MOS ESS101
November 2018

This exam contains 15 pages (including this cover page) and 5 problems.

You are allowed to use the following books:

- **Official copy of “Modelling & Simulation, Lecture Notes for ESS101”**
- **β-handbook**
- **“Physics handbook for science and engineering”**

and a calculator. Some formula specific to this course are provided in the end as an appendix

- Organize your work in a reasonably neat and coherent way. Work scattered all over the page without a clear ordering may receive less credit.
- Mysterious or unsupported answers will not receive credit, but an incorrect answer supported by substantially correct calculations and explanations will receive partial credit.
- Answers blindly copied from the Lecture Notes will not receive credits. Make sure that you develop your answers in your own words and show that you have understood the statements you are making.
- None of the proposed questions require extremely long computations. If you get caught in endless algebra, you have probably missed the simple way of doing it.
- The passing grade will a priori be given at 26 points, and the top grade at 39 points. These limits may be lowered (but not increased) depending on the outcome of the exam.

Problem	Points	Score
1	6	
2	11	
3	10	
4	6	
5	11	
Total:	44	

Best of luck to all !!

1. **Lagrange modelling** we will consider the ring problem illustrated in Fig. 1. A massless ring-shaped rail is rotating around its vertical axis, subject to a torque T. A mass m slides along the ring without friction.

   (a) (6 points) Choose a set of coordinates to describe the system, and write the model equations of the system. If the resulting equations are a high-index DAE, propose an index-reduced version.

   Hint: let us consider a point $\mathbf{p} \in \mathbb{R}^3$ *on the ring. The work* δW *produced by* T *for a displacement* $\delta \mathbf{p}$ *is then given by:*

$$\delta W = \frac{1}{R^2} \det \left(\begin{bmatrix} \mathbf{T} & \mathbf{p} & \delta \mathbf{p} \end{bmatrix} \right) \tag{1}$$

   where $\mathbf{T} = \begin{bmatrix} 0 & 0 & T \end{bmatrix}^\top$. *You may want to use this observation or not, depending on how you decide to approach the problem.*

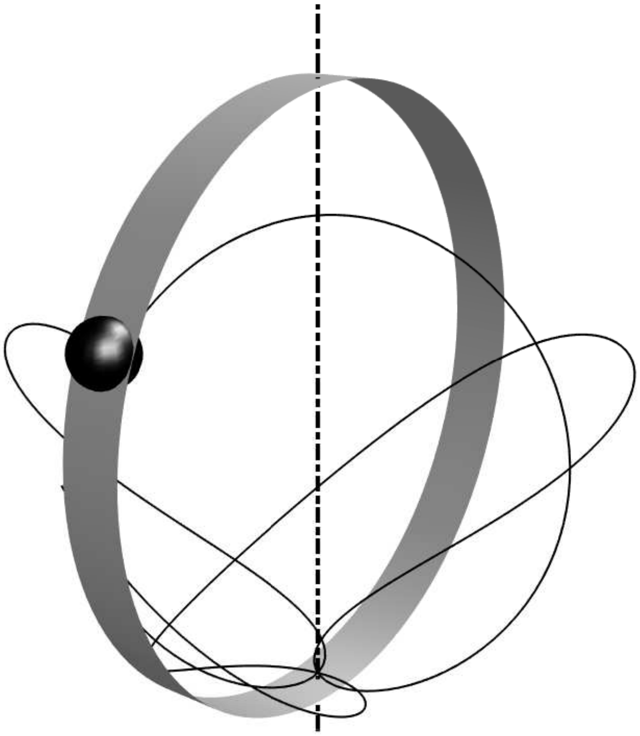

Figure 1: Illustration of the ring problem. The ring has a negligible inertia, a radius R, and is subject to a torque T on its vertical rotation axis (dashed-dotted line). The mass (black ball) slides without friction on the ring.

