Lectures - Week 11
General First Order ODEs & Numerical Methods for IVPs

In general, nonlinear problems are much more difficult to solve than linear ones. Unfortunately many phenomena exhibit nonlinear behavior. We want to look at the general form of a first order ODE (which includes the case of linear ODEs in the previous lecture), state an existence/uniqueness theorem, and consider some numerical methods for approximating the solution. Some nonlinear homogeneous equations can be solved by separation of variables as we see in the following example.

Example Determine if separation of variables can be used to find the general solution of each homogeneous nonlinear IVP; if so find it.

$$y' - 2ty^2 = 0 \qquad y' + \ln(t^2 y) = 0$$

For the first equation we have

$$\frac{1}{y^2}dy = 2tdt \Rightarrow \int \frac{1}{y^2}dy = \int 2tdt \Rightarrow -\frac{1}{y} = t^2 + C \Rightarrow y = \frac{-1}{t^2 + C}$$

For the second equation we are unable to separate the variables.

Example Use your results in the previous example to solve the two IVPs:

$$y' - 2ty^2 = 0, \quad y(0) = 1 \qquad y' - 2ty^2 = 0, \quad y(0) = 0$$

From the previous example our solution is $y = \frac{-1}{t^2+C}$ so if $y(0) = 1$ then $1 = \frac{-1}{C}$ which implies $C = -1$ and the solution is $y = \frac{-1}{t^2-1}$. For the second IVP, $y(0) = 0$ implies $0 = \frac{-1}{C}$ for which there is no solution. If we look at the ODE we see that the solution is simply $y = 0$. What happened? We thought we solved for the general solution when we obtained $y = \frac{-1}{t^2+C}$ but we were unable to use this formula to obtain the solution to this IVP. The difficulty is that for nonlinear equations we are not guaranteed a general formula for *all* solutions as in the case of linear equations.

We now want to write a general first order ODE which will encompass both linear and nonlinear equations and then state existence/uniqueness theorems. The general form of a first order IVP is

$$y'(t) = f(t, y) \quad t > t_0 \quad y(t_0) = y_0 \tag{6}$$

Note that both linear and nonlinear equations fit into this abstract formulation. For example, our general linear equation

$$y'(t) + p(t)y(t) = g(t) \Rightarrow y'(t) = g(t) - p(t)y(t) = f(t, y)$$

where $f(t, y) = g(t) - p(t)y(t)$. Any existence result we postulate must agree with our previous result for linear equations. Previously we required $g(t)$ and $p(t)$ to be continuous

but now our condition will be on $f(t, y)$ which is a function of two independent variables t, y. Recall from calculus that when we differentiate a function of more than one independent variable we have to specify which variable we are differentiating with respect to; we use the terminology "partial derivative" to indicate a derivative with respect to one of the independent variables. If our function $g(x, y)$ depends on two independent variables we use the notation $g_x(x, y)$ to denote the first partial derivative with respect to x and $g_y(x, y)$ to denote the first partial derivative with respect to y. Leibniz notation would be to replace dg with ∂g, i.e., $\frac{\partial g}{\partial x}$ or $\frac{\partial g}{\partial y}$. Partial derivatives in the directions of the coordinate axes are easy to determine; if you want to differentiate with respect to say x, then you hold all other variables fixed and differentiate in the usual fashion.

Example Find f_x, f_{yy}, f_{xy} if $f(x, y) = x^3 \sin y$.

To calculate f_x we simply hold y fixed, i.e., treat it the same way we treat constants. So $f_x = 3x^2 \sin y$. To find the second partial of f with respect to y we first find f_y and then take the partial of f_y with respect to y. We have

$$f_y = x^3 \cos y \Rightarrow f_{yy} = -x^3 \sin y$$

For f_{xy} we first take f_x and then take the partial of f_x with respect to y. We have

