-1})\right]$

for $i = 1,2,\ldots,N$. Thus, over each subinterval, the dependent variables are approximated by means of constant or linear functions of t.

Before discussing the order of accuracy of the three methods, the convergence as $k \to \infty$ of $\vec{y}_R^{(1)*}, \vec{y}_R^{(2)*}, \ldots, \vec{y}_R^{(N)*}$, as well as the convergence of the solutions $\vec{y}(t)$, will be considered.

IV. Convergence Properties

There is no assurance at this point that the modified methods of quasilinearization will converge. An example problem, to which Method 1 has been applied, is presented in Reference [1] in full detail. In the latter example convergence is demonstrated. Next the convergence problem will be discussed from a theoretical point of view. The approach will be to transform modified quasilinearization problems to standard problems which are identical in every respect to the modified problems except for certain terms which approach zero as the iteration proceeds.

The problems will be converted to standard TPBC problems by referencing all subintervals to a single interval [0,1]. To do so, let $\Delta t_i = t_i - t_{i-1}$ and

$$\vec{F}^{(i)}(\tau,\vec{y}) = \Delta t_i \vec{F}(t_{i-1} + \tau \Delta t_i , \vec{y})$$

for i = 1,2,...,N, where τ varies from 0 to 1 and $t = t_{i-1} + \tau \Delta t_i$ on the i-th subinterval. Let $\vec{y}^{(i)}(\tau)$ signify $\vec{y}(t_{i-1} + \tau \Delta t_i)$. Let $\dot{\vec{y}}^{(i)}(\tau) = d\vec{y}^{(i)}/d\tau$. Then $\dot{\vec{y}}^{(i)} = \Delta t_i \vec{y}' = \vec{F}^{(i)}(\tau,\vec{y})$ since $dt/d\tau = \Delta t_i$. Thus the O.D.E.'s (1) have been placed in the form

$$\dot{\vec{y}}^{(i)} = \vec{F}^{(i)}(\tau,\vec{y}) \qquad (i = 1, 2, \cdots, N) \qquad (2)$$

The boundary condition problems associated with equations (1) and (2) will be equivalent if it is required that

$$\vec{y}^{(1)}(0) = \vec{y}_0,\ \vec{y}^{(N)}(1) = \vec{y}_F$$

$$\vec{y}^{(i+1)}(0) = \vec{y}^{(i)}(1) \quad (i = 1,2,\ldots,N-1)$$

where $\vec{y}_0$ and $\vec{y}_F$ satisfy the original boundary conditions.

The solution to the TPBC problem associated with equations (2) will be compared in the limit (as $\max_i \Delta t_i \to 0$) to the solutions to the TPBC problems to be defined next. The O.D.E.'s are:

$$\left.\begin{array}{ll} \dot{\vec{y}}^{(i)} = \vec{F}^{(i)}(\tau,\vec{z}^{(i)}) + \vec{F}_y^{(i)}(\tau,\vec{z}^{(i)})(\vec{y}^{(i)} - \vec{z}^{(i)}) & \\ \dot{\vec{z}}^{(i)} = \vec{c} & \text{in the case of Method 1} \\ \ddot{\vec{z}}^{(i)} = \vec{c} & \text{in Methods 2 and 3} \\ \dot{\vec{w}}^{(i)} = \vec{g}^{(i)}(\tau,\vec{y}^{(i)}) & \text{in Method 3} \\ (i = 1, 2, \cdots, N) & \end{array}\right\} (3)$$

where $\vec{g}^{(i)} = \frac{d}{d\tau}\vec{F}^{(i)} = \vec{F}_{\tau}^{(i)}(\tau, \vec{y}^{(i)}) + \vec{F}_{y}^{(i)}(\tau, \vec{y}^{(i)})\vec{F}^{(i)}(\tau, \vec{y}^{(i)})$

In the case of Method 1 the following boundary conditions will be imposed:

$$\left.\begin{array}{ll} \vec{y}^{(1)}(0) = \vec{y}_0 \,, \quad \vec{y}^{(N)}(1) = \vec{y}_F & \\ \vec{z}^{(i)}(0) = \frac{1}{2}\left[\vec{y}^{(i)}(0) + \vec{y}^{(i)}(1)\right] & (i = 1, 2, \cdots, N) \\ \vec{y}^{(i+1)}(0) = \vec{y}^{(i)}(1) & (i = 1, 2, \cdots, N-1) \end{array}\right\} \tag{4-A}$$

where $\vec{y}_0$ and $\vec{y}_F$ are required to satisfy the original boundary conditions. In the case of Method 2 the following boundary conditions will be used:

$$\left.\begin{array}{ll} \vec{y}^{(1)}(0) = \vec{y}_0 \,, \quad \vec{y}^{(N)}(1) = \vec{y}_F & \\ \vec{z}^{(i)}(0) = \vec{y}^{(i)}(0) & (i = 1, 2, \cdots, N) \\ \dot{\vec{z}}^{(i)}(0) = \vec{y}^{(i)}(1) - \vec{y}^{(i)}(0) & (i = 1, 2, \cdots, N) \\ \vec{y}^{(i+1)}(0) = \vec{y}^{(i)}(1) & (i = 1, 2, \cdots, N-1) \end{array}\right\} \tag{4-B}$$

Since $\ddot{\vec{z}}^{(i)} \equiv 0$, we have $\dot{\vec{z}}^{(i)} \equiv \vec{y}^{(i)}(1) - \vec{y}^{(i)}(0)$ and $\vec{z}^{(i)} \equiv \vec{z}^{(i)}(0) + \tau\left[\vec{y}^{(i)}(1) - \vec{y}^{(i)}(0)\right]$ in the case of Method 2. In the case of Method 3 take

$$\left.\begin{array}{ll} \vec{y}^{(1)}(0) = \vec{y}_0 \,, \quad \vec{y}^{(N)}(1) = \vec{y}_F & \\ \vec{z}^{(i)}(0) = \vec{y}^{(i)}(0) & (i = 1, 2, \cdots, N) \\ \dot{\vec{z}}^{(i)}(0) = \vec{w}^{(i)}(0) & (i = 1, 2, \cdots, N) \\ \vec{y}^{(i+1)}(0) = \vec{y}^{(i)}(1) & (i = 1, 2, \cdots, N-1) \end{array}\right\} \tag{4-C}$$

Therefore, $\vec{z}^{(i)} = \vec{y}^{(i)}(0) + \tau\vec{F}^{(i)}\left[0, \vec{y}^{(i)}(0)\right]$ in the case of Method 3, because $\dot{\vec{w}}^{(i)} = \vec{F}^{(i)}$.

As N → ∞ and $\max_i \Delta t_i \to 0$, the solutions to the TPBC problem consisting of equations (3) and (4) will approach the solution to the TPBC problem associated with equations (2) because $\vec{z}_j^{(i)}(\tau) \to \vec{y}^{(i)}(\tau)$.

What is also important is that the linearized form of equations (3) (including boundary conditions) will be the same as that of equations (2) using the modified methods of quasilinearization except for some terms which — under certain assumptions — will approach zero as k → ∞, where k is the number of iterations.

Let $\vec{y}_k^{(i)*}(\tau)$ and $\vec{z}_{jk}^{(i)*}(\tau)$ $(i = 1, 2, \cdots, \nu)$ be the result of the k-th iteration with any one of the three modified quasilinearization methods which happens to be under consideration.

The linearization of equations (3) about $\vec{y}_{jk}^{(i)*}$ and $\vec{z}_{jk}^{(i)*}$ yields

$$\begin{aligned}
\dot{\vec{y}}_j^{(i)} = {} & \vec{F}^{(i)}(\tau, \vec{z}_{jk}^{(i)*}) + \vec{F}_j^{(i)}(\tau, \vec{z}_{jk}^{(i)*})(\vec{y}_{jk}^{(i)*} - \vec{z}_{jk}^{(i)*}) \\
& + \vec{F}_j^{(i)}(\tau, \vec{z}_{jk}^{(i)*})(\vec{y}_j^{(i)} - \vec{y}_{jk}^{(i)*}) + \vec{F}_{z_j}^{(i)}(\tau, \vec{z}_{jk}^{(i)*})(\vec{z}_j^{(i)} - \vec{z}_{jk}^{(i)*}) \\
& - \vec{F}_j^{(i)}(\tau, \vec{z}_{jk}^{(i)*})(\vec{z}_j^{(i)} - \vec{z}_{jk}^{(i)*}) + \vec{\alpha}_k^{(i)}
\end{aligned}$$

$$\dot{\vec{z}}_j^{(i)} = \vec{C} \quad \text{or} \quad \ddot{\vec{z}}_j^{(i)} = \vec{C}$$

$$\dot{\vec{w}}^{(i)} = \vec{G}_j^{(i)}(\tau, \vec{y}_{jk}^{(i)*}) + \vec{G}_{y_j}^{(i)}(\tau, \vec{y}_{jk}^{(i)*})(\vec{y}_j^{(i)} - \vec{y}_{jk}^{(i)*})$$

where

$$\vec{\alpha}_k^{(i)} = \left[\sum_{j=1}^{n} (z_j^{(i)} - z_{jk}^{(i)*}) \frac{\partial}{\partial z_j} \vec{F}_j^{(i)}(\tau, \vec{z}_{jk}^{(i)*})\right](\vec{y}_{jk}^{(i)*} - \vec{z}_{jk}^{(i)*})$$

The latter equations reduce to

$$\dot{\vec{y}}_j^{(i)} = \vec{F}^{(i)}(\tau, \vec{z}_{jk}^{(i)*}) + \vec{F}_j^{(i)}(\tau, \vec{z}_{jk}^{(i)*})(\vec{y}_j^{(i)} - \vec{y}_{jk}^{(i)*}) + \vec{\alpha}_k^{(i)}$$

$$\dot{\vec{z}}_j^{(i)} = \vec{C} \quad \text{or} \quad \ddot{\vec{z}}_j^{(i)} = \vec{C}$$

$$\dot{\vec{w}}^{(i)} = \vec{G}_j^{(i)}(\tau, \vec{y}_{jk}^{(i)*}) + \vec{G}_{y_j}^{(i)}(\tau, \vec{y}_{jk}^{(i)*})(\vec{y}_j^{(i)} - \vec{y}_{jk}^{(i)*})$$

$$(i = 1, 2, \cdots, \nu)$$

Also to be considered are the following O.D.E.'s obtained by setting $\vec{u}_k^{(i)} = \vec{0}$ in equations (5):

$$\left.\begin{aligned} \dot{\vec{y}}^{(i)} &= \vec{f}^{(i)}(\tau, \vec{y}_k^{(i)*}) + \vec{f}_y^{(i)}(\tau, \vec{y}_k^{(i)*})(\vec{y}^{(i)} - \vec{y}_k^{(i)*}) \\ \dot{\vec{z}}^{(i)} &= \vec{0} \quad \text{or} \quad \ddot{\vec{z}}^{(i)} = \vec{0} \\ \dot{\vec{w}}^{(i)} &= \vec{g}^{(i)}(\tau, \vec{y}_k^{(i)*}) + \vec{g}_y^{(i)}(\tau, \vec{y}_k^{(i)*})(\vec{y}^{(i)} - \vec{y}_k^{(i)*}) \\ & (i = 1, 2, \cdots, N) \end{aligned}\right\} \qquad (5)^*$$

Equations (4) and (5)* are equivalent to the linearized TPBC problems associated with the three modified quasilinearization methods. (It is assumed that all end conditions, as well as the O.D.E.'s have been linearized.) If the $\vec{u}_k^{(i)}$ terms were to be ignored, equations (3) and (4) would lead to the same quasilinearization algorithms as would the modified methods applied to equations (2) and the original boundary conditions.

It will be assumed that if for $i = 1,2,\ldots,N$ the initial approximation $(\vec{y}_0^{*(i)}, \vec{z}_0^{*(i)})$ is sufficiently close to the true solution $(\vec{y}_T^{(i)}, \vec{z}_T^{(i)})$ to the TPBC problem consisting of equations (3) and (4), then the standard quasilinearization method applied to the latter problem will converge quadratically to the solution. Such behavior is typical of the standard quasilinearization method applied to TPBC problems occurring in practice. It will be shown that - under certain conditions - the convergence assumed above implies the convergence to $(\vec{y}_T^{(i)}, \vec{z}_T^{(i)})$, of the modified quasilinearization methods which use the linear differential equations (5)* and the boundary conditions (4).