Solution:

(a) Let us pick

$$\mathbf{q} = \mathbf{p} \tag{2}$$

as our generalized coordinates, where $\mathbf{p} \in \mathbb{R}^3$ is the position of the mass in the cartesian reference frame attached to the axis of the ring (vertical axis aligned with the vertical ring axis, origin at the center of the ring). The kinetic and potential energies read as:

$$T = \frac{1}{2}m\dot{\mathbf{p}}^\top \dot{\mathbf{p}}, \qquad V = mg\mathbf{p}_3 \tag{3}$$

and we use the constraints:

$$\mathbf{c} = \frac{1}{2}\left(\mathbf{p}^\top \mathbf{p} - R^2\right) = 0 \tag{4}$$

that imposes the ball to be at distance R from the origin. We then write the Lagrange function:

$$\mathcal{L} = \frac{1}{2}m\dot{\mathbf{p}}^\top \dot{\mathbf{p}} - mg\mathbf{p}_3 - z\frac{1}{2}\left(\mathbf{p}^\top \mathbf{p} - R^2\right) \tag{5}$$

where $z \in \mathbb{R}$. We can then trivially compute:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{q}}^\top = -\begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} - z\mathbf{p}, \qquad \frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}}^\top = m\dot{\mathbf{p}}, \qquad \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}}^\top = m\ddot{\mathbf{p}} \tag{6}$$

and assemble the Lagrange equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}}^\top - \frac{\partial \mathcal{L}}{\partial \mathbf{q}}^\top = m\ddot{\mathbf{p}} + \begin{bmatrix} 0 \\ 0 \\ mg \end{bmatrix} + z\mathbf{p} = \mathbf{Q} \tag{7a}$$

$$\frac{1}{2}\left(\mathbf{p}^\top \mathbf{p} - R^2\right) = 0 \tag{7b}$$

We finally compute the generalized forces $\mathbf{Q}$. Using (1), we have:

$$\delta W = \frac{T}{R^2}\left(\delta\mathbf{p}_2\mathbf{p}_1 - \delta\mathbf{p}_1\mathbf{p}_2\right) \tag{8}$$

such that

$$\mathbf{Q} = \frac{T}{R^2}\begin{bmatrix} -\mathbf{p}_2 \\ \mathbf{p}_1 \\ 0 \end{bmatrix} \tag{9}$$

The resulting equations are an index-3 DAE (constrained Lagrange approach). The index-reduction is fairly straightforward. The constraint $\mathbf{c}$ is differentiated twice with respect to time, yielding:

$$\ddot{\mathbf{c}} = \mathbf{p}^\top \ddot{\mathbf{p}} + \dot{\mathbf{p}}^\top \dot{\mathbf{p}} = 0 \tag{10}$$

where the consistency conditions:

$$\mathbf{c} = 0, \qquad \dot{\mathbf{c}} = \mathbf{p}^\top \dot{\mathbf{p}} \tag{11a}$$

must hold at the initial conditions.

2. **System Identification**

(a) (2 points) Consider an estimation problem where the data are generated via:

$$y_k = \phi_k(\theta) + e_k \tag{12}$$

where e_k is uncorrelated and generated according to the distribution:

$$e_k \sim \sqrt{\frac{2}{\pi}}\frac{x^2 e^{-\frac{x^2}{2a^2}}}{a^3} \tag{13}$$

where $a > 0$ is a distribution parameter. Propose a penalty function ϕ supported by the MLE principle to use in the fitting problem:

$$\hat{\theta} = \mathrm{a}\min_{\theta} \sum_k \Phi(y_k - \phi_k(\theta)) \tag{14}$$

to estimate $\hat{\theta}$.

(b) (2 points) Consider the following system generating the data

$$y_k + 0.5y_{k-1} = u_{k-1} + 1.5u_{k-2} + e_k - 0.2e_{k-1} \tag{15}$$

Find the plant model $G(z)$ and the noise model $H(z)$. Find the one-step-ahead predictor for the system.