$$f_x = 3x^2 \sin y \Rightarrow f_{xy} = 3x^2 \cos y$$

An important vector space (or linear space) is the space of all continuous functions on some domain Ω, which we denote by $C^0(\Omega)$ or sometimes just $C(\Omega)$. For example if $\Omega = (-1, 1)$, we want to consider all possible continuous functions defined on this interval with the usual definitions of addition and scalar multiplication. There is a huge difference in this space and, e.g., the space $\mathbf{R}^n$. The vector space $\mathbf{R}^n$ has dimension n because the number of elements in its basis is n. Thus $\mathbf{R}^n$ is a finite dimensional space. However, we can NOT find a basis for $C^0(\Omega)$ and so it is an *infinite dimensional* vector or linear space. Because we are studying differential equations we are interested in the differentiability of the unknown. The infinite dimensional vector space of all functions which are continuous and have one continuous derivative on a domain Ω is denoted $C^1(\Omega)$. In general, $y(t) \in C^k(\Omega)$ means that y has k continuous derivatives defined in Ω. If we say $y(t) \in C^\infty(\Omega)$ we mean that it is infinitely differentiable. So for the solution of our IVP we seek a function y in at least C^1 which satisfies (6).

Example Decide which space $C^k(\Omega)$ each of the functions are in.

$$y(t) = \sin t \quad \Omega = (-\pi, \pi) \qquad w(t) = t^{3/2} \quad \Omega = [0, 1]$$

The function $y(t)$ is continuous and has its first derivative as $\cos t$ which is also continuous. Differentiating again we get $y''(t) = -\sin t$ which again is continuous. Thus we see that $y(t)$ is infinitely differentiable and thus $y(t) \in C^\infty(\Omega)$.

The function $w(t)$ is continuous on $[0,1]$. Its first derivative is $w'(t) = 1.5\sqrt{t}$ which is also continuous on $[0,1]$. However, $w''(t) = (1.5/2)t^{-1/2}$ is not continuous on $[0,1]$ and thus $w(t) \in C^1(\Omega)$ and $w(t)$ is not in C^2.

We can now state our first existence theorem for a general first order linear ODE. It is a *localization theorem* in the sense that it guarantees a solution in the neighborhood of the starting point (t_0, y_0).

Theorem Let R be a rectangle defined by

$$R: \qquad |t - t_0| \leq a \qquad |y(t) - y_0| \leq b\,.$$

If f and f_y are continuous in R (i.e., $f, f_y \in C^0(R)$) then there exists some interval $(t_0 - \delta, t_0 + \delta)$, $\delta < a$ where the first order ODE (6) has a unique solution.

The first thing we should look at is whether this result reduces to our theorem for the first order linear equation $y' + p(t)y = g(t)$ which required that $p(t), g(t)$ be continuous. Writing it in the form $y'(t) = f(t,y)$ we see that $f(t,y) = g(t) - p(t)y$. Now in this theorem we require f and f_y to be continuous and for our linear case $f_y = p(t)$ so we are really requiring f and p to be continuous for the linear case. Because $f = g(t) - p(t)y$ is continuous we know that $g(t)$ must be continuous.

This result requires $f_y \in C^0$ and thus requires that f_y exists. This is actually a stronger condition than is needed. We will look at a concept called *Lipschitz continuity* and see its relationship to continuity and differentiability. Recall from calculus that there is connection between continuity and differentiability. If a function is continuous in an interval, does this imply differentiability there or does the existence of a derivative everywhere in an interval imply continuity. To recall this relationship remember that the classic example of a function that is not differentiable at $x = 0$ is $f(x) = |x|$. This function is clearly continuous everywhere but its derivative does not exist at $x = 0$ so continuity does not imply differentiability. The function $f(x) = |x|$ is not differentiable at $x = 0$ because

$$\lim_{x \to 0^+} \frac{f(x) - f(0)}{x} \neq \lim_{x \to 0^-} \frac{f(x) - f(0)}{x}$$

i.e., the slope of the secant line to the right of $x = 0$ is +1 and to the left it is -1. Here the notation $\lim_{x \to 0^+}$ means that x is approaching 0 through values slightly larger than zero, i.e., from the right. In calculus the following theorem was presented that tells us that differentiability of a function implies continuity.

Theorem Suppose that the function $f(x)$ is defined in a neighborhood of $x = a$. If f is differentiable at $x = a$, then f is continuous at $x = a$. If f is differentiable at each point of an interval I then it is continuous on I.

We now want to look at a condition called *Lipschitz continuity* which is stronger than continuity but weaker than differentiability. We will state this for a function of one independent variable.

Definition A function $f(x)$ is called Lipschitz continuous on I if there exists a real constant $L > 0$ such that, for all $x_1, x_2 \in I$

$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2|\,.$$

The smallest such L is often called the "best" Lipschitz constant of the function $f(x)$ on I.

