Problem 4.1 (**Consistency of two-stage ERK**, 4 Points)
A general explicit Runge-Kutta method with two stages has the following form

$$y_n = y_{n-1} + h_n F(h_n; t_{n-1}, y_{n-1})$$
$$F(h_n; t_{n-1}, y_{n-1}) = b_1 f_{n-1} + b_2 f(t_{n-1} + h_n c_2, y_{n-1} + h_n a_{21} f_{n-1}),$$

where $f_{n-1} := f(t_{n-1}, y_{n-1})$.

a) Use Taylor-Expansion to find conditions on a_i, b_i, c_i such that they imply consistency of order 2.

Proof. We do a Taylor-Expansion of $u(t_1)$ around t_0 and obtain:

$$u(t_1) = u(t_0) + (t_1 - t_0)\frac{d}{dt}u(t)\Big|_{t=t_0} + \frac{(t_1-t_0)^2}{2}\cdot\frac{d^2}{dt^2}u(t)\Big|_{t=t_0} + \mathcal{O}(|t_1-t_0|^3)$$

As we have seen in the proof of 2.3.9 the Taylor-expansion of the exact sol^n up to $\mathcal{O}(h^2)$ is

$$u(t_1) = u(t_0) + hf(t_0,u_0) + \frac{h^2}{2}\big(\underbrace{\partial_t f(t_0,u_0)}_{:=f_t} + \underbrace{\partial_y f(t_0,u_0)}_{:=f_y} f(t_0,u_0)\big) + \mathcal{O}(h^3) \quad (*)$$

We now do a Taylor-Expansion of

$$f(t_0 + hc_2, y_0 + hc_{21}\underbrace{f(t_0,u_0)}_{:=f})$$ around (t_0,y_0). We get:

$$f(t_0+hc_2, y_0+hc_{21}f) = f + \partial_t f(t_0,u_0)(t_0+hc_2-t_0) + \partial_y f(t_0,y_0)(y_0+hc_{21}f - y_0) + \mathcal{O}(h^2)$$
$$= f + hc_2 f_t + f_y(hc_{21}f) + f + hc_2 f_t + hc_{21} f_y(f)$$

Thus, we obtain:

$$y_1 = y_0 + h\big(b_1 f + b_2\big(f + hc_2 f_t + hc_{21} f_y f + \mathcal{O}(h^2)\big)\big)$$
$$= y_0 + hb_1 f + hb_2 f + h^2 c_2 b_2 f_t + h^2 c_{21} b_2 f_y f + \mathcal{O}(h^3) \quad (**)$$

Comparing $(**)$ w/ $(*)$ we get:

$$y_1 - u_1 = y_0 - u_0 + hf(b_1+b_2-1) + h^2 f_t\left(c_2 b_2 - \tfrac{1}{2}\right) + h^2 f_y(f)\left(c_{21}b_2 - \tfrac{1}{2}\right) + \mathcal{O}(h^3)$$

For y_n to be of consistency order 2 it is sufficient therefore sufficient that:

$$b_1 + b_2 = 1, \quad c_2 b_2 = \tfrac{1}{2}, \quad c_{21} b_2 = \tfrac{1}{2}$$ □

b) Thus, verify that both the *modified Euler method* and the *Heun method* satisfy your conditions.

- Modified Euler: We have: $b_1 = 0, b_2 = 1 \Rightarrow b_1 + b_2 = 0 + 1 = 1$. ✓
  and $c_2 = \frac{1}{2}, b_2 = 1 \Rightarrow c_2 b_2 = \frac{1}{2}$ ✓
  and $a_{21} = \frac{1}{2}, b_2 = 1 \Rightarrow a_{21} b_2 = \frac{1}{2}$ ✓
- Heun-Method: We have: $b_0 = b_1 = \frac{1}{2} \Rightarrow b_1 + b_2 = 1$
  and $c_2 = 1, b_2 = \frac{1}{2} \Rightarrow c_2 b_2 = \frac{1}{2}$

and $a_{21} = 1$, $b_2 = \frac{1}{2}$ $\Rightarrow$ $a_{21} b_2 = \frac{1}{2}$

Thus both methods satisfy the derived conditions. □

c) Are there any other explicit two-stage Runge-Kutta methods with consistency of order 2?

