MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 137
JANUARY 1977, PAGES 139–147

Identifying Differential Equations by Galerkin's Method

By Jack W. Mosevich*

Abstract. A numerical technique based on Galerkin's method is presented for computing unknown parameters or functions occurring in a differential equation whose solution is known. Under certain conditions a solution can be shown to exist to the integral equation formulation of this problem. It is also shown that the resulting nonlinear system is nonsingular.

1. Introduction. In most mathematical modeling problems differential equations of specific forms are derived which describe a system. Values of the coefficients, which can be constants or functions, of the differential equations are usually specified, and the solutions are calculated or presented in closed form, with little, if any, indication of how the coefficients can be estimated from observations. Clearly, this inverse problem is very interesting and important but somewhat difficult. The purpose of this paper is to describe a numerical technique for calculating unknown functions in a differential equation (or system) supposing its solution to be known. That is, we assume a function y is given whose derivative $\dot{y}$ is continuous on $[0, T]$ and such that y satisfies the differential equation

$$\dot{y}(t) = f(t, y(t), c_1(t), \ldots, c_m(t)), \qquad y(0) = y_0$$

on $[0, T]$. Our goal is to compute the unknown functions (or constants), $c_1, \ldots, c_m$. We may sometimes only be given $\{y_i\}_{i=1}^n$ where $y_i = y(t_i)$, $0 = t_0 < t_1 < \cdots < t_n = T$, such as in a case where $\{y_i\}$ is a set of observed data. Thus we do not assume $\dot{y}$ is known accurately. Also, the case of a system of differential equations should be admissible too. We stress that f must have a prescribed form to ensure a well-defined problem.

2. Some Preliminary Examples. To illustrate our method we consider the simple growth problem of a population in an unlimited environment. At time $t \geqslant 0$ the differential equation governing the number of entities present is

$$\dot{y}(t) = cy(t), \qquad y(0) = y_0, \qquad t \in [0, T],$$

where c is constant and unknown, but we assume that $y(t)$ is known on $[0, T]$.

The exact solution to the differential equation is, of course, $y(t) = y_0 e^{ct}$ which can certainly be solved for c; but this is rarely possible in general and the expression for c, $c = (\ln y - \ln y_0)/t$, requires a specific value for t which ignores many other

Received March 2, 1976.
AMS (MOS) subject classifications (1970). Primary 65D15.
Key words and phrases. Galerkin's method, differential equation.
*This work was supported by the National Research Council of Canada under Grant A8864.

known points. Another possibility is to solve the differential equation for $c = \dot{y}/y$ which is very poor since numerical differentiation is required, and we must again choose a specific t. A third possibility is to express the solution in integral form

$$y(t) = y_0 + c\int_0^t y(x)\,dx$$

and solve for c, again requiring a value of t for which $y(t)$ may be inaccurate. If, however, we integrate once more we obtain

$$\int_0^T (y(t) - y_0)\,dt = c\int_0^T \int_0^t y(x)\,dx\,dt$$

whereupon

$$c = \int_0^T (y(t) - y_0)\,dt \Big/ \int_0^T \int_0^t y(x)\,dx\,dt.$$

This is the basis of our method, which utilizes all values of y on $[0, T]$, requires no specific value of t and smooths the data in the process. On substituting $y = y_0 e^{ct}$ into the right-hand side, we find that this formula does give the correct result. Note that the double integral can be simplified to $\int_0^T (T - x)y(x)\,dx$ by reversing the order of integration.

A more complicated example is the case of two competing species where the Volterra-Lotka equations are usually used to describe the populations:

$$\begin{aligned} \dot{x} &= Ax - Bxy, \qquad x(0) = x_0, \\ \dot{y} &= -Cy + Dxy, \qquad y(0) = y_0. \end{aligned} \tag{1}$$

The positive constants A, B, C and D are the birth, death and mixing rates of the species. In [1] and [3] methods are described for computing these unknowns in case they are constants (in [1]) or functions (in [3]). These techniques are iterative methods, quite different from the one described here, which give best l_2 norm fits. Their methods do appear to work quite well but can be rather complicated to program. The presently described method is not iterative and appears to work well in addition to being relatively simple to code.

