Modelling With ODEs

Andrew Martin

May 31, 2019

0.1 Introduction

0.1.1 Prelim

This course concerns DEs with only one independent variable, and varying numbers of dependent variables. We will consider scalar ODEs with a single dependent variable, vector ODEs, and equivalently, systems of ODEs having vector (or multiple) dependent variables.

Two types of problems:

- Dynamics (temporal problems): Independent variable is time. Unknown (dependent) function will be $\mathbf{u}(t)$. We will need an Initial condition, e.g. at $t = t_0$ $\mathbf{u}(t_0) = \mathbf{u}_0$

$$\text{ODE} + \text{ initial value } = \text{ IVP}$$

- Spatial problems (BVPs): Independent variable is space, and unknown dependent function will be $\mathbf{u}(x)$. We will need two boundary conditions, e.g. a value at $x = a$ giving $\mathbf{u}(a) = \mathbf{u}_a$ and at $x = b$ giving $u(b) = u_b$

$$\text{ODE} + \text{ Boundary Values } = \text{ BVP}$$

We can also have derivative boundary/initial conditions

0.1.2 Dynamics, Modelling and Computation

Modelling

Deriving solutions for real world problems, given a worded problem and associated ICs or BCs. Use to predict and understand behaviours and complex interactions.

Simple Pendulum

Let $\theta(t)$ be the angle away from directly down, l be the length of the pendulum, m be the mass of the pendulum, and g be the force due to gravity.

Assumptions:

Mass m is much greater than the mass of the string (neglect mass of string) $m >> m_{string}$.

Gravity is the only force acting.

Independent variable: t. Lastly assume m, g, l are constant.

Dependent variable $\theta(t)$.

Initial condition: $\theta(0) = \theta_0$

Mechanics problem: resolve motion of mass into the normal of the mass, and the tangent to the string. Note there is no motion in the normal. So we only consider the tangent. We find that these are directly related to $\theta(t)$.

Apply Newtons second law

$$\begin{aligned}
F &= ma \\
F_{norm} &= ma = 0 * m = 0 \\
F_{tan} &= ma = ml\frac{\partial^2\theta}{\partial t^2} \\
F_{tan} &= -mg\sin(\theta) \\
-mg\sin(\theta) &= ml\frac{\partial^2\theta}{\partial t^2} \\
-g\sin(\theta) &= l\frac{\partial^2\theta}{\partial t^2} \\
\frac{\partial^2\theta}{\partial t^2} + \frac{g}{l}\sin(\theta) &= 0
\end{aligned}$$

The F_{tan} lines are found by resolving forces and basic trig Initial Condition: assume the mass is released at $\theta(0) = \theta_0$ which also means $\frac{d\theta}{dt}|_{t=0} = 0$ For this problem the parameters were mass, the length of the string, gravity and the initial angle we release the string at. However we do find that the mass cancels out.

While this could be solved exactly (even though its non linear) by double integrating $\sin(\theta)$ that would go against the philosophy of the course...
Time to approximate. Lets use forward difference approximation (Euler's method).

$$\frac{df}{dt} \approx \frac{f(t+h) - f(t)}{h}$$

Where h is a small perturbation.

Turn the 2nd order scalar ODE into a first order system of

$$x(t) = \theta(t)$$
$$y(t) = \frac{d\theta}{dt} = \frac{dx}{dt}$$

The system becomes:

$$\frac{dx}{dt} = y, \quad \frac{dy}{dt} = -\frac{g}{l}\sin(x)$$

With the initial conditions

$$x(0) = x_0 = \theta_0, \quad y(0) = 0$$

Using the forward difference approximation to the system:

$$\frac{x(t+h) - x(t)}{h} = y(t)$$
$$\implies x(t+h) = x(t) + hy(t)$$
$$\frac{y(t+h) - y(t)}{h} = \frac{-g}{l}\sin(x(t))$$
$$\implies y(t+h) = y(t) + \frac{-gh}{l}\sin(x(t))$$

So what we are saying is the values of x and y at time step h into the future, it is equal to a function of the values at the present. These are known as update formulae Start with $t = t_0 = 0$:

$$(x_0, y_0) = (\theta_0, 0)$$

Then increment t to $t + h$.

$$(x_1, y_1) = \left(x(t) + hy(t), y(t) + \frac{-gh}{l}\sin(x(t)\right) = \left(x_0, \frac{-gh}{l}\sin(x_0)\right)$$

This method is known as **time stepping**.
We could have alternatively used finite differences (PDEs) to approximate the second order derivative, and then we wouldn't have had to convert the ODE into a system.
Alternative: assume θ is small, then use

$$\sin\theta \approx \theta$$

The ODE becomes linear:

$$\frac{d^2\theta}{dt^2} \approx -\frac{g}{l}\theta$$

Which is very easy to solve:

$$\theta = \theta_0 \cos(\sqrt{g/l}t)$$

Of course this is ONLY valid when θ is small (so θ_0 must be small).

Note that there are some very trivial solutions to this problem. $\theta(t) = 0$ and when $\theta(t) = \pi$, since $\sin(\theta) = 0$ if $\theta = 0, \pi$. I.e. when θ is vertical The fundamental difference between these two states is $\theta = \pi$ is **unstable**. By shifting θ by some δ such that $\theta = \pi \pm \delta$ we find it will drastically change the state.

Dynamics

Whoops i think this started earlier...

Computation

Ideally we would have exact solutions to ODEs (no errors and more useful as analytic). Unfortunately most ODEs cannot be solved exactly, so we use approximations (yay practical asymptotics). We find a trade-off between error and computational time. Remember that the models we use are approximations of the real world anyway (we might miss factors or something). Simpler models are more likely to be solvable, but they are less likely to be realistic and applicable.

Analysis

This is the most important part. Converting the problem into a format which is useful to both mathematicians and non-mathematicians. I.e. This analysis is more reflecting what sort of real behaviour we think. We want to interpret what the model predicts for the initial problem.

0.2 One-Dimensional autonomous ODE models

We want to be able to use **phase line analysis** to determine long-term behaviour of solutions, to be able to identify **fixed points** and determine their **stability**. The goal is to be able to apply suitable **non-dimensionalisations** of IVPs and to be able to identify and classify **bifurcations** in solutions as parameters are varied.

0.2.1 Autonomous Scalar Equations

A general first order scalar ODE has form

$$\frac{dx}{dt} = f(x(t), t)$$

If f does not explicitly on t. I.e:

$$\frac{dx}{dt} = f(x)$$

Then the equation is *autonomous* as it only implicitly depends on time (through $x(t)$). For this section we will only consider autonomous scalar ODEs.

0.2.2 Fixed Points, phase line analysis, and stability criteria

We are looking to consider the long term behaviour of these autonomous ODEs.

Fixed Point

For the first order, one dimensional, autonomous ODE:

$$\frac{dx}{dt} = f(x(t))$$

The *fixed points*, *steady states*, or *equilibria* of the ODE are the values of $x = x^*$ such that:

$$\frac{dx}{dt}\Big|_{x^*} = 0, \quad \Leftrightarrow \quad f(x^*) = 0$$

Example: Logistic equation

$$\frac{dx}{dt} = f(x) = rx(1 - \frac{x}{k})$$

$x(t)$ is the population, r is the growth rate, and k is the carrying capacity. We have 2 parameters (r, k).

Of course the assumption that $x(t)$ is discrete or alternatively that it is a proportion. If we consider it as a proportion then we assume that it is very large (this is where non-dimensionalisaion comes in). This is the continuum approximation.

Looking for fixed points gives:

$$x^* = 0, \quad x^* = k$$

Phase Line Analysis

If we take a perturbation for x^* to either side, what do we expect to happen?

1. if $\frac{dx}{dt} > 0$ then the population will increase.
2. if $\frac{dx}{dt} < 0$ then the population will decrease.
3. if $\frac{dx}{dt} = 0$ then the population will stay where it is.

Basically graphically if the arrows point towards the fixed point then it is stable, if they move away then it is unstable.

This is phase line analysis.

Stability

It shows that $x^* = 0$ is an unstable point as a perturbation will shift the system away from 0, and $x^* = k$ is stable as we would expect it to return to $x^* = k$

More rigorously: If we shift x^* by $\delta x \ll 1$ then (by using taylor's theorem)

$$\begin{aligned}\frac{dx}{dt}\Big|_{x=x^*+\delta x} &= f(x^* + \delta x) \\ &= f(x^*) + \delta x f'(x^*) + \dots \\ &= \delta x f'(x^*) + \dots\end{aligned}$$

The possibilities are:

1. $f'(x) < 0$:

$$\implies \begin{cases} \delta x f'(x^*) < 0 & \delta x < 0, \quad x \text{ decreasing for } \delta x > 0 \\ \delta x f'(x^*) > 0 & \delta x < 0, \quad x \text{ increasing for } \delta x < 0 \end{cases}$$

I.e. x returns to the steady state, and thus it is **stable**.

2. $f'(x) > 0$:

$$\implies \begin{cases} x \text{ increasing for } \delta x < 0 \\ x \text{ decreasing for } \delta x > 0 \end{cases}$$

I.e. x shifts away from the steady state. Therefore the steady state x is an **unstable** point

3. $f'(x) = 0$: We then need to consider the next term in the series: $\frac{1}{2}\delta x^2 f''(x^*)$.

$$\implies \begin{cases} x^* \text{ is stable if } f''(x^*) < 0 \\ x^* \text{ is unstable if } f''(x^*) > 0 \\ x^* \text{ is semi-stable if } f'(x) = 0, \quad \text{and } x^* \text{ is a turning point} \end{cases}$$

Semi-stable points are essentially considered unstable.

Alternatively we can check for these graphically: If there is a negative slope at x^* then it is stable, positive slope it is stable and then 0 slope then it is semi-stable.

Example:

$$\frac{dx}{dt} = x^2$$

The equilibria are $x = 0, x = 0$. We get

$$f'(x) = 2x \implies f'(x^*) = 0$$

$$f''(x) = 2 \implies f''(x^*) = 2$$

So we find that x^* is semi-stable.

Another example:

$$\frac{dx}{dt} = x(1-x)(2-x)$$

Fixed points at $x^* = 0, 1, 2$

$$f'(x) = 3x^2 - 6x + 2$$

$$f'(x) = \begin{cases} 2 & x = 0 \\ -1 & x = 1 \\ 2 & x = 2 \end{cases}$$

So $x = 0$ is unstable, $x = 1$ is stable, and $x = 2$ is unstable.

We would expect alternating between stable and unstable (unless it happens at a turning point [semi-stable point]).

0.2.3 Non Dimensionalisation

Consider the fishing example (2.5).

$$\frac{dN}{dt} = BN - DN^2 - g(N), \quad N(0) = N_0$$

Where N is the number of fish, B is a birth rate, D is the death rate, and $g(N)$ is the yield from fishing. Note that this is the logistic model with an extra term (the yield term). We have 3 visible parameters (B, D, N_0)Note that there could be more parameters in $g(N)$. We can reduce the number of parameters via non-dimensionalisation. If we define dimensionless terms:

$$\hat{t}, \quad \hat{N}(\hat{t})$$

Such that

$$t = t_c\hat{t}, \quad N = N_c\hat{N}$$

Where N_c and t_c are the (characteristic) scales for the fish population and time respectively. We want to choose N_c and t_c to reduce the number of parameters.

$$N_c = K = \frac{B}{D}, \quad t_c = \frac{1}{B}$$

Gives:

$$\frac{d\hat{N}}{d\hat{t}} = \hat{N}(1-\hat{N}) - \hat{g}(\hat{N}), \quad \hat{N}(0) = \hat{N}_0 = \frac{N_0}{N_c}$$

Note we now have parameters $\hat{N}_0$ and any in $\hat{g}$ which is less than we had originally.

To find N_c, t_c:

1. Substitute $N = N_c\hat{N}$ and $t = t_c\hat{t}$ into the dimensional ode.
   LHS: Note that $\frac{d\hat{t}}{dt} = \frac{1}{t_c}$

$$\begin{aligned}\frac{dN}{dt} &= \frac{d}{dt}(N_c\hat{N}) \\ &= \frac{d\hat{t}}{dt}\frac{d}{d\hat{t}}(N_c\hat{N}) \\ &= \frac{N_c}{t_c}\frac{d\hat{N}}{d\hat{t}}\end{aligned}$$

   RHS:

$$BN - DN^2 - g(N) = BN_c\hat{N} - DN_c^2\hat{N}^2 - g(N)$$

   Combine

$$\begin{aligned}\frac{N_c}{t_c}\frac{d\hat{N}}{d\hat{t}} &= BN_c\hat{N} - DN_c^2\hat{N}^2 - g(N) \\ \frac{d\hat{N}}{d\hat{t}} &= Bt_c\hat{N} - DN_ct_c\hat{N}^2 - \frac{t_c}{N_c}g(N) \\ \frac{d\hat{N}}{dt} &= Bt_c\hat{N}(1 - \frac{N_c}{K}\hat{N}) - \hat{g}(\hat{N})\end{aligned}$$

   Where $k = \frac{B}{D}$ and $\hat{g}(N) = \frac{t_c}{N_c}g(N)$

2. Choose the scales to simplify the ODE.
   Want to simplify the Bt_c and $\frac{N_c}{K}$ terms. So

$$t_c = \frac{1}{B}, \quad N_c = K = \frac{B}{D}$$

   Giving

$$\frac{d\hat{N}}{dt} = \hat{N}(1-\hat{N}) - \hat{g}(\hat{N})$$

If we now suppose $g(N) = Y = const$ Then

$$\hat{g}(\hat{N}) = \frac{t_c}{N_c} Y = \frac{1}{B}\frac{D}{B} = \frac{YD}{B^2} = y$$

Which is still constant.

This reduces the number of parameters to 2 since we have an initial condition and a parameter y.

Once we have solved the non-dimensional IVP, to get the dimensional answer out, we just convert back to the dimensional terms:

$$N = N_c \hat{N}, \quad t = t_c \hat{t}$$

But you must know the parameters within N_c and t_c to give this answer back.

So the reasons we non-dimensionalise:

1. Reduce the number of parameters
2. Improves analysis
   (a) Easier to see the type of equation (and solve it)
   (b) Easier to see how it changes with respect to parameters when there are less (especially if solving numerically)
   (c) Results are more compact
   (d) Solutions for one system can be applied to other systems with the same scaled equation but different parameters)
3. It helps 'experimental validation' using a smaller-scale version. We can see how to choose parameters for an experiment without changing the behaviour or the model (good for experimental validation)
4. We can see the relative importance of the different terms in the equations - even if we can't reduce the number of parameters!

This next bit isn't examinable

Point 4 in that list requires some care... Basically we can't use asymptotic relations in certain regions of the problem.

Its worth paying attention to small and large terms like in practical asymptotics and using that to approximate solutions.

0.2.4 Bifurcations

A bifurcation is a qualitative change in the behaviour of the system due to the effect of varying a parameter.

A bifurcation point is a point in which a parameter to the ODE can create or destroy fixed points depending on its value. We do not consider the initial condition to relate to bifurcation points.

