Ordinary Differential Equations

1 Introduction: first order ODE

We are given

- a function $f(t, y)$ which describes a "direction field" in the (t, y) plane
- an initial point (t_0, y_0)

We want to find a function $y(t)$ for $t \in [t_0, T]$ such that

- $y(t_0) = y_0$ "initial condition"
- $y'(t) = f(t, y(t))$ for $t \in [t_0, T]$ "ordinary differential equation" (ODE)

This is called an **initial value problem** (IVP).

The partial derivative $f_y(t, y) = \dfrac{\partial f}{\partial y}(t, y)$ is important for the behavior of the differential equation.

Theorem 1.1. *Assume that $f(t, y)$ and $f_y(t, y)$ are continuous for $t \in [t_0, T]$, $y \in \mathbb{R}$.*
For a given initial value y_0 there is a unique solution $y(t)$ of the initial value problem. Either the solution exists for all $t \in [0, T]$, or it only exists on a smaller interval $[t_0, t_*)$ with $t_0 < t_* < T$.

We can solve the IVP in Matlab with **ode45**:

```
f = @(t,y) ...                  % define function f(t,y)
[ts,ys] = ode45(f,[t0,T],y0);   % find column vectors ts,ys with values of solution
ys(end)                         % value y(T) of solution
plot(ts,ys)                     % plot solution
```

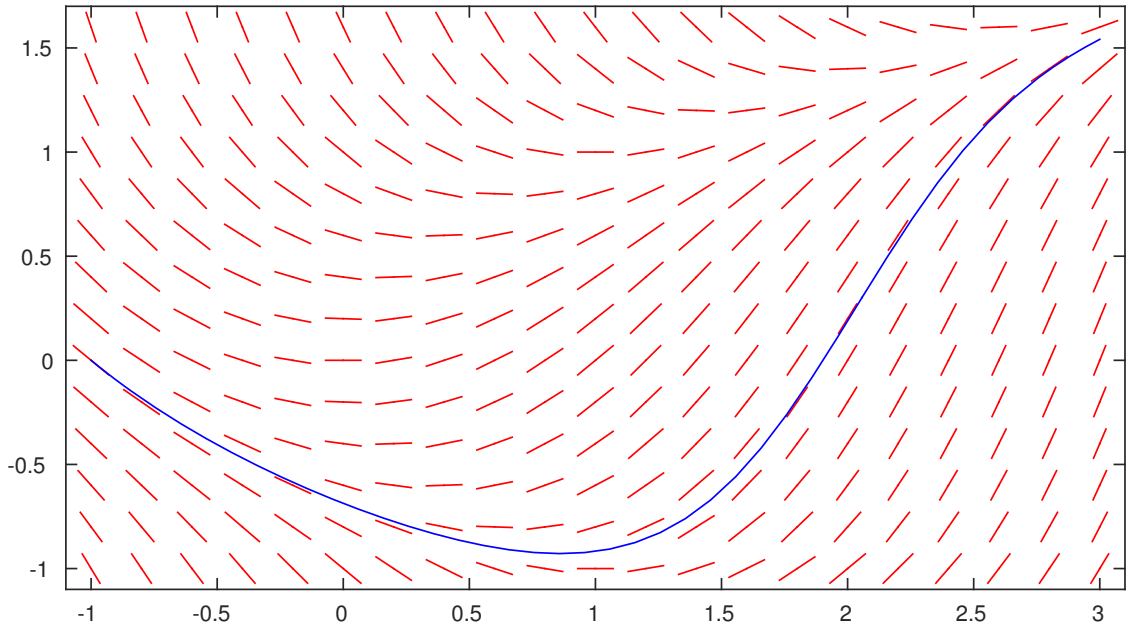
Example: Find a function $y(t)$ for $t \in [-1, 3]$ such that

$$\begin{aligned} y'(t) &= t - y(t)^2 \\ y(-1) &= 0 \end{aligned}$$

Here we have $t_0 = -1$, $y_0 = 0$, $f(t, y) = t - y^2$.

Numerical solution in Matlab: (using m-file `dirfield.m` from course web page)

```
f = @(t,y) t-y^2                          % define function f(t,y)
dirfield(f, -1:.2:3, -1:.2:1.6); hold on % plot direction field
[ts,ys] = ode45(f,[-1,3],0);              % solve IVP for t from -1 to 3, initial value 0
                                          % this gives vectors ts,ys
plot(ts,ys,'b'); hold off                 % plot solution
```



What happens if we perturb the initial value y_0?

Theorem 1.2. *Let $y(t)$ denote the solution of the IVP with initial condition $y(t_0) = y_0$, let $\tilde{y}(t)$ denote the solution of the IVP with initial condition $\tilde{y}(t_0) = \tilde{y}_0$. Assume $\boxed{f_y(t,y) \le M}$ for $t \in [t_0, T]$, $y \in \mathbb{R}$. Then*

$$|\tilde{y}(t) - y(t)| \le |\tilde{y}_0 - y_0|\, e^{M(t-t_0)} \qquad \text{for } t \in [t_0, T]$$

For $M < 0$ the difference $|\tilde{y}(t) - y(t)|$ decays exponentially for increasing t. For $M > 0$ the difference may increase exponentially.