Let $\left(\vec{y}_{k+1}^{(i)}(\tau), \vec{z}_{k+1}^{(i)}(\tau)\right)$ for $i = 1,2,\ldots,N$ be the solution obtained from a quasilinearization iteration employing equations (4) and (5). Similarly $\left(\vec{y}_{k+1}^{(i)*}(\tau), \vec{z}_{k+1}^{(i)*}(\tau)\right)$, $i = 1,2,\ldots,N$, will be the solution obtained using equations (4) and (5)*. [In both cases the linearizations are about $\left(\vec{y}_k^{(i)*}(\tau), \vec{z}_k^{(i)*}(\tau)\right)$.] The solutions must satisfy equations (5) and (5)*, respectively. Subtracting the latter equations, one obtains

$$\Delta \dot{\vec{y}}_{k+1}^{(i)} = \vec{u}_k^{(i)}$$

$$\Delta \dot{\vec{z}}_{k+1}^{(i)} = \vec{0} \quad \text{or} \quad \Delta \ddot{\vec{z}}_{k+1}^{(i)} = \vec{0}$$

$$\Delta \dot{\vec{w}}_{k+1}^{(i)} = \vec{0}$$

where $\Delta \vec{y}_{k+1}^{(i)} = \vec{y}_{k+1}^{(i)} - \vec{y}_{k+1}^{(i)*}$ and $\Delta \vec{z}_{k+1}^{(i)} = \vec{z}_{k+1}^{(i)} - \vec{z}_{k+1}^{(i)*}$.

The terms $\vec{u}_k^{(i)}$ may be written as

$$\vec{u}_k^{(i)} = R_k^{(i)} \left(\vec{z}_{k+1}^{(i)} - \vec{z}_k^{(i)*} \right)$$

where R_k is the matrix with j-th column equal to $\vec{F}_{yy_j}^{(i)} \; (y_k^{(i)*} - z_k^{(i)*})$ for $j = 1,2,\ldots,N$.

For $i = 1,2,\ldots,N$,

$$\Delta \vec{y}_{k+1}^{(i)}(\tau) = \Delta \vec{y}_{k+1}^{(i)}(0) + \int_0^{\tau} R_k^{(i)} \left(\vec{z}_{k+1}^{(i)} - \vec{z}_k^{(i)*} \right) d\tau'$$

$$\Delta \vec{z}_{k+1}^{(i)}(\tau) = \Delta \vec{z}_{k+1}^{(i)}(0)$$

<u>for case of Method 1</u>

$$\Delta \vec{z}_{k+1}^{(i)}(\tau) = \Delta \vec{z}_{k+1}^{(i)}(0) + \tau \Delta \dot{\vec{z}}_{k+1}^{(i)}(0)$$

$$\Delta \dot{\vec{z}}_{k+1}^{(i)}(\tau) = \Delta \dot{\vec{z}}_{k+1}^{(i)}(0)$$

<u>in cases of Methods 2 and 3</u>

$$\Delta \vec{w}_{k+1}^{(i)}(\tau) = \Delta \vec{w}_{k+1}^{(i)}(0)$$

<u>in the case of Method 3</u>

Let the superscript T indicate matrix transpose and define

$$\vec{p}_k^{\,T} = \left(\vec{y}_k^{(1)T}, \ldots, \vec{y}_k^{(N)T}, \vec{z}_k^{(1)T}, \ldots, \vec{z}_k^{(N)T} \right)$$

$$\vec{p}_k^{\,*T} = \left(\vec{y}_k^{(1)*T}, \ldots, \vec{y}_k^{(N)*T}, \vec{z}_k^{(1)*T}, \ldots, \vec{z}_k^{(N)*T} \right)$$

and $\Delta \vec{p}_k = \vec{p}_k - \vec{p}_k^{\,*}$ in the case of Method 1.

In the cases of Methods 2 and 3, $\dot{\vec{z}}_k^{(i)}$ and $\dot{\vec{z}}_k^{(i)*}$ for $i = 1,2,\ldots,N$ are also to be included in the vectors $\vec{p}_k$ and $\vec{p}_k^{\,*}$, respectively. In case of Method 3, $\vec{w}_k^{(i)}$ for $i = 1,2,\ldots N$ will also be included. Then

$$\Delta \vec{p}_{k+1}(\tau) = Q(\tau)\, \Delta \vec{p}_{k+1}(0) + \int_0^{\tau} S_k \left(\vec{p}_{k+1} - \vec{p}_k^{\,*} \right) d\tau' \tag{6}$$

where $Q(\tau)$ and $S_k(\tau)$ are matrices. Clearly $Q(\tau) \equiv I$ in the case of Method 1. In the cases of Methods 2 and 3, Q has some elements equal to τ which multiply the $\Delta \dot{\vec{z}}_{k+1}^{(i)}(0)$ components of $\Delta \vec{p}_{k+1}(0)$. Otherwise $Q = I$. The matrix S_k is made up of submatrices consisting of zero matrices and the matrices $R_k^{(i)}$.

The linearized boundary conditions on the k-th iteration have the form

$$B_k \vec{p}_{k+1}(0) + C_k \vec{p}_{k+1}(1) = \vec{J}_k \quad .$$

It is reasonable to assume that the matrix $B_k + C_k$ is nonsingular. For later reference it should be observed that the elements of C_k, which multiply the components of $\vec{z}^{(i)}_{k+1}(1)$ in the vector $\vec{p}_{k+1}(1)$ in the above equation are all zeros. Therefore $C_k Q \Delta \vec{p}_{k+1}(0) = C_k \Delta \vec{p}_{k+1}(0)$.

Clearly

$$B_k \Delta \vec{p}_{k+1}(0) + C_k \Delta \vec{p}_{k+1}(1) = \vec{0} \quad .$$

From equations (6) it follows that

$$B_k \Delta \vec{p}_{k+1}(0) + C_k \left[Q(1) \Delta \vec{p}_{k+1}(0) + \int_0^1 S_k (\vec{p}_{k+1} - \vec{p}_k^{\,*}) d\tau \right] = \vec{0}$$

Since $C_k Q \Delta \vec{p}_{k+1}(0) = C_k \Delta \vec{p}_{k+1}(0)$,

$$\Delta \vec{p}_{k+1}(0) = -(B_k + C_k)^{-1} C_k \int_0^1 S_k (\vec{p}_{k+1} - \vec{p}_k^{\,*}) d\tau \quad .$$

From equation (6) it follows that

$$\Delta \vec{p}_{k+1}(\tau) = -Q(\tau)(B_k + C_k)^{-1} C_k \int_0^1 S_k (\vec{p}_{k+1} - \vec{p}_k^{\,*}) d\tau + \int_0^\tau S_k (\vec{p}_{k+1} - \vec{p}_k^{\,*}) d\tau'$$

Let $\cdot$ signify the Euclidean norm of a matrix. Then, using the properties of norms,

$$|\Delta \vec{p}_{k+1}(\tau)| \le \|Q(\tau)\| \cdot \|(B_k + C_k)^{-1}\| \cdot \|C_k\| \cdot \int_0^1 |S_k (\vec{p}_{k+1} - \vec{p}_k^{\,*})| d\tau + \int_0^\tau |S_k (\vec{p}_{k+1} - \vec{p}_k^{\,*})| d\tau'$$

By the mean value theorem for integrals,

$$|\Delta \vec{p}_{k+1}(\tau)| \le \|Q(\tau)\| \cdot \|(B_k + C_k)^{-1}\| \cdot \|C_k\| \cdot |S_k(\tilde{\tau})[\vec{p}_{k+1}(\tilde{\tau}) - \vec{p}_k^{\,*}(\tilde{\tau})]| + \tau |S_k(\tau^*)[\vec{p}_{k+1}(\tau^*) - \vec{p}_k^{\,*}(\tau^*)]|$$

where $0 \le \tilde{\tau} \le 1$ and $0 \le \tau^* \le \tau$. Therefore,

Hence $|\Delta \vec{p}_{k+1}(\tau)| \le [\|Q(\tau)\| \cdot \|(B_k + C_k)^{-1}\| \cdot \|C_k\| + 1] \cdot \|S_k\|_{max} \cdot \max_{\tau'} |\vec{p}_{k+1}(\tau') - \vec{p}_k^{\,*}(\tau')|$

$$|\Delta \vec{p}_{k+1}(\tau)| \le c \max_{0 \le \tau' \le 1} |\vec{p}_{k+1}(\tau') - \vec{p}_k^{\,*}(\tau')| \qquad (7)$$

where c is a constant independent of k. The existence of c is predicated upon the boundedness of the elements of $(B_k + C_k)^{-1}$, C_k , and S_k within the region of interest. The elements of B_k , C_k, and S_k are first and second partial derivatives except that the elements of S_k contain factors Δt_i and $|y_{jk}^{(i)*}(\tau) - z_{jk}^{(i)*}(\tau)|$. In Section V it is shown that, under certain reasonable assumptions, for any $\beta^* > 0$ it is possible to take $\max\limits_i \Delta t_i$ sufficiently small that

$$\max_{\tau} |\vec{y}_k^{(i)*}(\tau) - \vec{z}_k^{(i)*}(\tau)| \leq \beta^*$$

for i = 1,2,...,N and k = 0,1,2,... . (Since $z_k^{(i)*}$ is an approximation to $y_k^{(i)*}$, the above result is not surprising provided Δt_i is chosen sufficiently small.) Under the assured conditions, it is therefore possible to choose the positive number ε, which contains $S_{k\,max}$ as a factor, as small in magnitude as desired by choosing β^* to be sufficiently small.

It has been assumed that the standard quasilinearization method applied to equations (3) and (4) will converge quadratically to the true solution $\vec{p}_T$. This means that there is a positiv. contant ξ such that

$$|\vec{p}_{k+1}(\tau) - \vec{p}_T(\tau)| < \xi |\vec{p}_k^*(\tau) - \vec{p}_T(\tau)|^2 \tag{8}$$

for all values of τ in the interval [0, 1], provided $\vec{p}_k^*$ is sufficiently close to $\vec{p}_T$. Under the same assumption ,

$$|\vec{p}_{k+1}(\tau) - \vec{p}_T(\tau)| < |\vec{p}_k^*(\tau) - \vec{p}_T(\tau)| .$$

Let β be any number in the interval (0, 1). It will be shown that if $\vec{p}_k^*$ is sufficiently close to $\vec{p}_T$, then

$$\max_{\tau} |\vec{p}_{k+1}^*(\tau) - \vec{p}_T(\tau)| \leq \beta \max_{\tau} |\vec{p}_k^*(\tau) - \vec{p}_T(\tau)| \tag{9}$$

implying that the modified quasilinearization algorithms are contraction mappings. Moreover, $\vec{p}_k^* \to \vec{p}_T$ as $k \to \infty$.

In order to verify inequality (9), the triangle inequality will be employed as follows:

$$|\vec{p}_{k+1}^*(\tau) - \vec{p}_T(\tau)| \leq |\vec{p}_{k+1}(\tau) - \vec{p}_{k+1}^*(\tau)| + |\vec{p}_{k+1}(\tau) - \vec{p}_T(\tau)|$$

By inequalities (7) and (8),

$$\left|\vec{p}^{\,*}_{k+1}(\tau) - \vec{p}_T(\tau)\right| \le c \max_{\tau'} \left|\vec{p}_{k+1}(\tau') - \vec{p}^{\,*}_k(\tau')\right| + \varepsilon \left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right|^2 \tag{10}$$

By means of inequality (10) it can be observed that if a modified method of quasilinearization converges, then the convergence becomes quadratic in character as $c \to 0$ (i.e., as $\max_i \Delta t_i \to 0$).

Provided $\vec{p}^{\,*}_k$ is sufficiently close to $\vec{p}_T$,

$$\varepsilon \left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right|^2 \le \frac{\beta}{2} \left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right| \tag{11}$$

For all values of τ in the interval [0, 1],

$$\begin{aligned} c\left|\vec{p}_{k+1}(\tau) - \vec{p}^{\,*}_k(\tau)\right| &\le c\left|\vec{p}_{k+1}(\tau) - \vec{p}_T(\tau)\right| + c\left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right| \\ &\le c\left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right| + c\left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right| \end{aligned}$$

Taking $c \le \beta/4$, we obtain

$$c\left|\vec{p}_{k+1}(\tau) - \vec{p}^{\,*}_k(\tau)\right| \le \frac{\beta}{2} \left|\vec{p}^{\,*}_k(\tau) - \vec{p}_T(\tau)\right| \tag{12}$$

for all values of τ in [0, 1].

By inequalities (10) - (12)

$$\left|\vec{p}^{\,*}_{k+1}(\tau) - \vec{p}_T(\tau)\right| \le \frac{\beta}{2} \max_{\tau} \left|\vec{p}^{\,*}_k(\tau') - \vec{p}_T(\tau')\right| + \frac{\beta}{2} \max_{\tau'} \left|\vec{p}^{\,*}_k(\tau') - \vec{p}_T(\tau')\right|$$

for all values of τ in [0, 1]. The latter inequality implies inequality (9) which was to be proven.