To solve for the constants A, B, C and D in (1) we write the solutions in integral form

$$\begin{aligned} x(t) &= x_0 + A\int_0^t x(s)\,ds - B\int_0^t x(s)y(s)\,ds, \\ y(t) &= y_0 - C\int_0^t x(s)\,ds + D\int_0^t x(s)y(s)\,ds. \end{aligned} \tag{2}$$

We shall now obtain four linear equations in the four unknowns by integrating (2) and repeating this after multiplying (2) through by a function ϕ which is a member of a basis for $C[0, T]$: The results are

$$\int_0^T (x(t) - x_0)\,dt = A\int_0^T \int_0^t x(s)\,ds\,dt - B\int_0^T \int_0^t x(s)y(s)\,ds\,dt, \tag{3a}$$

$$\int_0^T (y(t) - y_0)\,dt = -C\int_0^T \int_0^t y(s)\,ds\,dt + D\int_0^T \int_0^t x(s)y(s)\,ds\,dt, \tag{3b}$$

$$\text{(3c)}\quad \int_0^T \phi(t)(x(t)-x_0)dt = A\int_0^T \phi(t)\int_0^t x(s)\,ds\,dt - B\int_0^T \phi(t)\int_0^t x(s)y(s)\,ds\,dt,$$

$$\text{(3d)}\quad \int_0^T \phi(t)(y(t)-y_0)dt = -C\int_0^T \phi(t)\int_0^t y(s)\,ds\,dt + D\int_0^T \phi(t)\int_0^t x(s)y(s)\,ds\,dt.$$

The double integrals can be integrated once, for example

$$\int_0^T \phi(t)\int_0^t x(s)\,ds\,dt = \int_0^T x(s)[\Phi(T)-\Phi(s)]\,ds,$$

where $\Phi = \int\phi$. Note that (3) is really two sets (3a) and (3c), (3b) and (3d), of two equations in two unknowns.

The method just described gave excellent results with $\phi(t) = t$ for the example considered in [1]:

$$\text{(4)}\qquad \begin{aligned} \dot{x} &= x - xy, \quad & x(0) &= 1.2, \\ \dot{y} &= -y + xy, \quad & y(0) &= 1.1 \quad \text{so } A = B = C = D = 1. \end{aligned}$$

Solving (4) numerically by a Runge-Kutta routine and computing the integrals by quadrature with $T = 1$ resulted in

$$\int_0^T (x(t)-x_0)dt = .0910006, \qquad \int_0^T t(x(t)-x_0)dt = -.0626198,$$

$$\int_0^T (y(t)-y_0)dt = .0823068, \qquad \int_0^T t(y(t)-y_0)dt = .0519937,$$

$$\int_0^T\int_0^t x(s)\,ds\,dt = .5716192, \qquad \int_0^T t\int_0^t x(s)\,ds\,dt = .3783768,$$

$$\int_0^T\int_0^t y(s)\,ds\,dt = .5803130, \qquad \int_0^T t\int_0^t y(s)\,ds\,dt = .3890028,$$

$$\int_0^T\int_0^t x(s)y(s)\,ds\,dt = .6626198, \qquad \int_0^T t\int_0^t x(s)y(s)\,ds\,dt = .4409966;$$

with these values we find that the rearranged linear equations (3) are

$$\begin{aligned} A(.5716192) - B(.6626198) &= -.0910006, \\ A(.3783768) - B(.4409966) &= -.0626198, \\ -C(.5803130) + D(.6626198) &= \ .0823068, \\ -C(.3890028) + D(.4409966) &= \ .0519937, \end{aligned}$$

whose solutions are $A = B = C = D = 1$.

The integrations were performed to seven-place accuracy, and we see that A, B, C and D can be calculated to at least six places.

3. The General Method. In the above example we computed the moments of the solutions of the differential equations with respect to the functions $\phi_1(t) = 1$ and $\phi_2(t) = t$. Our general scheme for the case of not-necessarily constant unknowns is a natural extension of this concept known as Galerkin's method of undetermined coefficients for solving boundary value problems (see [2] or [6]). This method is based on the fact that in a Hilbert space H an element is zero if and only if it is orthogonal to every element of a basis of H, or in other words all its moments are

zero. In our case $H = L_2[0, T]$ and $\langle f, g\rangle = \int_0^T f(x)g(x)\,dx$.