We call the ODE **unstable** if we have $\boxed{f_y(t,y) > 0}$ for all $t \in [t_0, T]$, $y \in \mathbb{R}$.

We call the ODE **stable** if we have $\boxed{f_y(t,y) < 0}$ for all $t \in [t_0, T]$, $y \in \mathbb{R}$.

2 System of ODEs, higher order ODEs

We want to find n functions $y_1(t), \ldots, y_n(t)$ for $t \in [t_0, T]$ satisfying the differential equations

$$\begin{aligned} y_1'(t) &= f_1\left(t, y_1(t), \ldots, y_n(t)\right) \\ &\vdots \\ y_n'(t) &= f_n\left(t, y_1(t), \ldots, y_n(t)\right) \end{aligned}$$

and the initial conditions $y_1(t_0) = y_1^{(0)}, \ldots, y_n(t_0) = y_n^{(0)}$.

We use vector notation: E.g., for $n = 2$ we want to find $\vec{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}$ such that

$$\begin{bmatrix} y_1'(t) \\ y_2'(t) \end{bmatrix} = \begin{bmatrix} f_1\left(t, y_1(t), y_2(t)\right) \\ f_2\left(t, y_1(t), y_2(t)\right) \end{bmatrix}, \qquad \begin{bmatrix} y_1(t_0) \\ y_2(t_0) \end{bmatrix} = \begin{bmatrix} y_1^{(0)} \\ y_2^{(0)} \end{bmatrix}$$

$$\vec{y}\,'(t) = \vec{f}\left(t, \vec{y}(t)\right), \qquad \vec{y}(t_0) = \vec{y}^{(0)}$$

We will omit the vector arrows from now on.

We denote by $D_y f(t,y)$ the Jacobian of $f(t,y)$ with respect to y:

$$D_y f(t,y) = \begin{bmatrix} \frac{\partial f_1}{\partial y_1} & \cdots & \frac{\partial f_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial y_1} & \cdots & \frac{\partial f_n}{\partial y_n} \end{bmatrix}$$

It is important for the behavior of the differential equation.

Theorem 2.1. *Assume that $f(t,y)$ and $D_y f(t,y)$ are continuous for $t \in [t_0, T]$, $y \in \mathbb{R}^n$. For a given initial value $y^{(0)}$ there is a unique solution $y(t)$ of the initial value problem. Either the solution exists for all $t \in [0,T]$, or it only exists on a smaller interval $[t_0, t_*)$ with $t_0 < t_* < T$.*

We can solve the IVP in Matlab with **ode45**: For $n = 2$ we use

```
f = @(t,y) [ ... ; ... ]        % define function f(t,y) using t, y(1), y(2)
[ts,ys] = ode45(f,[t0,T],y0);   % find column vector ts, array ys with values of solution
ys(end,:)                       % values y1, y2 at final time T
plot(ts,ys(:,1))                % plot solution y1(t)
```

2nd order ODE

So far the differential equations only contained the first derivative $y'(t)$. But in many applications (e.g. Newton's law) we have differential equations containing $y''(t)$. We then need initial conditions for $y(t_0)$ and $y'(t_0)$.

We can **rewrite this as a first order system:** Let $y_1(t) := y(t)$ and $y_2(t) := y'(t)$. Then we have $y_1' = y_2$ and $y_2' = \cdots$ where we solve the 2nd order ODE for y''.

Example: Find a function $y(t)$ for $t \in [0,4]$ such that

$$y''(t) - y'(t) + 3y(t) = t \tag{1}$$
$$y(0) = 1, \quad y'(0) = -2 \tag{2}$$

This gives the first order system

$$\begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \begin{bmatrix} y_2 \\ t + y_2 - 3y_1 \end{bmatrix}, \qquad \begin{bmatrix} y_1(0) \\ y_2(0) \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

Numerical solution in Matlab: Print out $y(T)$ and plot the function $y(t)$

```
f = @(t,y) [y(2); t+y(2)-3*y(1)];   % define function f(t,y)
[ts,ys] = ode45(f,[0,4],[-1;2]);    % solve IVP for t from 0 to 4, initial value [-1;2]
finalval = ys(end,1)                % value of y1 at final time T
plot(ts,ys(:,1));                   % plot solution y1(t)
```

3 Euler method

Consider a first order system of ODEs: We want to find $y(t)$ for $t \in [t_0, T]$ such that

$$y'(t) = f\left(t, y(t)\right), \qquad y(t_0) = y^{(0)}$$

For the **Euler method** we divide the interval $[t_0, T]$ into N subintervals of equal length $h = (T - t_0)/N$ (we can also use subintervals of different length). Let $t_j = t_0 + jh$. We then want to find approximations $y^{(1)}, \ldots, y^{(N)}$ for $y(t_j)$.

