A SERIES OF CLASS NOTES FOR 2005-2006 TO INTRODUCE LINEAR AND NONLINEAR PROBLEMS TO ENGINEERS, SCIENTISTS, AND APPLIED MATHEMATICIANS

DE CLASS NOTES 1

A COLLECTION OF HANDOUTS ON
FIRST ORDER ORDINARY DIFFERENTIAL EQUATIONS (ODE's)

CHAPTER 7

An Introduction to Numerical Methods for

Linear and Nonlinear ODE's

1. An Introduction to Numerical Solutions for the IVP and Mapping Problems

2. Solution of IVP's: Euler's Method for a Scalar Equation

3. An Introduction to Error Analysis for the Mapping Problem

We consider the (possibly nonlinear) initial value problem:

$$\text{ODE} \quad \frac{d\vec{u}}{dt} = \vec{F}(t, \vec{u}) \tag{1}$$

IVP

$$\text{IC} \quad \vec{u}(0) = \vec{u}_0 \tag{2}$$

where $\vec{u}$ is the **state vector** containing all of the **state variables** for the system which vary with time t.

Time Invariant (Autonomous) (Possibly Nonlinear) System

We are particularly interested in the problem:

$$\text{ODE} \quad \frac{d\vec{u}}{dt} + T(\vec{u}) = \vec{b} + \vec{g}(t) \tag{3}$$

IVP

$$\text{IC} \quad \vec{u}(0) = \vec{u}_0 \tag{4}$$

where T maps a vector space V back into itself; that is, T:V→V where V is the (real or complex) vector space of all possible states of the system.

Steady State or Equilibrium Problem for the

Time Invariant (Autonomous) (Possibly Nonlinear) System

Assume that $\frac{d\vec{u}}{dt} = \vec{0}$ and that $\lim_{t\to\infty} \vec{g}(t) = \vec{0}$ (e.g., $\vec{g}(t) = \vec{0}$). Then we have $T(\vec{u}) = \vec{b}$.

We view this as a **mapping problem**; that is we wish to find those vectors $\vec{u}$ which T maps into $\vec{b}$. If we have a **well-posed problem**, then there is exactly one such vector. If this is true for all $\vec{b}$ then T is **one-to-one**. Then its inverse, T^{-1}, exists with domain the **range** of T.. Then $\vec{u} = T^{-1}(\vec{b})$. However, even when T^{-1} exists, it is rare that it is "computed".

(Possibly Nonlinear) Dynamical System

Now let $N[\vec{u}] = \frac{d\vec{u}}{dt} + T(\vec{u})$ so that N:V(t)→V(t) where V(t) is the vector space of time-varying vectors in V. That is, $V(t) = \{\vec{u}(t) : \vec{u}(t) : I \to V\}$ where I = (a,b) is an open interval in **R** to be

determined as the interval of validity of the solution $\vec{u}(t)$. Now let $D = \{ \vec{u}(t) \in V(t): \vec{u}(0) = \vec{u}_0 \}$ and N_0 be the **restriction** of N to D so that $N_0: D \to V(t)$. Viewed as a mapping problem, to solve $N_0[\vec{u}(t)] = \vec{g}(t)$, we simply wish to find all vectors $\vec{u}(t) \in V(t)$ that are mapped into $\vec{g}(t)$ by the operator N_0. If we have a **well-posed problem**, then there is exactly one such vector. If this is true for all $\vec{g}(t)$, then N_0 is **one-to-one** and its inverse, N_0^{-1}, exists and $\vec{u}(t) = N_0^{-1}(\vec{g}(t))$.

Although we can treat both the steady state and the dynamic (initial value) problem as mapping problems, numerically they are treated differently. If we wish to solve an IVP numerically we discretize both the state vector $\vec{u}$ and time t and solve using finite difference or finite element methods as a "marching" problem. That is, we find the time varying vector $\vec{u}(t)$ one step at a time, where as we find the vector that solves the steady state problem all at once.