Definition:

For the ODE

$$\frac{\partial x}{\partial t} = f(x : \mu)$$

Where μ is a scalar parameter, $\bar{x}$ is a bifurcation point, with bifurcation value $\bar{\mu}$ if

$$f(\bar{x} : \bar{\mu}) = 0, \quad \frac{\partial f}{\partial x}(\bar{x} : \bar{\mu}) = 0$$

I.e. a fixed point, with derivative equal to zero.

So if we consider

$$f(\hat{N} : y) = \hat{N}(1 - \hat{N}) - y$$

This is 0 for

$$\hat{N} = \frac{-1 \pm \sqrt{1 - 4y}}{2}$$

$$\begin{aligned} \frac{\partial f}{\partial \hat{N}} &= 1 - 2\hat{N} \\ &= 1 - (1 \pm \sqrt{1 - 4y}) \\ &= \mp\sqrt{1 - 4y} \\ \implies \bar{y} &= \frac{1}{4} \end{aligned}$$

And $\bar{\hat{N}} = 1/2$. So the bifurcation is at $(\bar{\hat{N}}, \bar{y}) = (\frac{1}{2}, \frac{1}{4})$ This is an example of a *saddle-mode bifurcation*. I.e. you have 0 fixed points, then at the bifurcation you have one fixed point, and then you have 2 fixed points. Its canonical form is

$$\frac{\partial x}{\partial t} = \mu - x^2$$

Consider

$$\frac{\partial x}{\partial t} = \mu - x^2 \implies \bar{x} = \pm\sqrt{\mu}$$

$$\begin{aligned} \frac{\partial f}{\partial x} &= -2x \\ \implies \bar{x} &= 0 \\ \implies \bar{\mu} &= 0 \end{aligned}$$

The bifurcation occurs at $(x, \mu) = (0, 0)$. We could see this immediately with the square root - if $\mu > 0$ there are 2 fixed points, if $\mu = 0$ then there is 1, and if $\mu < 0$ there are none (another saddle node fyi). You shouldn't always need to solve the system - just finding the fixed points can sometimes be enough.

0.2.5 Transcritical Bifurcation

IF we now consider the fishing model with a constant effort fishing term

$$\frac{dN}{dt} = BN - DN^2 - EN$$

Where the parameter E is the effort put into fishing. Same non-dimensionalisation, $t_c = \frac{1}{B}, N_c = \frac{B}{D}$

$$\frac{d\hat{N}}{dt} = \hat{N}(1 - \hat{N}) - e\hat{N}, \quad e = \frac{E}{B}$$

Here we have one parameter e instead of three. Note the equation is of form

$$\frac{\partial x}{\partial t} = \mu x - x^2 = f(x : \mu)$$

where $x \equiv \hat{N}$, $\mu \equiv 1 - e$.
Find the fixed points!

$$x^* = 0, \mu$$

The bifurcation pretty clearly is $(\bar{x}, \bar{\mu}) = (0, 0)$ For $\mu \neq 0$ there are two steady states, For $\mu < 0$

$$\begin{cases} x = 0 & stable \\ x = \mu & unstable \end{cases}$$

For $\mu > 0$

$$\begin{cases} x = 0 & unstable \\ x = \mu & stable \end{cases}$$

for $\mu = 0$ there is just one equilibrium, $\mu = 0, x = 0$, which is semi-stable.

This is a transcritical bifurcation. Where the branches of equilibria ($\mu < 0, \mu > 0$ swap their stabilities, as they pass through the bifurcation. Its canonical form is

$$\frac{\partial x}{\partial t} = \mu x - x^2$$

So now if we return to the original model

$$(\hat{N}, 1 - e) = (x, \mu)$$

So the fixed points are $\hat{N} = 0, 1 - e$. And the bifurcation occurs at $e = 1$,

$$\hat{N} = 0 \begin{cases} stable & e > 1 \\ unstable & e < 1 \end{cases}$$

$$\hat{N} = 1 - e \begin{cases} unstable & e > 1 \\ stable & e < 1 \end{cases}$$

Remember $e = \frac{E}{B}$. So: if $E > B$, the population survives, $if E < B$ the population dies out.

If $e < 1$, and we allow the population to reach the steady state then we expect $\hat{N}_* = 1 - e$ The yield at this state will be $\hat{y}_* = e\hat{N}_* = e(1 - e)$. This is a negative quadratic with zeros $0, 1$. If we want the yield to be $\hat{y}_* = 0.1$, this will be satisfied for $e(1 - e) = 0.1$ The maximum yield will be obtained at $\hat{y}_*^{\max} = 0.25$ for $e = 0.5$.

Remember that e is non-dimensional so this value isn't very useful. So we want to bring it back to a dimensional quantity. So recall the dimensional yield is $EN = \frac{N_c}{t_c} e\hat{N} = \frac{B^2}{D} e\hat{N}$. The maximum is

$$\frac{1}{4}\frac{B^2}{D}, \ at \ e\hat{N} = \frac{1}{4}$$

Note

$$\begin{aligned} \frac{dN}{dt} &= BN - DN^2 - EN \\ &= (B - E)N - DN^2 \\ &= (B - E)N(1 - \frac{N}{k}) \end{aligned}$$

Where $k = \frac{B-E}{D}$. Which is a logistic with a constant out the front $(B - E)$. Scaling for the logistic $t_c = \frac{1}{B-E}$, and $N_c = k$ (we are assuming $B > E$, and this means $k > 0$)

$$\frac{d\hat{N}}{d\hat{t}} = \hat{N}(1 - \hat{N})$$

Which is the non-dimensional logistic with no parameters (except the initial condition)!

If we want to find the yield (even though its not in the equation) we just continue how we would anyway. Fixed points at $\hat{N}_* = 0, 1$. So the yield is

$$Y = EN = Ek\hat{N} = \frac{E(B-E)}{D}\hat{N} = \frac{E(B-E)}{D}$$

at $\hat{N} = \hat{N}_* = 1$. Rearranging this will give us the same result as before.

0.2.6 Pitchfork Bifurcations

Supercritical Pitchfork Bifurcations

The canonical form for this is

$$\dot{x} = \mu x - x^3 = f(x : \mu)$$

Note that this is invariant under $x = -x$ (since both sides are odd powers of x) Fixed points occur at

$$x^* = 0, \pm\sqrt{\mu}$$

For the $\sqrt{\mu}$ fixed points to exist we require $\mu \geq 0$
Stability:

$$\frac{\partial f}{\partial x} = \mu - 3x^2 = \begin{cases} \mu & x = 0 \\ -2\mu & x = \pm\sqrt{\mu} \end{cases}$$

So the fixed point $x^* = 0$ is stable for $\mu < 0$ and unstable for $\mu > 0$. For $x^* = \pm\sqrt{\mu}$ it is stable for $\mu > 0$

- $\mu < 0$ $x^* = 0$ is stable
  $x^* = \pm\mu$ does not exist
- $\mu = 0$ $x^* = 0$ is stable
- $\mu > 0$ $x^* = 0$ is unstable
  $x^* = \pm\sqrt{\mu}$ is stable

We can deduce the bifurcation occurs at $(\bar{x}, \bar{\mu}) = (0, 0)$ Bifurcation diagram, plot x^*, μ and their stabilities. I.e. plot

$$x^* = 0 \ \forall \mu, x^* = \pm\sqrt{\mu}, \mu > 0$$

(where the first one is stable until $x^* = 0$.

This is a supercritical pitchfork bifurcation because the 'new' ($x^* = \pm\sqrt{\mu}$) fixed points generated are stable. So we say that the cubic term is stabilising.

Lets try solving $\dot{x} = -x + \mu\tanh(x)$ (use `matlab`). This looks a lot like a cubic. We find that the bifurcation value is $\mu = 1$. For all cases, there is a fixed point at $x_* = 0$. If $\mu > 1$ we get 3 fixed points.
To solve this, we could alternatively not use `Matlab` and use a different method. Write as

$$\tanh(x) = \frac{x}{\mu}$$

We could separate it as $C(x) = \tanh(x)$ and $S(x : \mu) = \frac{x}{\mu}$. Note that only S contains the parameter, and C contains the difficult function. Now we plot $C(x)$ and $S(x : \mu)$. Since S is easy to plot, we can repeatedly plot it for different values of μ and then look at the intersections to identify bifurcations.

Calculate the derivatives of C and S

$$C_x = sech^2 x = 1 \; at \; x = 0$$

$$S_x = \frac{1}{\mu} = \frac{1}{1} \; at \; \mu = 1$$

So $C_x(0) = S_x(0:1)$

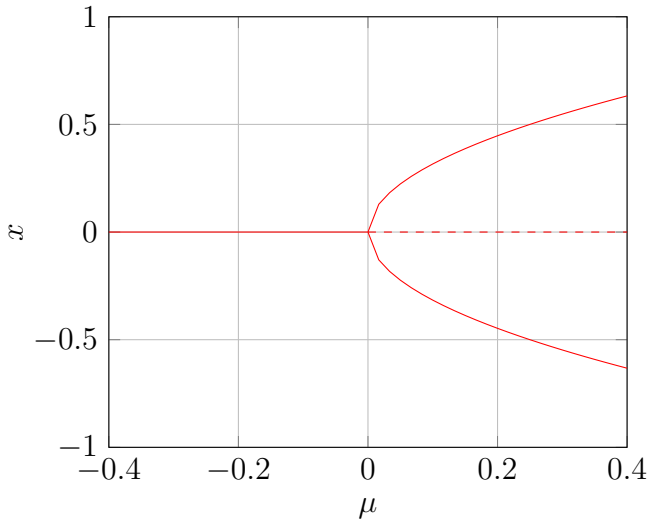


Figure 1: Supercritical case

Subcritical Pitchfork Bifurcations

Basically reverse the stabilities of the supercritical pitchfork bifurcation.
E.x

$$\dot{x} = f(x:\mu) = \mu x + x^3$$

Fixed points

$$\mu x + x^3 = 0 \implies x_* = 0, \pm\sqrt{-\mu}$$

So the second 2 fixed points will only exist for $\mu < 0$. Stability:

$$f'(x) = \mu + 3x^2$$

For $x = 0$, $f'(x) = \mu$, $x = \pm\mu$, $f(x) = -2\mu$. So the fixed point, $x_* = 0$, is stable for $\mu < 0$ and unstable for $\mu > 0$. For $x^* = \pm\sqrt{-\mu}$ it is unstable $\mu < 0$ and doesn't exist for $\mu > 0$.

Given an initial value of μ we know that in the long term we can tell what the system will do. e.g. if we choose $\mu = -1$. And the initial value of $x > \sqrt{-\mu}$. Since this lies above the fixed point, we expect the system to blow up to positive infinity. if $\mu > 0$, x will tend to $\pm\infty$ depending on the sign of the initial x.

We say that the cubic term is destabilising

Back to the $\tanh x$ example. The Taylor series ends up as

$$\tanh x = x - \frac{1}{3}x^3 + \frac{2}{15}x^5 - \frac{17}{315}x^7 + \mathcal{O}(x^9)$$

For $|x| < \frac{\pi}{2}$. So we can (very roughly) approximate this by saying $\tanh x \approx x$ (for very small x). Or by saying $\tanh x \approx x - \frac{1}{3}x^3$ (for slightly larger x). and then we can keep doing this as we want to bring $|x|$ closer to $\pi/2$.

The reasoning we are covering this is we can approximate solutions to

$$\dot{x} = \mu x - \tanh x$$

for small x by approximating with:

$$\dot{x} \approx \mu x - (x - \frac{1}{3}x^3) \implies \dot{x} \approx \lambda x + \frac{1}{3}x^3$$

$\lambda := \mu - 1$. This is the form of a subcritical bifurcation. So we know what will happen with the fixed points.

Note that this will ONLY hold for small $|x|$. So we can't say that x_* becomes unboundedly large under the same circumstances as before.

So if we then use the next larger term from the $\tanh x$ approximation, we can repeat the analysis and see how it changes the fixed points. Remember that our approximation will completely stop holding for $|x| > \pi/2$

If out approximation maintains $|x| < \pi/2$ then the approximation is (relatively) valid.

We showed that by increasing μ or decreasing μ will have different outcomes for the path taken by x. This lack of reversibility is known as hysteresis. I.e. hysteresis is a lack of reversibility as a parameter is varied.

The region where the lack of reversibility occurs is known as a hysteresis loop. They are always one directional (in this example it was an anticlockwise loop). When you sketch the loop you must show direction

Spruce budworm model

$$\frac{dN}{dt} = RN(1 - \frac{N}{K}) - p(N), \quad N(0) = N_0$$

Where $N(t)$ is the budworm population at t. And we assume it is logistic with a linear birth rate R, and carrying capacity K.

The $p(N)$ term is predation:

$$p(N) = \frac{BN^2}{A^2 + N^2}$$

For small $N = \epsilon$ we get

$$p(\epsilon) = \frac{B\epsilon^2}{A^2} \to 0$$

For large N

$$p(N) = B, \quad N \gg A$$

For $N \lessapprox A$, p is small, and so predation is low (refuge state). For $N \gtrapprox A$, predation is large so predation is high ().

Firstly we want to get the non-dimensional form

$$\frac{d\hat{N}}{d\hat{t}} = r\hat{N}\left(1 - \hat{N}/k\right) - \frac{\hat{N}^2}{1 + \hat{N}^2}$$

And $\hat{N}(0) = \hat{N}_0$ Where

$$k = \frac{K}{A},\ r = \frac{AR}{B},\ \hat{N} = \frac{N}{A},\ \hat{t} = \frac{Bt}{A},\ \hat{N}_0 = \frac{N_0}{A}$$

For notational convenience let $x := \hat{N}$

$$f(x : r, k) = rx(1 - x/k) - \frac{x^2}{1 + x^2}$$

So we have 2 parameters now. So for the analysis of bifurcations we have more possibilities!

We want to know

1. For which values of r, k will the budworm population be under control?

2. For which values of r, k will the budworm population outbreak?

What do we mean by outbreak? The budworm population suddenly jumps from low to high as r, k vary. Similarly, what do we consider a low population or a high population? Define $x \lessapprox 1$ to be a low population and $x \gg 1$ to be a high population.

To play around with analysis: First we could try setting one parameter constant, and vary the other (and then swap) By fixing k and varying r ew can see that we have an unstable fixed point at $x = 0$ always, and a stable fixed point near $x = rk$, and that there is some variation of the fixed points in the middle.

We have a region called a bistable interval, since here are 2 stable intervals in that interval. This will give us a hysteresis loop.

Recap

$$\dot{x} = f(x : r, k) = rx\left(1 - \frac{x}{k}\right) - \frac{x^2}{1 + x^2}$$

Since we are studying an outbreak, we want to look at jumps based on variations of parameters. Where a jump is $x \approx 1$ jumping to $x \gg 1$

Initially fix one of the parameters and then vary the other.

The bistable interval has 2 saddle-node bifurcations.