Lets look at what this says graphically. Let $I = [a, b]$ and setting $x_1 = x$, $x_2 = a$ in the definition gives

$$|f(x) - f(a)| \leq L|x - a| \Rightarrow -L|x - a| \leq f(x) - f(a) \leq L|x - a|$$

which implies

$$f(a) - L|x - a| \leq f(x) \leq f(a) + L|x - a|$$

This says that the graph of the function $f(x)$ has to lie between the lines $y = f(a) - L|x - a|$ and $y = f(a) + L|x - a|$. So the maximum the derivative of f (if it exists) can be is L. This tells us that if f is Lipschitz continuous then it can't be too steep. If $f(x)$ is differentiable, then we can use the maximum value of its derivative to serve as a Lipschitz constant.
To extend this definition to a function of several variables, i.e., $f(x_1, x_2, \ldots, x_n)$ we use norms instead of absolute values. Let $\vec{x} = (x_1, x_2, \ldots, x_n)^T$. The function $f(x_1, x_2, \ldots, x_n)$ is Lipschitz continuous with respect to the norm $\|\cdot\|$ on a domain $D \subset \mathbf{R}^n$ if for all $\vec{x}_1, \vec{x}_2 \in D$

$$\|f(\vec{x}_1) - f(\vec{x}_2)\| \leq L\|\vec{x}_1 - \vec{x}_2\|\,.$$

Example The function $f(x) = x^2$ is Lipschitz continuous on $[1, 4]$ with Lipschitz constant $L = 8$ because $f'(x) = 2x$ and its maximum value on $[1, 4]$ is 8. The function $f(x) = \sqrt{x}$ is continuous on $[0, 1]$ but it is not Lipschitz continuous there because its derivative is unbounded as $x \to 0$ since $\sqrt{x}$ becomes arbitrarily steep near 0. So $f(x) = \sqrt{x}$ is an example of a function which is continuous on $[0, 1]$ but is not Lipschitz continuous there.

Example This example demonstrates that Lipschitz continuous does not imply differentiability. Consider $f(x) = |x|$ which we know is continuous on $[-1, 1]$. We want to demonstrate that it is also Lipschitz continuous on $[-1, 1]$ but we know it is not differentiable at $x = 0$.

We must find L such that for any $x, y \in [-1, 1]$

$$\Big||y| - |x|\Big| \leq L|x - y|$$

Clearly $L = 1$ works for all points in $[-1, 1]$ so it is Lipschitz continuous but not differentiable in $[-1, 1]$.

We now want to state an existence theorem for our first order IVP which does not require differentiability of $f(x, y)$. Here $f(x, y)$ is a function of two variables. The requirements are

continuity of $f(x, y)$ in both variables and Lipschitz continuous in the y variable. Note that before we required continuity plus f_y continuous which means we required differentiability of f with respect to y which is stronger than just requiring Lipschitz continuity in y.

Theorem Let R be a rectangle defined by

$$R: \qquad |t - t_0| \leq a \qquad |y(t) - y_0| \leq b\,.$$

If $f \in C^{)}(R)$ and f is Lipschitz continuous on R, i.e., there exists a constant L such that

$$f(x, y_1) - f(x, y_2)| \leq L|y_1 - y_2| \quad \forall (x, y_1), (x, y_2) \in R$$

then there exists some interval $(t_0 - \delta, t_0 + \delta)$, $\delta < a$ where the first order ODE (6) has a unique solution.

Numerical Methods for Solving First Order IVPs

As we have seen, ODEs, especially nonlinear ones, can be difficult, if not impossible, to obtain an analytic solution; in addition, when we do find the solution it may be in terms of an integral of a function which we can not evaluate. For these reasons we seek methods for approximating their solution. One of the most useful techniques is the use of a Taylor series for approximation which is just a series expansion for a function. Recall the following two theorems from calculus; the first is the general formula for Taylor series and the second is Taylor series with remainder.

Theorem Let $y(t) \in C^\infty(\Omega)$ where $\Omega = (t - \Delta t, t + \Delta t)$. Then

$$y(t + \Delta t) = y(t) + \Delta t y'(t) + \frac{\Delta t^2}{2!} y''(t) + \frac{\Delta t^3}{3!} y'''(t) + \cdots + \frac{\Delta t^n}{n!} y^{[n]}(t) + \cdots$$

Note that this is an infinite series and if Δt is "small" then the terms are decreasing in size. If we truncate this series then we have an approximation to y in the neighborhood of t.