Time Invariant (Autonomous) Linear Discrete System

$$\text{ODE} \quad \frac{d\vec{u}}{dt} + T(\vec{u}) = \vec{b} + \vec{g}(t) \tag{3}$$

IVP

$$\text{IC} \quad \vec{u}(0) = \vec{u}_0 \tag{4}$$

We now assume that Tis a **linear operator**. We also assume that there are a finite number of state variables so that we have a **discrete** system. If we started with a continuous problem, this could be the discretized problem. It could also be a **lumped parameter** system (e.g., circuits, springs, and trusses). Again T is a linear operator from a vector space V into itself; that is, $T: V \to V$ where V is the (real or complex) vector space of all possible states of the system. However, since we have now assumed that there are only a finite number of state variables, V is **finite dimensional** and we assume $V = \mathbf{R}^n$. Since T is linear, it has a matrix representation, say A. We change notation and use $\vec{x}$ as our state variable since now the operator T may be represented using matrix multiplication as

$$T(\vec{x}) = \underset{nxn}{A}\ \underset{nx1}{\vec{x}} \tag{5}$$

and we may rewrite (3) and (4) as

$$\text{ODE} \quad \frac{d\vec{x}}{dt} + \underset{nxn}{A}\ \underset{nx1}{\vec{x}} = \vec{b} + \vec{g}(t) \tag{6}$$

IVP

$$\text{IC} \quad \vec{x}(0) = \vec{x}_0 \tag{7}$$

Steady State or Equilibrium Problem for the Time Invariant (Autonomous) Linear System

Assume that $\frac{d\vec{x}}{dt} = \vec{0}$ and that $\lim_{t\to\infty} \vec{g}(t) = \vec{0}$ (e.g., $\vec{g}(t) = \vec{0}$). Then we have

$$\underset{nxn}{A}\ \underset{nx1}{\vec{x}} = \underset{nx1}{\vec{b}}\ . \tag{8}$$

Thus the problem indicates that we wish the inverse of the operator (matrix) A where $A:\mathbf{R}^n \to \mathbf{R}^n$. Again we view this problem as a mapping problem; that is we wish to find those vectors $\vec{x}$ x which T maps into $\vec{b}$ (i.e., those column vectors $\vec{x}$ that when multiplied by A result in the column vector $\vec{b}$). The use of Gauss Elimination (row reduction) to solve linear algebraic equations is discussed in Chapter 2-4. The purpose here is to introduce such problems as mapping problems.

Linear Dynamical Systems

Now assume T:V→V is linear ($L[\vec{u}] = \frac{d\vec{u}}{dt} + T(\vec{u})$ is also linear), let $\vec{u}_{ss}$ be the solution to the steady state problem and replace $\vec{u}$ with $\vec{u}_{ss} + \vec{u}_d$ so that $\vec{u}_d$ is the displacement from equilibrium. Since T is linear we have $T(\vec{u}_{ss} + \vec{u}_d) = T(\vec{u}_{ss}) + T(\vec{u}_d)$. Also $d\vec{u}/dt = d(\vec{u}_{ss} + \vec{u}_d)/dt$. Substituting into (3) and letting $T(\vec{u}_{ss}) - \vec{b} = 0$ we obtain.

$$\text{ODE} \quad \frac{d\vec{u}_d}{dt} = T(\vec{u}_d) + \vec{g}(t) \tag{3}$$

IVP

$$\text{IC} \quad \vec{u}_d(0) = \vec{u}_0 - \vec{u}_{ss} \tag{4}$$

Now let $L[\vec{u}] = \frac{d\vec{u}}{dt} + T(\vec{u})$ so that L:V(t)→V(t) where V(t) is the vector space of time-varying vectors in V. That is, $V(t) = \{\vec{u}(t) : \vec{u}(t) : I \to V\}$ where I = (a,b) is an open interval in **R** to be

determined as the interval of validity of the solution $\vec{u}(t)$. Now let $D = \{\vec{u}(t) \in V(t): \vec{u}(0) = \vec{u}_0\}$ and L_0 be the restriction of L to D so that $L_0:D \to V(t)$. Theoretically, to solve $L_0[\vec{u}(t)] = g(t)$, we simply wish to invert the operator L_0. However, numerically we would discretize and solve using **finite difference** or **finite element** methods and solve as a marching problem. An introduction to numerical techniques for first order systems is given in the next section by considering Euler's method for a scalar equation (only one state variable).