The lower stable branch is a refuge state, whereas the jump at after the unstable branch is an outbreak.

Should relate this to the dimensional model. Recall

$$r = \frac{AR}{B}, \quad k = \frac{K}{A} = 6$$

So if we increase R (increased birth rate) or decrease B (amount of predation), or increase A while decreasing K to keep k fixed. Then we can cause an outbreak

If we want to try varying both parameters Go back to the initial problem

$$f(x : r, k) = rx\left(1 - \frac{x}{k}\right) - \frac{x^2}{1 + x^2}$$
$$x\{r\left(1 - \frac{x}{k}\right) - \frac{x}{1 + x^2}\}$$

So we know $x^* = 0$ is a fixed point.

Note that for small x,

$$f(x : r, k) \sim rx \implies f_x \sim r > 0$$

So $x^* = 0$ is unstable since $f_x > 0$. The other fixed points are given by

$$r\left(1 - \frac{x}{k}\right) = \frac{x}{1 + x^2}$$
$$S(x : r, k) = C(x)$$

Note that we have a simple function containing parameters, and a complicated function with no parameters. Plot these and use the same logic as in the tute

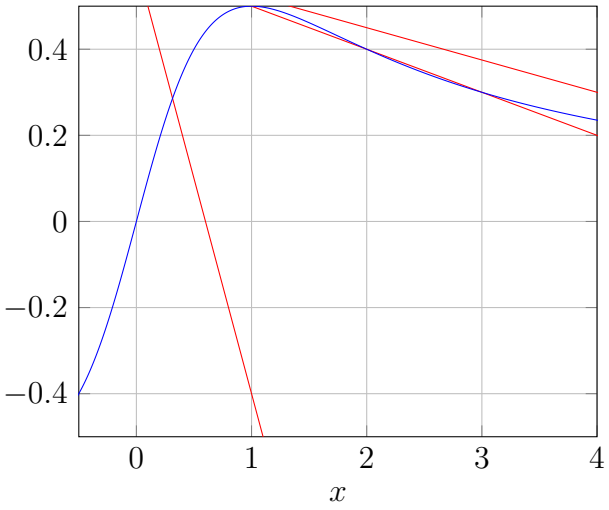


Figure 2: Varying k for the two functions, blue is $C(x)$, red are the $S(x : r, k)$

Note that $f = xg \implies f_x = g + xg_x$. Since we want the fixed point at $g = 0$ we can cross out the g term in f_x.

$$\begin{aligned}
f_x = 0 & \\
&\Leftrightarrow g + xg_x = 0 \\
&\Leftrightarrow g_x = 0 \\
&\Leftrightarrow (S - C)_x = 0 \\
&\Leftrightarrow S_x = C_x \\
\frac{\partial r\left(1 - \frac{x}{k}\right)}{\partial x} &= \frac{\partial \frac{x}{1+x^2}}{\partial x} \\
-r/k = \frac{(1+x^2) - 2x^2}{(1+x^2)^2} &= \frac{1 - x^2}{(1+x^2)^2} \\
r = \frac{k(x^2-1)}{(1+x^2)^2} &
\end{aligned}$$

Now we have to get the fixed point

$$\begin{aligned}
r(1 - \frac{x}{k}) &= \frac{x}{1+x^2} \\
\frac{k(x^2-1)}{(1+x^2)^2}(1 - \frac{x}{k}) = \frac{x}{1+x^2} \frac{k(x^2-1)}{1+x^2}(1 - \frac{x}{k}) &= x \\
k - x &= \frac{x(1+x^2)}{x^2-1} \\
k &= \frac{x(1+x^2)}{x^2-1} + x \\
k &= \frac{x(1+x^2) + x(1+x^2)}{x^2-1} \\
k &= \frac{x(x^2-1) + x(1+x^2)}{x^2-1} \\
k &= \frac{2x^3}{x^2-1}
\end{aligned}$$

Substituting into the equation for r:

$$\begin{aligned} r &= \frac{k(x^2-1)}{(1+x^2)^2} \\ &= \frac{2x^3}{x^2-1} \times \frac{(x^2-1)}{(1+x^2)^2} \\ r &= \frac{2x^3}{(1+x^2)^2} \end{aligned}$$

So the expressions for r and k are the bifurcation curves.

Since $k > 0$ we must have $x > 1$. So bifurcations only occur for $x > 1$. Bifurcations will only occur when the bifurcation curves cross.

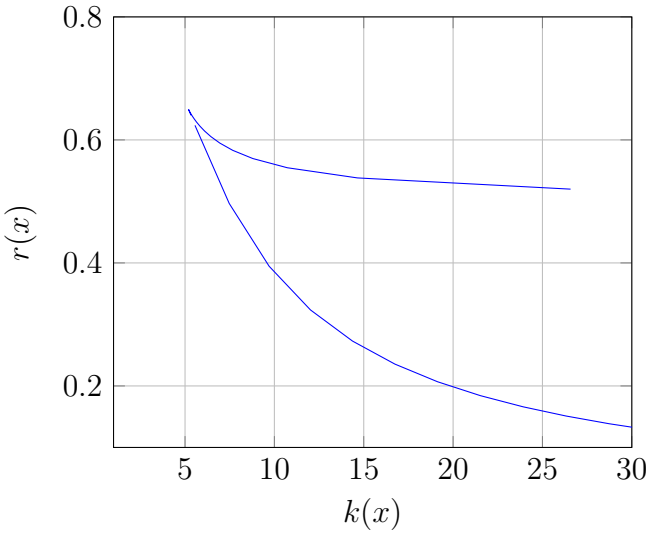


Figure 3: Bifurcation curves $k(x), r(x)$

Bifurcation diagram is now x^* against (k, r). Fixed points

$$r(1 - \frac{x}{k}) - \frac{x}{1+x^2} = 0$$

Introduce $y = x/k > 0$ as $x, k > 0$. Then

$$r = \frac{x}{(1+x^2)(1-y)}$$

In practical terms our analysis says that an outbreak of the pest corresponds to a small change in the environment (slight change in parameter values), which can result in the population jumping from the lower stable branch to the upper (outbreak). Refuge happens with the opposite

0.3 2D Autonomous Systems

So now we will consider the fixed points, stabilities and bifurcations in two-dimensional (second-order) ODE models, expressed as systems of two first-order ODEs with two unknown functions.

0.3.1 Qualitative Analysis of 2D Systems

Consider

$$\frac{dx}{dt} = f(x, y), \quad \frac{dy}{dt} = g(x, y)$$

Or in shorthand

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$$

Where $\mathbf{x} = (x, y)$ and $\mathbf{f}(\mathbf{x}) = (f, g)$

And recall this is autonomous since f, g do not explicitly depend on time.

The goal is to be able to perform analysis including:

1. Phase *plane* analysis
2. Sketching *phase portraits*
3. Identifying steady states and *nullclines*
4. Determining stability
5. Identifying *periodic orbits* and *limit cycles*
6. Identifying bifurcations

Definitions

- **Phase plane:** xy-space
- **Vector Field:** phase plane containing vectors $\mathbf{f}(\mathbf{x})$
- **Direction field:** vector field in which all vectors have same length
- **Trajectory/orbit:** solution curve $\mathbf{x}(t)$ (i.e. one path from the vector field)
- **Phase portrait:** phase plane with 'typical' solutions shown (i.e. realistic trajectories)

Note that we're working with $(x(t), y(t))$ as parametric equations.

To help with phase portraits, define **Nullclines**:

The x-**nullcline** is $n_x = \{(x, y) : f(x, y) = 0\}$.

The y-**nullcline** is $n_y = \{(x, y) : g(x, y) = 0\}$.

On n_x, vectors $\mathbf{f}$ are vertical. On n_y, vectors $\mathbf{f}$ are horizontal (since we have ignored g, f respectively.

Steady states/fixed points/equilibria are now points $(x, y) = (x^*, y^*)$ or $\mathbf{x} = \mathbf{x}^*$ such that

$$f(x^*, y^*) = g(x^*, y^*) = 0, \quad or, \quad \mathbf{f}(\mathbf{x}^*) = \mathbf{0}$$

They occur where x and y nullclines (i.e. η_x, η_y) intersect.

0.3.2 Linear Systems

The archetypical linear system:

$$\dot{x} = \alpha x + \beta y$$
$$\dot{y} = \gamma y + \delta x$$

Or in vector notation

$$\dot{\mathbf{x}} = A\mathbf{x} \implies \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \delta & \gamma \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Solutions of this system have form:

$$\mathbf{x} = c_1 \mathbf{v}_1 e^{\lambda_1 t} + c_2 \mathbf{v}_2 e^{\lambda_2 t}$$

Where λ_i, v_i are the eigenvalues and eigenvectors of A respectively, c_i are constants which depend on the ICs.

We assume that eigenvalue and eigenvector pairs exist here.

Real Eigenvalues

Consider the degenerate case where we set $\beta = \delta = 0$:

$$\dot{x} = \alpha x$$
$$\dot{y} = \gamma y$$

The x-nullcline: setting $\alpha x = 0$, which is through $x = 0$. (if $\alpha \neq 0$).
The y-nullcline: setting $\gamma y = 0$, which is through $y = 0$.

So the the steady state is $\mathbf{x}^* = (0, 0)$

The general form of the solution to this system is

$$x(t) = x(0)e^{\alpha t}$$
$$y(t) = y(0)e^{\gamma t}$$

So $\lambda_1 = \alpha$, $v_1 = (1, 0)^T$ and $\lambda_2 = \gamma$, $v_2 = (0, 1)^T$

What happens when we change α, γ?

1. $\alpha, \gamma > 0$ we see that the steady state is unstable - a kick in any direction will be **repelled** from the steady state. We call this a source. I.e. everything comes out of the fixed point

2. $\alpha, \gamma < 0$ the steady state is stable- any initial value will be absorbed in the steady state. We call this a sink.

3. $\alpha > 0, \gamma < 0$ the steady state is unstable for $x \neq 0$. We call this a saddle. The y nullcline causes the instability.

4. $\alpha < 0, \gamma > 0$ will be the same as (3) except swapping x, and y.

Complex Eigenvalues

Now if we instead set $\gamma = -\beta$ and $\delta = \alpha$

$$\dot{\mathbf{x}} = A\mathbf{x} \implies \begin{pmatrix} \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

And assume $\beta \neq 0$. The x-nullcline is

$$\alpha x + \beta y = 0 \implies y = -\alpha x/\beta$$

The y-nullcline is

$$-\beta x + \alpha y = 0 \implies y = \beta x/\alpha$$

Where the fixed point is $\mathbf{x}^* = (0, 0)$.

Eigenvalues and eigenvectors:

$$\lambda_1 = \alpha + i\beta, \quad \lambda_2 = \alpha - i\beta$$

$$\mathbf{v}_1 = (1, i), \quad \mathbf{v}_2 = (1, -i)$$

I should probably relearn how to calculate these.

So the general solution is

$$\begin{aligned} \mathbf{x}(t) &= c_1 e^{\lambda_1 t}\mathbf{v}_1 + c_2 e^{\lambda_2 t}\mathbf{v}_2 \\ &= c_3 e^{\alpha t} \begin{pmatrix} \cos(\beta t + \phi) \\ -\sin(\beta t + \phi) \end{pmatrix} \end{aligned}$$

We did this since the original problem didn't have complex numbers. ϕ is introduced as the phase. So now our constants are c_3, ϕ rather than c_1, c_2.

Parameter cases:

1. $\alpha = 0$. This gives solution
$$\mathbf{x} = c_3 \begin{pmatrix} \cos(\beta t + \phi) \\ -\sin(\beta t + \phi) \end{pmatrix}$$
This is a rotating field with only circular solutions - which are known as closed trajectories. What is the stability of the steady state for this case?

2. $\alpha > 0$

3. $\alpha < 0$ This gives a spiral sink - this is also known as a stable spiral. So it spirals towards the steady state at 0. This is stable.

General Linear system

now the general system

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

Assuming $\det A \neq 0$ then we can diagonalise A as

$$A = PBP^{-1}$$

Where $B = diag\{\lambda_1, \lambda_2\}$ and $P = [\mathbf{v}_1, \mathbf{v}_2]$

This gives

$$\begin{aligned} \dot{x} = Ax = PBP^{-1}x \\ \implies P^{-1}\dot{x} = BP^{-1}x \\ \frac{d}{dt}(P^{-1}x) = B(P^{-1}x) \\ \dot{w} = Bw \end{aligned}$$

Where $\mathbf{w} = P^{-1}\mathbf{x}$, or $\mathbf{x} = P\mathbf{w}$. I.e. the derivative of $\mathbf{w}$ is just a linear transformation of $\mathbf{w}$. Note that B is diagonal. We know how diagonal systems behave (look back to the example).

1. $\lambda_1, \lambda_2 \in \mathbb{R}$ gives source,sink or saddle
2. $\lambda_1 = \bar{\lambda_2} \in \mathbb{C}$ gives spiral or centre

We will find that $\mathbf{x}$ has the same **type** of behaviour as $\mathbf{w}$.

Asymptotic Stability

Solutions converge to the steady state when both eigenvalues are $\lambda_1, \lambda_2 < 0$ (if they're real). The origin becomes a stable node. Or when the real **parts** of the eigenvalues are negative $\alpha < 0$. Then we get a stable spiral.

When solutions converge to the steady state, then we say that the steady state is *asymptotically stable*

Theorem: For a two-dimensional linear system $\dot{\mathbf{x}} = A\mathbf{x}$, the following are equivalent:

1. The equilibrium $(0, 0)$ is asymptotically stable
2. all eigenvalues of A have negative real part
3. $\det A > 0$ and $trA < 0$

Proof: By definition (1) $\Leftrightarrow$ (2). For (3), consider the characteristic equation of A:

$$\begin{aligned} (a-\lambda)(d-\lambda) - bc = 0 \\ ad - (a+d)\lambda + \lambda^2 - bc = 0 \\ \lambda^2 - (a+d)\lambda + (ad - bc) = 0 \end{aligned}$$

Note that $(a+d) = trA$, and $(ad - bc) = \det A$

$$\lambda = \frac{1}{2}trA \pm \sqrt{(trA)^2 - 4\det A}$$

For the complex λ case, we require the real part to be negative: $trA < 0$.

For real λ, we require the discriminant to be positive. For asymptotic stability we need the real part of both to be 0. So we require $\det A > 0$.

Hand wave-y argument but we'll leave it as that

Basically we just need to calculate the trace and the determinant of A to determine whether or not the system is asymptotically stable.

- $\det A > 0$
  - $trA > 0$ Unstable region:

    - $detA > \frac{1}{4}(trA)^2$ unstable spiral
    - $detA < \frac{1}{4}(trA)^2$ unstable node
  - $trA < 0$ Asymptotically stable region.
    - $detA > \frac{1}{4}(trA)^2$ stable spiral
    - $detA < \frac{1}{4}(trA)^2$ stable node
  - $trA = 0$ centre behaviour
- $\det A < 0$ not stable.