Theorem Let $y(t) \in C^\infty(\Omega)$ where $\Omega = (t - \Delta t, t + \Delta t)$. Then

$$y(t+\Delta t) = y(t)+\Delta t y'(t)+\frac{\Delta t^2}{2!} y''(t)+\frac{\Delta t^3}{3!} y'''(t)+\cdots+\frac{\Delta t^{n+1}}{n!} y^{[n+1]}(\eta) \qquad \eta \in (t-\Delta t, t+\Delta t)$$

This last expression for y in the neighborhood of t is no longer an infinite series. How can it be possible that we have represented $y(t + \Delta t)$ as an infinite series and then expressed it exactly in a finite number of terms? The reason is that we don't know the point η, just that one exists. This remainder theorem will provide us with a bound for the error if we truncate the series using only the first n terms.

So a simple way to approximate the solution of our ODE is to replace $y'(t)$ with an approximation using a fixed number of terms in the Taylor series. For example, keeping only the first two terms we have

$$y(t+\Delta t) \approx y(t) + y'(t)\Delta t \Rightarrow y'(t) \approx \frac{y(t+\Delta t) - y(t)}{\Delta t}$$

This approximation should already be familiar to you because we know that from the definition of derivative

$$y'(t) = \lim_{\Delta t \to 0} \frac{y(t+\Delta t) - y(t)}{\Delta t}$$

and of course it is just the slope of the secant line joining the points $(t, y(t))$ and $(t + \Delta t, y(t+\Delta t)$.

The basic idea of approximating the solution to our IVP is to use the initial condition as a starting point, then approximate the solution at $y(t+\Delta t)$ and then use this to approximate the solution at $y(t+2\Delta t)$ and continue in this way until we get an approximation for $y(T)$. We call this "marching in time". Using this approximation to $y'(t)$ leads us to the well known (forward) Euler Method. To make it clear which is our approximation and which is the exact solution we will denote our approximation at $t_n = (t_0 + n\Delta t)$ to the exact solution $y(t^n)$ as Y^n. All we do is replace the derivative in the exact equation by its approximation to obtain the *difference equation*

$$y'(t_n) \approx \frac{Y^{n+1} - Y^n}{\Delta t} = f(t_n, Y^n)\,.$$

It is called a difference equation because we have replaced the differential equation with differences in function values to approximate the derivative. It is important to realize that we are approximating our ODE at $t = t_n$ and looking forward in time to approximate the derivative.

Forward Euler Method for the IVP (6).

Given $Y^0 = y(t_0)$.

For $n = 0, 1, 2,$

$$Y^{n+1} = Y^n + \Delta t f(t^n, Y^n)$$

Example Use Forward Euler to approximate the solution of the IVP

$$y'(t) + y(t) = 4, \quad y(0) = 1$$

using $\Delta t = .1$. Obtain the approximation at $T = 0.2$ and compare with the exact solution. We have $Y^0 = 1$. Using our difference quotient to approximate the derivative at t_0 gives

$$\frac{Y^1 - Y^0}{.1} + Y^0 = 4 \Rightarrow Y^1 = (0.1)4 + Y^0 - .1Y^0 = .4 + 1 - .1 = 1.3$$

Now to find Y^2 we use our difference quotient to approximate the derivative at t_1

$$\frac{Y^2 - Y^1}{.1} + Y^1 = 4 \Rightarrow Y^2 = (0.1)4 + Y^1 - .1Y^1 = .4 + 1.3 - .13 = 1.57$$

The ODE is a first order linear inhomogeneous equations so the exact solution is found by using an integrating factor which in this case is $e^{\int\ dt} = e^t$.

$$e^t y'(t) + e^t y(t) = 4e^t \Rightarrow \int d(e^t y) = \int 4e^t\ dt \Rightarrow e^t y = 4e^t + C \Rightarrow y(t) = 4 + Ce^{-t}$$

and satisfying the initial condition $y(0) = 1$ yields $C = -3$ and thus $y(t) = 4 - 3e^{-t}$. Now $y(.2) = 4 - 3e^{-.2} = 1.5438$ so our actual error is 0.0262.