EULER'S METHOD FOR A SCALAR EQUATION

Recall the Initial Value Problem (IVP)

$$\text{IVP} \quad \begin{array}{ll} \text{ODE} & y' = f(x,y) \quad (1) \\ \text{IC} & y(x_0) = y_0. \quad (2) \end{array}$$

If f(x,y) is such that we can not solve this IVP, we can use numerical techniques. Although there are many numerical techniques, we consider a simple one known as Euler's method or the tangent line method. Since it is an IVP and since the independent variable is often time, we think in terms of starting at x_0 and determining the solution for $x > x_0$. Since we are proceeding numerically, we wish to find (approximations) for the solution at a finite number of points $x_1, x_2, \ldots x_{n-1}, x_n$, where $x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n$. Recall that $y' = f(x,y)$. We consider the approximation:

$$\frac{y(x_1) - y(x_0)}{x_1 - x_0} \approx f(x_0, y_0).$$

That is, we approximate the slope of the solution between the points x_0 and x_1 using two methods:

1) As the difference quotient $(y(x_1) - y(x_0))/(x_1 - x_0)$ and
2) By the value of f(x,y) at the initial point $(x_0, y(x_0))$.

To simply the notation, we let $y_k = y(x_k)$ for k = 0, 1, 2, ... , n. Hence we get

$$\frac{y_1 - y_0}{x_1 - x_0} \approx f(x_0, y_0).$$

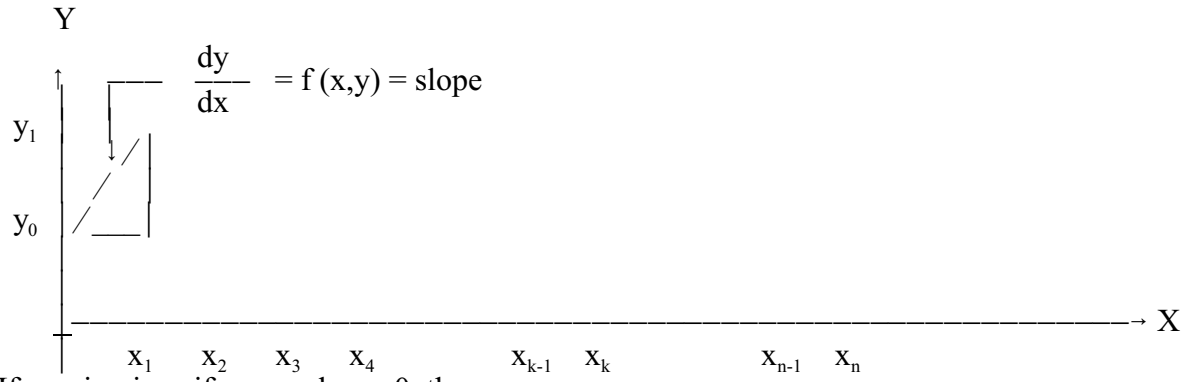
We use this approximation to define our numerical value of y_1 (see sketch on next page). Hence we obtain:

$$y_1 - y_0 = f(x_0, y_0)\,(x_1 - x_0)$$
$$y_1 = y_0 + f(x_0, y_0)\,(x_1 - x_0).$$

Repeating the process to obtain $y_2, y_3, \ldots, y_{k-1}, y_k, \ldots y_{n-1}, y_n$:

$$y_2 = y_1 + f(x_1, y_1)\,(x_2 - x_1)$$
$$y_3 = y_2 + f(x_2, y_2)\,(x_3 - x_2)$$
$$\vdots$$

$$y_{k-1} = y_{k-2} + f(x_{k-2}, y_{k-2})\,(x_{k-1} - x_{k-2})$$
$$y_k = y_{k-1} + f(x_{k-1}, y_{k-1})\,(x_k - x_{k-1})$$
$$\vdots$$
$$y_{n-1} = y_{n-2} + f(x_{n-2}, y_{n-2})\,(x_{n-1} - x_{n-2})$$
$$y_n = y_{n-1} + f(x_{n-1}, y_{n-1})\,(x_n - x_{n-1})$$