- start at the initial value $t_0, y^{(0)}$
- for $k = 0, \ldots, N-1$ do
  $s := f(t_k, y^{(k)})$
  $y^{(k+1)} := y^{(k)} + hs$
  $t_{k+1} := t_k + h$

Errors for Euler method

We consider the case $n = 1$. At time t_k the Euler method gives an approximation y_k for the exact value $y(t_k)$. We denote the **error** by

$$e_k := y_k - y(t_k)$$

By Taylor's theorem we have with a remainder term $r_k = \frac{1}{2}y''(\tau_k)h^2$

$$\begin{aligned} y(t_{k+1}) &= y(t_k) + h \cdot \overbrace{f\left(t_k, y(t_k)\right)}^{y'(t_k)} + r_k \\ y_{k+1} &= y_k \quad + h \cdot f(t_k, y_k) \end{aligned}$$

The second equation is just the definition of the Euler approximation y_{k+1}. Subtracting the first from the second equation gives

$$e_{k+1} = e_k + h \cdot [f(t_k, y_k) - f(t_k, y(t_k))] - r_k$$

Using the mean value theorem for $g(y) := f(t_k, y)$

$$f(t_k, y_k) - f(t_k, y(t_k)) = f_y(t_k, \eta_k) \cdot [y_k - y(t_k)],$$

hence the new error is

$$e_{k+1} = \underbrace{[1 + hf_y(t_k\eta_k)]}_{a_k} e_k - r_k$$

with the **amplification factor** $a_k = 1 + hf_y(t_k, \eta_k)$ and the local **truncation error** $r_k = \frac{1}{2}y''(\tau_k)h^2$**.**

For an unstable ODE we have $f_y(t, y) > 0$ and hence $a_k > 1$.

For a stable ODE we have $f_y(t, y) < 0$ and hence $a_k < 1$. However, in this case we want $|a_k| < 1$, i.e.,

$$-1 < 1 + hf_y(t, y) < 1$$

The right inequality is true for any $h > 0$. The left inequality is true if the following **stability condition** holds:

$$\boxed{h < \frac{2}{-f_y}}$$

(1) General case:

We assume

$$\begin{aligned} |f_y(t, y)| &\le C_1 \quad \text{for } t \in [t_0, T],\ y \in \mathbb{R} \\ \left|y''(t)\right| &\le C_2 \quad \text{for } t \in [t_0, T] \end{aligned}$$

Then we get bounds $|a_k| \le A$ for the amplification factor and $|r_k| \le R$ for the local truncation error:

$$\begin{aligned} |a_k| &= |1 + hf_y(t_k, \eta_k)| \le 1 + hC_1 =: A \\ |r_k| &= \left|\frac{1}{2}y''(\tau_k)h^2\right| \le \frac{C_2}{2}h^2 =: R \end{aligned}$$

yielding

$$|e_{k+1}| \le A\,|e_k| + R$$

Since $|e_0| = 0$ we obtain

$$\begin{aligned}
|e_1| &\le R \\
|e_2| &\le AR + R = (1+A)\,R \\
|e_3| &\le A(1+A)R + R = \left(1 + A + A^2\right) R \\
&\vdots \\
|e_k| &\le \left(1 + A + \cdots + A^{k-1}\right) R
\end{aligned} \tag{3}$$

We have for the geometric series

$$1 + A + \cdots + A^{k-1} = \frac{A^k - 1}{A - 1} \le \frac{A^k}{A-1} = \frac{(1+hC_1)^k}{hC_1}$$

The function e^x satisfies $1 + x \le e^x$, hence with $x = hC_1$ we get $1 + hC_1 \le e^{hC_1}$. Using this and $R = \frac{C_2}{2}h^2$ in (3) gives the **error bound**

$$\boxed{|y_k - y(t_k)| \le \frac{C_2}{2C_1} e^{C_1(t_k - t_0)} h}$$

This shows:

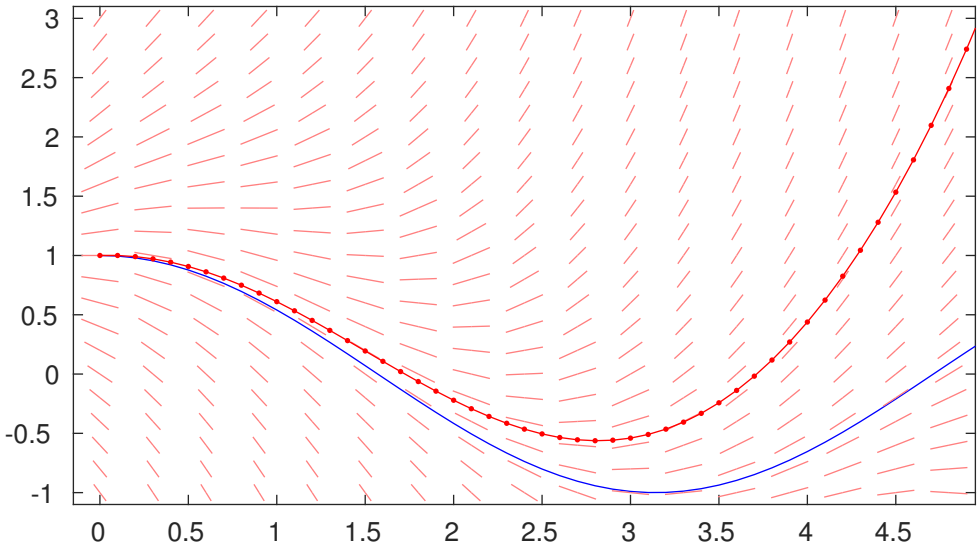
- if we keep taking Euler steps with a fixed value h for $t \to \infty$ the error can increase exponentially. This is not surprising: For an unstable ODE any tiny initial error can cause an exponentially increasing error for $t \to \infty$.
- if we only want to find the solution for $t \in [t_0, T]$: We use $h = \frac{T - t_0}{N}$ and obtain errors bounded by $ch = c'N^{-1}$: **The Euler method is a method of order 1**.