Example Use Forward Euler to approximate the solution of the nonlinear IVP

$$y'(t) + 2ty^2(t) = 0, \quad y(0) = 1$$

using $\Delta t = .1$. Obtain the approximation at $T = 0.2$ and compare with the exact solution. Using our difference quotient to approximate the derivative at t_0 gives

$$\frac{Y^1 - Y^0}{.1} + 2t_0\left[Y^0\right]^2 = 0 \Rightarrow Y^1 = Y^0 - .2(0)(1) = 1$$

Using our difference quotient to approximate the derivative at t_1 gives

$$\frac{Y^2 - Y^1}{.1} + 2t_1\left[Y^1\right]^2 = 0 \Rightarrow Y^2 = Y^1 - .2(.1)(1)^2 = 1 - .02 = 0.98$$

The equation is separable so its exact solution can be found by

$$\frac{dy}{dt}\frac{1}{y^2} = -2t \Rightarrow \int \frac{dy}{y^2} = -2\int t\ dt$$

and thus

$$-y^{-1} = -t^2 + C \Rightarrow y(t) = \frac{1}{C + t^2}$$

Satisfying $y(0) = 1$ gives $1 = \frac{1}{C}$ which implies $y(t) = \frac{1}{1+t^2}$. Thus the exact solution at $t = .2$ is $y(.2) = \frac{1}{1+.04} = 0.9615$ and the actual error is 0.0185.

Alternate derivation of Euler's Method

We have seen that the Forward Euler Method can be derived by using Taylor series. If we include more terms in the series we can get a better approximation to $y'(t)$ but at the price of having to provide derivatives of $f(t, y)$ which can become complicated in addition to requiring more smoothness on $f(t, y)$ then what we need. For these reasons, higher order Taylor series methods (such as the one you derived in your homework) are not

popular. There is an alternate approach to deriving Euler's method which works for other methods as well. The approach uses a rule for numerical integration (also called numerical quadrature) instead of Taylor series

.
If we integrate our IVP given in (6) from t_n to t_{n+1} we obtain

$$\int_{t_n}^{t_{n+1}} y'(t)\, dt = \int_{t_n}^{t_{n+1}} f(t,y)\, dt$$

Now the left hand side can be integrated exactly to give

$$\int_{t_n}^{t_{n+1}} y'(t)\, dt = y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t,y)\, dt \tag{7}$$

Unless $f(t,y)$ is a function of t only, we won't be able to integrate it but we can approximate the integral using a numerical integration rule. If you recall from calculus when integrals were first introduced to calculate the area under a curve, one started by approximating the area by summing up areas of rectangles and the integral was defined as the limit of these areas as the number of rectangles approached infinity. This procedure used Riemann sums to approximate the integral. For example, if you want to approximate $\int_a^b g(t)\, dt$ you divide the interval $[a,b]$ into n subintervals of length Δx. These subintervals form the base for each rectangle. The height of the rectangle is determined by evaluating g at some point in the subinterval; for example at the left endpoint of each interval for a left Riemann sum or the right endpoint of each interval for a right Riemann sum. Another method introduced in calculus is the Midpoint Rule, where you evaluate g at the midpoint of each subinterval.

If we use a left Riemann sum to approximate the integral then we obtain the Forward Euler Method. For the interval $[t_n, t_{n+1}]$ we evaluate $f(t,y)$ at the left endpoint t_n to get

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t,y)\, dt \approx f(t_n, y(t_n))\Delta t \Rightarrow y(t_{n+1}) \approx y(t_n) + \Delta t f(t_n, y(t_n))$$

This is the same as our Euler's method when we derived it by approximating $y'(t_n)$ by $\big(y(t_{n+1}) - y(t_n)\big)/\Delta t$, i.e.,

$$\big(y(t_{n+1}) - y(t_n)\big)/\Delta t \approx f(t_n, y(t_n)) \Rightarrow y(t_{n+1}) \approx y(t_n) + \Delta t f(t_n, y(t_n))$$

Both derivations lead to the Forward Euler scheme

$$Y^{n+1} = Y^n + \Delta t f\big(t_n, Y^n\big)$$

This approach gives us a better way to derive schemes because it does not require repeated differentiation of $f(t,y)$ which the higher order Taylor series required.