(c) (4 points) Consider the ARX model:

$$y_k + a_1 y_{k-1} + a_2 y_{k-2} = b_0 u_k + e_k \tag{16}$$

and the associated data $y_{0,\dots,N}$ and $u_{0,\dots,N}$ obtained from applying the input sequence $u_{0,\dots,N}$ to the real system, started with $y_{k<0} = 0$.

Write the problem delivering the maximum-likelihood estimator of $\boldsymbol{\theta} = \begin{bmatrix} a_1 & a_2 & b_0 \end{bmatrix}^\top$ according to the one-step ahead prediction when the additive noise e_k is uncorrelated and uniformly distributed in the interval [0, 1], i.e. the probability density function of e_k is:

$$f(e_k) = \begin{cases} 1 & \text{if} \quad e_k \in [0,\, 1] \\ 0 & \text{if} \quad e_k \notin [0,\, 1] \end{cases} \tag{17}$$

Describe the solution to the max-likelihood problem, in particular, is the solution unique?

(d) (3 points) Consider a linear least-squares problem delivering a parameter estimation $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ \frac{1}{2}\left(A\boldsymbol{\theta} - \mathbf{y}\right) W \left(A\boldsymbol{\theta} - \mathbf{y}\right) \tag{18}$$

1. What are we assuming about the data when using a least-squares fitting problem like (18).
2. Explain in detail the meaning of formula (85) in the Formula Sheet
3. How should matrix W be selected?

Solution:

(a) We can apply the transformation classically used to obtain MLE → LS. Let us write the Likelihood function:

$$L(\theta) = c \cdot \sqrt{\frac{2}{\pi}}\frac{(y_k - \phi_k(\theta))^2 e^{-\frac{(y_k - \phi_k(\theta))^2}{2a^2}}}{a^3} \tag{19}$$

for some $c > 0$ and the MLE problem:

$$\hat{\theta} = \mathrm{a}\max_{\theta} \; L(\theta) \tag{20}$$

We turn the MLE problem into a fitting problem by using:

$$\hat{\theta} = \mathrm{a}\min_{\theta} \; -\log(L(\theta)) = \mathrm{a}\min_{\theta} \left[\frac{(y_k - \phi_k(\theta))^2}{2a^2} - \log 2 \log(y_k - \phi_k(\theta)) \right] \tag{21}$$

hence the penalty function can be selected as:

$$\Phi(y_k - \phi_k(\theta)) = \frac{(y_k - \phi_k(\theta))^2}{2a^2} - \log 2 \log(y_k - \phi_k(\theta)) \tag{22}$$

or more simply:

$$\Phi(x) = \frac{x^2}{2a^2} - \log 2 \log(x) \tag{23}$$

(b) The plant and noise model are

$$G(z) = \frac{z^{-1} + 1.5z^{-2}}{1 + 0.5z^{-}1} \quad H(z) = \frac{1 - 0.2z^{-1}}{1 + 0.5z^{-}1}, \tag{24}$$

and the one-step-ahead predictor can be found using $H(z)\hat{y}(t) = G(z)u(t) + (H(z) - 1)y(t)$.

(c) The one-step ahead predictor reads as:

$$\hat{y}_k = -a_1 y_{k-1} - a_2 y_{k-2} + b_0 u_k \tag{25}$$

and the mismatch between the data and the predictor is given by:

$$e_k = \hat{y}_k - y_k = -y_k - a_1 y_{k-1} - a_2 y_{k-2} + b_0 u_k = \mathbf{d}_k^\top \boldsymbol{\theta} - y_k \tag{26}$$

where $\mathbf{d}_k = \begin{bmatrix} -y_{k-1} & -y_{k-2} & u_k \end{bmatrix}$. According to our model, e_k is uniformly distributed. As the noise is uncorrelated, the probability density of observing a sequence $e_{0,\ldots,N}$ is given by:

$$\mathbb{P}\left[e_{0,\ldots,N}\right] = \prod_{k=0}^{N} f(e_k) = \begin{cases} 1 & \text{if} \quad e_{0,\ldots,N} \in [0,\,1] \\ 0 & \text{if} \quad e_k \notin [0,\,1] \text{ for some } k \end{cases} \tag{27}$$