V. Examination of $|\vec{y}^{(i)*}_k - \vec{z}^{(i)*}_k$

In order for the analysis of Section IV to be valid it must be argued that, for any given positive number β^*,

$$\max_{\tau} \left|\vec{y}^{(i)*}_K(\tau) - \vec{z}^{(i)*}_K(\tau)\right| \le \beta^* \tag{13}$$

for $i = 1,2,\ldots,N$ and $K = 0,1,2,\ldots$ provided the maximum interval size is taken small enough.

Take N sufficiently large and $\max\limits_{i} \Delta t_i$ sufficiently small that

$$\max_{\tau} \left| \vec{y}_0^{(i)*}(\tau) - \vec{z}_0^{(i)*}(\tau) \right| \leq \beta_0 \tag{14}$$

for i = 1,2,...,N, where β_0 will be chosen sufficiently small that inequality (13) will hold. It is reasonable to assume that inequality (14) can be satisfied because the boundary conditions imposed upon $\vec{z}_0^{(i)*}(\tau)$ have been chosen in such a way that $\vec{z}_0^{(i)*}(\tau)$ will be a constant or linear approximation to $\vec{y}_0^{(i)*}(\tau)$. The approximation would be expected to improve as $\max\limits_{i} \Delta t_i$ is decreased.

Also assume that $\vec{p}_0^{\,*}$ is sufficiently close to $\vec{p}_T$ that

$$\max_{\tau} \left| \vec{p}_0^{\,*}(\tau) - \vec{p}_T(\tau) \right| \leq \alpha \tag{15}$$

where α will be selected later.

Mathematical induction will be employed. Assume that it is known that inequality (13) is true for K = 0,1,2,...,k and i = 1,2,...,N. It will be shown that it is true for K = k + 1. It was shown in Section IV that, under the given assumptions, inequality (9) would hold true as long as $\max\limits_{\tau} \vec{y}_K^{(i)*} - \vec{z}_K^{(i)*} \leq \beta^*$. Therefore

$$\max_{\tau} \left| \vec{p}_{K+1}^{\,*}(\tau) - \vec{p}_T(\tau) \right| \leq \beta \max_{\tau} \left| \vec{p}_K^{\,*}(\tau) - \vec{p}_T(\tau) \right| \tag{16}$$

for K = 0,1,2,...,k.

From inequalities (15) and (16) it follows that

$$\max_{\tau} \left| \vec{p}_K^{\,*}(\tau) - \vec{p}_T(\tau) \right| \leq \beta^K \alpha$$

for K = 0,1,2,...,k + 1. Therefore,

$$\max_{\tau} \left| \vec{z}_K^{(i)*}(\tau) - \vec{y}_T^{(i)}(\tau) \right| \leq \beta^K \alpha$$
$$(K = 0, 1, 2, \cdots k+1)$$

Now in all of the modified methods of quasilinearization, $\vec{z}_K^{(i)*}$ is of the form

$$\vec{z}_K^{(i)*}(\tau) = \vec{h}(\tau, \vec{u}, \vec{v})$$

where $\vec{u} = \vec{y}_K^{(i)*}(c)$ and $\vec{v} = \vec{y}_K^{(i)*}(1)$.

Hence

$$\vec{h}\left[\tau, \vec{y}_K^{(i)*}(0), \vec{y}_K^{(i)*}(1)\right] - \vec{h}\left[\tau, \vec{y}_T^{(i)}(0), \vec{y}_T^{(i)}(1)\right]$$

$$= \frac{\partial \vec{h}}{\partial \vec{u}} \cdot \left[\vec{y}_K^{(i)*}(0) - \vec{y}_T^{(i)}(0)\right] + \frac{\partial \vec{h}}{\partial \vec{v}} \cdot \left[\vec{y}_K^{(i)*}(1) - \vec{y}_T^{(i)}(1)\right]$$

where the partial derivatives are evaluated at intermediate values of the variables. Assuming the partial derivatives are bounded within the region of interest, we obtain

$$\left|\vec{z}_K^{(i)*}(\tau) - \vec{z}_T^{(i)}(\tau)\right| \le \left\|\frac{\partial \vec{h}}{\partial \vec{u}}\right\|_{max} \cdot \left|\vec{y}_K^{(i)*}(0) - \vec{y}_T^{(i)}(0)\right| + \left\|\frac{\partial \vec{h}}{\partial \vec{v}}\right\|_{max} \cdot \left|\vec{y}_K^{(i)*}(1) - \vec{y}_T^{(i)}(1)\right|$$

$$\le \left\|\frac{\partial \vec{h}}{\partial \vec{u}}\right\| \beta^K \alpha + \left\|\frac{\partial \vec{h}}{\partial \vec{v}}\right\| \beta^K \alpha = \mu \beta^K \alpha$$

where

$$\mu = \left\|\frac{\partial \vec{h}}{\partial \vec{u}}\right\|_{max} + \left\|\frac{\partial \vec{h}}{\partial \vec{v}}\right\|_{max}$$

Now

$$\left|\,\left|\vec{y}_K^{(i)*}(\tau) - \vec{z}_K^{(i)*}(\tau)\right| - \left|\vec{y}_T^{(i)}(\tau) - \vec{z}_T^{(i)}(\tau)\right|\,\right|$$

$$\le \left|\left[\vec{y}_K^{(i)*}(\tau) - \vec{z}_K^{(i)*}(\tau)\right] - \left[\vec{y}_T^{(i)}(\tau) - \vec{z}_T^{(i)}(\tau)\right]\right|$$

$$\le \left|\vec{y}_K^{(i)*}(\tau) - \vec{y}_T^{(i)}(\tau)\right| + \left|\vec{z}_K^{(i)*}(\tau) - \vec{z}_T^{(i)}(\tau)\right|$$

$$\le \beta^K \alpha + \mu \beta^K \alpha$$

Hence

$$\left|\,\left|\vec{y}_K^{(i)*}(\tau)-\vec{z}_K^{(i)*}(\tau)\right|-\left|\vec{y}_T^{(i)}(\tau)-\vec{z}_T^{(i)}(\tau)\right|\,\right| \leq (1+\mu)\beta^K\alpha \; .$$

With K = k + 1 and K = k , the latter inequality implies the following inequalities respectively:

$$\left|\vec{y}_{k+1}^{(i)*}(\tau)-\vec{z}_{k+1}^{(i)*}(\tau)\right|-\left|\vec{y}_T^{(i)}(\tau)-\vec{z}_T^{(i)}(\tau)\right| \leq (1+\mu)\beta^{k+1}\alpha$$

$$\left|\vec{y}_T^{(i)}(\tau)-\vec{z}_T^{(i)}(\tau)\right|-\left|\vec{y}_k^{(i)*}(\tau)-\vec{z}_k^{(i)*}(\tau)\right| \leq (1+\mu)\beta^k\alpha$$

Adding the two latter inequalities, we obtain

$$\left|\vec{y}_{k+1}^{(i)*}(\tau)-\vec{z}_{k+1}^{(i)*}(\tau)\right| \leq \left|\vec{y}_k^{(i)*}(\tau)-\vec{z}_k^{(i)*}(\tau)\right| + 2(1+\mu)\beta^k\alpha$$

using the fact that $0 < \beta < 1$. Repeated application of the above inequality gives

$$\left|\vec{y}_{k+1}^{(i)*}(\tau)-\vec{z}_{k+1}^{(i)*}(\tau)\right| \leq \left|\vec{y}_0^{(i)*}(\tau)-\vec{z}_0^{(i)*}(\tau)\right| + 2(1+\mu)\alpha(\beta^0+\beta^1+\cdots+\beta^k)$$

Therefore

$$\left|\vec{y}_{k+1}^{(i)*}(\tau)-\vec{z}_{k+1}^{(i)*}(\tau)\right| \leq \beta_0 + 2(1+\mu)\alpha S \tag{17}$$

where $S = \sum_{m=0}^{\infty}\beta^m$.

Inequality (17) holds for $0 \leq \tau \leq 1$ and $i = 1,2,\ldots,N$. In order to prove inequality (13) for $K = k + 1$, the numbers β_0 and α (see inequality (15)) must be chosen sufficiently small that $\beta_0 + 2(1 + \mu)\alpha \cdot S \leq \beta^*$; i.e., such that

$$\beta_0 + 2\left(\sum_{m=0}^{\infty}\beta^m\right)\left(1 + \left\|\frac{\partial\vec{h}}{\partial u}\right\|_{max} + \left\|\frac{\partial\vec{h}}{\partial v}\right\|_{max}\right)\alpha \leq \beta^* \quad .$$

VI. Orders of Accuracy of the Modified Methods

The question arises concerning the orders of accuracy of the converged solutions resulting from the applications of the three modified methods of quasilinearization. An interval of length Δt will be considered. For simplicity of notation it will be the first subinterval.

Define the operators

$$D = \frac{\partial}{\partial t} + \sum_j f_j \frac{\partial}{\partial y_j}$$

$$D_y = \sum_j f_j \frac{\partial}{\partial y_j} , \qquad D_{yy} = \sum_i \sum_j f_i f_j \frac{\partial}{\partial y_i} \frac{\partial}{\partial y_j}$$

The Taylor series solution of the differential equations $\vec{y}' = \vec{F}(t, \vec{y})$ through fourth-order terms is

$$\vec{y}(t_0 + \Delta t) = \vec{y}_0 + \Delta t \vec{F}_0 + \frac{1}{2}\Delta t^2 (D\vec{F}_0) + \frac{1}{6}\Delta t^3 (D^2\vec{F} + \vec{F}_y D\vec{F})_0$$
$$+ \frac{1}{24}\Delta t^4 \left[D^3\vec{F} + 3(D\vec{F}_y)D\vec{F} + \vec{F}_y D^2\vec{F} + \vec{F}_y^2 D\vec{F}\right]_0 + O(\Delta t^5)$$

where the subscript "0" indicates evaluation at $(t_0, \vec{y}_0)$. Expanding further, we obtain

$$\vec{y}(t_0 + \Delta t) = \vec{y}_0 + \Delta t \vec{F}_0 + \frac{1}{2}\Delta t^2 (\vec{F}_t + D_y \vec{F})_0$$
$$+ \frac{1}{6}\Delta t^3 (\vec{F}_{tt} + 2D_y \vec{F}_t + D_{yy}\vec{F} + \vec{F}_y \vec{F}_t + \vec{F}_y D_y \vec{F})_0$$
$$+ \frac{1}{24}\Delta t^4 \Big[\vec{F}_{ttt} + 3D_y \vec{F}_{tt} + 3D_{yy}\vec{F}_t + D_{yyy}\vec{F}$$
$$+ (3\vec{F}_{ty} + 3D_y \vec{F}_y + \vec{F}_y^2)(\vec{F}_t + D_y \vec{F})$$
$$+ \vec{F}_y (\vec{F}_{tt} + 2D_y \vec{F}_t + D_{yy}\vec{F})\Big]_0 + O(\Delta t^5) \qquad (16)$$

Now suppose $\vec{z}(t)$ is some approximation to $\vec{y}(t)$ over the interval $[t_o, t_o + \Delta t]$. Suppose $\vec{y}(t)$ is a solution to the differential equations

$$\vec{y}\,' = \vec{F}(t,\vec{z}) + \vec{F}_y(t,\vec{z})(\vec{y}-\vec{z}) \quad .$$

We will think of $\vec{y}(t)$ as the converged solution resulting from one of the modified methods of quasilinearization. The Taylor series expansion of $\vec{y}(t_o + \Delta t)$ as determined from the above differential equations will be compared to expansion (18). Let

$$\vec{g}(t,\vec{y}) = \vec{F}[t,\vec{z}(t)] + \vec{F}_y[t,\vec{z}(t)][\vec{y}-\vec{z}(t)] \quad .$$

Then in the same manner that expansion (18) was obtained, we have

$$\begin{aligned}\vec{y}(t_o+\Delta t) = \vec{y}_o + \Delta t\,\vec{g}_o &+ \tfrac{1}{2}\Delta t^2(\vec{g}_t + D_y\vec{g})_o \\ &+ \tfrac{1}{6}\Delta t^3(\vec{g}_{tt} + 2D_y\vec{g}_t + \vec{g}_y\vec{g}_t + \vec{g}_y D_y\vec{g})_o \\ &+ \tfrac{1}{24}\Delta t^4[\vec{g}_{ttt} + 3D_y\vec{g}_{tt} + (3\vec{g}_{ty} + \vec{g}_y^{\,2})(\vec{g}_t + D_y\vec{g}) \\ &+ \vec{g}_y(\vec{g}_{tt} + 2D_y\vec{g}_t)]_o + \mathcal{O}(\Delta t^5)\end{aligned} \qquad (19)$$

where certain terms have been omitted because $\partial^2\vec{g}/\partial y_i\partial y_j = \vec{0}$.