0.3.3 Non-linear Systems and Linearisation

General non-linear system

Consider the two dimensional non-linear system:

$$\dot{x} = f(x, y)$$
$$\dot{y} = g(x, y)$$

Where we will assume f, g are continuously differentiable functions. Each point $\mathbf{x}^* = (x^*, y^*)$ satisfying $f(x^*, y^*) = g(x^*, y^*) = 0$ is a fixed point/equilibrium/steady state.

Definitions

1. A steady state $\mathbf{x}^*$ is stable if a solution which starts nearby stays nearby
2. A steady state $\mathbf{x}^*$ is unstable if it is not stable.
3. A steady state $\mathbf{x}^*$ is asymptotically stable if $\mathbf{x}^*$ is stable, and all solutions near $\mathbf{x}^*$ converge to $\mathbf{x}^*$.

We can determine stability through linearisation

Linearisation

Suppose we are close to a steady state $\mathbf{x} = \mathbf{x}^*$ then

$$\mathbf{x}(t) = \mathbf{x}^* + \mathbf{w}(t), \quad \mathbf{w}(t) = \begin{pmatrix} w(t) \\ z(t) \end{pmatrix}$$

And $||\mathbf{w} \ll 1$

Take a Taylor series (we will have to do this in x and y). We will only do this to the linear terms

$$\begin{aligned} f(x, y) = f(x^* + w, y^* + z) &= f(x^*, y^*) + w f_x(x^*, y^*) + z f_y(x^*, y^*) + \dots \\ &= w f_x(x^*, y^*) + z f_y(x^*, y^*) + \dots \\ g(x, y) &= w g_x(x^*, y^*) + z g_y(x^*, y^*) + \dots \end{aligned}$$

So we get

$$\mathbf{f}(\mathbf{x}) = J(\mathbf{x}^*)\mathbf{w} + \dots$$

Where

$$J(\mathbf{x}) = \begin{pmatrix} f_x(\mathbf{x}) & f_y(\mathbf{x}) \\ g_x(\mathbf{x}) & g_y(\mathbf{x}) \end{pmatrix}$$

Is the Jacobian

Hence close to $\mathbf{x}^*$

$$\dot{x} = \frac{\partial}{\partial t}(\mathbf{x}^* + \mathbf{w}) \implies \dot{\mathbf{w}} \approx J(\mathbf{x}^*)\mathbf{w}$$

This is a linear system. Which can be solved easily.

Definition: The steady state $\mathbf{x}^*$ is called *hyperbolic* if **all** eigenvalues of the Jacobian $J(x^*, y^*)$ have non-zero real part

Theorem: Hartman-Grobman Theorem

Assume that $\mathbf{x}^*$ is a *hyperbolic equilibrium*. Then, in a small neighbourhood of $\mathbf{x}^*$, the phase portrait of the non-linear system

$$\dot{x} = f(x, y)$$
$$\dot{y} = g(x, y)$$

Is equivalent to that of the linearised system.

I.e. near a *hyperbolic equilibrium*, $\mathbf{x}^*$, we can use the linearised problem.

$$\dot{\mathbf{x}} = A\mathbf{x}$$

Where $A = J(\mathbf{x}^*)$

So if we calculate the trace and determinant of A we know about the stability of the state and its type.

$$\lambda_\pm = \frac{1}{2}\{trA \pm \sqrt{(trA)^2 - 4\det A}\}$$

We require the real part of BOTH λs to be non-zero. So we **cant** have:

- $trA = 0$ with $\det A > 0$
- $\det A = 0$

0.3.4 Application to models

Interaction model for 2 populations

Consider 2 interacting populations. Each has exponential growth with some interaction term proportional to both of them

$$\frac{dx}{dt} = \alpha x + \beta xy$$
$$\frac{dy}{dt} = \gamma y + \delta xy$$

This is clearly *not* a linear model.

Write

$$\begin{cases} f(x, y) = \alpha x + \beta xy \\ g(x, y) = \gamma y + \delta xy \end{cases}$$

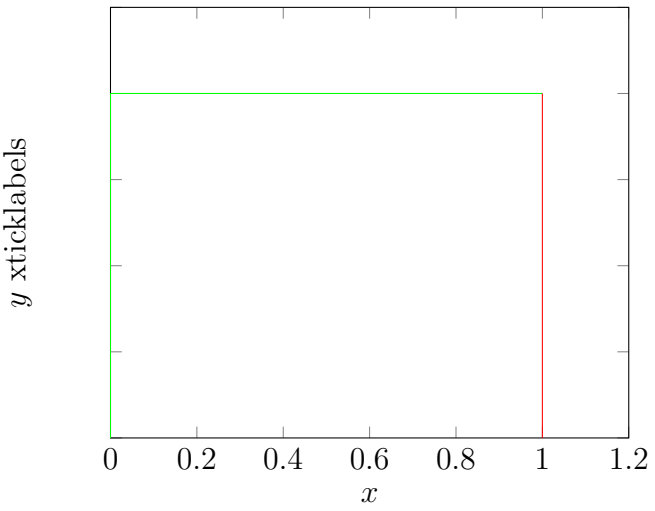
Nullclines

$$\begin{cases} n_x : \alpha x + \beta xy = 0 & \implies x = 0, y = -\alpha/\beta \\ n_y : \gamma y + \delta xy = 0 & \implies y = 0, x = -\delta/\gamma \end{cases}$$

Note this is a population model, so we require "biologically relevant solutions" $x, y \geq 0$

We only care about 2 intersections: $x, y = (0, 0)$ and the top right intersection.

α	β	γ	δ	
+	+	+	-	Predator(x), prey(y) models
+	+	-	-	Predator(x), prey(y) models
-	+	+	-	Predator(x), prey(y) models
-	+	-	-	Predator(x), prey(y) models
+	-	+	+	Predator(y), prey(x) models
+	-	-	+	Predator(y), prey(x) models
-	-	+	+	Predator(y), prey(x) models
-	-	-	+	Predator(y), prey(x) models
+	+	+	+	Mutualism / symbiosis models
+	+	-	+	Mutualism / symbiosis models
-	+	+	+	Mutualism / symbiosis models
-	+	-	+	Mutualism / symbiosis models
+	-	+	-	Competition Models
+	-	-	-	Competition Models
-	-	+	-	Competition Models
-	-	-	-	Competition Models



Calculate the jacobian: (I HAVEN"T FINISHED THIS PART)

$$J(0,0) = \begin{pmatrix} f_x(0,0) & f_y(0,0) \\ g_x(0,0) & g_y(0,0) \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \gamma \end{pmatrix}$$

Which has

$$\begin{cases} trace = \alpha + \gamma \\ det = \alpha\gamma \end{cases}$$

Hence

$$\begin{cases} \lambda_1 = \alpha \\ \lambda_2 = \gamma \end{cases}$$

λ's are real. We don't know if its a source, sink or saddle until we know the signs of α, γ.

For the second steady state:

$$J(-\gamma/\delta, -\alpha/\beta) = \begin{pmatrix} \alpha - \beta\alpha/\beta & -\beta\gamma/\delta \\ -\delta\alpha/\beta & \gamma - \delta\gamma/\delta \end{pmatrix}$$
$$= \begin{pmatrix} 0 & -\beta\gamma/\delta \\ -\alpha\delta/\beta & 0 \end{pmatrix}$$
$$=: A$$

We have

$$\begin{cases} trA = 0 \\ \det A = -\alpha\gamma \end{cases}$$

$$\lambda_{\pm} = \frac{1}{2}\{trA \pm \sqrt{(tr(A))^2 - \frac{1}{4}\det A}\}$$
$$= \pm\sqrt{\det A}$$
$$= \pm\sqrt{\alpha\gamma}$$

Which will be a centre if $\alpha\gamma < 0$ or a saddle if $\alpha\gamma > 0$

Lutka-Voltera predator-prey model:

$\alpha < 0$, $\beta > 0$, $\gamma > 0$ and $\delta < 0$. From the table, x is the predator while y is the prey. Since $\alpha < 0$ the predator while die out without prey, and $\gamma > 0$ the prey will grow out of control without the predator.

At the steady state $P_1 = (0, 0)$

$$\lambda = \begin{cases} \alpha < 0 \\ \gamma > 0 \end{cases} \implies P_1 \text{ is a saddle}$$

At $P_2 = (-\gamma/\delta, -\alpha/\beta)$: Is the steady state biologically relevant? $-\gamma/\delta > 0$ and $-\alpha/\beta > 0$ so yes.

$$detA = -\alpha\gamma > 0$$

So P_2 is a centre (in the linearised problem) And

$$\lambda_{\pm} = \pm i\sqrt{|\alpha\gamma|}$$

Which is purely imaginary. The real part is zero - so P_2 is not hyperbolic, and so the linearised problem cannot be used.

Instead write the system in a different form (it comes out as a separable ODE)

$$\begin{aligned}
\frac{dy}{dx} &= \frac{\frac{dy}{dt}}{\frac{dx}{dt}} \\
\frac{dy}{dx} &= \frac{g}{f} \\
\frac{dy}{dx} &= \frac{y(\gamma+\delta x)}{x(\alpha+\beta y)} \\
\int \frac{\alpha+\beta y}{y} dy &= \int \frac{\gamma+\delta x}{x} dx \\
\alpha \int \frac{1}{y} dy + \beta \int 1 dy &= \gamma \int \frac{1}{x} dx + \delta \int 1 dx \\
\alpha \log y + \beta y &= \gamma \log x + \delta x + C \\
\text{let } \alpha = -1 = \delta, \quad \beta = 1 = \gamma & \\
\log y + \log x &= -C + x + y \\
xy &= e^{-C+x+y} \\
\frac{e^{x+y}}{xy} &= B = e^C
\end{aligned}$$

Where B is determined by the initial conditions

This gives a non-linear centre. The corresponding solutions for $x(t)$ and $y(t)$ oscillate around the steady state.

Mutualism model

$\alpha < 0$, $\beta > 0$, $\gamma < 0$, $\delta > 0$ I.e. neither can survive alone.

At $P_1(0,0)$

$$\begin{cases} \lambda_1 = \alpha < 0 \\ \lambda_2 = \gamma < 0 \end{cases}$$

Hence P_1 is a sink (stable node).

At P_2 check if it is biologically relevant: $-\gamma/\delta > 0$ and $-\alpha/\beta > 0$ so it is!

$detA = -\alpha\gamma < 0$ so P_2 is a saddle

$$\lambda_\pm = \pm\sqrt{\alpha\gamma}$$

Since $Re(\lambda_\pm) \neq 0$ the HG theorem applies (the linearised problem works here).

To sketch a phase portrait (by hand) start with the nullclines: On η_x where $y = -\alpha/\beta > 0$,

$$g = \gamma y + \delta x y = -\alpha/\beta(\gamma+\delta x) \begin{cases} > 0 & \gamma + \delta x > 0 \implies x > -\gamma/\delta \\ < 0 & \gamma + \delta x < 0 \implies x < -\gamma/\delta \end{cases}$$

On η_y, where $x = -\gamma/\delta$

$$f = \alpha x + \beta x y = -\gamma/\delta(\alpha+\beta x) \begin{cases} > 0 & \alpha + \beta x > 0 \implies y > -\alpha/\beta \\ < 0 & \alpha + \beta x < 0 \implies y < -\alpha/\beta \end{cases}$$

The other piece of information we can use is the steady and unsteady directions from the saddle. For P_2 we have

$$V_\pm = \begin{pmatrix} \pm 1 \\ 1 \end{pmatrix}$$

Noting that V_+ will be the unstable direction: since $\lambda_+ > 0$ and V_- will be stable since it is related to the negative eigenvalue

Competition model

$\alpha > 0, \beta < 0, \gamma < 0, \delta < 0$

At $P_1 = (0,0)$

$$\begin{cases} \lambda_1 = \alpha > 0 \\ \lambda_2 = \gamma < 0 \end{cases} \implies P_1 \text{ is a saddle}$$

$$J(0,0) = \begin{pmatrix} \alpha & 0 \\ 0 & \gamma \end{pmatrix} \implies v_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Since its a diagonal matrix the eigenvectors are really simple. Unstable direction is v_1, since $\lambda_1 = \alpha > 0$ and $\lambda_2 = \gamma < 0$ gives the stable v_2.

The second steady state $P_2 = (-\gamma/\delta, -\alpha/\beta)$ is it biologically relevant?

$$-\alpha/\beta > 0, \quad -\gamma/\delta < 0$$

So no this is not biologically relevant.

0.3.5 The Kermack-McKendrick epidemic model

Start with the $SIRS$ model.

$$\frac{dS}{dt} = -\beta IS + \gamma R$$
$$\frac{dI}{dt} = \beta IS - \alpha I$$
$$\frac{dR}{dt} = \alpha I - \gamma R$$

The model has units people per unit time. It contains a nonlinear term, βIS.

$\beta I > 0$ is the rate at which S become I.

$\alpha > 0$ is the rate at which I become R.

$\gamma > 0$ is the rate at which R become S.

This is a $3D$ model. So lets make it $2D$.

First we will just consider the SIR model, so $\gamma = 0$

We can drop $\frac{dR}{dt}$ since it now gives no information and isn't in the other equations. The SIR model is the KM model. If we know $S(t)$ and $I(t)$ then we can find

$$R(t) = N - S(t) - I(t)$$

Where N is the fixed population size.

$$\frac{dS}{dt} = -\beta IS$$
$$\frac{dI}{dt} = \beta IS - \alpha I$$

We won't bother non-dimensionalising because it makes life easier. This is effectively a predator prey model, where the prey is constant if there are no infected.

The S- nullcline $\beta IS = 0$, so either $I = 0$ or $S = 0$

The I- nullcline $\beta IS - \alpha I = 0$. So either $I = 0$ or $S = \alpha/\beta$

The steady states are along the S axis. I.e. whenever $I = 0$ we have a steady state. We will call this a **ray** of steady states.

The Jacobian:

$$J(S, I) = \begin{pmatrix} -\beta I & -\beta S \\ \beta I & \beta S - \alpha \end{pmatrix}$$

Along the ray

$$J(S^*, 0) = \begin{pmatrix} 0 & -\beta S_* \\ 0 & \beta S_* - \alpha \end{pmatrix} = A$$

$$\begin{cases} trA & = \beta S_* - \alpha \\ \det A & = 0 \end{cases}$$

Hence $\lambda_1 = 0$ and $\lambda_2 = \beta S_* - \alpha$

Thus all the steady states are non-hyperbolic so we cannot use the linearised problem. Note that $\lambda_2 < 0$ if $S_* < \alpha/\beta$

$$f = -\beta IS$$
$$g = BI(S - \frac{\alpha}{\beta})$$

So f is everywhere pointing left g points down for $S < \frac{\alpha}{\beta}$ and points up for $S > \frac{\alpha}{\beta}$

Hence $S < \frac{\alpha}{\beta}$ with $I = 0$ is a steady state, and $S > \frac{\alpha}{\beta}$ with $I = 0$ is an unsteady state. To really construct the phase portrait use the trick we used for the predator prey model:

$$\frac{dI}{dS} = \frac{\dot{I}}{\dot{S}} = \frac{I(\beta S - \alpha)}{-\beta IS} = -1 + \frac{\alpha}{\beta S}$$

Hence

$$I = -S + \frac{\alpha}{\beta} \log S + C$$

Where

$$C = I_0 + S_0 - \frac{\alpha}{\beta} \log S_0$$

Where $I_0 = I(0)$ and $S_0 = S(0)$.