For a given set of parameters $\boldsymbol{\theta} = \begin{bmatrix} a_1 & a_2 & b_0 \end{bmatrix}^\top$, the probability of obtaining a given noise sequence is then:

$$\mathbb{P}\left[e_k = \mathbf{d}_k^\top \boldsymbol{\theta} - y_k \quad \text{for} \quad k = 0, \ldots, N \,\middle|\, \boldsymbol{\theta} \right] = \begin{cases} 1 & \text{if} \quad \mathbf{d}_k^\top \boldsymbol{\theta} - y_k \in [0,\,1] \quad \forall\, k \\ 0 & \qquad\qquad \text{otherwise} \end{cases} \tag{28}$$

The maximum-likelihood problem is then:

$$\max_{\boldsymbol{\theta}} \quad \begin{cases} 1 & \text{if} \quad \mathbf{d}_k^\top \boldsymbol{\theta} - y_k \in [0,\,1] \quad \forall\, k \\ 0 & \qquad\qquad \text{otherwise} \end{cases} \tag{29}$$

If there exists $\boldsymbol{\theta}$ such that $\mathbf{d}_k^\top \boldsymbol{\theta} - y_k \in [0,\,1]$ for all k, then the solution set is

$$\boldsymbol{\theta}^\star = \left\{ \boldsymbol{\theta} \quad \text{s.t} \quad \mathbf{d}_k^\top \boldsymbol{\theta} - y_k \in [0,\,1] \quad \forall\, k \right\} \tag{30}$$

and the solution is (possibly) not unique. If the solution set $\boldsymbol{\theta}^\star$ is empty, then any $\boldsymbol{\theta}$ has probability 0.

(d)

1. We are assuming that the data $\mathbf{y}$ are corrupted by normal centred (not necessarily white) noise, and by that only. I.e. the data sequence $\mathbf{y}_{0,\ldots,N}$ reads as:

$$\mathbf{y} = A\boldsymbol{\theta}_{\mathrm{true}} + \mathbf{e} \tag{31}$$

   where $\boldsymbol{\theta}_{\mathrm{true}}$ is the true model parameters.

2. The expression:

$$\Sigma_{\hat{\boldsymbol{\theta}}} = \left(A^\top \Sigma^{-1} A\right)^{-1} \tag{32}$$

   describes the covariance of the parameter estimation, labelled $\Sigma_{\hat{\boldsymbol{\theta}}}$. The formula can be understood as follows. The normal centred noise sequence $\mathbf{e}$ is a vector or random variables. In that sense, the data we feed into the least-squares problem $\mathbf{y} = A\boldsymbol{\theta}_{\mathrm{true}} + \mathbf{e}$ is also a vector or random variables. Hence the outcome $\hat{\boldsymbol{\theta}}$ of the least-squares problem (**??**) (see formula (84))

$$\hat{\boldsymbol{\theta}} = \left(A^\top \Sigma^{-1} A\right)^{-1} A^\top \Sigma^{-1}\mathbf{y} \tag{33}$$

   is itself a random variable (it is in fact a linear function of the random vector $\mathbf{y}$). Equation (32) provides the covariance of that random variable, i.e.

$$\mathbb{E}\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}}\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}}\right)^\top\right] \tag{34}$$

   where $\boldsymbol{\mu}_{\hat{\boldsymbol{\theta}}} = \mathbb{E}\left[\hat{\boldsymbol{\theta}}\right]$.

3. The weighting matrix W ought to be selected as the inverse of the covariance matrix $\Sigma = \mathbb{E}\left[\mathbf{e}\mathbf{e}^\top\right]$ (assuming $\mathbb{E}\left[\mathbf{e}\right] = 0$). Then equation (85) holds, and the LS problem corresponds to the MLE problem.