The terms of the righthand side of equations (19) will be expressed in terms of $\vec{F}$ and its derivatives. Thus

$$\begin{aligned}\vec{g}_t = \vec{F}_t(t,\vec{z}) &+ \vec{F}_y(t,\vec{z})\vec{z}\,' - \vec{F}_y(t,\vec{z})\vec{z}\,' \\ &+ \vec{F}_{ty}(t,\vec{z})(\vec{y}-\vec{z}) + [\vec{D}_y\vec{F}_y(t,\vec{z})](\vec{y}-\vec{z})\end{aligned}$$

where

$$\tilde{D}_y = \sum_j z_j' \frac{\partial}{\partial y_j}$$

$$\vec{y}_{tt} = \vec{F}_{tt}(t,\vec{z}) + \cancel{\vec{F}_{ty}}(t,\vec{z})\vec{z}' + \vec{F}_{tty}(t,\vec{z})(\vec{y}-\vec{z})$$
$$+2[\tilde{D}_y \vec{F}_{ty}(t,\vec{z})](\vec{y}-\vec{z}) - \cancel{\vec{F}_{ty}}(t,\vec{z})\vec{z}'$$
$$+[\tilde{D}_{yy}\vec{F}_y(t,\vec{z})](\vec{y}-\vec{z}) + [\tilde{D}_y'\vec{F}_y(t,\vec{z})](\vec{y}-\vec{z})$$
$$-[\tilde{D}_y \vec{F}_y(t,\vec{z})]\vec{z}'$$

where

$$\tilde{D}_{yy} = \sum_i \sum_j z_i' z_j' \frac{\partial}{\partial y_i}\frac{\partial}{\partial y_j} \quad , \quad \tilde{D}_y' = \sum_i z_i'' \frac{\partial}{\partial y_i}$$

$$\vec{y}_{ttt} = \vec{F}_{ttt}(t,\vec{z}) + \vec{F}_{tty}(t,\vec{z})\vec{z}' - \cancel{\vec{F}_{tty}}(t,\vec{z})\vec{z}'$$
$$-3[\tilde{D}_y \vec{F}_{ty}(t,\vec{z})]\vec{z}' - 2[\tilde{D}_{yy}\vec{F}_y(t,\vec{z})]\vec{z}'$$
$$-2[\tilde{D}_y'\vec{F}_y(t,\vec{z})]\vec{z}' - [\tilde{D}_y \vec{F}_y(t,\vec{z})]\vec{z}'' + o(|\vec{y}-\vec{z}|)$$

$$D_y\vec{y} = \vec{F}_y(t,\vec{y})\vec{F}(t,\vec{y}) \quad , \quad \vec{y}_y = \vec{F}_y(t,\vec{z})$$

$$D_y\vec{y}_t = \vec{F}_{ty}(t,\vec{z})\vec{F}(t,\vec{y}) + [\tilde{D}_y \vec{F}_y(t,\vec{z})]\vec{F}(t,\vec{y})$$

$$D_y\vec{y}_{tt} = \vec{F}_{tty}(t,\vec{z})\vec{F}(t,\vec{y}) + 2[\tilde{D}_y \vec{F}_{ty}(t,\vec{z})]\vec{F}(t,\vec{y})$$
$$+[\tilde{D}_{yy}\vec{F}_y(t,\vec{z})]\vec{F}(t,\vec{y}) + [\tilde{D}_y'\vec{F}_y(t,\vec{z})]\vec{F}(t,\vec{y}) + o(|\vec{y}-\vec{z}|)$$

$$\vec{g}_{ty} = \vec{F}_{ty}(t,\vec{z}) + \tilde{D}_y \vec{F}_y(t,\vec{z})$$

Expansion (19) may now be written as follows:

$$\begin{aligned}
\vec{y}(t_0+\Delta t) = {} & \vec{y}_0 + \Delta t\left[\vec{F}(t,\vec{z}) + \vec{F}_y(t,\vec{z})(\vec{y}-\vec{z})\right]_0 + \tfrac{1}{2}\Delta t^2\Big\{\vec{F}_t(t,\vec{z}) \\
& + \vec{F}_y(t,\vec{z})\vec{F}(t,\vec{y}) + \left[\vec{F}_{ty}(t,\vec{z}) + \tilde{D}_y\vec{F}_y(t,\vec{z})\right](\vec{y}-\vec{z})\Big\}_0 \\
& + \tfrac{1}{6}\Delta t^3\Big\{\vec{F}_{tt}(t,\vec{z}) - \left[\tilde{D}_y\vec{F}_y(t,\vec{z})\right]\vec{z}' + \Big[\vec{F}_{tty}(t,\vec{z}) \\
& + 2\tilde{D}_y\vec{F}_{ty}(t,\vec{z}) + D_{yy}\vec{F}_y(t,\vec{z}) + \tilde{D}_y'\vec{F}_y(t,\vec{z})\Big](\vec{y}-\vec{z}) \\
& + 2\left[\vec{F}_{ty}(t,\vec{z}) + \tilde{D}_y\vec{F}_y(t,\vec{z})\right]\vec{F}(t,\vec{y}) + \vec{F}_y(t,\vec{z})\Big[\vec{F}_t(t,\vec{z}) \\
& + \vec{F}_{ty}(t,\vec{z})(\vec{y}-\vec{z}) - (\tilde{D}_y\vec{F}_y(t,\vec{z}))(\vec{y}-\vec{z})\Big] + \vec{F}_y^2(t,\vec{z})\vec{F}(t,\vec{y})\Big\}_0 \qquad (20) \\
& + \tfrac{1}{24}\Delta t^4\Big\{\vec{F}_{ttt}(t,\vec{z}) - 3\left[\tilde{D}_y\vec{F}_{ty}(t,\vec{z})\right]\vec{z}' - 2\left[\tilde{D}_{yy}\vec{F}_y(t,\vec{z})\right]\vec{z}' \\
& - 2\left[\tilde{D}_y'\vec{F}_y(t,\vec{z})\right]\vec{z}' - \left[\tilde{D}_y\vec{F}_y(t,\vec{z})\right]\vec{z}' + 3\Big[\vec{F}_{tty}(t,\vec{z}) \\
& + 2\tilde{D}_y\vec{F}_{ty}(t,\vec{z}) + \tilde{D}_{yy}\vec{F}_y(t,\vec{z}) + \tilde{D}_y'\vec{F}_y(t,\vec{z})\Big]\vec{F}(t,\vec{y}) \\
& + \left[3\vec{F}_{ty}(t,\vec{z}) + 3\tilde{D}_y\vec{F}_y(t,\vec{z}) + \vec{F}_y^2(t,\vec{z})\right]\left[\vec{F}_t(t,\vec{z}) + \vec{F}_y(t,\vec{z})\vec{F}(t,\vec{y})\right] \\
& + \vec{F}_y(t,\vec{z})\Big[\vec{F}_{tt}(t,\vec{z}) - (\tilde{D}_y\vec{F}_y(t,\vec{z}))\vec{z}' + 2\vec{F}_{ty}(t,\vec{z})\vec{F}(t,\vec{y}) \\
& + 2(\tilde{D}_y\vec{F}_y(t,\vec{z}))\vec{F}(t,\vec{y})\Big] + \mathcal{O}(|\vec{y}-\vec{z}|)\Big\}_0 + \mathcal{O}(\Delta t^5)
\end{aligned}$$

We must now expand expressions such as $\vec{f}(t_0, \vec{z}_0)$ about $\vec{y}_0$. Thus

$$\vec F(t_0,\vec z_0) = \vec F(t_0,\vec y_0) + \vec F_y(t_0,\vec y_0)(\vec z_0-\vec y_0) + \tfrac12[D_y^* \vec F_y(t_0,\vec y_0)](\vec z_0-\vec y_0)$$

$$+ \tfrac16[D_{yy}^* \vec F_y(t_0,\vec y_0)](\vec z_0-\vec y_0) + O(|\vec z_0-\vec y_0|^4)$$

where

$$D_y^* = \sum_j (z_{j0}-y_{j0})\frac{\partial}{\partial y_j} \quad , D_{yy}^* = \sum_i\sum_j (z_{i0}-y_{i0})(z_{j0}-y_{j0})\frac{\partial^2}{\partial y_i \partial y_j}$$

$$\vec F_y(t_0,\vec z_0) = \vec F_y(t_0,\vec y_0) + D_y^*\vec F_y(t_0,\vec y_0) + \tfrac12 D_{yy}^*\vec F_y(t_0,\vec y_0) + O(|\vec z_0-\vec y_0|^3)$$

$$\vec F_t(t_0,\vec z_0) = \vec F_t(t_0,\vec y_0) + \vec F_{ty}(t_0,\vec y_0)(\vec z_0-\vec y_0) + \tfrac12[D_y^*\vec F_{ty}(t_0,\vec y_0)](\vec z_0-\vec y_0) + O(|\vec z_0-\vec y_0|^3)$$

$$\vec F_{ty}(t_0,\vec z_0) = \vec F_{ty}(t_0,\vec y_0) + D_y^*\vec F_{ty}(t_0,\vec y_0) + O(|\vec z_0-\vec y_0|^2)$$

$$\tilde D_y \vec F_y(t_0,\vec z_0) = \tilde D_y \vec F_y(t_0,\vec y_0) + D_y^*\tilde D_y \vec F_y(t_0,\vec y_0) + O(|\vec z_0-\vec y_0|^2)$$

$$\vec F_{tt}(t_0,\vec z_0) = \vec F_{tt}(t_0,\vec y_0) + \vec F_{tty}(t_0,\vec y_0)(\vec z-\vec y_0) + O(|\vec z_0-\vec y_0|^2)$$

Substituting into expansion (20), we obtain

$$\vec y(t_0+\Delta t) = \vec y_0 + \Delta t[\vec F + \vec F_y(\vec z-\vec y) - \tfrac12(D_y^*\vec F_y)(\vec z-\vec y) - \tfrac16(D_{yy}^*\vec F_y)(\vec z-\vec y)$$

$$- \vec F_y(\vec z-\vec y)]_0 + \tfrac12\Delta t^2[\vec F_t + \vec F_{ty}(\vec z-\vec y) + \tfrac12(D_y^*\vec F_{ty})(\vec z-\vec y)$$

$$+ (\vec F_y + D_y^*\vec F_y + \tfrac12 D_{yy}^*\vec F_y)\vec F - (\vec F_{ty} + D_y^*\vec F_{ty} + \tilde D_y\vec F_y$$

$$+ D_y^*\tilde D_y\vec F_y)(\vec z-\vec y)]_0 + \tfrac16\Delta t^3\{\vec F_{tt} + \vec F_{tty}(\vec z-\vec y)$$

$$\begin{aligned}
&-(\tilde{D}_y \vec{F}_y + D_y^* \tilde{D}_y \vec{F}_y)\vec{z}' - (\vec{F}_{tty} + 2\tilde{D}_y \vec{F}_{ty} + \tilde{D}_{yy} \vec{F}_y + \tilde{D}_y' \vec{F}_y)(\vec{z} - \vec{y}) \\
&+2(\vec{F}_{ty} + D_y^* \vec{F}_{ty} + \tilde{D}_y \vec{F}_y + D_y^* \tilde{D}_y \vec{F}_y)\vec{F} + (\vec{F}_y + D_y^* \vec{F}_y)[\vec{F}_t \\
&+ \cancel{\vec{F}_{ty}(\vec{z} - \vec{y})} - \cancel{\vec{F}_{ty}(\vec{z} - \vec{y})} - \tilde{D}_y \vec{F}_y(\vec{z}\vec{y}) + (\vec{F}_y + D_y^* \vec{F}_y)^2 \vec{F}\}_0 \qquad (21) \\
&+ \frac{1}{24}\Delta t^4 \{\vec{F}_{ttt} - 3(\tilde{D}_y \vec{F}_{ty})\vec{z}' - 2(\tilde{D}_{yy} \vec{F}_y)\vec{z}' - 2(\tilde{D}_y' \vec{F}_y)\vec{z}' \\
&- (\tilde{D}_y \vec{F}_y)\vec{z}'' + 3(\vec{F}_{tty} + 2\tilde{D}_y \vec{F}_{ty} + \tilde{D}_{yy} \vec{F}_y + \tilde{D}_y' \vec{F}_y)\vec{F} \\
&+ (3\vec{F}_{ty} + 3\tilde{D}_y \vec{F}_y + \vec{F}_y^2)(\vec{F}_t + \vec{F}_y \vec{F}) + \vec{F}_y[\vec{F}_{tt} - (\tilde{D}_y \vec{F}_y)\vec{z}' \\
&+ 2\vec{F}_{ty}\vec{F} + 2(\tilde{D}_y \vec{F}_y)\vec{F}]\}_0 + \sigma(\Delta t^5)
\end{aligned}$$

where $\vec{z}_0 - \vec{y}_0$ is considered to be equal to $O(\Delta t)$ because $\vec{z}_0$ will be of the form $\vec{z}_0 = \vec{y}_0 + O(\Delta t)$. (See below)

Suppose $\vec{z} \equiv \vec{y}_0 + \Delta t\, \vec{\eta} + \sigma(\Delta t^2)$. Then $\vec{z}' = \vec{z}'' = \vec{0}$ if $\vec{\eta}$ is constant.