So we can write I in terms of S.

$$I(S) = I_0 + S_0 - S + \frac{\alpha}{\beta} \log \frac{S}{S_0}$$

And then we can plot!

0.3.6 Big Recap

2D Autonomous Systems.

$$\dot{x} = f(x, y)$$
$$\dot{y} = g(x, y)$$

or alternatively

$$\dot{\mathbf{x}} = f(\mathbf{x})$$

Phase portraits require:

- nullclines
  - x nullcline, η_x where $f(x, y) = 0$
  - y nullcline, η_y where $g(x, y) = 0$
- Steady states - intersection of an x and a y nullcline.
- Vector field
- Trajectories/orbits

Linear systems: unique fixed point at origin.

$$\dot{x} = \alpha x + \beta y$$
$$\dot{y} = \gamma y + \delta x$$

Solutions have form

$$\mathbf{x} = c_1 \mathbf{v}_1 e^{\lambda_1 t} + c_2 \mathbf{v}_2 e^{\lambda_2 t}$$

They display one of the following behaviours:

- Unstable behaviour (source or unstable node)
- Stable behaviour (sink or stable node)
- Saddle behaviour
- Centre (closed orbits)
- Unstable spiral
- Stable spiral

The behaviour is determined by the trace and determinant of the matrix.
For non-linear systems

$$\dot{x} = f(x, y) \dot{y} = g(x, y)$$

HG theorem: near a hyperbolic steady state (linearised system only has non-zero eigenvalues), the linear system is equivalent to the linear system.

0.3.7 Limit Cycles

In 1D/scalar systems, solutions either tend to fixed points or become unbounded.
In 2D autonomous systems, we can get periodic solutions (closed trajectories around centre).
If we now consider

$$\dot{r} = r(1 - r^2)$$
$$\dot{\theta} = 1$$

In polar coordinates (r, θ) (convert back with $x = r\cos\theta, y = r\sin\theta$)

For the r equation, the steady states are $r_* = 0, \pm 1$ And we get $r_* = 0$ is unstable, $r_* = 1$ is stable. In Cartesian coordinates this will move us into the fixed circle $r = 1$. I.e. the circle $x^2 + y^2 = 1$

If we solve the second equation, we get:

$$\theta = t + \theta_0$$

So θ changes continuously.

So we expect to spiral until we reach the unit circle (whether we come from inside or outside) I.e. $r_* = 0$ is an unstable fixed point (an unstable spiral), The circle $r_* = 1$ is known as a **limit cycle**. It is an isolated closed orbit.

If we plot this for x and t, we find all solutions tend towards a single periodic solution. This is the sign of a limit cycle.

0.3.8 Bifurcations in 2D systems

We already did this for $1D$. Same thing occurs in 2D: fixed points can be created or destroyed or destabilised with parameter changes. The difference is that closed orbits can do the same thing as fixed points. In 2D we say a bifurcation has occurred if the phase portrait changes its topological structure as a parameter is varied.

Saddle-node, transcritical and pitchfork bifurcations

The bifurcations of all 1D fixed points have 2D analogues. We find that one of the two dimensions will do what we saw in 1D, while in the second dimension there will simply by attraction or repulsion.

Example

$$\begin{aligned}\dot{x} &= \mu - x^2 = f(x, y)\\ \dot{y} &= -y = g(x, y)\end{aligned}$$

These are decoupled, so they will be easy to work with, and μ only appears in the x equation.

The x equation is the standard form for a saddle node bifurcation. $\mu > 0$ gives 2 fixed points ($x_* = \pm\sqrt{\mu}$, positive is stable, negative is unstable), $\mu = 0$ gives 1 fixed point and $\mu < 0$ gives 0 fixed points.

The y equation has the fixed point $y_* = 0$, which will be stable.

Thus in the 2D system, for $\mu > 0$ there are two fixed points $\mathbf{x_*} = \left(-\sqrt{\mu}, 0\right)$ and $\left(\sqrt{\mu}, 0\right)$. The positive one is a sink, the negative one is a saddle.

for $\mu < 0$ there are no fixed points.

For $\mu = 0$ the fixed point is non-hyperbolic interestingly enough.

We won't bother with the other canonical examples because they're the same analysis as this.

Hopf Bifurcation

Consider the system

$$\begin{aligned}\dot{x} &= -y + x(\mu - x^2 - y^2)\\ \dot{y} &= x + y(\mu - x^2 - y^2)\end{aligned}$$

With $\mu \in \mathbb{R}$

We can rewrite this system in polar coordinates:

$$\begin{aligned}\dot{r} &= r(\mu - r^2)\\ \dot{\theta} &= 1\end{aligned}$$

Earlier, we saw that when $\mu = 1$ we get a limit cycle.

Recall $\dot{r}$ is the equation for a pitchfork bifurcation (in $\dot{r}$). Fixed points $r_* = 0, \sqrt{\mu}$ for $\mu > 0$ which is stable. Note that $r \geq 0$ we cannot have $r < 0$ due to the use of polar coordinates. $r_* = 0$ is stable for $\mu < 0$ and unstable for $\mu > 0$

The Jacobian (in terms of x, y):

$$J = \begin{pmatrix} \mu - 3x^2 - y^2 & -1 - 2xy \\ 1 - 2xy & \mu - x^2 - 3y^2 \end{pmatrix}$$

At the fixed point $x_* = y_* = 0$

$$J(0,0) = \begin{pmatrix} \mu & -1 \\ 1 & \mu \end{pmatrix}$$

Referring to 3.2.2 this has eigenvalues $\mu \pm i$. Meaning the fixed point is a stable spiral for $\mu < 0$, an unstable spiral for $\mu > 0$ and for $\mu = 0$ it is non-hyperbolic.

At $r_* = \sqrt{\mu}$ ($\mu > 0$) we get the limit cycle behaviour with radius $\sqrt{\mu}$ so we get an 'attracting limit cycle'

This is a **Hopf bifurcation**. I.e. the change of stability of a fixed point....

0.4 Numerics!

Easy shit!

We will look at numerical solutions of first order IVPs.

$$\frac{du}{dt} = f(u,t), \quad u(0) = u_0$$

However most concepts wlil extend to higher dimensions very straightforwardly. We don't always need numerical solutions. We can get exact solutions thus far for:

- Separable ODEs
- Linear ODEs
- Exact ODEs
- Homogeneous ODEs

Goals:

1. To understand the concept of a 'well posed problem'
2. How to apply the Picard-Lindelof theorem
3. To be able to measure error, determine conditioning of a problem, and the stability of a numerical algorithm
4. To know some important finite difference numerical algorithms and how to choose them

0.4.1 Existence and Uniqueness

Canonical Example:

Consider the first order IVP of form:

$$\dot{x} = x^{1/3}, \quad x(0) = 0$$

And assuming $x(t) \in \mathbb{R}$ for all $t \geq 0$.

Note this is a non-linear ODE since we have a fractional power of the dependent variable.

Clearly the trivial solution works: $x(t) = 0$

Note the ODE is separable:

$$x^{1/3}\dot{x} = 1 \implies \int x^{-1/3} dx = \int dt \implies \frac{3}{2}x^{2/3} = t + c$$

IC $x(0) = 0 \implies c = 0$

$$\implies x(t) = \left(\frac{2t}{3}\right)^{3/2}$$

Hence

$$x(t) = \begin{cases} 0 & t < \hat{c} \\ \left(\frac{2(t-\hat{c})}{3}\right)^{3/2} & t \geq \hat{c} \end{cases}$$

Is a solution also, for any $\hat{c} \geq 0$. This is a 'family' of solutions. Also the square root has positive and negative branches.

If we attempted to obtain a numerical solution to this:

Using Euler's method:

$$\frac{dx}{dt} = \frac{x(t+h) - x(t)}{h}$$

$$\implies x_{n+1} = x_n + hx_n^{1/3}$$

Where $x_n \approx x(t_n)$ and $t_n = nh$.

Start with $x_0 = x(0) = 0$

$$x_1 = x_0 + hx_0^{1/3} = 0$$

And hence via iteration we get $x_n = 0$, i.e. the trivial solution.

Well and Ill posed Problems

For mathematical models of physical phenomena:

1. A solution must exist
2. The solution must be unique
3. The solution's behaviour must change continuously with the initial conditions

If we have these 3 properties, the problem is well posed. Otherwise it is ill posed.

A well posed problem should give numerical solutions (it has a 'good chance'), ill posed problems will not, and need to be reformulated.

0.4.2 Existence and Uniqueness Theorem

The Picard-Lindelöf theorem

Suppose $I = [t_-, t_+]$ is an interval in t with $t_- \leq t_0 < t_+$ and $J = [u_-, u_+]$ is an interval in u with $u_- < u_0 < u_+$. If $f : I \times J \to \mathbb{R}$ is continuous on its domain, and Lipschitz continuous (shown below) on J, then there exists $\epsilon > 0$ such that the IVP:

$$\begin{aligned} \frac{du}{dt} &= f(t,u), \quad t \in [t_0, t_0 + \epsilon] \\ u(t_0) &= u_0 \end{aligned}$$

Has a unique solution $u \in C^1[t_0, t_0 + \epsilon]$

Lipschitz continuity:

Let $J = [x_0, x_1]$ be an interval on the real line. We say $f : J \to \mathbb{R}$ is **Lipschitz continuous** on J if there exists $L > 0$ such that

$$|f(x) - f(y)| \leq L|x - y|, \quad \text{for all } x, y \in J$$

Examples:

Show $f(x) = \sin x$ is Lipschitz continuous on $\mathbb{R}$. Let $x, y \in \mathbb{R}$, and assume $y < x$ (wlog)

$$\sin x - \sin y = (x - y)\cos c, \quad c \in (y, x)$$

Using the Mean Value Theorem (MVT) So

$$\begin{aligned} |f(x) - f(y)| &= |\sin x - \sin y| \\ &= |(x - y)\cos c| \\ &= |(x - y)||\cos c| \leq |x - y| \end{aligned}$$

Since $|\cos c| \leq 1$ for all $c \in \mathbb{R}$

Thus $f(x) = \sin x$ is Lipschitz continuous, with $L = 1$.

We call the smallest possible value of L the best Lipschitz constant.

We say $f = \sin x$ is globally Lipschitz continuous, since it is continuous over its entire domain, where not all functions will be.

Example 2:

$$g(x) = x^2$$

We could use the MVP as above, but there is a simpler approach:

$$\begin{aligned} |g(x) - g(y)| &= |x^2 - y^2| \\ &= |(x + y)(x - y)| \\ &= |x + y||x - y| \end{aligned}$$

Using the fact that $|x + y| \leq |x| + |y|$ by the triangle inequality, consider g on a bounded interval, say $J = [-1, 1]$.

$g(x) = x^2$ Say we consider an odd interval J, e.g. $J = [-1, 1]$, then $|x| + |y| \leq 2$ So that

$$|g(x) - g(y)| \leq 2|x - y|$$

So g is Lipschitz continuous with best const $L = 2$ on $J = [-1, 1]$ However if J is unbounded, e.g. $J = \mathbb{R}$. Then there doesn't exist an L. But for any finite interval, then we have Lipschitz continuity.

If we had $f(x) = |x|$, we can use the reverse triangle inequality

$$|x| = |(x-y)+y| \leq |x-y| + |y| \implies |x| - |y| \leq |x-y|$$

$$|y| = |(y-x)+x| \leq |y-x| + |x| \implies -|y| - |x| \leq |x| - |y|$$

$$-|x-y| \leq |x| - |y| \leq |x-y| \implies ||x| - |y|| \leq |x-y|$$

This last statement is

$$|f(x) - f(y)| \leq |x-y|$$

And hence $f(x) = |x|$ is globally Lipschitz continuous with $L = 1$

We couldn't have used the MVT for this since the derivative of $|x|$ is not continuous. Hence, Lipschitz continuity does not require differentiability.

If we have $J = [x_0, x_1] \subset \mathbb{R}$, and $f : J \to \mathbb{R}$. Suppose $f(x)$ is continuously differentiable on J. We write this as $f \in C^1(J)$.

Then f is Lipschitz continuous on J. For a proof of this:

We use the MVT, for every $x, y \in J$, $\exists c \in J$:

$$f(x) - f(y) = f'(c)(x-y)$$

And therefore we set $L = \max_{c \in J}\{|f'(c)|\}$ and deduce

$$|f(x) - f(y)| \leq L|x-y|$$

This means

$$\textit{continuous differentiability} \implies \textit{Lipschitz continuity}$$

But the converse is not true.

If we have the ODE $\dot{u} = f(t,u)$ where $t \in I$ and $u \in J$, so we get $f : I \times J \to \mathbb{R}$. So we can plot this system in 3D as a surface plot. I.e. plotting u, t, f (literally half of what I'm doing in my thesis...)

0.4.3 Picard Iteration

Start with ODE

$$\begin{aligned} \dot{u}(t) &= f(t, u(t)) \\ \int_{t_0}^{t} \dot{u}(s)ds &= \int_{t_0}^{t} f(s, u(s))ds \\ LHS = [u(s)]_{t_0}^{t} &= u(t) - u(t_0) (= u_0) \\ \implies u(t) &= u_0 + \int_{t_0}^{t} f(s, u(s))ds \end{aligned}$$

This is the **integral** equation form of the IVP. Note it contains the initial condition, and there are no derivatives.

Example:

$$\frac{du}{dt} + u \tan t = \sin 2t$$

With $u(0) = 1$.

$$\begin{aligned}
\frac{u}{t} &= \sin 2t - u \tan t \\
u(t) &= 1 + \int_0^t \sin 2s - u \tan s ds \\
&= 1 + \int_0^t \sin 2s ds - \int_0^t u \tan s ds \\
&= 1 - \frac{1}{2} \cos 2t - \int_0^t u \tan s ds
\end{aligned}$$

Note that the integral equation allows us to form series approximations to the solution. This is actually used in the Picard-Lindelöf theorem proof

Picard iteration method:

1. Take the initial approximation $n \approx u^{(0)} = u_0$ (which is constant)
2. Iterate
$$u^{(k)}(t) = u_0 + \int_{t_0}^t f(s, u^{(k-1)}(s)) ds$$
3. Terminate when $|u^{(k)} - u^{(k-1)}| < tol$ where tol is some tolerance.