If we approximate the right hand side of equation (7) by using the midpoint rule, then we are lead to another method which we will appropriately call the Midpoint Rule for approximating our IVP. Approximating the integral using the midpoint rule gives

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t,y)\,dt \Rightarrow y(t_{n+1}) \approx y(t_n) + \Delta t f\big(t_n + \frac{\Delta t}{2}, y(t_n + \frac{\Delta t}{2})\big)$$

We want to write our discretization of this approximation but there is a small problem. We can evaluate f at $t_n + \frac{\Delta t}{2}$ but what do we use for approximating $y(t_n + \frac{\Delta t}{2})$ because we only have our approximation Y^n to $y(t_n)$. One thing we could do is take an Euler step of length $\Delta t/2$ starting at t_n, i.e., with Y^n. This would give us an approximation to $Y^{n+1/2}$. In particular we would have

$$y(t_n + \frac{\Delta t}{2}) \approx Y^n + \frac{\Delta t}{2} f(t_n, Y^n))$$

Using this in our scheme gives the Midpoint rule for our IVP (6)

$$Y^{n+1} = Y^n + \Delta t f\big(t_n + \frac{\Delta t}{2}, Y^n + \frac{\Delta t}{2} f(t_n, Y^n)\big)$$

This is sometimes written as

$$\begin{aligned} k_1 &= \Delta t f(t_n, Y^n) \\ k_2 &= \Delta t f\Big(t_n + \frac{\Delta t}{2}, Y^n + \frac{1}{2}k_1\Big) \\ Y^{n+1} &= Y^n + k_2 \end{aligned}$$

to clearly demonstrate how many function evaluations (i.e., evaluations of f) are required.

In calculus you probably learned other methods for approximating integrals such as the Trapezoidal Rule and Simpson's rule. These can be used to obtain other schemes. Basically there are two general groups of methods for approximating the solution to our IVP (6). These are motivated by requiring more accuracy than Euler's method can produce. The first group of methods is called *single-step* methods for which Euler's and the Midpoint Rule are examples. Notice that in the Midpoint Rule we are doing an extra function evaluation in the interval $[t_n, t_{n+1}]$ to improve our accuracy; if we perform two extra function evaluations in $[t_n, t_{n+1}]$ we would expect to get a better approximation than the Midpoint Rule. Single step methods use the previous solution Y^n and intermediate steps in the interval $[t_n, t_{n+1}]$ to improve our accuracy. Another approach would be to use the solution that we obtained at previous points, e.g., Y^{n-2}, Y^{n-1} as well as Y^n to approximate the solution at Y^{n+1}. These are called *multistep methods*. The derivation of all methods are motivated by finding a scheme which has better accuracy than Euler's method. Before looking at these schemes, we want to investigate the error in Euler's Method.

Local truncation errors and global errors

The *global error* in our approximation to the IVP at time t_n is simply

$$|y(t_n) - Y^n|.$$

We want to investigate what contributes to this error so we can derive schemes to control it. First of all we have a contribution from approximating the derivative $y'(t)$ by a difference quotient if we use Taylor series. If we approximate the right hand side of equation (7) by a numerical integration scheme such as the midpoint rule, we are making an error. Is this the only error that is contributing to the global error, neglecting roundoff error? The answer is no; we also have a propagating error which must be controlled. When we take our first step we are starting with the exact initial condition $Y^0 = y_0$ and the only error we make in approximating Y^1 is due to the particular discretization we use whether it is a difference quotient or numerical integration scheme. However when we approximate Y^2 we have two sources of error – one from the discretization and another due to the fact that we are using Y^1 and it is not the same as $y(t_1)$; thus this error will be propagated through our approximation.

We call the error that is made in taking one step of our scheme starting with the exact value $y(t_n)$ (instead of Y^n) the *local truncation error*. This should be straightforward to quantify because it is due to our choice of discretization. We will look at this local error for Euler's Method and then use this to get a bound for the global error. We want to look at the error between $y(t_{n+1})$ and our approximation found by starting with $y(t_n)$, i.e.,

$$\hat{Y}^{n+1} = y(t_n) + \Delta t f(t_n, y(t_n))$$

Here we used the notation $\hat{Y}^{n+1}$ to make it clear that it is not the usual approximation by Euler's scheme. The local truncation error is given by $|y(t_{n+1}) - \hat{Y}^{n+1}|$. Expanding $y(t_{n+1})$ in terms of a Taylor series gives

$$y(t_{n+1}) - \hat{Y}^{n+1} = y(t_n) + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(t_n) + \mathcal{O}(\Delta t^3) - \hat{Y}^{n+1}$$