Substituting into expansion (21) and rewriting the expansion through third-order terms, we obtain

$$\vec{y}(t_0+\Delta t) = \vec{y}_0 + \Delta t\left[\vec{F} - \tfrac{1}{2}\Delta t^2 (D_\eta \vec{F}_y)\vec{\eta}\right]_0$$
$$+ \tfrac{1}{2}\Delta t^2\left[\vec{F}_t + (\vec{F}_y + \Delta t\, D_\eta \vec{F}_y)\vec{F}\right]_0$$
$$+ \tfrac{1}{6}\Delta t^3\left(\vec{F}_{tt} + 2\vec{F}_{ty}\vec{F} + \vec{F}_y\vec{F}_t + \vec{F}_y^2\vec{F}\right)_0 + O(\Delta t^4)$$

where $D_\eta = \sum_j \eta_j \frac{\partial}{\partial y_j}$. Therefore

$$\vec{y}(t_0+\Delta t) = \vec{y}_0 + \Delta t\,\vec{F}_0 + \tfrac{1}{2}\Delta t^2(\vec{F}_t + \vec{F}_y\vec{F})_0$$
$$+ \tfrac{1}{6}\Delta t^3\left[-3(D_\eta \vec{F}_y)\vec{\eta} + 3(D_\eta \vec{F}_y)\vec{F} + \vec{F}_{tt}\right.$$
$$\left. + 2\vec{F}_{ty}\vec{F} + \vec{F}_y\vec{F}_t + \vec{F}_y^2\vec{F}\right]_0 + O(\Delta t^4) \qquad (22)$$

In order for the above expansion to agree with expansion (18) through third-order terms, it would only be necessary that

$$3(D_\eta \vec{F}_y)_0(\vec{F}_0 - \vec{\eta}) = (D_y\vec{F}_y)_0\vec{F}_0$$

Suppose $\vec{\eta} = \alpha\vec{F}_0$. Then we desire

$$3\alpha(1-\alpha)(D_y\vec{F}_y)_0\vec{F}_0 = (D_y\vec{F}_y)_0\vec{F}_0$$

This is equivalent to $3\alpha(1-\alpha) = 1$ or $3\alpha^2 - 3\alpha + 1 = 0$. Minimizing $3\alpha^2 - 3\alpha + 1$ with respect to α, we find $\alpha = \frac{1}{2}$. Therefore, we let $\vec{\eta} = \frac{1}{2}\vec{F}_o$ and we obtain the term $\frac{3}{4}(D_y\vec{F}_y)_o\vec{F}_o$ in expansion (22) rather than $(D_y\vec{F}_y)_o\vec{F}_o$. Consequently, expansion (22) nearly" compares to expansion (18) through third-order terms.

Hence, if y(t) is approximated by means of $\vec{z} \equiv \vec{y}_0 + \frac{1}{2}\Delta t \vec{F}_0 + O(\Delta t^2)$, we obtain $\vec{y}$ to nearly third-order accuracy from modified quasilinearization.

Suppose we take $\vec{z} \equiv \frac{1}{2}[\vec{y}_0 + \vec{y}_a(t_0 + \Delta t)]$ as in Method 1, where $\vec{y}_a$ is the solution to $\vec{y}_a{}' = \vec{F}(t, \vec{z}) + \vec{F}_y(t, \vec{z})(\vec{y}_a - \vec{z})$. Then

$$\vec{y}_a(t_0 + \Delta t) = \vec{y}_0 + \Delta t[\vec{F}(t_0, \vec{z}) + \vec{F}_y(t_0, \vec{z})(\vec{y}_0 - \vec{z})] + O(\Delta t^2)$$

We see that $\vec{z} = \vec{y}_0 + O(\Delta t)$. Therefore

$$\vec{y}_a(t_0 + \Delta t) = \vec{y}_0 + \Delta t \vec{F}(t_0, \vec{y}_0) + O(\Delta t^2)$$

$$\vec{z} \equiv \vec{y}_0 + \frac{1}{2}\Delta t \vec{F}_0 + O(\Delta t^2)$$

Hence nearly third-order accuracy is obtained either from Method 1 or the method in which $\vec{z} \equiv \vec{y}_0 + \frac{1}{2}\Delta t \vec{F}_0$ directly (assuming the latter method converges).

Now take $\vec{z} = \vec{y}_0 + (t - t_0)[\vec{y}_a(t_0 + \Delta t) - \vec{y}_0]/\Delta t$ as in Method 2. Then $\vec{z}_0 - \vec{y}_0 = \vec{0}$, $\vec{z}'' = \vec{0}$, and

$$\vec{z}' = \frac{1}{\Delta t}[\vec{y}_a(t_0 + \Delta t) - \vec{y}_0]$$

$$= \frac{1}{\Delta t}\Big\{\Delta t[\vec{F}(t, \vec{z}) + \vec{F}_y(t, \vec{z})(\vec{y}_a - \vec{z})]_0$$

$$+ \frac{1}{2}\Delta t^2[\vec{F}_t(t, \vec{z}) + \cancel{\tilde{D}_y}\vec{F}(t, \vec{z}) + \vec{F}_y(t, \vec{z})(\vec{y}_a{}' - \cancel{\vec{z}'})] + O(\Delta t^3)\Big\}$$

$$= [\vec{F}(t, \vec{z})]_0 + \frac{1}{2}\Delta t[\vec{F}_t(t, \vec{z}) + \vec{F}_y(t, \vec{z})\vec{F}(t, \vec{z})]_0 + O(\Delta t^2)$$

$$= \vec{F}(t_0, \vec{y}_0) + \frac{1}{2}\Delta t[\vec{F}_t(t_0, \vec{y}_0) + \vec{F}_y(t_0, \vec{y}_0)\vec{F}(t_0, \vec{y}_0)] + O(\Delta t^2)$$

$$\vec{z}' = \vec{F}_0 + \frac{1}{2}\Delta t(\vec{F}_t + \vec{F}_y\vec{F})_0 + O(\Delta t^2)$$

In the case of Method 3, $\vec{z} = \vec{y}_0 + (t - t_0)\vec{f}_0$ so that again $\vec{z}_0 - \vec{y}_0 = \vec{0}$ and $\vec{z}'' = \vec{0}$. However $\vec{z}' = \vec{f}_0$. Therefore, in the cases of Methods 2 and 3,

$$\vec{z}_0 - \vec{y}_0 = \vec{0}\,, \qquad \vec{z}'' = \vec{0}$$
$$\vec{z}' = \vec{F}_0 + \tfrac{1}{2}\beta\Delta t(\vec{F}_t + \vec{F}_y\vec{F})_0 + \mathcal{O}(\Delta t^2)$$

where $\beta = 1$ and $\beta = 0$ in the cases of Methods 2 and 3 respectively.

Now substituting into expansion (21), we obtain

$$\begin{aligned}
\vec{y}(t_0 + \Delta t) = {} & \vec{y}_0 + \Delta t\vec{F}_0 + \tfrac{1}{2}\Delta t^2(\vec{F}_t + \vec{F}_y\vec{F})_0 \\
& + \tfrac{1}{6}\Delta t^3\Big\{\vec{F}_{tt} - (\tilde{D}_y\vec{F}_y)\big[\vec{F} + \tfrac{1}{2}\beta\Delta t(\vec{F}_t + \vec{F}_y\vec{F})\big] \\
& + 2(\vec{F}_{ty} + \tilde{D}_y\vec{F}_y)\vec{F} + \vec{F}_y\vec{F}_t + \vec{F}_y^{\,2}\vec{F}\Big\}_0 \\
& + \tfrac{1}{24}\Delta t^4\Big\{\vec{F}_{ttt} - 3(\cancel{D_y}\vec{F}_{ty})\vec{F} - 2(\cancel{D_{yy}}\vec{F}_y)\vec{F} \\
& + (3\vec{F}_{tty} + \overset{3}{6}D_y\vec{F}_{ty} + \overset{1}{3}D_{yy}\vec{F}_y)\vec{F} \\
& + (3\vec{F}_{ty} + 3D_y\vec{F}_y + \vec{F}_y^{\,2})(\vec{F}_t + \vec{F}_y\vec{F}) \\
& + \vec{F}_y\big[\vec{F}_{tt} - (\cancel{D_y}\vec{F}_y)\vec{F} + 2\vec{F}_{ty}\vec{F} + \tfrac{1}{2}(D_y\vec{F}_y)\vec{F}\big]\Big\}_0 + \mathcal{O}(\Delta t^5)
\end{aligned}$$

But

$$\big[(\tilde{D}_y\vec{F}_y)\vec{F}\big]_0 = \big[(D_y\vec{F}_y)\vec{F} + \tfrac{1}{2}\beta\Delta t(D_y\vec{F}_y)(\vec{F}_t + \vec{F}_y\vec{F})\big]_0$$

because
$$\Big[\Big(\sum_j f_j'\frac{\partial}{\partial y_j}\Big)\vec{F}_y\Big]\vec{F} = \Big[\Big(\sum_j f_j\frac{\partial}{\partial y_j}\Big)\vec{F}_y\Big]\vec{F}'$$

where $\vec{F} = \vec{F}_t + \vec{F}_y\vec{F}$. $\{$If $\vec{a}$ and $\vec{b}$ are constant, then

$$\Big(\sum_j a_j \frac{\partial}{\partial y_j}\vec{F}_y\Big)\vec{b} = \sum_j a_j \frac{\partial}{\partial y_j}(\vec{F}_y\vec{b}) = \Big[\frac{\partial}{\partial y}(\vec{F}_y\vec{b})\Big]\vec{a} = \Big(\sum_j b_j \frac{\partial}{\partial y_j}\vec{F}_y\Big)\vec{a}.\Big\}$$

Therefore the expansion becomes

$$\begin{aligned}
\vec{y}(t_0+\Delta t) &= \vec{y}_0 + \Delta t\,\vec{F}_0 + \tfrac{1}{2}\Delta t^2(\vec{F}_t + \vec{F}_y\vec{F})_0 \\
&+ \tfrac{1}{6}\Delta t^3\Big[\vec{F}_{tt} - (\cancel{D_y}\vec{F}_y)\vec{F} - \tfrac{1}{2}\cancel{\beta\Delta t}\,(D_y\vec{F}_y)(\vec{F}_t + \vec{F}_y\vec{F}) \\
&- \tfrac{1}{2}\beta\cancel{\Delta t}\,(D_y\vec{F}_y)(\vec{F}_t + \vec{F}_y\vec{F}) + 2\vec{F}_{ty}\vec{F} + \overset{1}{\cancel{2}}(D_y\vec{F}_y)\vec{F} \\
&+ \beta\cancel{\Delta t}\,(D_y\vec{F}_y)(\vec{F}_t + \vec{F}_y\vec{F}) + \vec{F}_y\vec{F}_t + \vec{F}_y^2\vec{F}\Big]_0 \\
&+ \tfrac{1}{24}\Delta t^4\Big\{\vec{F}_{ttt} + (3\vec{F}_{tty} + 3D_y\vec{F}_{ty} + D_{yy}\vec{F}_y)\vec{F} \\
&+ (3\vec{F}_{ty} + 3D_y\vec{F}_y + \vec{F}_y^2)(\vec{F}_t + \vec{F}_y\vec{F}) \\
&+ \vec{F}_y\big[\vec{F}_{tt} + 2\vec{F}_{ty}\vec{F} + (D_y\vec{F}_y)\vec{F}\big]\Big\}_0 + O(\Delta t^5)
\end{aligned}$$

The latter expansion compares term-by-term with expansion (18). This implies that Methods 2 and 3 give rise to fourth-order numerical accuracy.

REFERENCES

1. Andrus, J. F., "A Guidance Method for Air-to-Air Missiles Using Quasilinearization," Tech. Rept. No. 68, May 1979, Mathematics Dept., University of New Orleans.