If spacing is uniform: and $x_0 = 0$, then

h 2h 3h 4h (k-1)h k h (n-1)h nh

←–h–→ ←–h–→ ←–h–→ ←–h–→

Hence we obtain the general formula:

$$y_k = y_{k-1} + f(x_{k-1}, y_{k-1})\,(x_k - x_{k-1}) \quad k = 1, 2, ..., n.$$

If the spacing is uniform: $h = \Delta x = x_k - x_{k-1}$, $k = 1, 2, ..., n$, then we obtain:

$$y_k = y_{k-1} + h\, f(x_{k-1}, y_{k-1}) \qquad k = 1, 2, ..., n$$

or

$$y_{k+1} = y_k + h\, f(x_k, y_k). \qquad k = 0, 1, 2, ..., n-1.$$

MEMORIZE THIS FORMULA AND BE ABLE TO USE IT.

EXAMPLE. Using Euler's Method with $h = 0.1$, find the first two iterates (i.e. y_1 and y_2) to obtain a numerical approximation of $y(0.2)$ if y is the solution to the Initial Value Problem (IVP)

IVP

ODE $y' = 2xy$

IC $y(0) = 1$

Be sure you 1) Give the general formula for Euler's method
2) Develop a table to display your results.

Solution.

General formula for Euler's method: $y_k = y_{k-1} + h\, f(x_{k-1}, y_{k-1})$. $k = 1, 2, ..., n$.

Table for $h = 0.1$

TABLE

n	x_n	y_n	$f_n = f(x_n, y_n) = 2\, x_n\, y_n$	$y_{n+1} = y_n + h\, f_n$
0	0	1	2 (0) (1) = 0	1 + (0.1) (0) = 1
1	0.1	1	2 (0.1) (1) = 0.2	1 + (0.1)(0.2) = 1.02
2	0.2	1.02		

"Clearly" the method extends directly to the **discrete system**

ODE $$\frac{d\vec{u}}{dt} = \vec{F}(t, \vec{u}) \tag{3}$$

IVP

IC $$\vec{u}(0) = \vec{u}_0 \tag{4}$$

where $\vec{u}$ has only a finite number of state variables.

EXERCISES on Numerical Solution of IVP'S: Euler's Method for a Scalar Equation

EXERCISE #1. Using Euler's Method with h = 0.1, find the first two iterates (i.e. y_1 and y_2) to obtain a numerical approximation of y(0.2) if y is the solution to the Initial Value Problem (IVP)

.

IVP
ODE y' = xy
IC y(0) = 1

Be sure you 1) Give the general formula for Euler's method
2) Develop a table to display your results.

EXERCISE #2. Using Euler's Method with h = 0.05, find the first four iterates (i.e. y_1 and y_2) to obtain a numerical approximation of y(0.2) if y is the solution to the Initial Value Problem (IVP)

.

IVP
ODE y' = xy
IC y(0) = 1

Be sure you 1) Give the general formula for Euler's method
2) Develop a table to display your results.

Let T:V→W be a mapping (i.e., **operator**) from a **vector space** V to another vector space W. (It is assumed that you have some knowledge of vectors. See Chapter 2-4 for the mathematical definition of a vector space. It is sufficient at this point to understand that a vector space is a set with structure. Examples are $\mathbf{R}^n$ and the function spaces C(**R**,**R**) and A(**R**,**R**).) We wish to solve the (vector) equation