We begin with the case of one unknown function c occurring on the right-hand side of a differential equation

$$\dot{y} = f(t, y(t), c(t)), \qquad y(0) = y_0 \quad \text{for } t \in [0, T]. \tag{5}$$

If f is linear in c, $f(t, y, c) = c(t)g(t, y) + h(t, y)$, then c is a coefficient in the usual sense; but for nonlinear f this is not the case.

Let $\{\phi_i\}$ be a basis for H, and suppose there exists an L_2 integrable c satisfying (5), so we have $c = \Sigma_{i=1}^{\infty}\alpha_i\phi_i(t)$. Our goal is to approximate c by the partial sum $c_n(t) = \Sigma_{i=1}^{n}\alpha_i\phi_i(t)$, and we hope that $c_n \longrightarrow c$ as $n \longrightarrow \infty$. In general, α_i depends on n; but we elect to keep the notation simple by not using superscripts.

To solve for α_i, $i = 1, \ldots, n$, we express the solution to (5) in integral form

$$y(t) - y_0 = \int_0^t f(s, y(s), c_n(s))\,ds,$$

multiply through by $\phi_k(t)$ in which case

$$\phi_k(t)(y(t) - y_0) = \phi_k(t)\int_0^t f(s, y(s), c_n(s))\,ds$$

and integrate to get the system of n nonlinear equations in the n unknowns α_i:

$$\int_0^T \phi_k(t)(y(t) - y_0)\,dt = \int_0^T \phi_k(t)\int_0^t f\left(s, y(s), \sum_{i=1}^{n}\alpha_i\phi_i(s)\right)ds\,dt, \qquad k = 1, 2, \ldots, n. \tag{6}$$

If f is linear in c, $f(t, y, c) = cg(t, y) + h(t, y)$, this becomes the linear system

$$\mathbf{Ax} = \mathbf{b} \quad \text{with } \mathbf{x} = (\alpha_1, \ldots, \alpha_n)^T, \qquad \mathbf{b} = (b_1, \ldots, b_n)^T$$

and $\mathbf{A} = (a_{kj})$, where

$$b_k = \int_0^T \phi_k(t)\left(y(t) - \int_0^t h(s, y(s))\,ds - y_0\right)dt,$$

$$a_{kj} = \int_0^T \phi_k(t)\int_0^t g(s, y(s))\phi_j(s)\,ds\,dt.$$

Note that the double integrals can be integrated once to yield

$$a_{kj} = \int_0^T \phi_j(s)g(s, y(s))(\Phi_k(T) - \Phi_k(s))\,ds,$$

where $\Phi_k = \int\phi_k$, with a similar expression for b_k.

It is evident from the Volterra-Lotka example that the problem of several unknown constants and several differential equations is very similar to (6), but with the sum $\Sigma\alpha_i\phi_i(s)$ on the right side replaced by the vector $(\alpha_1, \alpha_2, \ldots, \alpha_n)$. In the case of several unknown functions $c_1(t), \ldots, c_m(t)$ and one differential equation

$$\dot{y}(t) = f(t, y, c_1(t), \ldots, c_m(t))$$

we express each c_i as a sum

$$c_i^n = \sum_{k=1}^{n}\alpha_{ki}\phi_k(t)$$

using only $\phi_1, \ldots, \phi_n$ as coordinate functions. In order to obtain enough nonlinear equations in the α_{ki} we must require orthogonality to mn of the ϕ_i. For example,

$$\dot{y} = f(t, y, c_1, c_2, c_3) \quad \text{with } c_i^n = \sum_{k=1}^{n} \alpha_{ki}\phi_k(t)$$

will require $3n$ equations

$$\int_0^T \phi_k(t)(y(t) - y_0)\,dt = \int_0^T \phi_k(t)\int_0^t f\left(s, y, \sum_{j=1}^{n} \alpha_{j1}\phi_j, \sum_{j=1}^{n} \alpha_{j2}\phi_j, \sum_{j=1}^{n} \alpha_{j3}\phi_j\right) ds\,dt,$$

$k = 1, 2, \ldots, 3n$.

For systems of differential equations with several unknown functions c_i the details are more complicated and will not be exhibited here.