2. Bellman, R. E. and Kalaba, R. E., Quasilinearization and Nonlinear Boundary-Value Problems, Amer. Elsevier, New York, 1965.

A GUIDANCE METHOD FOR AIR-TO-AIR MISSILES USING QUASILINEARIZATION

by

Jan F. Andrus

Technical Report No. 68

May 1979

A GUIDANCE METHOD FOR AIR-TO-AIR MISSILES USING QUASILINEARIZATION

J. F. Andrus
University of New Orleans, New Orleans, La.

ABSTRACT

The problem is real-time guidance of an air-to-air missile engaging an accelerating target whose velocity vector can be predicted as a function of time. A Zermelo-type model, consisting of the nonlinear translational equations of motion, is employed. It is assumed that the velocity magnitude of the missile is a known function of time on each guidance cycle. The performance index is a weighted sum of squares of the miss-distance and the time-rates-of-change, u_A and u_E.of the flight path angles. It is to be minimized with respect to $u_A(t)$, $u_E(t)$, and the final time. A modified quasilinearization method is used to solve the two-point boundary-condition problem associated with the necessary conditions of optimality. It is applied to several missile engagements and is shown to be considerably more accurate over a single guidance cycle than a method based upon the linearized translational equations of motion.

Research sponsored by the Air Force Office of Scientific Research, Air Force Systems Command, USAF, under Grant No. AFOSR 78-3641 - The United States Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation hereon.

I. INTRODUCTION

There is an acknowledged gap between the theory and the application of nonlinear optimal control theory - especially in the case of real-time applications. Since the necessary conditions of optimality frequently take the form of a two-point boundary-condition problem, the quasilinearization (or generalized Newton-Raphson) method[1] comes to mind because of its favorable convergence properties. However, there are difficulties associated with the latter method because of the need to solve sequences of linear two-point boundary-value problems: One must deal with the problem of storing intermediate solutions or must numerically integrate a large number of differential equations. Methods based upon parameterization of the control and use of nonlinear programming bring forth similar problem.

Modified quasilinearization methods[2] have been developed in which the problems of storage and numerical integration have been reduced, especially for problems in which high-order accuracy in the numerical integration is uncalled for. As a matter of fact, in most real-time applications it is unnecessary and unwise to spend computer time integrating with high-order accuracy when the physical constants can not be determined with comparable accuracy.

In the modified quasilinearization method which is applied in this paper, the solutions to the differential equations are divided into several subarcs. Intermediate approximate solutions are taken to be constant over each subarc, thereby permitting the linearized differential equations to be integrated in closed form over each subarc in the case of the air-to-air missiles problem. It can be argued[2] that this quasilinearization method will converge under certain assumptions and that the converged solution will have nearly third-order accuracy in the sense that it is equivalent to integrating the differential equations with an approximate third-order formula of numerical integration.

The development of microprocessor technology and improvements in estimation methods now make it feasible to incorporate nonlinear real-time guidance algorithms, such as the one developed in this paper, into the control systems of air-to-air missiles.

The new guidance method is expected to lead to smaller miss-distances than the commonly used proportional navigation method which is non-optimal for the cases of accelerating targets and nonlinear equations of motion.

It should be mentioned that there is a system[3] of finite, nonlinear equations, involving elliptic integrals, whose solution provides the optimum solution to the air-to-air missile problems of the type considered in this paper. However, the method is restricted to two dimensions and constant missile velocity magnitude. It also requires a good initial approximation to the solution.

Models of the physical problem, similar to that to be employed in this paper, have been used in the past; for example, one paper[4] makes use of a linear expansion about a solution to a simple linear control problem in order to obtain an approximate solution to the nonlinear equations. The latter paper makes use of several simplifying assumptions such as constant closing rate.

II. THE MODEL

The space-fixed coordinate system to be employed is depicted in Figure 1. The angles, γ_A and γ_E, define the direction of the instantaneous missile velocity vector $\vec{V}_M$.

The equations of relative motion of the target and missile are

$$\dot{\vec{R}} = \vec{V}_T(t) - V_M(t)\vec{p}(\vec{\gamma}) \qquad (1)$$

where $\vec{R}(t)$ is the line-of-sight vector from the missile to the target, $\vec{V}_T(t)$ is the predicted velocity vector of the target, $V_M(t)$ is the velocity magnitude of the missile,

$$\vec{\gamma} = \begin{bmatrix} \gamma_A \\ \gamma_E \end{bmatrix} = \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \qquad \vec{p} = \begin{bmatrix} \cos\gamma_E \cos\gamma_A \\ \cos\gamma_E \sin\gamma_A \\ -\sin\gamma_E \end{bmatrix} .$$

Here $V_M(t)$ is assumed to be a known function of time. However, it may be adjusted from one guidance cycle to the next. The initial values, $\vec{\gamma}_0$ and $\vec{R}_0$, of $\vec{\gamma}$ and $\vec{R}$ are assumed to be given.

The control is $\vec{u} = \dot{\vec{\gamma}}$ which may be bounded according to the inequalities

$$\vec{u}_{min} \leq \vec{u} \leq \vec{u}_{max}$$

where $\vec{u}_{min} < 0$ and $\vec{u}_{max} > 0$.

The performance index is

$$J = \frac{1}{2}\kappa \|\vec{R}(t_F)\|^2 + \frac{1}{2}\int_{t_0}^{t_F} \vec{u}^T W(t)\vec{u}\, dt$$

where $W(t)$ is a given continuous positive-definite weighting matrix and κ is a given positive constant. The index J is to be minimized with respect to $\vec{u}(t)$ and the final time, t_F.

III. NECESSARY CONDITIONS OF OPTIMALITY

It is necessary that the Hamiltonian

$$H = \frac{1}{2}\vec{u}^T W \vec{u} + \vec{\lambda}^T(\vec{V}_T - V_M\vec{p}) + \vec{\mu}^T\vec{u}$$

be a minimum with respect to $\vec{u}$ at each instant of time. Therefore, when $\vec{u}$ is not on a boundary, $\partial H/\partial\vec{u} = \vec{0}^T$ so that

$$\vec{u} = -W^{-1}\vec{\beta} \quad .$$

The adjoint equations are

$$\dot{\vec{\lambda}} = \vec{0}$$

$$\dot{\vec{\beta}} = V_M \vec{p}_\gamma^T \vec{\gamma} \qquad (2)$$

where $\vec{\lambda}$ and $\vec{\beta}$ are the adjoint variables and $\vec{p}_\gamma$ is the partial derivative matrix $\partial\vec{p}/\partial\vec{\gamma}$.

The transversality conditions are

$$\vec{\lambda}(t_F) = k\vec{R}(t_F), \quad \vec{\beta}(t_F) = \vec{0}$$

$$\dot{\vec{R}}^T(t_F)\vec{R}(t_F) = 0 \qquad (t_F \text{ free})$$

Clearly $\vec{u}(t_F) = 0$ and $\vec{\lambda} = k\vec{R}(t_F)$.

Now consider a time interval on which u_i is on a boundary. Later arguments indicate that, if u_i is on a boundary, it is likely to be during the initial portion of the flight. Recall that $H^* = \frac{1}{2}\vec{u}^T W \vec{u} + \vec{\beta}^T \vec{u}$ must be a minimum with respect to $\vec{u}$. At any given time the surface $z = H^*(\vec{u})$ must be concave upwards because W is positive definite. Let $\vec{u}^* = -W^{-1}\vec{\beta}$. Then we have the following necessary condition:

$$\left.\begin{aligned} &\text{If } \ u_{i\,\min} \le u_i^* \le u_{i\,\max}\,, \text{ then } \ u_i = u_i^* \,. \\ &\text{If } \ u_i^* < u_{i\,\min}, \text{ then } \ u_i = u_{i\,\min} \,. \\ &\text{If } \ u_i^* > u_{i\,\max}\,, \text{ then } \ u_i = u_{i\,\max} \,. \end{aligned}\right\} \qquad (3)$$

Since $\vec{\beta}$ and W are continuous, $\vec{u}^*$ and (hence) $\vec{u}$ will be continuous.

At least in the case of two-dimensional problems with $v_E \equiv 0$, one would expect the flight path angle γ_A to vary monotonically with respect to time. This point will be argued more precisely as follows. If t_F is free (not given), then $\vec{R}(t_F) \cdot \dot{\vec{R}}(t_F) = 0$. Therefore, if $V_M(t_F)$ is large in comparison to $\vec{V}_T(t_F)$, equations (1) imply $\vec{p}[\vec{\gamma}(t_F)] \cdot \vec{R}(t_F) \doteq 0$; i.e. the optimal flight path of the missile is approximately perpendicular to the line-of-sight vector at the final time. Now suppose that at an instant, $\dot{u}_A(\tau) = 0$ for some τ in the intervₐl (t_o, t_F). Therefore, $\dot{\beta}_1(\tau) = 0$ and equations (2) imply $\vec{p}_{\gamma_A}[\gamma_A(\tau)] \cdot \vec{R}(t_F) = 0$. However, $\vec{p}$ is perpendicular to $\vec{p}_{\gamma_A}$ so that $\vec{p}[\gamma_A(\tau)]$ is parallel to $\vec{R}(t_F)$ and hence nearly perpendicular to $\vec{p}[\gamma_A(t_F)]$. This situation can not be possible if γ_A varies by less than 90° over the course of the flight. Consequently, in such a case it can not be true that $\dot{u}_A(\tau) = 0$. It follows that u_A is monotonic and, since $u(t_F) = 0$, u can not change sign. Therefore, γ_A is also monotonic under the assumed conditions.

In summary, the necessary conditions consist of the differential equations

$$\left.\begin{aligned} \dot{\vec{\gamma}} &= \vec{u} \\ \dot{\vec{\beta}} &= V_M \vec{P}_\gamma^T \vec{\lambda} \\ \dot{\vec{R}} &= \vec{V}_T - V_M \vec{p}(\vec{\gamma}) \\ \dot{\vec{\lambda}} &= \vec{0} \end{aligned}\right\} \qquad (4)$$

and the boundary conditions

$$\vec{\gamma}(t_o) = \vec{\gamma}_o$$

$$\vec{\beta}(t_F) = \vec{0}$$

$$\vec{R}(t_o) = \vec{R}_o$$

$$\vec{\lambda}(t_F) = k\vec{R}(t_F)$$

$$\vec{R}^T(t_F)\vec{R}(t_F) = 0$$

where $\vec{u}$ is determined by means of equations (3).

IV. MODIFIED QUASILINEARIZATION METHOD

A modification of the quasilinearization method has been developed by the author[2]. The method will now be described briefly.

Consider the differential equations

$$\dot{\vec{y}} = \vec{f}(t,\vec{y}) \qquad (5)$$

with some associated boundary conditions applying at times t_o and t_F. Here $\vec{y}$ signifies a vector of n components. The time interval $[t_o,t_F]$ is divided into subintervals $[t_{i-1},t_i]$ where $i = 1,2,\ldots,N$ and $t_N = t_F$. The linearized equations are

$$\dot{\vec{y}} = \vec{f}(t,\vec{y}_s) + \vec{f}_y(t,\vec{y}_s)(\vec{y}-\vec{y}_s) \qquad (6)$$

where $\vec{y}_s(t)$ is an approximate solution to the two-point boundary-condition (TPBC) problem. The usual quasilinearization method calls for the solution to the TPBC problem associated with equations (6). The solution $\vec{y}$ is the new approximation $\vec{y}_s$ to be used in equations (6) for the next iteration. The solution to the TPBC problem during each iteration calls for several integrations of equations (6) or for the solution to a large system of finite difference equations.

In the modified method to be applied in this paper, the solution $\vec{y}$ to the TPBC problem is obtained at times $t_o,t_1,\ldots,t_N$. Then one lets

$$\vec{y}_s = \vec{y}_{si} \equiv \frac{1}{2}[\vec{y}(t_{i-1}) + \vec{y}(t_i)]$$

on the i-th time interval for $i = 1,2,\ldots,N$. As we will see, the resulting method converges in the missile encounters to be considered. Moreover, it is known[2] that the numerical accuracy of the converged solution is nearly of the third order.

Over each subinterval, $\vec{y}_s$ in equations (6) will be constant. Only $\vec{y}(t_o)$, $\vec{y}(t_1),\ldots,\vec{y}(t_N)$ must be stored after each iteration. Moreover, the integration of equations (6) over each interval has been simplified. In fact, it can often be carried out in closed form.