$$T(\vec{u}) = \vec{b} \tag{1}$$

where $\vec{b}$ is a given vector in W. That is we wish to find all of the vectors $\vec{u}$ that map into $\vec{b}$. It is assumed that we know how to compute T($\vec{u}$) for any $\vec{u} \in$V. Thus we can check to see if any particular vector $\vec{u}$ is indeed a solution If T provides a **one-to-one correspondence** between V and W, then for every $\vec{b} \in$W there is exactly one vector $\vec{u} \in$V that maps into $\vec{b}$. Thus the **inverse operator** T^{-1} exists and the formal solution to the problem is $\vec{u} = T^{-1}(\vec{b})$. Knowing that T^{-1} exists proves the existence and uniqueness of the solution but may do little to help us compute $\vec{u}$ as we must first compute T^{-1}. Even when T^{-1} is known to exist, it is rare to actually compute it. Since we generally are only solving (1) for one (or a few) specific $\vec{b}$'s and not for all $\vec{b}$'s, we do not really need T^{-1}. Generally, computation of T^{-1} is not cost effective.

More importantly, if we are using a computer, we expect to find **approximate solutions** using **approximate arithmetic** rather than **exact solutions** using **exact arithmetic**. We assume that V and W are not just mathematical vector spaces but that all vectors have lengths or **norms**. That is, we assume that V and W are **normed linear spaces**. (Vector spaces are also called **linear spaces**.) If $||\vec{u}||$ is the length of a vector $\vec{u}$ in V, then the **metric** $\rho(\vec{u},\vec{v}) = ||\vec{u} - \vec{v}||$ gives a measure of the distance between $\vec{u}$ and $\vec{v}$. (In $\mathbf{R}^3$, $||\vec{u} - \vec{v}||$ is the distance between the ends of the **position vectors** $\vec{u}$ and $\vec{v}$.)

Now let $\vec{u}_a$ be an approximate solution to (1) and $\vec{u}_e$ be its exact solution (which we assume to exist). A measure of how good $\vec{u}_a$ is given by the norm of the **error vector**

$$\vec{E}_V = \vec{u}_e - \vec{u}_a . \tag{2}$$

Let

$$E_v = ||\vec{E}_V|| = ||\vec{u}_e - \vec{u}_a||. \tag{3}$$

Suppose T is invertible. Let $T(\vec{u}_a) = \vec{b}_a$ b_a so that $\vec{u}_a = T^{-1}(\vec{b}_a)$ and $\vec{u}_e = T^{-1}(\vec{b})$. Substituting into (3) we obtain

$$E_v = ||\vec{E}_V|| = ||T^{-1}(\vec{b}_a) - T^{-1}(\vec{b})||. \tag{4}$$

If T is a **linear operator**, then so is T^{-1} and we have

$$E_v = \| \vec{E}_V \| = \| T^{-1}(\vec{b}_a - \vec{b}) \|. \tag{5}$$

If T^{-1} is a **bounded linear operator**, then

$$E_v = \| \vec{E}_V \| = \| T^{-1}(\vec{b}_a - \vec{b}) \| \leq \| T^{-1} \| \, \| \vec{b}_a - \vec{b} \|. \tag{6}$$

If an a priori "estimate" (i.e., bound, but analysts use the term "estimate") of $\| T^{-1} \|$ can be obtained, then we see that an "estimate" of (i.e., bound for) E_v can be obtained by first computing

$$E_W = \| \vec{b}_a - \vec{b} \| = \| T(\vec{u}_a) - \vec{b} \| \tag{7}$$

Even without an estimate of $\| T^{-1} \|$, and indeed without T even being linear, we can still use E_W as a measure of the error in $\vec{u}_a$. Amazingly, this is possible even if (1) does not have a solution. That is, we can look for a **least error solution** that minimizes E_W. (In $\mathbf{R}^n$ with the **Euclidean norm**, this is called a **least squares solution**.) If (1) has one or more solutions, then these are all least error solutions since they all give $E_W = 0$. On the other hand (OTOH), choosing $\vec{u}_a$ to minimize E_W, gives , in some sense, the "best possible" solution to a problem with **no solution** even though it may or may not minimize E_V. Also viewing the solution of (1) as a **minimization problem** often has a physical interpretation (e.g., minimizing potential energy).