3. **Differential-Algebraic and Implicit Differential Equations**

(a) (2 points) Consider a fully-implicit DAE

$$\mathbf{F}\left(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{z}, \mathbf{u}\right) = 0 \tag{35}$$

rewrite it in a semi-explicit form. What are the differential and algebraic variables then, and what are their dimension, assuming that $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{z} \in \mathbb{R}^m$?

(b) (3 points) Consider the differential equation:

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \dot{\mathbf{x}} = \mathbf{x} + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u \tag{36}$$

1. Is (36) an implicit ODE or a DAE? Justify.
2. If it is an implicit ODE, what is its solution for $u = 0$? If it is a DAE, what is its differential index?

(c) (3 points) Consider the semi-explicit DAE:

$$\dot{\mathbf{x}} = A\mathbf{x} + \mathbf{b}z \tag{37a}$$

$$0 = \frac{1}{2}\left(\mathbf{x}^\top \mathbf{x} - L(t)^2\right) \tag{37b}$$

What is the size of $\mathbf{z}$? Perform an index-reduction of (37). What are the consistency conditions? What happens for $L = 0$?

(d) (2 points) Consider the semi-explicit DAE:

$$\dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \mathbf{z}, \mathbf{u}\right) \tag{38a}$$

$$0 = \mathbf{g}\left(\mathbf{x}, t\right) \tag{38b}$$

Provide the condition for (38) to be of index 2.

Solution:

(a) The semi-explicit form corresponding to this DAE can be generally written as:

$$\dot{\mathbf{x}} = \mathbf{v} \tag{39a}$$

$$0 = \mathbf{F}\left(\mathbf{v}, \mathbf{x}, \mathbf{z}, \mathbf{u}\right) \tag{39b}$$

The differential variables are still $\mathbf{x} \in \mathbb{R}^n$, but we have introduced new algebraic variables $\mathbf{v}$ such that the complete set of algebraic variable is

$$\mathbf{w} = \begin{bmatrix} \mathbf{v} \\ \mathbf{z} \end{bmatrix} \in \mathbb{R}^{m+n} \tag{40}$$

(b)

1. It is a DAE as the matrix “E” multiplying $\dot{\mathbf{x}}$ is rank deficient (one column of zero)
2. The DAE reads as:

$$\dot{\mathbf{x}}_2 = \mathbf{x}_1 + u \tag{41}$$

$$\dot{\mathbf{x}}_3 = \mathbf{x}_2 \tag{42}$$

$$\dot{\mathbf{x}}_2 + \dot{\mathbf{x}}_3 = \mathbf{x}_3 \tag{43}$$

which we can rewrite as:

$$\begin{aligned}\dot{\mathbf{x}}_2 &= \mathbf{x}_1 + u && (44)\\ \dot{\mathbf{x}}_3 &= \mathbf{x}_2 && (45)\\ 0 &= \mathbf{x}_3 - \mathbf{x}_1 - u - \mathbf{x}_2 && (46)\end{aligned}$$

We can time-differentiate the last equation to get:

$$0 = \dot{\mathbf{x}}_3 - \dot{\mathbf{x}}_1 - \dot{u} - \dot{\mathbf{x}}_2 \tag{47}$$

and replace $\dot{\mathbf{x}}_3, \dot{\mathbf{x}}_2$ by the expressions in (44), (45) to get:

$$\dot{\mathbf{x}}_1 = \mathbf{x}_2 - \dot{u} - \mathbf{x}_1 - u \tag{48}$$

We then can collect:

$$\begin{aligned}\dot{\mathbf{x}}_1 &= \mathbf{x}_2 - \dot{u} - \mathbf{x}_1 - u && (49)\\ \dot{\mathbf{x}}_2 &= \mathbf{x}_1 + u && (50)\\ \dot{\mathbf{x}}_3 &= \mathbf{x}_2 && (51)\end{aligned}$$

which is an ODE. We needed a single time-differentiation, hence the original DAE is of index 1.