A few questions / problems:

- Why is this a series approximation?
- How does it compare to a Taylor series?
- Most importantly: do we have convergence? I.e.
$$\lim_{k \to \infty} u^{(k)}(t) = u(t)$$

We will answer these with some examples

$$\dot{u} = u, \quad u(0) = 1$$

The exact solution is

$$u(t) = e^t = \sum_{n=0}^{\infty} \frac{t^n}{n!} = 1 + t + \frac{1}{2} t^2 + \dots$$

Picard iteration gives, with $u^{(0)} = u_0 = 1$:

$$\begin{aligned}
u^{(k)}(t) &= u_0 + \int_0^t u^{(k-1)}(s)ds \\
u^{(k)}(t) &= 1 + \int_0^t u^{(k-1)}(s)ds \\
u^{(1)}(t) &= 1 + \int_0^t ds = 1 + t \\
u^{(2)}(t) &= 1 + \int_0^t 1 + sds = 1 + t + \frac{1}{2}t^2 \\
&\vdots \\
u^{(k)}(t) &= \sum_{n=0}^{k} \frac{t^n}{n!} \\
&\qquad \text{(this is a truncated taylor series)} \\
\implies \lim_{k\to\infty} u^{(k)}(t) &= \sum_{n=0}^{\infty} \frac{t^n}{n!} = e^t
\end{aligned}$$

Which is the exact solution.

For another example, consider

$$\dot{u} = u^2, \quad u(0) = 1$$

This has exact solution $u(t) = (1-t)^{-1}$. In terms of uniqueness and existence, this will only be valid for $|t| < 1$ The Taylor series is

$$(1-t)^{-1} = \sum_{n=0}^{\infty} t^m$$

Which converges for $|t| < 1$.

Picard solution

$$\begin{aligned}
u^{(k)}(t) &= 1 + \int_0^t \left(u^{(k-1)}(s)\right)^2 ds \\
u^{(1)}(t) &= 1 + \int_0^t 1^2 ds = 1 + t \\
u^{(2)}(t) &= 1 + \int_0^t (1+s)^2 ds = 1 + t + t^2 + \frac{1}{3}t^3 \\
u^{(3)}(t) &= 1 + \int_0^t (1 + s + s^2 + \frac{1}{3}s^3)^2 ds \\
&= 1 + \int_0^t 6^{th} \text{ order polynomial} ds \\
&= 7th \text{ order polynomial}
\end{aligned}$$

This gets grotty real quick - so we use `Matlab`.

To prove the Picard- Lindelöf theorem:

- Show Picard iterates exist, i.e.

$$u_0 + \int_{t_0}^t f(s, u(s))ds \in J$$

For $u(t) \in J$

- Show Picard iterates converge
- Show they converge to a solution of the IVP.
- Show that the solution is unique

It is in the lecture notes, but we won't go through it in lectures (and we will not be tested on the proof).

We will be expected to **know** the theorem and **apply** it.

In proving the first step, we find the maximum length of time we can guarantee a unique solution after the initial time is

$$\epsilon = \min\left\{\alpha, \frac{\delta}{M}\right\}$$

Where α is taken from $I = [t_0 - \alpha, t_0 + \alpha]$. I.e. α is the radius of the interval, δ is taken from the spatial interval; $J = [u_0 - \delta, u_0 + \delta]$, and we define M as

$$M = ||f|| = \sup_{I\times J} |f|, \ i.e. \ |f| \leq M \ on \ I \times J$$

Consider $\dot{u} = u$ with $u(0) = 1$. Noting the exact unique solution $u(t) = e^t$.

Choose $\alpha > 0$, $\delta > 0$, and define $I = [-\alpha, \alpha]$ and $J = [1 - d, 1 + \delta]$.

We know that $f(t, u) = u$ is continuous and globally Lipschitz continuous. Hence we can apply the PL theorem on any domain.

By the PL thm, $\exists \epsilon > 0$ such that the IVP has a unique solution for $t \ in [-\epsilon, \epsilon]$, where

$$\epsilon = \min\left\{\alpha, \frac{\delta}{M}\right\}$$

Where

$$M = ||u|| = \sup_{I\times J} |u|$$

Since the ODE is autonomous, we only care about the interval J, and sup will be the end point, i.e.

$$M = ||u|| = \sup_{J} |u| = 1 + \delta$$

Thus

$$\frac{\delta}{M} = \delta/(1 + \delta)$$

Increases with δ, and

$$\lim_{\delta\to\infty} \frac{\delta}{1 + \delta} = 1$$

Since α doesn't appear here and we have no restrictions on α, we can pick ANY α. Hence

$$\epsilon = \min\{\alpha, \delta/(1 + \delta)\} = \min\{\infty, 1\} < 1$$

For $\delta < \infty$.

It follows that the PL theorem establishes existence and uniqueness, but only for time $t \in [-\epsilon, \epsilon] = (-1, 1)$.

Consider the ODE

$$\frac{dx}{dt} = x^{1/3} = f(t, x)$$

With $x(0) = 0$.

Set $I = [t_0 - \alpha, t_0 + \alpha] = [-\alpha, \alpha]$ and $J = [x_0 - \delta, x_0 + \delta] = [-\delta, \delta]$

Is $f(t,x)$ lipschitz continuous?

$$\begin{aligned}|f(t,x_2)-f(t,x_1)| &= |x_2^{1/3}-x_1^{1/3}| \\ &= \frac{1}{3}|c^{-2/3}|x_2-x_1|\end{aligned}$$

By blindly applying the MVT, where $c \in (x_1, x_2)$ wlog.

Note that

$$\lim_{c\to 0} c^{-2/3} = \infty$$

I.e. if $c = 0$ then we cannot use the MVT for a bound on Lipschitz continuity.

Hence $f(t,x)$ is not Lipschitz continuous on any interval J containing $x_0 = 0$. We cannot apply the PL theorem now, and so we cannot guarantee there is a unique solution.

0.4.4 Error, Conditioning and Stability

A numerical solution always contains error. We need this to be small enough that the numerical solution is useful.

Error

Two types of error

1. Truncation error (e_T), associated with the mathematical process - typically by truncating an infinite series (Taylor series).
2. Round-off or Machine error (e_R), associated with computer arithmetic.

The local error e is the sum of the two

$$e = e_T + e_R$$

We want to keep this as small as possible.

We can also measure absolute and relative errors in an approximation. These are (respectively)

$$\begin{aligned}e_{abs} &= |x-\bar{x}| \\ e_{rel} &= \left|\frac{x-\bar{x}}{x}\right|, \quad x \neq 0\end{aligned}$$

Where x is the exact solution and $\bar{x}$ is an approximate solution. Note that e_{rel} is a non-dimensional value.

Example of truncation error: Consider evaluating the function $f(t)$ some small time step h in the future using only data at time t. If f has $N+1$ continuous derivatives

$$f^{(n)}(t) = \frac{d^n f}{dt^n}$$

For $n = 1, \ldots, N+1$.

Then

$$f(t+h) = f(t) + hf'(t) + \frac{h^2}{2}f^{(2)}(t) + \ldots + \frac{h^N}{N!}f^{(N)}(t) + R_N(\xi)$$

Where

$$R_N(\xi) = \frac{h^{N+1}}{N+1!} f^{(N+1)}(\xi) \quad \text{For } \xi \in [t, t+h]$$

Truncate the series by dropping $R_N(\xi)$.

$$f(t+h) \approx \sum_{n=0}^{N} \frac{h^n}{n!} f^{(n)}(t)$$

We say the truncation error is

$$e_T = R_N(\xi)$$

Which has order $\mathcal{O}(h^{N+1})$. If there exists a constant k such that

$$|e_T| < Kh^{N+1}$$

For $h \in (0, h_{max})$ Hence

$$K = \max_{\xi \in (t, t+h_{max})} \left| f^{(N+1)}(N+1)! \right|$$

We say the approximation is N^{th} order accurate.

Conditioning

A problem is *well conditioned* if small changes in the input data only makes small changes in the solution. Otherwise it is ill or poorly conditioned.

Consider a scalar function $f(x)$. A small change/perturbation to the input data x (say to $x + e = x(1+\epsilon)$ where $\epsilon = e/x$). With $\epsilon \ll 1$. This yields the solution $f(x+e)$. To determine whether $f(x)$ is well conditioned, we need to check if

$$f(x+e) - f(x)$$

has small magnitude. Assuming f is differentiable, we can use taylor series:

$$f(x+e) = f(x) + ef'(\xi), \quad x \leq \xi \leq x+e$$

For small e and well behaved f, we have $f'(\xi) \approx f'(x)$, such that

$$f(x+e) - f(x) \approx ef'(x)$$

Then $f(x)$ is well conditioned if the absolute error $ef'(x)$ is small magnitude for all possible inputs x. Otherwise it is ill-conditioned.

Ideally we would measure this smallness as relative error.

The condition number is defined as

$$c = \left| \frac{\text{relative error in solution}}{\text{relative error in data}} \right|$$

A well conditioned problem has c close to unity, large c means it is ill-conditioned. For IVPs, conditioning is often measured in terms of absolute error.

An example; considering

$$\dot{x} = -x, \quad x(0) = 1$$

The exact solution is

$$x(t) = e^{-t} \equiv x(t:1)$$

We are considering the initial value as a parameter.

Now consider the perturbed problem

$$x(0) = 1 + \epsilon, \quad \epsilon \ll 1$$

Then

$$x(t) = (1+\epsilon)e^{-t} \equiv x(t : 1+\epsilon)$$

The absolute error:

$$x(t : 1+\epsilon) - x(t : 1 = \epsilon) = e^{-t}$$

With max value ϵ at $t = 0$ and is decreasing monotonically as $t \to \infty$. We deduce the IVP is well conditioned

For this problem the relative error would be

$$e_{rel} = \left| \frac{\epsilon e^{-t}}{e^{-t}} \right| = |\epsilon| \ll 1$$

So the condition number is

$$c = \left| \frac{e_{rel}}{input_{rel}} \right| = \left| \frac{\epsilon}{\epsilon} \right| = 1$$

So it is definitely well conditioned.

Another example

$$\frac{dx}{dt} = x, \quad x(0) = 1$$

Exact sol

$$x(t) = e^t \equiv x(t : 1)$$

Perturbed solution $x(0) = 1 + \epsilon$

$$x(t) = (1+\epsilon)e^t \equiv x(t : 1+\epsilon)$$

The absolute error would be

$$\left|(1+\epsilon)e^t - e^t\right| = |\epsilon|e^t$$

Which grows unboundedly as $t \to \infty$. Hence the problem is ill-conditioned under this measure

The relative error is

$$\left| \frac{\epsilon e^t}{e^t} \right| = |\epsilon| \ll 1$$

And the condition number is

$$\frac{\epsilon}{\epsilon} = 1$$

Which indicates this is a well conditioned problem under this measure.

So there is a bias depending on which measure you use!

Stability

An algorithm is (numerically) stable if it always produces the solution to a nearby problem, otherwise it is unstable.

Consider a scalar function $f(x)$ and suppose that we evaluate it using an algorithm $\text{alg}(x)$. The absolute error in the solution is

$$|\text{alg}(x) - f(x)|$$

Suppose that we can show that

$$\text{alg}(x) = f(x + e)$$

for some small e, i.e. the algorithm yields the solution to the nearby problem of evaluating $f(x+e)$. Then the algorithm is numerically stable.

At each step an approximation gains an amount of error $e = (e_T + e_R)$. This error then affects the accuracy of the approximation at subsequent steps. We say the error propagates. **If the error grows as it propagates then the algorithm is unstable, if it decays then it is stable.**

Conditioning and Stability

We will hence use an *algorithm* to solve a *problem*. It is important to distinguish between **difficult** problems and **bad** algorithms. A numerical solution my be bad because

1. The problem is ill-conditioned;
2. The numerical method is not stable;
3. Or both

To get a good solution, we require a well-conditioned problem and a stable numerical method. The solution we obtain will be to a problem near the one we are trying to solve (so it is stable) and the error will be small (well conditioned).

We can choose a new algorithm, but we can't quite pick a new problem. It is hard to improve conditioning, but it is easy to change algorithm to a more stable one. As we consider numerical solutions for IVPs, we will **assume that we have well conditioned problems**, and focus mainly on numerical stability.

0.4.5 Finite Differences

We start with a familiar IVP

$$\frac{dx}{dt} = f(t,x), \quad x(t_0) = x_0$$

We know everything on the RHS, but we need to fix up the derivative. The idea behind Finite Difference methods is to use a Taylor series approximation.

We don't know the slope at $x(t)$ but we do know the value of $x(t)$ and $x(t+h)$. We can take the slope with

$$slope = \frac{x(t+h) - x(t)}{h}$$

Using Taylor series:

$$\begin{aligned} x(t+h) &= x(t) + hx'(t) + \frac{h^2}{2!}x''(\xi) \\ hx'(t) &= x(t+h) - x(t) \\ x'(t) &= \frac{x(t+h) - x(t)}{h} - \frac{h}{2}x''(\xi) \end{aligned}$$

Where $\xi \in [t, t+h]$

Since the first term is $\mathcal{O}(1/h)$ and the second is $\mathcal{O}(h)$:

$$\frac{x(t+h) - x(t)}{h} \gg \frac{h}{2}x''(\xi) \quad as\ h \to 0$$

And so the truncation error $e_T = \frac{h}{2}x''(\xi) = \mathcal{O}(h)$.

Now substitute this into the ODE

$$\frac{x(t+h)-x(t)}{h}+\mathcal{O}(h)=f(t,x)$$
$$\frac{x_{j+1}-x_j}{h}+\mathcal{O}(h)=f(t_j,x_j)$$
$$x_{j+1}=x_j+h(f(t_j,x_j))+\mathcal{O}(h^2)$$

Where $t_j=t_j+jh$ and $x_j\approx x(t_j)$

This is forward Euler's method (forward finite differences).

It has local discretisation error $\ell(h)=\mathcal{O}(h^2)=he_T$. When we say "local" we refer to the error in taking 1 step using this method.

The global error from taking multiple steps is obtained by:

Suppose we take N steps. $t_N=t_0+Nh$, then the global discretisation error is

$$g(h)=N\times\mathcal{O}(h^2)=\frac{t_N-t_0}{h}\mathcal{O}(h^2)$$
$$=\mathcal{O}(h)$$

Hence Euler's method is 1st order accurate.

Another method: backwards finite differences, so using $t-h$ instead of $t+h$. The slope of the tangent

$$\frac{x(t)-x(t-h)}{h}$$

Taylor series in $x(t-h)$

$$x(t-h)=x(t)-hx'(t)+\frac{h^2}{2}x''(\xi),\quad \xi\in[(-h,t)$$
$$x'(t)=\frac{x(t)-x(t-h)}{h}+\mathcal{O}(h)$$

With $e_T=\mathcal{O}(h)$, substitute into the ODE

$$\frac{x_j-x_{j-1}}{h}+\mathcal{O}(h)=f(t_j,x_j)$$
$$x_j=x_{j-1}+hf(t_j,x_j)+\mathcal{O}(h^2)$$

This is the backwards Euler method (surprise surprise). Just like the forward method,

$$\ell(h)=\mathcal{O}(h^2),\quad g(h)=\mathcal{O}(h)$$

I.e. the same as the forward Euler method.