Example: The initial value problem

$$y' = y - \sin t - \cos t, \qquad y(0) = 1$$

has the solution $y(t) = \cos t$. Here $f_y(t, y) = 1 > 0$. We use the Euler method with $h = 0.1$:

```
f = @(t,y) y-sin(t)-cos(t)
tv = 0:.1:5; plot(tv,cos(tv),'b'); hold on    % plot exact solution
dirfield(f,0:.3:5,-1:.2:3);
[ts,ys] = Euler(f,[0,5],1,50);                % use 50 steps of size 5/50
plot(ts,ys,'r.-'); hold off
```



We see that the Euler values go exponentially to $+\infty$ as t gets larger than 4, whereas the exact solution $y(t) = \cos t$ stays bounded.

(2) Special case: Stable ODE where h satisfies stability condition:

We assume $f_y(t, y) < 0$: We have $C_1 \geq C_0 > 0$ such that

$$-C_1 \leq f_y(t, y) \leq -C_0 \quad \text{for } t \in [t_0, T],\ y \in \mathbb{R}$$
$$\left|y''(t)\right| \leq C_2 \quad \text{for } t \in [t_0, T]$$

We now want to have an amplification factor with $|a_k| \leq 1 - C_0 h$, i.e.,

$$-(1 - C_0 h) \leq 1 + h f_y(t_k, \eta_k) \leq 1 - C_0 h$$

The right inequality holds for any $h > 0$. We have $1 - hC_1 \leq 1 + h f_y(t_k, \eta_k)$, therefore the left inequality holds if $-(1 - C_0 h) \leq 1 - hC_1$ or

$$\boxed{h \leq \frac{2}{C_0 + C_1}} \tag{4}$$

If h satisfies this stability condition we have $|a_k| \leq 1 - C_0 h =: A < 1$, hence

$$1 + A + \cdots + A^{k-1} = \frac{1 - A^k}{1 - A} \leq \frac{1}{1 - A}$$

and (3) now gives

$$\boxed{|y_k - y(t_k)| \leq \frac{C_2}{2C_0} h}$$

This shows:

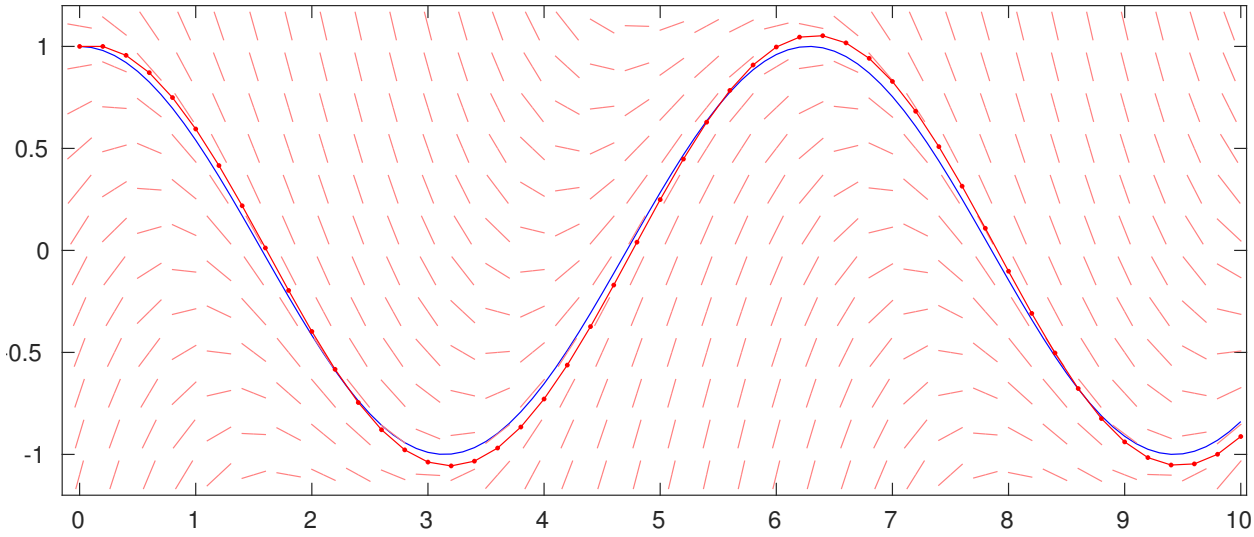
- if we keep taking Euler steps with a fixed value h for $t \to \infty$ the error is bounded by Ch with a fixed constant C.

Example: The initial value problem

$$y' = -y - \sin t + \cos t, \qquad y(0) = 1$$

has the solution $y(t) = \cos t$. Here $f_y(t, y) = -1 < 0$. We use the Euler method with $h = 0.2$:

```
f = @(t,y) -y-sin(t)+cos(t)
tv = 0:.1:10; plot(tv,cos(tv),'b'); hold on   % plot exact solution
dirfield(f,0:.3:10,-1.1:.2:1.1);
[ts,ys] = Euler(f,[0,10],1,50);               % use 50 steps of size 10/50
plot(ts,ys,'r.-'); hold off
```



Here we have an error of size Ch, but the error stays bounded as $t \to \infty$.