If there are then the coefficients must obey:

$$\begin{aligned} & b_1 + b_2 = 1 \\ & c_2 b_2 = \tfrac{1}{2} \overset{c_2 \neq 0}{\Leftrightarrow} b_2 = \frac{1}{2c_2} \\ & a_{21} b_2 = \tfrac{1}{2} \overset{a_{21} \neq 0}{\Leftrightarrow} b_2 = \frac{1}{2a_{21}} \end{aligned}$$

$$\Rightarrow \frac{1}{2c_2} = \frac{1}{2a_{21}} \Leftrightarrow c_2 = a_{21} \quad (1.1)$$

$$\Rightarrow b_1 + b_2 = \frac{1}{2c_2} + b_1 = 1 \Rightarrow b_1 = 1 - \frac{1}{2c_2}, \text{ and thus } b_2 = 1 - \left(1 - \frac{1}{2c_2}\right) = \frac{1}{2c_2}$$

Thus our coefficients now become:

$$\begin{cases} b_1 = 1 - \frac{1}{2c_2}, \; b_2 = \frac{1}{2c_2} \\ a_{21} = c_2 \\ c_2 \in [0,1] \text{ arbitrary.} \end{cases}$$

Thus we conclude that there are indeed infinitely many two-stage ERKs w/ consistency order two.

□

Problem 4.2 (**Maximal consistencty order of ERK**, 4 Points)
Let $\lambda, u_0 \in \mathbb{R}$ and consider the initial value problem

$$u'(t) = \lambda u(t) \text{ for all } t \in \mathbb{R},$$
$$u(0) = u_0.$$

Now let s be a natural number and assume we apply an explicit s-stage Runge-Kutta method to this problem. (See Definition 2.3.2.)

a) Let k_i for $i = 1, .., s$ be defined as in Definition 2.3.2. Show that

$$hk_i = p_i(\lambda h) y_0$$

with polynomials p_i with $\deg(p_i) \leq i$.

Proof. As in 2.3.2 we have

$k_i = f(t_0 + c_i h, g_i)$ where $g_i = y_0 + h \sum_{j=1}^{i-1} a_{ij} k_j$, for $i = 1, \dots, s$. In this particular case $f(t,y) = \lambda y$, thus, $k_i = \lambda g_i$. We show the statement using induction on i for a fixed s.

$i=1$: Then $g_1 = y_0$ and $k_1 = f(t_0 + c_1 h, g_1) = \lambda y_0 \Rightarrow hk_1 = h\lambda y_0$, thus $p_1 = x$ does the job.

Let $i \geq 1$ and $i \leq s-1$. Then we obtain

$$hk_{i+1} = \lambda g_{i+1} = \lambda \left(y_0 + \sum_{j=1}^{i} a_{ij} k_j \right) = h\lambda \left(y_0 + \sum_{j=1}^{i} a_{ij} p_j(\lambda h) y_0 \right) = h\lambda y_0 \left(1 + \sum_{j=1}^{i} a_{ij} p_j(\lambda h) \right)$$

We define

$$p_{i+1}(T) = T \left(1 + \sum_{j=1}^{i} a_{ij} p_j(T) \right).$$ Since $\deg p_j \leq j, \forall j = 1, \dots, i$. We obtain

$\deg \left(1 + \sum_{j=1}^{i} a_{ij} p_j(T) \right) \leq i$ ($\deg \leq i$), thus $\deg \left(T \left(1 + \sum_{j=1}^{i} a_{ij} p_j(T) \right) \right) = \deg(T) + \deg \left(1 + \sum_{j=1}^{i} a_{ij} p_j(T) \right) \leq i+1$ ($\deg(T) \leq 1$, $\deg \left(1 + \sum_{j=1}^{i} a_{ij} p_j(T) \right) \leq i$).

Thus we obtain: $hk_{i+1} = y_0 h \left(1 + \sum_{j=1}^{i} a_{ij} p_j(h) \right) = y_0 p_{i+1}(h)$, with $\deg p_{i+1} \leq i+1$.

which finishes the proof. □

b) Thus, show that an explicit s-stage Runge-Kutta method cannot be consistent of order $p > s$.