Now using the expression for $\hat{Y}^{n+1}$ we have

$$y(t_{n+1}) - \hat{Y}^{n+1} = y(t_n) + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(t_n) + \mathcal{O}(\Delta t^3) - \Big[y(t_n) + \Delta t f(t_n, y(t_n))\Big]$$

We see that the terms $y(t_n)$ cancel. To simplify farther we note that from the ODE $y'(t_n) = f(t_n, y(t_n))$ so we have

$$y(t_{n+1}) - \hat{Y}^{n+1} = \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(t_n) + \mathcal{O}(\Delta t^3) - \Delta t y'(t_n) = \frac{\Delta t^2}{2} y''(t_n) + \mathcal{O}(\Delta t^3)$$

In this case the local truncation error is $(\frac{\Delta t^2}{2} M)$ where M is a bound on $|y''(t)|$. We say that the local truncation error for Euler's method is "order $(\Delta t)^2$" or $\mathcal{O}(\Delta t^2)$. This should

be easy to remember for Euler's method because we kept terms through Δt in the Taylor series and thus the leading term we didn't include is $\mathcal{O}(\Delta t^2)$.

We now turn to approximating our global error $E^{n+1} = |y(t_{n+1}) - Y^{n+1}|$. We will see that typically if the local truncation error is $\Delta t \mathcal{O}(\Delta t^r)$ then the global error will be $\mathcal{O}(\Delta t^r)$. So for Euler's Method we expect the global error to be $\mathcal{O}(\Delta t)$ and we call it a *first order scheme*.

Expanding $y(t_{n+1})$ in terms of a Taylor series with remainder yields

$$y(t_{n+1}) - Y^{n+1} = y(t_n) + \Delta t y'(t_n) + \frac{\Delta t^2}{2} y''(\xi_n) - \Big[Y^n + \Delta t f(t_n, Y^n)\Big] .$$

We group the terms as

$$y(t_{n+1}) - Y^{n+1} = \Big[y(t_n) - Y^n + \Delta t\Big(y'(t_n) - f(t_n, Y^n)\Big)\Big] + \frac{\Delta t^2}{2} y''(\xi_n) .$$

where $\xi_n \in (t_n, t_{n+1})$. Now the term inside the square brackets is the propagated error and the term that is $\mathcal{O}(\Delta t^2)$ is the local truncation error. Letting E^n denote the error at time t_n and taking absolute values of both sides and using the triangle inequality yields

$$\begin{aligned} E^{n+1} = |y(t_{n+1}) - Y^{n+1}| &\le |y(t_n) - Y^n| + \Delta t \Big|\Big(y'(t_n) - f(t_n, Y^n)\Big)\Big| + \tau \\ &= E^n + \Delta t |f(t_n, y(t_n)) - f(t_n, Y^n)| + \tau , \end{aligned}$$

where for simplicity we have written a bound for the truncation error $\frac{\Delta t^2}{2}|y''(\xi_n)|$ as $\tau = \frac{\Delta t^2}{2} M$ where M is an upper bound for $y''(t)$; in the last step we have also replaced $y'(t_n)$ by using the ODE. We would like to bound the term $|f(t_n, y(t_n)) - f(t_n, Y^n)|$ by a constant times $E^n = |y(t_n) - Y^n|$ because then we would have

$$E^{n+1} \le (1 + C\Delta t) E^n + \tau$$

With this simple formula we can apply it repeatedly. Why can we bound $|f(t_n, y(t_n)) - f(t_n, Y^n)| \le C|y(t_n) - Y^n|$? If $f(t, y)$ satisfies a Lipschitz condition in y then

$$|f(t, y_1) - f(t, y_2)| \le L|y_1 - y_2|$$

so if we take $y_1 = y(t_n)$ and $y_2 = Y^n$ we have $|f(t_n, y(t_n)) - f(t_n, Y^n)| \le C|y(t_n) - Y^n|$ with $C = L$. Using this result and applying the formula repeatedly yields

$$\begin{aligned} E^{n+1} &\le (1 + C\Delta t) E^n + \tau \le (1 + C\Delta t)\Big[(1 + C\Delta t) E^{n-1} + \tau\Big] + \tau \\ &= (1 + C\Delta t)^2 E^{n-1} + \tau\Big[1 + (1 + C\Delta t)\Big] \\ &\le \cdots \\ &\le (1 + C\Delta t)^{n+1} E^0 + \tau\Big[1 + (1 + C\Delta t) + (1 + C\Delta t)^2 + \cdots + (1 + C\Delta t)^n\Big] \end{aligned}$$