Assume that, with appropriate substitutions, equations (5) take an equivalent form

$$\dot{\vec{y}} = \vec{f}(\vec{y}) + \vec{g}(t)$$

Then on each iteration equations (6) will take the form

$$\dot{\vec{y}} = A_i\vec{y} + \vec{b}_i(t) \qquad (7)$$

on the i-th subarc, where A_i is a constant matrix. It can be shown that

$$\vec{y}(t_i) = B_i\vec{y}(t_{i-1}) + \vec{\vartheta}_i \qquad (8)$$

for some matrix B_i and some vector $\vec{\vartheta}_i$. Let $M_N = I$ and $M_i = M_{i+1}\ B_{i+1}$ for $i = N-1, N-2, \ldots, 0$. It can easily be shown that

$$\vec{y}(t_N) = M_0\vec{y}(t_0) + \sum_{i=1}^{N} M_i\vec{\vartheta}_i \ .$$

Therefore, using the latter expression which gives $\vec{y}(t_F)$ in terms of $\vec{y}(t_o)$, one can write all the boundary conditions in terms of $\vec{y}(t_o)$. One determines $\vec{y}(t_o)$ in order to satisfy the boundary conditions and employs equations (8) in order to obtain $\vec{y}(t_1)$, $\vec{y}(t_2), \ldots, \vec{y}(t_N)$.

Here we have assumed t_F is given. However, it is possible to adjust it in an "outside loop" as will be explained later.

The quasilinearization method can also be modified as follows: Rather than accepting the solution $\vec{y}$ to the linearized equations as the new approximate solution $\vec{y}_s$, let

$$\vec{y}_s^{\,(new)}(t_i) = \vec{y}_s^{\,(old)}(t_i) + \alpha(\vec{y}(t_i) - \vec{y}_s^{\,(old)}(t_i))$$

for $i = 1, 2, \ldots, N$, where α is a positive constant chosen such that the maximum change in $\vec{y}$ will not become "dangerously" large. This is analogous to the damping method sometimes used in connection with the Newton-Raphsor method used to solve a system of nonlinear. finite equations. The damping technique has enabled the quasilinearization method to converge in some cases in the air-to-air missile problem when it would not have converged otherwise. (In some cases it might be desirable to take $\alpha > 1$, but we have not done so.)

Now we will apply the modified quasilinearization method to the air-to-air missile problem. Although it is applicable to the three-dimensional problem, it will be applied herein only to the two-dimensional problem with $\gamma_E \equiv 0$. Let $\gamma = \gamma_A$. Take $J = \frac{1}{2} k|\vec{R}(t_F)|^2 + \frac{1}{2}\int_0^{t_F} u^2 dt$, where $u = \dot{\gamma}$. Equations (4) reduce to the equations

$$\left.\begin{aligned}
\dot{\gamma} &= u \\
\dot{u} &= -V_M \vec{p}_\gamma^{\,T} \vec{\lambda} \\
\dot{\vec{R}} &= \vec{V}_T(t) - V_M \vec{p} \\
\dot{\vec{\lambda}} &= \vec{0} \\
\dot{V}_M &= q(t)
\end{aligned}\right\} \qquad (9)$$

where $\vec{p}^{\,T} = (\cos\gamma, \sin\gamma)$ and - in order to place the equations in the form of equations (7) - we introduce the differential equation $\dot{V}_M = q(t)$ with $q(t)$ being a given function.

The linearized equations on the i-th subarc are

$$\left.\begin{aligned}
\dot{\gamma} &= u \\
\dot{u} &= -\gamma V_{Mi}\vec{p}_{\gamma\gamma i}^{\,T}\vec{\lambda}_i - V_{Mi}\vec{p}_{\gamma i}^{\,T}\vec{\lambda} - V_M\vec{p}_{\gamma i}^{\,T}\vec{\lambda}_i + V_{Mi}\vec{p}_{\gamma i}^{\,T}\vec{\lambda}_i + \gamma_i V_{Mi}\vec{p}_{\gamma\gamma i}^{\,T}\vec{\lambda}_i \\
\dot{\vec{R}} &= -\gamma V_{Mi}\vec{p}_{\gamma i} - V_M\vec{p}_i + \vec{V}_T(t) + \gamma_i V_{Mi}\vec{p}_{\gamma i} \\
\dot{\vec{\lambda}} &= \vec{0} \\
\dot{V}_M &= q(t)
\end{aligned}\right\} \qquad (10)$$

where all quantities with the subscript i are constant over the i-th subinterval. The quantity $\vec{p}_i$, for example.stands for $\vec{p}_\gamma(\gamma_{si})$. Thus the subscript i signifies evaluation on the approximate solution $\vec{y}_{si}$ on the i-th interval.

Letting $\vec{\eta}_i = -V_{Mi}\vec{p}_{\gamma i}$, $\alpha_i = V_{Mi}\vec{p}_i^{\,T}\vec{\lambda}_i$, and taking advantage of the knowledge that $\vec{\lambda}$ is a constant and $\vec{p}_{\gamma\gamma} = -\vec{p}$, the system (10) can be written as

$$\begin{bmatrix} \dot{\gamma} \\ \dot{u} \\ \dot{\vec{R}} \end{bmatrix} = \begin{bmatrix} 0 & 1 & \vec{0} \\ \alpha_i & 0 & \vec{0} \\ -\vec{\eta}_i & \vec{0} & \vec{0} \end{bmatrix} \begin{bmatrix} \gamma \\ u \\ \vec{R} \end{bmatrix} + \begin{bmatrix} 0 \\ \vec{\eta}_i^{\,T}\vec{\lambda}+[V_{Mi}-V_M(t)]\vec{p}_{\gamma i}^{\,T}\vec{\lambda}_i-\gamma_i\alpha_i \\ -V_M(t)\vec{p}_i+\vec{V}_T(t)-\gamma_i\vec{\eta}_i \end{bmatrix}$$

i.e., on the i-th subinterval

$$\dot{\vec{y}} = A_i\vec{y} + \vec{b}_i(t) \ .$$

Let $\xi_i = \sqrt{|\alpha_i|}$. Let $cs_i = \cos(\xi_i\Delta t)$, $sn_i = \sin(\xi_i\Delta t)$, and $\delta_i = 1$ in the case of $\alpha_i \leq 0$. Let $cs_i = \cosh(\xi_i\Delta t)$, $sn_i = \sinh(\xi_i\Delta t)$, and $\delta_i = -1$ when $\alpha_i > 0$. Here we are assuming a constant interval size Δt. Let $\phi_{1i} = sn_i/\xi_i$, $\phi_{2i} = (1-cs_i)/\xi_i^2$, $\phi_{3i} = \delta_i(\xi_i\Delta t-sn_i)/\xi_i^3$, and $\phi_{4i} = (-1+cs_i+3\delta_i\xi_i\Delta t\,sn_i/2-\delta_i\xi_i^2\Delta t^2)/\xi_i^4$. When $\xi_i\Delta t$ is small, it is necessary to evaluate ϕ_{1i}, ϕ_{2i}, ϕ_{3i}, and ϕ_{4i} by means of series. For example, when $\xi_i\Delta t \doteq 0$, $cs_i \doteq 1$ and - in order to avoid subtraction of nearly equal numbers - one expands cs_i and evaluates ϕ_{2i} using the infinite series

$$\phi_{2i} = \Delta t^2\delta_i \sum_{k=0}^{\infty} \frac{1}{(2k+2)!}(-\delta_i)^k(\xi_i\Delta t)^{2k}$$

In this formulation it is assumed that, on each subinterval $V_M(t) = V_M(t_{i-1}) + [(t-t_{i-1})/\Delta t][V_M(t_i)-V_M(t_{i-1})]$ and similarly for $\vec{V}_T(t)$.

It can be shown (using Laplace transforms or matrix eigenvalues and eigenvectors) that

$$\begin{bmatrix} \gamma(t_i) \\ u(t_i) \\ \vec{R}(t_i) \end{bmatrix} = \begin{bmatrix} B_i^* & 0 \\ C_i^* & I \end{bmatrix} \begin{bmatrix} \gamma(t_{i-1}) \\ u(t_{i-1}) \\ \vec{R}(t_{i-1}) \end{bmatrix} + (\vec{\eta}_i^{\,T}\vec{\lambda})\vec{\psi}_i^{\,*} + \vec{\theta}_i^{\,*} \qquad (11)$$

where

$$B_i^\star = \begin{bmatrix} cs_i & \nu_{1i} \\ -\delta_i \zeta_i sn_i & cs_i \end{bmatrix}, \quad C_i^\star = \vec{\eta}_i[\nu_{1i}, \delta_i \nu_{2i}] \ ,$$

$$\vec{\psi}_i^{\,\star} = \begin{bmatrix} \delta_i \nu_{2i} \\ \nu_{1i} \\ \delta_i \nu_{3i} \vec{\eta}_i \end{bmatrix} , \quad \vec{\emptyset}_i^{\,\star} = \begin{bmatrix} \delta_i \beta_i \nu_{2i} - \delta_i \zeta_i \nu_{3i} \\ \delta_i \zeta_i \nu_{2i} - (\Delta t \zeta_i - \beta_i) \nu_{1i} \\ (\zeta_i \nu_{4i} - \delta_i \gamma_i \alpha_i \nu_{3i}) \vec{\eta}_i + \Delta t(\vec{\beta}_i^{\,\star} - \Delta t \vec{\zeta}_i^{\,\star}/2) \end{bmatrix}$$

where

$$\zeta_i = -\frac{1}{\Delta t}[V_M(t_i) - V_M(t_{i-1})] \vec{p}_{\gamma i}^{\,T} \vec{\lambda}_i \ , \quad \beta_i = -\frac{\Delta t}{2} \zeta_i - \gamma_i \alpha_i \ ,$$

$$\vec{\zeta}_i^{\,\star} = -\frac{1}{\Delta t}[V_M(t_i) - V_M(t_{i-1})] \vec{p}_i + \frac{1}{\Delta t}[\vec{V}_T(t_i) - \vec{V}_T(t_{i-1})] \ ,$$

$$\vec{\beta}_i^{\,\star} = \vec{V}_T(t_{i-1}) - V_M(t_{i-1}) \vec{p}_i - \gamma_i \vec{\eta}_i$$

We will let the M_i matrix of the general formulation be written in terms of 2x2 matrices as

$$M_i = \begin{bmatrix} M_i^\star & 0 \\ N_i^\star & I \end{bmatrix}$$

Let $M_N^\star = I$ and $N_N^\star = 0$. Therefore

$$M_{i-1}^\star = M_i^\star B_i^\star, \quad N_{i-1}^\star = N_i^\star B_i^\star + C_i^\star$$

for i = N,N-1,...,0. Then the approximate solution to the linearized equations (10) is

$$\begin{bmatrix} \gamma(t_F) \\ u(t_F) \\ \vec{F}(t_F) \end{bmatrix} = \begin{bmatrix} M_0^\star & 0 \\ N_0^\star & I \end{bmatrix} \begin{bmatrix} \gamma_0 \\ u(0) \\ \vec{R}_0 \end{bmatrix} + \sum_{i=1}^{N} \begin{bmatrix} M_i^\star & 0 \\ N_i^\star & I \end{bmatrix} (\vec{\psi}_i^{\,\star} \vec{\eta}_i^{\,T} \vec{\lambda} + \vec{\emptyset}_i^{\,\star}) \ .$$

In order to solve the two-point boundary-value problem, one must set $u(t_F) = 0$, $\vec{\lambda} = k\vec{R}(t_F)$, and must solve the above system *(of linear equations)* for u(0) and $\vec{R}(t_F)$. Then one calculates the values of γ, u, and $\vec{R}$ at other times using equations (11).

Here it has been assumed that t_F is given. However, in practice we have started with an initial approximation to t_F and, after the iteration with quasilinearization seems to have stabilized somewhat, we have corrected t_F after each iteration using the condition $g(t_F) = 0$, where $g = \dot{\vec{R}}(t_F) \cdot \vec{R}(t_F)$. We let

$$t_F^{(New)} = t_F^{(Old)} - g(t_F^{(Old)})/\dot{g}(t_F^{(Old)})$$

with $\dot{g}$ being computed only approximately.

No bounds have been placed upon u in the numerical examples to be presented. However, the procedure can be extended in order to include the case of bounded u. The subintervals may be adjusted (using the function u* of conditions (3) in a manner similar to that used to predict t_F from g) in order that u will not be bounded over only part of an interval. On each interval on which u is on a boundary, the *second* of equations (9) would be replaced by $\dot{u} = 0$. Equations (11), etc. would have to be modified accordingly.

V. A LINEAR METHOD

For the sake of comparison, we will develop a rather elaborate linear method for optimizing the air-to-air missile problem.