4 Improved Euler method (aka RK2 method)

For the **Euler method** we divide the interval $[t_0, T]$ into N subintervals of equal length $h = (T - t_0)/N$ (we can also use subintervals of different length). Let $t_j = t_0 + jh$. We then want to find approximations $y^{(1)}, \ldots, y^{(N)}$ for $y(t_j)$.

- start at the initial value $t_0, y^{(0)}$
- for $k = 0, \ldots, N-1$ do
  $s^{(1)} := f(t_k, y^{(k)})$
  $y^E := y^{(k)} + hs^{(1)}$
  $s^{(2)} := f(t_k + h, y^E)$
  $y^{(k+1)} := y^{(k)} + \frac{1}{2}\left[s^{(1)} + s^{(2)}\right]$
  $t_{k+1} := t_k + h$

The local truncation error of the improved Euler method is of order $O(h^3)$. Hence the error at a time $t = T$ is of order $O(h^2) = O(N^{-2})$: **The improved Euler method is a method of order 2**.

5 Stiff ODE and ode15s

Consider a "very stable" ODE where $f_y(t, y)$ is very negative. Then the Euler method only works if the step size h satisfies the stability condition $h < \dfrac{2}{-f_y}$. This can force use to use very tiny steps even if the solution $y(t)$ is almost constant. This is called a **stiff ODE**. In this case `ode45` uses many tiny steps and takes a long time.

Example: A flame propagation model gives the following IVP:

$$y' = y^2 - y^3 \qquad \text{for } t \in [0, \frac{2}{\delta}]$$
$$y(0) = \delta$$

Here δ is very small, e.g., $\delta = 10^{-4}$. In this case we want to solve the problem for $t \in [0, \frac{2}{\delta}] = [0, 2 \cdot 10^4]$. The solution approaches $y = 1$. But there the problem becomes stiff: We have near $y = 1$ that $f_y(t, y) = 2y - 3y^2 \approx -1$, so the stability condition for the Euler method requires $h < \dfrac{2}{-f_y} = 2$. This means that we need $N = 10^4$ steps of size $h = 2$ to get from $t_0 = 0$ to $T = 2/\delta$, despite the fact that the solution is almost constant for most of $[0, T]$.

The adaptive method **`ode45`** (with default settings) also requires about 10^4 steps:

```
delta = 1e-4;
f = @(t,y) y^2-y^3;
y0 = delta;
t0 = 0;  T = 2/delta;
[ts,ys] = ode45(f,[0,T],y0);
length(ts)                      % print number of steps
```

This prints out 12113 for the number of steps.

Matlab has a special ode solver **ode15s** for stiff ODEs: We try this for our problem

```
[ts,ys] = ode15s(f,[0,T],y0);
length(ts)                      % print number of steps
```

and get 108 for the number of steps.

6 Backward Euler method (aka implicit Euler method)

For stable problems the Euler method gives magnification factors $|a_k| = |1 + hf_y| < 1$ for small h, but $|a_k| = |1 + hf_y| > 1$ for large h.

If I look at a problem with $f_y < 0$ from right to left with decreasing t, then an Euler method in decreasing t direction always has a magnification factor $|a| > 1$, for any step size $h > 0$.

This suggests to use a "**backward Euler step**":

At time t_k we have the value $y^{(k)}$.

For time $t_{k+1} = t_k + h$ we want to find a value $y^{(k+1)}$ such that an Euler step to the left takes us to t_k and $y^{(k)}$:

Find $y^{(k+1)}$ such that

$$\boxed{y^{(k+1)} - h \cdot f\left(t_{k+1}, y^{(k+1)}\right) = y^{(k)}} \tag{5}$$

Note that $y^{(k+1)}$ occurs also inside the function f on the left hand side. If the function $f(t, y)$ is linear in y this gives linear equations for y, see the example below.

If the function $f(t, y)$ is nonlinear in y this gives nonlinear equations for y. We can e.g. use 1 or 2 steps of the Newton method (note that we have a local truncation error of size $O(h^2)$).

Claim: Let $n = 1$. For a stable problem with $f_y(t, y) < 0$ the nonlinear equation has a unique solution.
Proof: The left hand side $F(y_{k+1}) := y_{k+1} - h \cdot f(t_{k+1}, y_{k+1})$ is strictly increasing for increasing y_{k+1}, with $F(y) \to -\infty$ for $y \to -\infty$ and $F(y) \to \infty$ for $y \to \infty$.

7 Theory for systems of ODEs (you can skip this section)

So far we explained most ideas for the case $n = 1$ of one differential equation. But everything generalizes to the case of a system. Instead of the scalar $f_y(t, y)$ we now have to use the Jacobian matrix $D_y f(t, y)$.

We will always use the 2-norm and denote it by $\|\cdot\|$. For vectors $u, v \in \mathbb{R}^n$ we have denote the dot product by

$$u^\top v = u \cdot v$$

Tools for the Jacobian

In the case $n = 1$ we applied the mean value theorem $g(\tilde{y}) - g(y) = g'(\eta)$ to $g(y) := f(t_k, y)$ yielding

$$f\left(t_k, y_k\right) - f\left(t_k, y(t_k)\right) = f_y\left(t_k, \eta_k\right)\left[y_k - y(t_k)\right].$$

In the case of a system we have a function $g(y) := f(t_k, y)$ which maps a vector y to a vector $g(y)$. We want to use the Jacobian $Dg(y) \in \mathbb{R}^{n \times n}$ to find bounds for $g(\tilde{y}) - g(y)$.