The initial error E^0 is known and can be zero. The term multiplying the truncation error τ is the finite series

$$1 + (1 + C\Delta t) + (1 + C\Delta t)^2 + \cdots + (1 + C\Delta t)^n$$

Recall from calculus that a geometric series is $1 + r + r^2 + \cdots + r^m$ and there is an explicit formula for its sum given by $\dfrac{r^{m+1} - 1}{r - 1}$. So in our case $r = 1 + C\Delta t$ and thus the sum is

$$\frac{(1 + C\Delta t)^n - 1}{C\Delta t}$$

Also the term $(1 + C\Delta t)^n$ can be bounded by $e^{(n)C\Delta t}$ because the Taylor series expansion of e^x about $x = 0$ is

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \cdots$$

So for $x > 0$, we have that $e^x > 1 + x$. Because this is true we also have that $e^{2x} > (1 + x)^2$ and in general $e^{nx} > (1 + x)^n$. In our case $x = C\Delta t$ so we have

$$(1 + C\Delta t)^n \leq e^{(n)C\Delta t}$$

Using this for the term involving the truncation error gives

$$\frac{(1 + C\Delta t)^n - 1}{C\Delta t}\frac{\Delta t^2}{2}M \leq \frac{e^{nC\Delta t} - 1}{C\Delta t}\frac{\Delta t^2}{2}M = [e^{nC\Delta t} - 1]\frac{\Delta t}{2C}M$$

We can also use this bound for the term multiplying E^0. Now our final time is say $T = t_0 + n\Delta t$ so that $n\Delta t = T - t_0$ and

$$E^{n+1} \leq \Delta t\Big[e^{(n+1)\Delta t C}E^0 + \big(e^{Cn\Delta t} - 1\big)\frac{1}{2C}M\Big] = \Delta t\Big[e^{(T - t_0 + \Delta t)C}E^0 + e^{C(T-t_0)-1}\frac{1}{2C}M]$$

Now at time $T = n\Delta t$ the terms inside the brackets are constant so we have that the error is bounded by a constant times Δt and the method is first order.

So we have seen that Forward Euler scheme for our IVP (6) is first order, i.e., the global error is $\mathcal{O}(\Delta t)$. Does this mean that we can choose a large Δt and still get reasonable answers? To see that this is not the case consider a specific IVP

$$y'(t) = -ay(t) \quad y(0) = 1 \quad \text{where } a > 0$$

whose exact solution is $y(t) = e^{-at}$ and as t grows the solution y approaches 0. If we apply Euler's scheme repeatedly for this equation we have

$$Y^{n+1} = Y^n - a\Delta t Y^n = (1 - a\Delta t)Y^n = (1 - a\Delta t)\Big[Y^{n-1} - a\Delta t Y^{n-1}\Big] = (1 - a\Delta t)^2 Y^{n-1}$$

Continuing in this manner we get

$$Y^{n+1} = (1 - a\Delta t)^{n+1} Y^0$$

Now $a > 0$ is fixed so our approximate solution will "blow up" if Δt is large enough, i.e., $|1 - a\Delta t| > 1$. In fact, we must require

$$-1 < 1 - a\Delta t < 1 \Rightarrow -2 < -a\Delta t < 0 \Rightarrow \Delta t < \frac{2}{a}$$

This is a *stability* requirement for the method.

Example Consider the IVP

$$y'(t) = -10y(t) \quad y(0) = 1$$

What is the stability criterion for this problem for Euler's Method? Compute the solution using a Δt greater than the stability criterion and explains what happens.

Here $|1 - 10\Delta t| < 1$ implies $-1 < 1 - 10\Delta t < 1$ which implies $-2 < -10\Delta t < 0$ which says that $\Delta t < \frac{1}{5}$. If we compute with a Δt larger than this, our solution will blow up.