Since the point we try to calculate depends on the point we are trying to calculate i.e. $x_j=\ldots f(x_j,x_j)$, it can be a bit harder to use backwards Euler.

If we combine the two methods, we get Central Differences

$$x(t+h)=x(t)+hx'(t)+\frac{h^2}{2}x''(t)+\frac{h^3}{3!}x'''(\xi^+)$$
$$x(t-h)=x(t)-hx'(t)+\frac{h^2}{2}x''(t)-\frac{h^3}{3!}x'''(\xi^-)$$

By subtracting the two, we get

$$x(t+h) - x(t-h) = 2hx'(t) + \mathcal{O}(h^3)$$
$$x'(t) = \frac{x(t+h) - x(t-h)}{2h} + \mathcal{O}(h^2)$$

Hence $e_T = \mathcal{O}(h^2)$. Which is better than using forward/backwards.

This is the central difference method for x'.

Substituting into the ODE gives

$$\frac{x_{j+1} - x_{j-1}}{2h} + \mathcal{O}(h^2) = f(t_j, x_j)$$
$$x_{j+1} = x_{j-1} + 2hf(t_j, x_j) + \mathcal{O}(h^3)$$

Hence

$$\ell(h) = \mathcal{O}(h^3), \quad g(h) = \mathcal{O}(h^2)$$

This is known as the leapfrog method, which is second order accurate. The dilemma here is that this extra order of accuracy does not inherently imply a better approximation.

So we have forward, backward and central differences, which are (respectively):

$$x'_j = \frac{x_{j+1} - x_j}{h} + \mathcal{O}(h)$$
$$x'_j = \frac{x_j - x_{j-1}}{h} + \mathcal{O}(h)$$
$$x'_j = \frac{x_{j+1} - x_{j-1}}{2h} + \mathcal{O}(h^2)$$

These are the **finite difference formulae**.

With corresponding Euler methods for the ODE $\dot{x} = f(t, x)$:

$$x_{j+1} = x_j + hf(t_j, x_j) + \mathcal{O}(h^2)$$
$$x_{j+1} = x_j + hf(t_{j+1}, x_{j+1}) + \mathcal{O}(h^2)$$
$$x_{j+1} = x_{j-1} + 2hf(t_j, x_j) + \mathcal{O}(h^3)$$

We also call these finite difference formulae, but we will call them Update/time stepping formulae. The $\mathcal{O}()$ are local truncation errors, which we denote $\ell(h)$. We are moreso interested in the Global truncation error

$$g(h) = \frac{1}{h}\ell(h)$$

TIL - Josh is bad at latex and is slow.

The generalised finite difference formulae for $\frac{d^r x}{dt^r}$

$$x^{(r)}(t_j) \equiv x_j^{(r)}$$

$$x_j^{(r)} = \sum_{i=-p}^{q} a_i x_{j+i} + \mathcal{O}(h^m)$$

For some constants a_i to derive an order m.

E.g.

Let $r = 1$, $p = 0$ and $q = 2$

$$\begin{aligned} x'_j &= \sum_{i=0}^{2} a_i x_{j+i} + \mathcal{O}(h^m) \\ &= a_0 x_j + a_1 x_{j+1} + a_2 x_{j+2} + \mathcal{O}(h^m) \\ &= a_0 x_j + a_1 \left(\sum_{k=0}^{\infty} \frac{h^k}{k!} x_j^{(k)} \right) + a_2 \left(\sum_{k=0}^{\infty} \frac{(2h)^k}{k!} x_j^{(k)} \right) + \mathcal{O}(h^m) \end{aligned}$$

Equate coefficients for the derivatives

$$\begin{aligned} x_j : \ & a_0 + a_1 + a_2 = 0 \\ x'_j : \ & h a_1 + 2h a_2 = 0 \\ x''_j : \ & \frac{h^2}{2} a_1 + \frac{4h^2}{2} a_2 = 0 \end{aligned}$$

Now solve the system: Use (3), then sub into (2), then sub into (1).

$$\begin{aligned} a + 4a_2 = 0 \implies a_1 &= -4a_2 \\ a_1 + 2a_2 &= \frac{1}{h} \\ -4a_2 + 2a_2 &= \frac{1}{h} \\ a_2 &= \frac{-1}{2h} \\ \implies a_1 = -4a_2 &= \frac{4}{2h} \\ \implies a_0 = -a_1 - a_2 &= -\frac{3}{2h} \end{aligned}$$

Alternatively just use matrices in matlab right?
Hence

$$x'_j = \frac{1}{2h} \left(-3x_j + 4x_{j+1} - x_{j+2} \right) + \mathcal{O}(h^m)$$

Note that the coefficients sum up to 0. This is a handy check for consistency. What is the interpretation of this? This holds for any function, and we require that a constant function must have zero derivative.

How accurate is this? I.e. what is m? To obtain this, look at the next term for those Taylor series, x'''_j.

$$\frac{h^3}{6} a_1 + \frac{8h^3}{6} a_2 = -\frac{h^3}{8} \frac{2}{h} = \frac{-h^2}{4}$$

Hence $m = 2$. I.e. The error is $\mathcal{O}(h^2)$, and so the formula is second order accurate. If the coefficient of x'''_j was 0, we would then look at $x_j^{(4)}$. And iterate until we get a non-zero coefficient.

Note, we have $p + q + 1 = 3$ unknowns (a_0, a_1, a_2), and so we needed 3 equations (corresponding to x_j, x'_j, x''_j). We need $r + 1 = 2$ equations for x_j and x'_j. So we actually had an oversubscribed system. In Luke's words, the equation for x''_j is a *bonus*.

Explicit versus Implicit methods

Forwards Euler method (a.k.a Explicit Euler's method)

$$x_{j+1} = x_j + hf_j$$

Where $f_j := f(t_j, x_j)$.

This is an explicit method, since the new term does not appear on the RHS. I.e.

$$\begin{aligned} x_0 &= x(t_0) \\ x_1 &= x_0 + hf_0 \\ x_2 &= x_1 + hf_1 \end{aligned}$$

Which is explicit, since x_1 doesn't appear on the RHS and we know x_0. Same thing for x_2 and so on.

Backwards Euler (a.k.a Implicit Euler's method)

$$x_{j+1} = x_j + hf_{j+1}$$

Since x_{j+1} appears on both sides and may not necessarily be explicitly extracted, it is implicit. So given x_0

$$\begin{aligned} x_1 &= x_0 + hf_1 \\ x_1 - hf_1 &= x_0 \end{aligned}$$

In general we don't know how to solve this, i.e. we don't know if we can find x_1 exactly. Lets say $f(t, x) = x$ then

- Explicit Euler:

$$\begin{aligned} x_{j+1} &= x_j + hx_j \\ x_{j+1} &= x_j(1 + h) \end{aligned}$$

- Implicit Euler:

$$\begin{aligned} x_{j+1} &= x_j + hx_{j+1} \\ (1 - h)x_{j+1} &= x_j \\ x_{j+1} &= \frac{x_j}{1 - h} \\ x_{j+1} &= (1 - h)^{-1}x_j \\ x_{j+1} &= (1 + h + h^2 + \ldots)x_j \end{aligned}$$

The last step shows the relationship between the two.

So this one was possible, though multiplication is nicer.

What if $f = e^{-x}$

- Explicit Euler

$$x_{j+1} = x_j + he^{-x_j}$$

- Implicit Euler

$$\begin{aligned} x_{j+1} &= x_j + he^{-x_{j+1}} \\ x_{j+1} - he^{-x_{j+1}} &= x_j \end{aligned}$$

Not easy to find x_{j+1} explicitly here.

Consistency

A finite difference formula is considered *consistent* with an ODE if the local discretization error $\ell(h) \to 0$ as $h \to 0$. This is equivalent to saying that at any fixed point t_i in the domain, the finite difference formula becomes the same as the ODE as $h \to 0$. We can use *consistency analysis* to find this.

Consider the FDF

$$x_{j+1} = x_j + hf_j$$

(Euler's method)

To derive this we took the derivative and approximated it with a taylor series to get the finite differences formula for the derivative. Substitute this into the ODE and then FDF with timestep.

We want to reverse the process. If we instead wrote the (example) FDF as

$$x(t_j + h) = x(t_j) + hf(t_j, x(t_j))$$

And applied a Taylor series to $x(t_j + h)$

$$x(t_j + h) = x(t_j) + hx'(t_j) + \frac{h^2}{2}x''(\xi), \quad \xi \in (t_j, t_j + h)$$

Substitute this back into the example FDF

$$\begin{aligned} x(t_j + h) &= x(t_j) + hf(t_j, x(t_j)) \\ x(t_j) + hx'(t_j) + \frac{h^2}{2}x''(\xi) &= x(t_j) + hf(t_j, x(t_j)) \\ hx'(t_j) + \frac{h^2}{2}x''(\xi) &= hf(t_j, x(t_j)) \\ x'(t_j) &= f(t_j, x(t_j)) - \frac{h}{2}x''(\xi) \\ x'(t_j) &= f(t_j, x(t_j)), \quad h \to 0 \end{aligned}$$

Hence this FDF is consistent with the ODE

$$\frac{dx}{dt} = f(t, x)$$

And we find $e_T = \mathcal{O}(h)$ and $\ell(h) = \mathcal{O}(h^2)$

Convergence

We have *convergence* to the true solution if the global discretisation error $g(h) \to 0$ as $h \to 0$. This is a stronger requirement than consistency, and is essential for accurate solutions.

Lax's Equivalence theorem states

$$Consistency + Stability = Convergence$$

Stability analysis

We want to determine if an algorithm (an FDF) gives a solution to a nearby problem (i.e. if it is stable).

For example if we applied Euler's method to $\frac{dx}{dt} = -100x$ with $x(0) = 1$, we would hope for

$$x(t) = e^{-100t}$$

always

Euler's method gives

$$\begin{aligned} x_0 &= 1 \\ x_1 = x_0 + h(-100x_0) &= (1-100h)x_0 \\ &\vdots \\ x_n = (1-100h)^n x_0 &= (1-100h)^n \end{aligned}$$

If say $h = 0.1$ then $x_n = (-9)^n$ which GROWS in n, and oscillates (we wanted exponential decay!).

If we however had $h = 0.001$

$$x_n = (1 - (100 * 0.001))^n x_0 = (1-0.1)^n = 0.9^n$$

Which does in fact decay, and we lose the oscillation. Hence Euler's method is stable for $h = 0.001$. More rigorously,

$$\begin{aligned} |1-100h| < 1 \implies 100h &< 2 \\ h &< 0.02 \end{aligned}$$

Since for stability we require $x_n \to 0$ as $n \to \infty$ To remove the oscillation we also require

$$0 < 1 - 100h < 1 \implies h < 0.01$$

(Given h is positive)

So we would say that Euler's method is **conditionally stable** for this IVP.

If we used backwards(implicit) Euler for the same problem, we get

$$\begin{aligned} x_0 &= 1 \\ x_1 = x_0 - 100hx_1 \implies x_1 &= \frac{x_0}{1+100h} \\ &\vdots \\ x_n &= \frac{x_0}{(1+100h)^n} \end{aligned}$$

And $1 + 100h > 0$ for $h > 0$, and hence $x_n \to 0$ as $n \to \infty$ for all $h > 0$ and we have $x_n > 0$ for all n

So backwards Euler is unconditionally stable (its always stable)

We want to generalise the analysis for general IVPs

$$\frac{dx}{dt} = f(t,x), \quad x(t_0) = x_0$$

For Eulers method

$$x_{n+1} = x_n + hf(t_n, x_n)$$

So

$$x(t_n + h) = x(t_n) + hx'(t_n) + \frac{1}{2}h^2 x''(\xi)$$

Combine this with Euler's method (subscripts for approximation, brackets for true val)

$$\begin{aligned} x(t_n+h) - x_{n+1} &= x(t_n) - x_n + h(x'(t_n) - f(t_n, x_n)) + \frac{1}{2}h^2 x''(\xi) \\ &= \frac{1}{2}h^2 x''(\xi), \text{ if } x_n = x(t_n) \end{aligned}$$

Using the fact that $x'(t_n) = f(t_n, x(t_n))$.
But, in general $x_n \neq x(t_n)$.
From the MVT

$$f(t_n, x(t_n)) - f(t_n, x_n) = (x(t_n) - x_n) f_x(t_n, \eta)$$

Where $\eta \in [x_n, x(t_n)]$

$$\implies x(t_{n+1}) - x_{n+1} = (x(t_n) - x_n)\,(1 + h f_x(t_n, \eta)) + \frac{1}{2} h^2 x''(\xi)$$

So what we have is

$$\text{Global error at } n+1 = \text{Global error at } n \times \text{something} + \text{Local error}$$

The 'something' term, $1 + h f_x(t_n, \eta)$ corresponds to the growth of global error. So the global errors grow if

$$|1 + h f_x| > 1$$

Hence, Euler's method is stable if

$$|1 + h f_x| < 1 \implies -2 < h f_x < 0$$

For backwards Euler, we find (we won't bother deriving it) it is stable if

$$|1 - h f_x| > 1 \implies h f_x < 0 \; or \; h f_x > 2$$

Consider the 'test' problem

$$\frac{dy}{dx} = \lambda y, \quad y(0) = 1, \quad \lambda \in \mathbb{C}$$

The stable region of a FDF is the set of values (regions) of $Z = \lambda h \in \mathbb{C}$ for which the method is stable.
Forwards:

$$\begin{aligned} |1 + h f_y| &< 1 \\ |1 + \lambda h| &< 1 \\ |1 + z| &< 1 \end{aligned}$$

So for values of z captured in the unit disc centered at $Re(z) = -1$.
For backwards Euler

$$\begin{aligned} |1 - h f_y| &> 1 \\ |1 - z| &> 1 \end{aligned}$$

This region is for the values of z **outside** of the disc centered at $Re(z) = 1$.
Since the second region is so big, we would prefer backwards.
Backwards Euler is $A - stable$. because it covers $Re|z| < 0$. This would mean $Re(\lambda) < 0$. And hence forwards Euler's method is not $A - stable$.
E.g. Heun's method (improved/modified Euler) id

$$x_{n+1} = x_n + \frac{1}{2} h \{ f(t_n, x_n) + f(t_{n+1}, x^*_{n+1}) \}$$

Where

$$x^*_{n+1} = x_n + hf(t_n, x_n)$$

So we get the simplicity of forwards Euler, with the improved stability of backwards.
Applying Heun's method to the test problem:

$$\begin{aligned} x^*_{n+1} &= x_n(1+\lambda h) \\ x_{n+1} &= x_n + \frac{1}{2}h\{\lambda x_n + \lambda x^*_{n+1}\} \\ x_{n+1} &= x_n\left(1 + \lambda h + \frac{1}{2}(\lambda h)^2\right) \end{aligned}$$

By looking at the Taylor series, we can see how much of an improvement this is:

$$\begin{aligned} x(t_{n+1}) = x(t_n + h) &= x(t_n) + hx'(t_n) + \frac{1}{2}h^2x''(t_n) + \frac{h^3}{6}x'''(\xi) \\ &= x(t_n)\{1 + \lambda h + \frac{1}{2}(\lambda h)^2\} + \frac{1}{6}h^3x'''(\xi) \end{aligned}$$

Looking at the error bit

$$x(t_{n+1}) - x_{n+1} = (x(t_n) - x_n)\left(1 + \lambda h + \frac{1}{2}(\lambda h)^2\right) + \frac{h^3}{6}x'''(\xi)$$

Hence for stability

$$|1 + \lambda h + \frac{1}{2}(\lambda h)^2| < 1$$

$$|1 + z + \frac{1}{2}z^2| < 1, \quad z = \lambda h$$

Which is a slight improvement on Euler's method.