We have for a matrix $A \in \mathbb{R}^{n \times n}$ and vectors $u, v \in \mathbb{R}^n$

$$\left|u^\top A v\right| \le \|u\| \, \|Av\| \le \|u\| \, \|A\| \, \|v\| \tag{6}$$

Since $u^\top A u = u^\top B u$ where $B = \frac{1}{2}(A + A^\top)$ we have

$$u^\top A u = u^\top B u \le \lambda_{\max}(B) \, \|u\|^2 \tag{7}$$

Here B is symmetric and has therefore real eigenvalues.

Definition 7.1. We use the notation

$$\boxed{\mu(A) := \lambda_{\max}\left(\tfrac{1}{2}(A + A^\top)\right)}$$

Remarks:

1. $\mu(A)$ can be negative, e.g., for $A = \begin{bmatrix} -2 & 0 \\ 0 & -3 \end{bmatrix}$ we have $\|A\| = 3$ and $\mu(A) = -2$.
2. We have $|\mu(A)| \le \|A\|$ since $\lambda_{\max}(B)$ is the smallest possible constant C with $u^\top Au \le C\|u\|^2$.
3. We have $\max_j \operatorname{Re}\lambda_j(A) \le \mu(A)$. **Proof:** Let $v^H := \overline{v}^\top$, then $v^H v = \|v\|^2$. For an eigenvalue $\lambda \in \mathbb{C}$ with eigenvector $v \in \mathbb{C}^n$ we have $Av = \lambda v$, taking $(\cdot)^H$ gives $v^H A^\top = \overline{\lambda} v^H$, hence $v^H Av = \lambda v^H v$ and $v^H A^\top v = \overline{\lambda} v^H v$ yielding
$$v^H \tfrac{1}{2}\left(A^\top + A\right) v = \tfrac{1}{2}\left(\lambda + \overline{\lambda}\right) v^H v$$
If λ is the eigenvalue where $\operatorname{Re}\lambda$ is maximal we obtain $(\operatorname{Re}\lambda)\|v\|^2 = v^H Bv \le \lambda_{\max}(B)\|v\|^2$.
4. Therefore $\mu(A) < 0$ implies $\operatorname{Re}\lambda_j(A) < 0$ for the eigenvalues of A. We can have $\operatorname{Re}\lambda_j(A) < 0$ for all j but $\mu(A) > 0$. **Example:** $A = \begin{bmatrix} -1 & 0 \\ 4 & -1 \end{bmatrix}$ has eigenvalues $\lambda_1 = \lambda_2 = -1$, but $B = \frac{1}{2}(A + A^\top) = \begin{bmatrix} -1 & 2 \\ 2 & -1 \end{bmatrix}$ has eigenvalues $-3, 1$. Hence $\mu(A) = 1 > 0$.

Now we consider a function $g(y) = \begin{bmatrix} g_1(y_1, \ldots, y_n) \\ \vdots \\ g_n(y_1, \ldots, y_n) \end{bmatrix}$ with Jacobian matrix $Dg(y) = \begin{bmatrix} \frac{\partial g_1}{\partial y_1} & \cdots & \frac{\partial g_1}{\partial y_n} \\ \vdots & & \vdots \\ \frac{\partial g_n}{\partial y_1} & \cdots & \frac{\partial g_n}{\partial y_n} \end{bmatrix}$.

Lemma 7.2. *Assume that the Jacobian $Dg(y)$ exists and is continuous for y in a convex set $\mathcal{B}$.*

1. If $\|Dg(y)\| \le L$ for all $y \in \mathcal{B}$ then
$$\|g(\tilde{y}) - g(y)\| \le L\|\tilde{y} - y\| \qquad \text{for all } \tilde{y}, y \in B \tag{8}$$
2. If $\mu\left(DG(y)\right) \le M$ for all $y \in \mathcal{B}$ then
$$[g(\tilde{y}) - g(y)] \cdot [\tilde{y} - y] \le M\|\tilde{y} - y\|^2 \qquad \text{for all } \tilde{y}, y \in B \tag{9}$$

Proof. For $y, \tilde{y} \in \mathcal{B}$ the points $y + s(\tilde{y} - y)$, $s \in [0, 1]$ on the straight line connnecting $y, \tilde{y}$ are also contained in $\mathcal{B}$. Let $g(s) := g\left(y + s(\tilde{y} - y)\right)$. Then the chain rule gives $g'(s) = Dg\left(y + s(\tilde{y} - y)\right)(\tilde{y} - y)$ and hence

$$\begin{aligned} g(1) - g(0) &= \int_0^1 g'(s)ds \\ g(\tilde{y}) - g(y) &= \int_0^1 Dg\left(y + s(\tilde{y} - y)\right)(\tilde{y} - y)\, ds \\ \|g(\tilde{y}) - g(y)\| &\le \int_0^1 \underbrace{\|Dg\left(y + s(\tilde{y} - y)\right)(\tilde{y} - y)\|}_{\le \|Dg\left(y + s(\tilde{y} - y)\right)\| \|\tilde{y} - y\| \le L\|\tilde{y} - y\|} ds \end{aligned}$$

Similarly we obtain

$$[g(\tilde{y}) - g(y)] \cdot [\tilde{y} - y] = \int_{s=0}^1 \underbrace{[Dg\left(y + s(\tilde{y} - y)\right)(\tilde{y} - y)] \cdot [\tilde{y} - y]}_{\le M\|\tilde{y} - y\|^2} ds$$

□

Perturbation of initial condition $y^{(0)}$

What happens if we perturb the initial value $y^{(0)}$ of the ODE?