Proof. Let $p > s$. A method is consistent of order p if (using that for ERKs we have $h_k = h$)

$\max_{k=1,\dots,n} |\tau_k| \leq C h^p$ where $\tau_k = \frac{u_k - u_{k-1}}{h} - F_h(t_{k-1}, y_{k-1})$ (*). We have

$$F_h(t_{k-1}, y_{k-1}) = \sum_{i=1}^{s} b_i f(t_0 + c_i h, g_i) \overset{f(t,y) = \lambda y}{=} \lambda \sum_{i=1}^{s} b_i g_i \text{ where } g_i = y_{k-1} + \sum_{j=1}^{i-1} a_{ij} k_j \quad (*)$$

An exact sol^n of the IVP is given by: $u(t) = u_0 e^{\lambda t}$ (*), because $u(0) = u_0 e^{\lambda \cdot 0} = u_0$ ($e^{\lambda \cdot 0} = 1$) and $u'(t) = \lambda u_0 e^{\lambda t} = \lambda u(t)$.

$$\tau_1 h^{-p} = \frac{u_1 - u_0}{h^{p+1}} - \frac{\lambda}{h^p} \sum_{i=1}^{s} b_i \left(y_0 + \sum_{j=1}^{i-1} a_{ij} \frac{p_j(\lambda h) y_0}{h} \right)$$

$$= \frac{u_1 - u_0}{h^{p+1}} - \lambda \sum_{i=1}^{s} \frac{b_i y_0}{h^p} + \sum_{j=1}^{i-1} a_{ij} \underbrace{\frac{p_j(\lambda h) y_0}{h^{p+1}}}$$

$p_j(\lambda h)$ are Poly's w/ degree $\leq j \leq s < p$, thus $\frac{p_j(\lambda h)}{h^{p+1}}$ diverges $\forall j = 1, \dots, i$.

Problem 4.3 (Local error control of ERK methods)
In Algorithm 2.4.2 of the lecture notes the optimal step size is determined by

$$h_{opt} = h_k \sqrt[p+1]{\frac{\varepsilon}{||y_k - \hat{y}_k||}}$$

where ε is a tolerance given by the user, y_k is the approximation for u_k obtained from a method with consisteny of order p using the time step h_k and $\hat{y}_k$ likewise, but obtained from a method with consisteny of order $\geq p+1$.

a) In the algorithm, if $||y_k - \hat{y}_k|| > \varepsilon$, then y_k is rejected. Why?

b) If y_k is rejected as above, step k is repeated with step size $h_k^{new} = h_{opt}$. Why is this a good choice?

c) If the step is accepted, i.e.

$$||y_k - \hat{y}_k|| \leq \varepsilon,$$

the step size for the next step is set to $h_{k+1} = h_{opt}$. Why is this a good choice?

Note: You will have to make reasonable assumptions for this algorithm to make sense. Extracting them from the paragraph preceding Algorithm 2.4.2 (or coming up with them yourself) is part of the exercise!

a) If our threshold $\varepsilon > 0$ (for $\|y_k - \tilde{y}_k\|$) is greater than or equals to $\|y_k - \hat{y}_k\|$, then we know that the allowed error (ε) was to big and we choose a smaller time-step (h_{opt}) and recompute $y_k, \hat{y}_k$. Thus we guarantee that $\|y_k - \tilde{y}_k\| < \varepsilon$, for all k.

b) If the time-step h_k and the computations for $y_k, \hat{y}_k$ are rejected, h_{opt} is a good choice for h_k^{new}, because $h_k^{new} < h_k$ and as the local error $|u_k - y_k| \leq C e^{LT} h^p$ (2.25) is bounded by (2.25), it is reasonable to choose a smaller h.

c) This is only a good choice if $h_{opt} \leq T - t_k$.
Thus we use a time-step s.t. $t_{k+1} \leq T$.
Since $\|y_k - \tilde{y}_k\| \leq \varepsilon$, and thus $h_{opt} \geq h_k$, it is reasonable to assume that h_{opt} is the maximal time-step we can make (using $\tilde{y}_k$ as u_0) without making a local error greater than ε.