Round-off error

$$e = e_T + e_R$$

Where e_T is truncation and e_R is round-off error.

e_R becomes important as h gets small

Consider the forwards difference equation

$$x'_j = \frac{x_{j+1} - x_j}{h} + e_T, \quad e_T = -\frac{1}{2}hx''(\xi),\ \xi \in (t_j, t_{j+1})$$

Define the round-off error

$$\epsilon_j = x(t_j) - x_j$$

The total error

$$\begin{aligned} e = x'(t_j) - x_j &= x'(t_j) - \frac{x_{j+1} - x_j}{h} \\ &= x'(t_j) - \frac{1}{h}\left\{(x(t_{j+1}) - \epsilon_{j+1}) - (x(t_j) - \epsilon_j)\right\} \\ &= \left\{x'(t_j) - \frac{x(t_{j+1}) - x(t_j)}{h}\right\} + \frac{\epsilon_{j+1} - \epsilon_j}{h} \end{aligned}$$

The term in the curly braces is the truncation error (apparently)

$$e_T = x'(t_j) - \frac{x(t_{j+1}) - x(t_j)}{h} = -\frac{1}{2}hx''(\xi)$$

And the other term is the round off error

$$e_R = \frac{\epsilon_{j+1} - \epsilon_j}{h}$$

Now

$$|e| = |e_T + e_R| \leq |e_T| + |e_R|$$

$$\begin{cases} |e_T| \leq \frac{1}{2}h \max|x''(\xi)| = \frac{1}{2}hk \\ |e_R| \leq \frac{|\epsilon_{j+1}|+|\epsilon_j|}{h} \leq \frac{2\epsilon}{h} \end{cases}$$

Where $\epsilon = \max_j |\epsilon_j|$ Thus

$$|e| \leq \frac{Kh}{2} + \frac{2\epsilon}{h} \to 0 + \infty \; as \; h \to 0$$

So the optimal step size (which minimises the error bound), h_m

$$\frac{d}{dh}\left(\frac{Kh}{2} + \frac{2\epsilon}{h}\right) = 0$$

$$\frac{K}{2} - \frac{2\epsilon}{h_m^2} = 0$$

$$h_m = 2\sqrt{\frac{\epsilon}{K}}$$

Stiff systems

A example stiff problem

$$\frac{dx}{dt} = -\alpha x, \quad x(0) = 1, \; and \; \alpha > 0$$

The exact solution is (astonishingly)

$$x = e^{-\alpha t}$$

We would expect to require small step sizes for small values of t, and large step sizes for larger values of t. This is to accommodate for the rapid change in x for small t, and near zero change for large t.

One issue is we have made the assumption that we have only one step size which doesn't change - so we can't do that.

Another issue - lets try using Euler's method on this. We would require for stability:

$$|1 - \alpha h| < 1 \implies h < \frac{2}{\alpha}$$

This actually is a boundary layer problem - there is a boundary layer near $t = 0$ of size $\approx \frac{1}{\alpha}$. The problem is stiff due to this boundary layer.

Linearisation

Just about everything is non-linear.

Consider the IVP

$$\frac{dx}{dt} = -x^2, \quad x(0) = 1$$

Lets ignore the fact that this is separable and can be solved easily explicitly. If we want to use a solver, e.g. Euler's method

$$x_{n+1} = x_n - hx_n^2$$

Which is stable for

$$|1 - 2hx| < 1$$

Note this stability actually depends on x.

In this case x is decreasing with time and is always positive (since there is a fixed point at $x^* = 0$, hence $0 < x \leq 1$. Hence

$$|1 - 2hx| < 1 \implies 0 < h < 1$$

Is the interval of stable step sizes.

Implicit methods are not straightforward to apply, e.g. backwards Euler

$$\begin{aligned} x_{n+1} &= x_n - hx_{n+1}^2 \\ hx_{n+1}^2 + x_{n+1} - x_n &= 0 \\ x_{n+1} &= \frac{-1 \pm \sqrt{1 + 4hx_n}}{2h} \end{aligned}$$

Which is an explicit expression in terms of x_n. Do we want the positive or negative branch of the square root? For small h, the square root becomes

$$\sqrt{1 + 4hx_n} = 1 + \frac{1}{2}4hx_n + \mathcal{O}(h^2)$$

$$\implies x_{n+1} = \frac{1}{2h}\{-1 \pm (1 + 2hx_n + \mathcal{O}(h^2))\}$$

$$= \begin{cases} \frac{1}{2h}(-2 - 2hx_n + \mathcal{O}(h^2) & +ve\ branch \\ \frac{1}{2h}(2hx_n + \mathcal{O}(h^2) = x_n + \mathcal{O}(h) & -ve\ branch \end{cases}$$

So we want the positive branch.

We could alternatively linearise the problem, by writing (read: cheating)

$$\frac{dx_n}{dt} = -x_n x_{n+1}$$

To give

$$x_{n+1} = x_n - hx_n x_{n+1} \implies x_{n+1} = \frac{x_n}{1 + hx_n}$$

Which is explicit, but we've kinda cheated not only the game, but ourselves.

This is unconditionally stable for $x > 0$ (that sounds like a condition to me).

0.4.6 Predictor-Corrector (PECE) methods

Because PC is too PC. The E's mean evaluate.

The general PECE algorithm is

1. Step from x_n to x_{n+1}
2. Predict: Using an explicit method to give x^p_{n+1}
3. $$x_{n+1^p} = x_n + hf_n$$
4. Evaluate $f^p_{n+1} = f(t_{n+1}, x^p_{n+1})$
5. Correct: Using an implicit method to give x^c_{n+1}
$$x^c_{n+1} = x_n + hf^p_{n+1}$$
6. Evaluate: $f^c_{n+1} = f(t_{n+1}, x^c_{n+1})$
7. Set $x_{n+1} = x^c_{n+1}$ or, we can repeat the CE steps.
8. If we repeat m times, we get $PE(CE)^m$.

We have already met a PECE method - Heun's method.

Accuracy is usually judged based on the size of $|x^p_{n+1} - x^c_{n+1}|$ since (for exmaple with Heun's method). Forwards Euler, the local error is

$$x_{n+1} - x(t_{n+1}) = -\frac{1}{2}h^2x''(\xi_1)$$

Backwards Euler, the local error is

$$x_{n+1} - x(t_{n+1}) = -\frac{1}{2}h^2x''(\xi_2)$$

$$\begin{aligned} x^p_{n+1} - x^c_{n+1} &= \left(x^p_{n+1} - x(t_{n+1})\right) - \left(x^c_{n+1} - x(t_{n+1})\right) \\ &= -\frac{1}{2}h^2x''(\xi_1) - \frac{1}{2}h^2x''(\xi_2) \\ &= -\frac{1}{2}h^2\left(x''(\xi_1) + x''(\xi_2)\right) \\ &\approx -h^2x''(\xi) = 2\ell(h) = \mathcal{O}(h^2) \end{aligned}$$

0.5 Extensions of numerical methods

0.5.1 Higher-order IVPS

Consider an n^{th} order ODE

$$\frac{d^nx}{dt^n} = f(t, x, x', \ldots, x^{(n-1)})$$

With ICs

$$x(0) = c_0, x'(0) = c_1, x^{(n-1)}(0) = c_{n-1}$$

To solve it, convert it into a system of first order ODEs

$$y_0 = x,\ y_1 = x', \ldots, y_{n-1} = x^{(n-1)}$$

Such that $y_i' = y_{i+1}$ with IC $y_i(0) = c_i$

For the last one it gets interesting

$$y_{n-1}' = x^{(n)} = f(t, y_0, \ldots, y_{n-1})$$

Is a system of n first order ODEs, with n initial conditions.

Suppose f is linear in all y's, i.e.

$$f = \alpha_0(t)y_0 + \alpha_1(t)y_1 + \ldots + \alpha_{n-1}(t)y_{n-1}$$

(Note we have made this homogeneous)

Then we can write the system in matrix vector form

$$\mathbf{y}' = M\mathbf{y}$$

Where

$$y = \begin{pmatrix} y_0 \\ \ldots \\ y_{n-1} \end{pmatrix}, \quad M = \begin{pmatrix} 0 & 1 & 0 & & \ldots & 0 \\ 0 & 0 & 1 & 0 \ldots 0 & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ldots & & & 1 \\ \alpha_0 & \alpha_1 & & \ldots & & \alpha_{n-1} \end{pmatrix}$$

E.g. consider

$$x'' = x, \quad x(0) = 1, \quad x'(0) = -1$$

Write $y_0 = x$ and $y_1 = x'$, and

$$f(t, x, x') = x \implies f(t, y_0, t_1) = y_0$$

$$y_0' = y_1, \quad y_1' = f = y_0$$

With $y_0(0) = 1, \quad y_1(0) = -1$

Since f is linear, we can use the matrix vector form, with $\alpha_0 = 1$ and $\alpha_1 = 0$

$$\mathbf{y}' = M\mathbf{y} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\mathbf{y}, \quad \mathbf{y}(0) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Suppose we want to solve the general system using backwards Euler, we have

$$y_0^{(m)} = y_0^{(m-1)} + hy_1^{(m)}$$

And in general

$$\begin{aligned} y_0^{(m)} - hy_1^{(m)} &= y_0^{(m-1)} \\ y_1^{(m)} - hy_2^{(m)} &= y_1^{(m-1)} \\ &\vdots \\ y_i^{(m)} - hy_i^{(m)} &= y_{i-1}^{(m-1)} \\ &\vdots \\ y_{n-1}^{(m)} - hf_m &= y_{n-1}^{(m-1)} \end{aligned}$$

If f is linear, we write this as the matrix system $M\mathbf{y}^{(m)} = \mathbf{y}^{(m-1)}$

Where

$$M = \begin{pmatrix} 1 & -h & 0 & \dots & 0 & \\ 0 & 1 & -h & 0 & \dots & 0 \\ & & \ddots & \ddots & & \\ & & & & 1 & -h \\ \beta_0^m & \beta_1^m & \dots & & \beta_{n-2}^m & 1+\beta_{n-1}^m \end{pmatrix}$$

Where $\beta_j^m = -h\alpha_j(t_m)$

We require the matrix to be 'diagonally dominant' for stability. so we would require $h < 1$.

There is an alternative method - a more direct approach to solve n^{th} order IVPs, but will be awful for large n.

Consider a second order IVP

$$x'' = f(t, x, x'), \quad x(0) = c_0,\ x'(0) = c_1$$

Derive an FDF for the IVP.

Write the central difference formula for the second derivative

$$x_n'' \approx \frac{x_{n+1} - 2x_n + x_{n-1}}{h^2} = f(t_n, x_n, x_n')$$

$$x_{n+1} = 2x_n - x_{n-1} + h^2 f_n$$

Where $f_n = f(t_n, x_n, x_n')$

If x' appears in f, we can use FDF on x' too.

$$x_n' = \frac{x_{n+1} - x_{n-1}}{2h}$$

The errors in the derivative approximations x', x'' have

$$e_T = \mathcal{O}(h^2)$$

If we used the forwards/backwards approximations for x', we would have

$$e_T = \mathcal{O}(h)$$

Then we risk compromising the accuracy of the FDF.

Even if x' didn't appear in f, we require it for the initial condition.

$$x'(0) = c_1 \approx \frac{x_1 - x_{-1}}{2h} \implies x_{-1} = x_1 - 2hc_1$$

Lets consider

$$x'' + a(t)x' + b(t)x = r(t)$$

Gives the FDF

$$\frac{x_{n+1} - 2x_n - x_{n-1}}{h^2} + a_n \frac{x_{n+1} - x_{n-1}}{2h} + b_n x_n = r_n$$

Where $a_n = a(t_n)$ and likewise for $b_n = b(t_n)$.

Simplification (rearrange and multiply by h^2 gives)

$$A_n x_{n+1} - B_n x_n + C_n x_{n-1} = D_n$$

Where

$$A_n = 1 + \frac{h}{2}a_n, \quad B_n = 2 - h^2 b_n, \quad C_n = 1 - \frac{h}{2}a_n, \quad D_n = h^2 r_n$$

Hence

$$x_{n+1} = \frac{B_n x_n - C_n x_{n-1} + D_n}{A_n}$$

Use $n = 0$ for the first step, i.e.

$$x_1 = \frac{B_0 x_0 - C_0 x_{-1} + D_0}{A_n}$$

What does x_{-1} mean? Appeal to our ICs

$$x_0 = c_0, \quad x_{-1} = x_1 - 2hc_1$$

$$\implies x_1 = \frac{B_0 c_0 + 2hc_1 C_0 + D_0}{A_0(1 + C_0/A_0)}$$

0.5.2 Boundary Value Problems (BVPs)

Use `bvp4c()` problem solved.

Consider the second order ODE-BVP

$$\frac{d^2y}{dx^2} = f(x, y, \frac{dy}{dx})$$

For $x \in [a, b]$ with BCs $y(a) = \alpha, y(b) = \beta$

The Dirichlet BCs since they are both for the function rather than derivatives.

Numerical solution is almost the same process as before.

$$y'' + a(x)y' + b(x)y = r(x)$$

Let $h = \frac{b-a}{N}$. Where N is the number of points we want to consider, and let $x_n = x_0 + nh$. With $x_0 = a$, $x_N = b$.

$$y_{n+1} = \frac{B_n y_n - C_n y_{n-1} + D_n}{A_n}$$

$$A_n y_{n+1} - B_n y_n + C_n y_{n-1} = D_1$$

Now the BCs become $y_0 = \alpha$, $y_N = \beta$.

We have

$$\begin{aligned}
n = 1 &: A_1 y_2 - B_1 y_1 + C_1 y_0 = D_1 \\
&\implies A_1 y_2 - B_1 y_1 = D_1 - \alpha C_1 \\
n = 2 &: A_2 y_3 - B_2 y_2 + C_2 y_1 = D_2 \\
&\vdots \\
n = N - 1 &: A_{N-1} y_n - B_{N-1} y_{N-1} + C_N y_{N-2} = D_{N-1} \\
&\implies - B_{N-1} y_{N-1} + C_N y_{N-2} = D_{N-1} - \beta A_{N+1}
\end{aligned}$$

This is a system of equations that have to be solved simultaneously. Note that the system is sparse (tridiagonal) so its not that hard to work with.

0.5.3