Theorem 7.3. *Let $y(t)$ denote the solution of the IVP with initial condition $y(t_0) = y^{(0)}$, let $\tilde{y}(t)$ denote the solution of the IVP with initial condition $\tilde{y}(t_0) = \tilde{y}^{(0)}$. Assume $\boxed{\mu\left(D_y f(t,y)\right) \le M}$ for $t \in [t_0, T]$, $y \in \mathbb{R}^n$. Then*

$$\|\tilde{y}(t) - y(t)\| \le \left\|\tilde{y}^{(0)} - y^{(0)}\right\| e^{M(t-t_0)} \qquad \text{for } t \in [t_0, T] \tag{10}$$

For $M < 0$ the difference $\|\tilde{y}(t) - y(t)\|$ decays exponentially for increasing t. F

We call the ODE **dissipative** if we have $\boxed{\mu\left(D_y f(t,y)\right) < 0}$ for all $t \in [t_0, T]$, $y \in \mathbb{R}$.

We call an ODE $y' = f(t,y)$ **autonomous** if the function $f(t,y)$ does not depend on t.

Example 1: the linear autonomous ODE $y' = Ay$ with $A = \begin{bmatrix} -1 & 0 \\ 4 & -1 \end{bmatrix}$ has only exponentially decaying solutions, but $\mu(A) = 1$. Therefore it is NOT dissipative. We see that (10) is not sharp in this case.
Example 2 (Vinograd): linear ODE $y' = A(t)y$ with

$$A(t) = \begin{bmatrix} -1 - 9\cos^2(6t) + 6\sin(12t), & 12\cos^2(6t) + \frac{9}{2}\sin(12t) \\ -12\sin^2(6t) + \frac{9}{2}\sin(12t), & -1 - 9\sin^2(6t) - 6\sin(12t) \end{bmatrix}$$

$A(t)$ has eigenvalues $-1, -10$ for all $t \in \mathbb{R}$. But the general solution of the ODE is

$$y(t) = C_1 e^{2t} \begin{bmatrix} \cos(6t) + 2\sin(6t) \\ 2\cos(6t) - \sin(6t) \end{bmatrix} + C_2 e^{-13t} \begin{bmatrix} \cos(6t) - 2\sin(6t) \\ 2\cos(6t) + \sin(6t) \end{bmatrix}$$

The eigenvalues of $B(t) := \frac{1}{2}\left[A(t) + A(t)^\top\right]$ are 2 and -13 for all $t \in \mathbb{R}$. Therefore (10) is sharp in this case.

This shows: The effect of perturbations depends on the Jacobian $A = D_y f(t,y)$. But we need to use $\mu(A) = \lambda_{\max}\left(\frac{1}{2}(A + A^\top)\right)$ rather than the eigenvalues $\lambda_j(A)$ in order to obtain upper bounds.

Therefore we also need to use $\mu\left(D_y f(t,y)\right)$ to understand the propagation of errors for the Euler method.

Errors for the Euler method

We consider the case $n = 1$. At time t_k the Euler method gives an approximation $y^{(k)}$ for the exact value $y(t_k)$. We denote the **error** by

$$e^{(k)} := y^{(k)} - y(t_k)$$

By Taylor's theorem we have with a remainder term $r^{(k)}$ with $\left\|r^{(k)}\right\| \le \frac{1}{2}h^2 \max_{\tau \in [t_k, t_{k+1}]} \left\|y''(\tau)\right\|$

$$\begin{aligned} y(t_{k+1}) &= y(t_k) + h \cdot \overbrace{f\left(t_k, y(t_k)\right)}^{y'(t_k)} + r^{(k)} \\ y^{(k+1)} &= y^{(k)} \ + \ h \cdot f(t_k, y^{(k)}) \end{aligned}$$

The second equation is just the definition of the Euler approximation $y^{(k+1)}$. Subtracting the first from the second equation gives

$$e^{(k+1)} = \underbrace{e^{(k)} + h \cdot \left[f(t_k, y^{(k)}) - f(t_k, y(t_k))\right]}_{\text{propagated old error}} - \underbrace{r^{(k)}}_{\text{local truncation error}}$$

Let $d := f(t_k, y^{(k)}) - f(t_k, y(t_k))$. If the Jacobian satisfies $\|D_y f(t_k, y)\| \le L$ and $\mu\left(D_y f(t_k, y)\right) \le M$ for y on the line segment from $y^{(k)}$ to $y(t_k)$ we get from Lemma 7.2

$$\begin{aligned}\left\|e^{(k)} + hd\right\|^2 &= e^{(k)} \cdot e^{(k)} + 2he^{(k)} \cdot d + h^2 d \cdot d \\ &\le \left\|e^{(k)}\right\|^2 + 2hM \left\|e^{(k)}\right\|^2 + h^2 L^2 \left\|e^{(k)}\right\|^2\end{aligned}$$

Therefore we have

$$\left\|e^{(k+1)}\right\| \le A \left\|e^{(k)}\right\| + R, \qquad A = \left(1 + 2hM + h^2 L^2\right)^{1/2}, \quad R = \tfrac{1}{2} C_2 h^2$$

if $\|y''(t)\| \le C_2$ for $t \in [t_k, t_{k+1}]$.