4. Existence Proofs. There are two basic existence problems which we must examine: the existence of a solution c to (5) and a solution $\mathbf{x} = (\alpha_1, \ldots, \alpha_n)^T$ to (6).

Writing (5) in integral form yields an implicit nonlinear integral equation in c similar to a Volterra equation of the first type:

$$y(t) = y_0 + \int_0^t f(s, y(s), c(s))\,ds. \tag{7}$$

This can be expressed as the operator equation

$$Ac = y \quad \text{where } Ac = \int_0^t f(s, y(s), c(s))\,ds,$$

and y has been transformed so as to satisfy the initial condition $y(0) = 0$. Since we assume that f is continuous, A is an operator on $L_2[0, T]$ into $L_2[0, T]$. The only topological existence proof which we could arrange is the following:

THEOREM 1. *If $f(t, y, c)$ is such that the operator $Pc = \lambda(Ac - y) + c(t)$ is completely continuous (for λ any nonzero real number), then (7) has at least one solution in $L_2[0, T]$.*

Proof. We first observe that P is defined on all of $L_2[0, T]$ and hence on any sphere $S \subset L_2[0, T]$. Also, with $\langle f, g\rangle = \int_0^T fg\,dt$, we have

$$\langle Pc, c\rangle = \langle \lambda(Ac - y), c\rangle + \langle c, c\rangle = \lambda\langle Ac - y, c\rangle + \langle c, c\rangle$$

so that choosing λ positive if $\langle Ac - y, c\rangle \geqslant 0$ or λ negative if $\langle Ac - y, c\rangle < 0$ gives $\langle Pc, c\rangle \leqslant \langle c, c\rangle$. By Theorem 1-27, p. 53 of [5], there exists a fixed point $c^*(P)$. Hence $c^* = Pc^* = \lambda(Ac^* - y) + c^*(t)$ so that $Ac^* - y = 0$ as required.

Since $\dot{y}$ is assumed to exist, we therefore have a c^* which satisfies (5) under the above assumption on f. Note that if f is of the linear form $f(t, y, c) = h(t, y)c + g(t, y)$ and if $h(t, y) \neq 0$ for all $t \in [0, T]$, then (5) can be solved for

$$c = (\dot{y}(t) - g(t, y))/h(t, y).$$

Another existence proof based directly on (5) is

THEOREM 2. *Suppose that $\dot{y}(t)$ is continuous on $[0, T]$, $y(0) = 0$ and that f is continuous on $[0, T] \times \text{Range}\, y \times [A, B]$. Suppose further that there exists a*

$c_0 \in [A, B]$ *such that* $f(0, 0, c_0) = \dot{y}_0, f_3(0, 0, c_0) \neq 0$. *Then there exists a continuous* $c = c(t)$ *in some neighborhood of* c_0 *such that* $\dot{y} = f(t, y(t), c(t))$.

Proof. Let $F(t, c) = f(t, y(t), c) - \dot{y}(t)$. Then by hypothesis $F(0, c_0) = 0$ and $F_2(0, c_0) = f_3'(0, 0, c_0) \neq 0$. By the Implicit Function Theorem there exists a neighborhood of c_0 and a continuous function $c = c(t)$ such that $F(t, c) = 0$, on this neighborhood of c_0.

The other question, whether or not (6) has a unique solution $(\alpha_1, \ldots, \alpha_n)$ if (5) does, can be partially answered for both linear and nonlinear f by the following:

THEOREM 3. *If there exists a unique solution* c *of* (5) *then for a suitable initial guess* $\alpha = (\alpha_1, \ldots, \alpha_n)$ *and subset* $\{\phi_i\}_{i=1}^n$ *of the basis* $\{\phi_i\}$, *the Jacobian matrix* $J(\alpha)$ *of* (6) *is nonsingular.*