Let $\vec{\gamma}_s(t)$ be an initial approximation to a solution, where $\vec{\gamma}_s(t) = \vec{\gamma}_{si}(t)$ = constant on the i-th subinterval. The linearization of equations (1) on the i-th interval is

$$\dot{\vec{R}} = \vec{V}_T(t) - V_M(t)\vec{p}_s - V_M(t)\vec{p}_{\gamma s}^{\,T}\,\delta\vec{\gamma}$$

where i = 1,2,...,N, $\delta\vec{\gamma} = \vec{\gamma} - \vec{\gamma}_s$, $\vec{p}_s = \vec{p}(\vec{\gamma}_s)$, and $\vec{p}_{\gamma s} = \vec{p}_\gamma(\vec{\gamma}_s)$. The idea is to choose $\vec{u} = \delta\dot{\vec{\gamma}}$ which minimizes

$$J = \frac{1}{2} k|\vec{R}(t_F)|^2 + \frac{1}{2} \int_{0}^{t_F} (\dot{\vec{\gamma}}_s+\vec{u})^T W(t)(\dot{\vec{\gamma}}_s+\vec{u})\ dt\ .$$

A development similar to the derivation of the necessary conditions of optimality of the nonlinear problem gives the necessary conditions

$$\dot{\vec{B}} = kV_M\vec{P}_{\gamma s}^T\ \vec{R}(t_F)$$

$$\ddot{\vec{\gamma}} = -W^{-1}\ \vec{B}$$

$$\vec{B}(t_F) = \vec{0}$$

Clearly

$$\vec{B}(t) = k(\int_{t_F}^{t} V_M\vec{P}_{\gamma s}^T\ dt)\vec{R}(t_F)$$

$$\ddot{\vec{\gamma}}(t) = -kW^{-1}(\int_{t_F}^{t} V_M\vec{P}_{\gamma s}^T\ dt)\vec{R}(t_F)$$

$$\vec{\gamma}(t) = \vec{\gamma}_o + \int^{t}\ \dot{\vec{\gamma}}\ dt$$

$$\vec{\gamma}(t) = \vec{\gamma}_o + kC(t,t_F)\vec{R}(t_F)$$

where

$$C(t,t_F) = \int_{o}^{t} W^{-1} \int_{\tau}^{t_F} V_M\vec{P}_{\gamma s}^T\ dt\ d\tau\ .$$

Therefore

$$\vec{R}(t_F) = \vec{R}_o + \int_{o}^{t_F}(\vec{V}_T-V_M\vec{P}_s + V_M\vec{P}_{\gamma s}^T\vec{\gamma}_s)dt - \int_{t_o}^{t_F} V_M\vec{P}_{\gamma s}^T\ \vec{\gamma}\ dt$$

$$\vec{R}(t_F) = \vec{S} - k[\int_{t_o}^{t_F} V_M\vec{P}_{\gamma s}\ C(t,t_F)dt]\ \vec{R}(t_F)$$

where

$$\vec{S} = \vec{R}_o + \int_{t_o}^{t_F} [\vec{V}_T - V_M\vec{P}_s + V_M\vec{P}_{\gamma s}^T\ (\vec{\gamma}_s-\vec{\gamma}_o)]\ dt$$

so that

$$[I + k\int_{t_o}^{t_F} V_M\vec{P}_{\gamma s}\ C(t,t_F)dt]\ \vec{R}(t_F) = \vec{S}\ .$$

The latter linear equations must be solved for $\vec{R}(t_F)$ in order to obtain the desired solution. This is a three-dimensional formulation and is based on the assumption that t_F is given. Observe that V_M, $\vec{V}_T$, and W may be given functions of time.

VI. NUMERICAL RESULTS

This section contains numerical results for the two-dimensional encounters now to be described. At the initial time, $t_o = 0$, the missile is located at the origin and the target is 914.402 meters (3000 ft) away on the positive x-axis. The missile velocity magnitude is

$$V_M = 217.712\ t + 295.549 \quad \text{m/sec.}$$

There are no bounds upon $\dot{\gamma}$. The target makes a 9g turn. Its velocity magnitude is $|\bar{V}_T| = 295.352$ m/sec. Specifically,

$$V_{T1} = |\vec{V}_T| \cos(\dot{\omega}t), \quad V_{T2} = |\vec{V}_T| \sin(\dot{\omega}t)$$

where $\dot{\omega} = 17.1217^\circ/\text{sec}$. We will consider four cases, corresponding to $\gamma_o = -30^\circ, 0^\circ, 45^\circ, 90^\circ$. By trial it was found that the weignting factor $\kappa = .000129166$ rad^2-sec/m^2 gives reasonable results in all four cases. A starting value of $t_F = 2$ seconds was used in all cases.

No full simulation of the guidance methods has been carried out. It is realized that a final choice of method for a particular missile must depend upon the results of a complete simulation of a wide range of encounters.

The initial (starting) solution for both the linear and quasilinear guidance methods makes use of the knowledge that $\dot{\gamma}(t_F) = 0$ and that $\dot{\gamma}$ is monotone. It is assumed that $\ddot{\gamma} \equiv$ constant and that $\gamma(t_F)$ is approximately equal to the angle between the positive x-axis and a line connecting the origin to the position of the target at time t_F. This angle was taken to be 40° in all cases. The constant $\ddot{\gamma}$ is chosen such that $\gamma(t_F) = 40^\circ$.

In most cases this initial solution gives a value for $\gamma(t_o)$ which is rather close to that obtained from the converged quasilinearization method. In all four cases the latter method converged with $N = 15$ when it was started with the solution just described.

Figures 2 and 3 depict the initial solution and the converged quasilinearization for the cases of $\gamma_o = -30^o$ and $\gamma_o = 90^o$. The other two cases are similar. A value of $N = 15$ was employed. The final time t_F as well as $\gamma(t)$ converged.

Table 1 shows the effect of the number N of subintervals for the case of $\gamma_o = 0^o$ as an example. There is little change in $\gamma(0)$ and $\vec{R}(t_F)$ as N increases. However, fewer iterations are required as N increases. It has been found in other cases that there is a high risk of nonconvergence when $N < 15$. The iterations were carried out until $R_1(t_F)$ and $R_2(t_F)$ seemed to have converged within about one meter. Calculations, however, were carried out in terms of feet, radians, and seconds.

Table 1. Quasilinearization results for $\gamma_o = 0$

N		6	10	15	20
$\dot{\gamma}(0)$	(deg/sec)	41.6	42.9	43.6	43.6
$R_1(t_F)$	(meters)	-2.1	-2.5	-2.6	-3.0
$R_2(t_F)$	(meters)	5.5	5.5	5.8	5.6
No. Iterations		19	15	11	11

Table 2. $\dot{\gamma}(0)$ (deg/sec), comparison of linear and quasilinear methods

γ_o (deg)	-30^o	0^o	45^o	90^o
Linear guidance	54.5	36.3	-8.7	-48.2
Initial solution	70.0	40.0	-5.0	-50.0
Quasi. (5th iteration)	90.7	45.3	-10.1	-68.8
Quasi. (10th iteration)	78.2	43.7	-11.2	-56.1
Quasi. (converged)	72.1	43.6	-11.2	-51.3

In Table 2 the linear and quasilinear methods with N = 15 are compared for all four cases. The linear guidance method is started with the same solution as the quasilinearization method. It could be observed that the value $\dot{\gamma}(t_o)$ of the guidance command obtained from the linear method differ appreciably from that of the quasilinearization method which gives the optimal solution to the nonlinear translational equations of motion.

It was found that when k was increased to .00015, the quasilinearization algorithm did not usually converge. It can be seen that the value .0000129156 used for k led to miss-distances which may be a bit large for some purposes. If one wishes to decrease the miss-distance it is possible to increase k gradually during the iteration.

A constant subinterval size Δt was used during each iteration. However, since $\dot{\gamma}$ is much larger at the beginning of the flight than at the end, the number of intervals required can be reduced if the interval size is variable. As a rule of thumb, the subinterval should be sufficiently small that γ does not vary by more than 10^o over the interval.

An alternate method for computing t_F would be to solve for it simultaneously with $\dot{\gamma}(t_o)$ and $\vec{R}_F$, rather than treating t_F separately. This alternate way has not been studied.

REFERENCES

[1] Bellman, R. E. and Kalaba, R. E., *Quasilinearization and Nonlinear Boundary-Value Problems*, American Elsevier, New York, 1965.

[2] Andrus, J. F., "Modified Quasilinearization Methods," Rept. No. 69, Math. Dept., Univ. of New Orleans, 1978. (Revised 1979) To be submitted for publication in a journal.

[3] Andrus, J. F., "Nonlinear Guidance for Air-To-Air Missiles," 1977 USAF-ASEE Summer Faculty Research Reports (AFOSR-TR-78-0348), Vol. I., Rept. No. 7. Accession No. ADA051624.

[4] Axelband, E. I. and Hardy, F. W., "Quasi-Optimum Proportional Navigation," *IEEE Transactions on Automatic Control*, Vol. AC-15, No. 6, Dec. 1970, pp. 620-626.

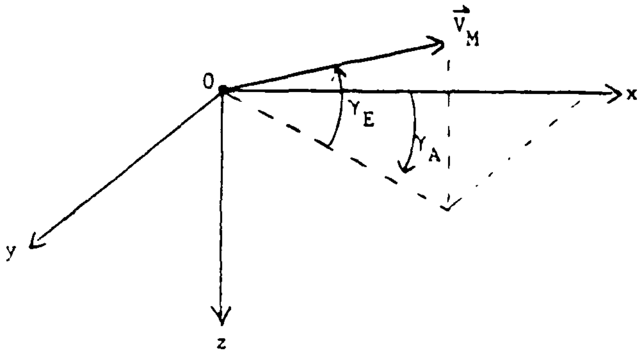


Fig. 1

Fig. 1 The coordinate system

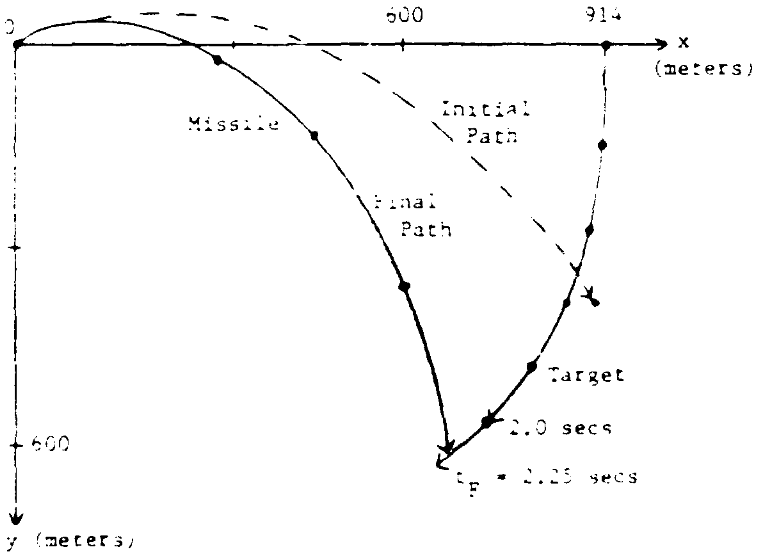


Fig. 2 Encounter with γ_o = -30°

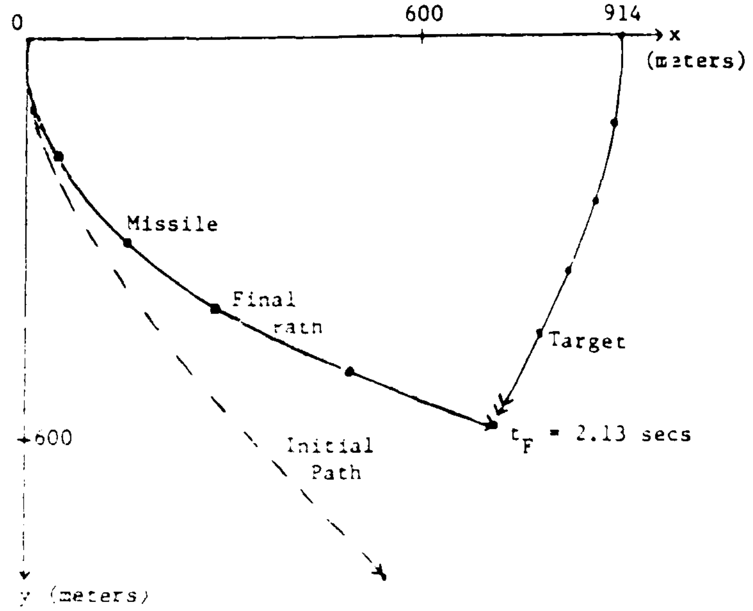


Fig. 3 Encounter with $\gamma_0 = 90^\circ$