We always have $|M| \le L$ and hence $A \le 1 + hL$ if a bound $\|D_y f(t, y)\| \le L$ holds.

For a dissipative ODE we have $\mu\left(D_y f(t, y)\right) \le M < 0$. Since

$$A^2 = 1 + 2hB \quad \text{with } B := M + \tfrac{1}{2} hL^2$$

we have $A < 1$ for $B < 0 \iff \boxed{h < \dfrac{-2M}{L^2}}$ and $A > 1$ for $B > 0 \iff h > \dfrac{-2M}{L^2}$.

(1) General case:

We assume

$$\begin{aligned}\|D_y f(t, y)\| &\le L \quad \text{for } t \in [t_0, T],\ y \in \mathbb{R}^n \\ \left\|y''(t)\right\| &\le C_2 \quad \text{for } t \in [t_0, T]\end{aligned} \tag{11}$$

Then we get the bounds $A := 1 + hL$ for the amplification factor and $\left\|r^{(k)}\right\| \le R := \frac{1}{2} C_2 h^2$ for the local truncation error yielding

$$\left\|e^{(k+1)}\right\| \le A \left\|e^{(k)}\right\| + R$$

Since $\left\|e^{(0)}\right\| = 0$ we obtain

$$\left\|e^{(k)}\right\| \le \left(1 + A + \cdots + A^{k-1}\right) R \tag{12}$$

We have for the geometric series

$$1 + A + \cdots + A^{k-1} = \frac{A^k - 1}{A - 1} \le \frac{A^k}{A - 1} = \frac{(1 + hL)^k}{hL}$$

The function e^x satisfies $1 + x \le e^x$, hence with $x = hL$ we get $1 + hL \le e^{hL}$. Using this and $R = \frac{C_2}{2} h^2$ in (12) gives the **error bound**

$$\boxed{\left\|y^{(k)} - y(t_k)\right\| \le \frac{C_2}{2L} e^{L(t_k - t_0)} h} \tag{13}$$

Remark: Often the bound for $\|D_y f(t, y)\|$ in (11) does not hold for all $y \in \mathbb{R}^n$. Actually, **we only need that the bound for the Jacobian holds near the exact solution $y(t)$ of the IVP**: assume we have $\delta > 0$ and L such that

$$\|D_y f(t, y)\| \le L \quad \text{for } t \in [t_0, T],\ \|y - y(t)\| \le \delta \tag{14}$$

Then we can choose $h_0 > 0$ such that the right hand side of (13) satisfies $\frac{C_2}{2L} e^{L(t_k - t_0)} h_0 \le \delta$. We then obtain $\left\|y^{(1)} - y(t_1)\right\| \le \delta$. For the 2nd step we need $\|D_y f(t_1, y)\| \le L$ only on the line segment from $y(t_1)$ to $y^{(1)}$. Therefore we can use L for $\left\|e^{(2)}\right\|$ and obtain $\left\|y^{(2)} - y(t_2)\right\| \le \delta$, etc.

Result: If we only have (14) instead of (11) then we have for $h \le h_0 := \delta \left[\frac{C_2}{2L} e^{L(t_k - t_0)}\right]^{-1}$ that (13) holds.

(2) Special case: Dissipative ODE where h satisfies stability condition:

We assume

$$\begin{aligned} \|D_y f(t,y)\| \le L, \quad \mu\left(D_y f(t,y)\right) \le M < 0 \qquad & \text{for } t \in [t_0, T],\ y \in \mathbb{R}^n \\ \left\|y''(t)\right\| \le C_2 \qquad & \text{for } t \in [t_0, T] \end{aligned} \tag{15}$$

We want to have $A \le 1 - \varepsilon h$ with some $\varepsilon > 0$ which gives $h \le \dfrac{2\left(-M-\varepsilon\right)}{L^2 - \varepsilon^2}$.

For $\boxed{H < \dfrac{-2M}{L^2}}$ we have a value $\varepsilon > 0$ such that $H = \dfrac{2\left(-M-\varepsilon\right)}{L^2-\varepsilon^2}$. We then have for all $h \le H$

$$1 + A + \cdots + A^{k-1} = \frac{1-A^k}{1-A} \le \frac{1}{1-A} = \frac{1}{\varepsilon h}$$

and (12) now gives the **error bound** which is independent of t

$$\boxed{\left\|y^{(k)} - y(t_k)\right\| \le \frac{C_2}{2\varepsilon} h} \tag{16}$$

Remark: We only need that the bounds (15) hold near the exact solution: instead of $y \in \mathbb{R}^n$ we only need the bounds for $\|y - y(t)\| \le \delta$. Then the error bound (16) holds if $h \le h_0 := \frac{2\varepsilon}{C_2}\delta$.