Proof. Let α be chosen so that $f_3(s, y(s), \Sigma_{i=1}^n \alpha_i \phi_i \, ds)$ is not identically zero (for sufficiently large n). Such an α must exist by the assumption that (5) has a solution $c(t) = \Sigma_{i=1}^{\infty} \alpha_i \phi_i(t)$, for if $f_3 \equiv 0$ then f is independent of c. Let $J(\alpha)$ be the Jacobian matrix of (6):

$$J(\alpha) = \frac{\partial(G_1, G_2, \ldots, G_n)}{\partial(\alpha_1, \alpha_2, \ldots, \alpha_n)}$$

where $G_k(\alpha) = 0$ represents (6) with

$$G_k(\alpha) = \int_0^T \phi_k(t) \int_0^t f\left(s, y(s), \sum_{i=1}^n \alpha_i \phi_i(s)\right) ds\, dt - \int_0^T \phi_k(t)(y(t) - y_0)\, dt$$

and

$$\frac{\partial G_k}{\partial \alpha_i} = \int_0^T \phi_k(t) \int_0^t \phi_i(s) f_3\left(s, y, \sum_{i=1}^n \alpha_i \phi_i\right) ds\, dt.$$

If the columns of $J(\alpha)$ are dependent, then there exist $A_1, A_2, \ldots, A_n$, not all zero, such that

$$A_1 \left\langle \phi_k, \int_0^t \phi_1 f_3 \right\rangle + \cdots + A_n \left\langle \phi_k, \int_0^t \phi_n f_3 \right\rangle = 0, \qquad k = 1 \text{ to } n.$$

Thus

$$\left\langle \phi_k, \int_0^t \left(\sum_{i=1}^n A_i \phi_i(s) \right) f_3 \, ds \right\rangle = 0, \qquad k = 1 \text{ to } n.$$

If this system holds for all n, we have

$$\int_0^t \left(\sum_{i=1}^n A_i \phi_i \right) f_3 \, ds = 0$$

(where we can define $A_i = 0$ for $i >$ the original n) which gives

$$\left(\sum_{i=1}^n A_i \phi_i \right) f_3 = 0.$$

But since $f_3 \neq 0$, we must have $\Sigma A_i \phi_i = 0$ for not all $A_i = 0$, which is impossible since $\{\phi_i\}$ is an independent set. Thus the Jacobian is nonsingular. A similar argument holds for several unknowns $c_i(t)$, which can also be assumed constants.

6. Examples. The success of Galerkin's method in a particular problem depends mainly on the choice of the coordinate functions ϕ_i. In linear problems it is usual to choose an orthonormal basis, but it is not clear that such a choice is best in nonlinear cases where one desires orthogonality of the operator with the ϕ_i's. We found that the method outlined in this paper worked quite well in sample problems and would now like to indicate some interesting points with some examples.

The simplest problem of an unknown function $c(t)$ is the one of growth in an unlimited environment $\dot{y} = c(t)y(t), y(0) = y_0$. We decided to try a case where the generated numbers were quite varied so as to get some idea as to the conditioning of the linear system which results. We chose to generate data from the test problem $\dot{y} = (1 - t + t^2)y, y(0) = 1$, whose solution is $y(t) = e^{t-t^2/2+t^3/3}$ and $c(t) = 1 - t + t^2$ for $T = 1$ or 5. The coordinate functions taken were the Legendre polynomials orthogonal on $[0, T]$ whose weight function is $W(x) \equiv 1$.

The resulting linear system for $n = 6$ and $T = 1$ gave a solution good to six places when the integration was performed to 8. When T was increased to 5 the conditioning of the resulting system became poorer giving only three-place accuracy. The conditioning was imporved moderately by equilibrating the data to keep the calculated numbers relatively reasonable in magnitude; the solutions improved one decimal place in accuracy.

Another test case involved two unknowns $c_1(t)$ and $c_2(t)$ in the problem of growth in a limited environment

$$\dot{y}(t) = c_1(t)y(t) + c_2(t)y^2(t), \qquad y(0) = 9,$$

where the data was generated from $y(t) = 90e^{10t}/(1 + 9e^{10t})$ so that $c_1(t) \equiv 10$, $c_2(t) \equiv -1$. Thus we assumed c_1 and c_2 to be functions even though the data came from constants. The results using Legendre polynomials with $T = 3$ and $N = 3$ gave

$$c_1(t) = 10.00002 - .00034t + .000686t^2,$$

$$c_2(t) = -1.000002 + .000034t - .0000686t^2.$$

Notice the interesting fact that $c_1 = -10c_2$ as is the case in the exact solutions.

Another question is what happens when the data contain random errors. We tried the Volterra-Lotka equations with constant coefficients $A = B = C = D = 1$ as before but added a random number $r, -.01 \leqslant r \leqslant .01$, to $x(t)$ and $y(t)$ as they were calculated. The resulting linear systems are

$$7.810149589A - 8.922227075B = -1.10365313,$$

$$20.363751096A - 23.02170886B = -2.63189518,$$

$$-9.077398776C + 8.922227075D = -.16104135,$$

$$-23.939409757C + 23.02170886D = -.91548973,$$

whose solutions are good to three places, with the data good only to two places.

The systems without any noise are

$$7.81317483A - 8.91937981B = -1.10620497,$$

$$20.36959213A - 23.007586878B = -2.63799472,$$

$$-9.07969595C + 8.91937981D = -.160316128,$$

$$-23.92854735C + 23.007586878D = -.92096046,$$

whose solutions are accurate to six places. In this example $T = 5$.

5. A More General Class of Problems. In this paper, as well as in [1] and [3], it was assumed that the solution y of the differential equation was known. A new class of problems can be formulated where y is also unknown and where $c = c(y)$ is sought so that if y and c solve $\dot{y} = f(t, y, c)$, then c and/or y will satisfy certain conditions.

The simplest example of such a problem is: determine a vertical force $f(y)$ in the $x - y$ plane so that all projectiles starting from the origin with arbitrary nonzero initial velocities and nonvertical directions will, after 1 unit of time, be travelling horizontally. Mathematically, we seek a function c so that if y satisfies

$$\ddot{y} = c(y), \quad y(0) = 0, \quad \dot{y}(0) = y_0 \quad \text{(arbitrary)}$$

then $\dot{y}(1) = 0$ regardless of y_0. Note that $y(t)$ is unknown but we really desire only c.

To solve such a problem one must assume some form for $c(y)$, say $c(y) = \lambda y$. Under this assumption we can solve $\ddot{y} - \lambda y = 0$ and consider the cases $\lambda > 0$, $\lambda = 0$ and $\lambda < 0$. In this problem the condition $\dot{y}(1) = 0$ can only be met if $\lambda = -\lambda_n^2$ where $\lambda_n = \pi/2 + n\pi$. Thus, we have $c(y) = -\lambda_n^2 y$ and $y(t) = y_0 \sin(\lambda_n t)/\lambda_n$ for any fixed n.

A general class of first order problems of this type is as follows: Find $c = c(y)$ so that the functional equation $G(y_1, y_2, c(y)) = 0$ holds for all y_1 in some domain $[a, b]$ where y solves $\dot{y} = f(t, y, c(y))$ subject to $y(0) = y_1, y(1) = y_2$. For example, G could be

$$G(y_1, y_2, c(y)) = \int_{y_1}^{y_2} c(y)\, dy.$$

An example of a second order problem of this type first occurred in [4] where f was the differential equation of geodesics on a surface S and the unknown c was a directrix curve which generated S. The problem there was to determine c so that all geodesics starting from a fixed point on S were parallel by the time they reached the edge of S. In that case G was a focussing condition and Galerkin's method was used to compute an approximation of c.

Department of Mathematics
Mount Allison University
Sackville, New Brunswick, Canada

1. R. BELLMAN, H. KAGIWADA & R. KALABA, "Inverse problems in ecology," *J. Theoret. Biol.*, v. 11, 1966, pp. 164–167.

2. L. V. KANTOROVIČ & V. I. KRYLOV, *Approximate Methods of Higher Analysis*, 3rd ed., GITTL, Moscow, 1950; English transl., Interscience, New York, 1958. MR **13**, 77; **21** # 5268.

3. R. J. LERMIT, "Numerical methods for the identification of differential equations," *SIAM J. Numer. Anal.*, v. 12, 1975, pp. 488–500.

4. J. W. MOSEVICH, "Analytical and numerical approximations of a functional differential equation arising in the design of a microwave antenna," *Utilitas Math.*, v. 4, 1973, pp. 129–145.

5. T. L. SAATY, *Modern Nonlinear Equations*, McGraw-Hill, New York, 1967. MR **36** # 1249.

6. Yu. V. VOROBYEV, *Methods of Moments in Applied Mathematics*, Fizmatgiz, Moscow, 1958; English transl., Gordon and Breach, New York, 1965. MR **21** # 7591; **32** # 1872.