(c) Since we have a scalar constraint (38b), the algebraic variables are scalar as well, i.e. $\mathbf{z} \in \mathbb{R}$. Let us perform the index reduction. We perform a time-differentiation of (38b) to obtain

$$\mathbf{x}^\top \dot{\mathbf{x}} - L(t)\dot{L}(t) = 0 \tag{52}$$

Using (38a), we can rewrite:

$$\mathbf{x}^\top A\mathbf{x} + \mathbf{x}^\top B\mathbf{z} - L(t)\dot{L}(t) = 0 \tag{53}$$

The latter constraint is "solvable" for $\mathbf{z}$ for $\mathbf{x}^\top B \neq 0$. If this holds, we can form the index-1 DAE:

$$\begin{aligned}\dot{\mathbf{x}} &= A\mathbf{x} + \mathbf{b}z && (54a)\\ 0 &= \mathbf{x}^\top A\mathbf{x} + \mathbf{x}^\top B\mathbf{z} - L(t)\dot{L}(t) && (54b)\end{aligned}$$

The consistency condition is simply $\frac{1}{2}\left(\mathbf{x}(0)^\top\mathbf{x}(0) - L(0)^2\right) = 0$. For $L = 0$, constraint (38b) entails that $\mathbf{x} = 0$, where $\mathbf{x}^\top B \neq 0$ fails. In fact, any index reduction would yield a meaningless DAE.

(d) In order for (38) to be of index 2, a single time-differentiation ought to deliver an index-1 DAE, i.e. a DAE that can be solved for $\mathbf{z}$. We perform that time differentiation on the algebraic part:

$$\dot{\mathbf{g}}(\mathbf{x}, t) = \frac{\partial \mathbf{g}}{\partial \mathbf{x}}\dot{\mathbf{x}} + \frac{\partial \mathbf{g}}{\partial t} = 0 \tag{55}$$

We then use the differential part to obtain:

$$\dot{\mathbf{g}}(\mathbf{x}, t) = \frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x}, \mathbf{z}, \mathbf{u}) + \frac{\partial \mathbf{g}}{\partial t} = 0 \tag{56}$$

The latter algebraic equation is solvable for $\mathbf{z}$ if

$$\frac{\partial}{\partial \mathbf{z}}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{x}}\mathbf{f}\right) \tag{57}$$

is full rank.

4. **Newton** The Newton methods aims at solving a set of equation $\mathbf{r}\left(\mathbf{x}\right) = 0$ numerically. To that end, iterates the recursion:

$$M\Delta\mathbf{x} + \mathbf{r}\left(\mathbf{x}\right) = 0 \tag{58a}$$

$$\mathbf{x} \leftarrow \mathbf{x} + \alpha\Delta\mathbf{x} \tag{58b}$$

where $\alpha \in]0,\, 1]$ is the step-size.

(a) (1 point) How should matrix M be ideally chosen?

(b) (2 points) Explain in words what condition(s) is (are) required for Newton to converge with $\alpha = 1$.

(c) (1 point) Why do we need α and how should it be chosen?

(d) (2 points) The local convergence rate of an exact, full-step Newton method can be summarized as:

$$\left\|\mathbf{x}_+ - \mathbf{x}_\star\right\| \leq c\left\|\mathbf{x} - \mathbf{x}_\star\right\|^2 \tag{59}$$

where $\mathbf{x}_\star$ is a solution of $\mathbf{r}\left(\mathbf{x}_\star\right)$. What is the meaning of this formula? When does it (doesn't it) occur?

Solution:

(a) Ideally, we ought to chose M as the Jacobian matrix $\frac{\partial \mathbf{r}}{\partial \mathbf{x}}$. In practice, approximations are often used.

(b) Full Newton steps are guaranteed to converge in a neighborhood of a solution only. The “size” of that neighborhood depends on how nonlinear $\mathbf{r}\left(\mathbf{x}\right)$ is, and the Jacobian $\frac{\partial \mathbf{r}(\mathbf{x})}{\partial \mathbf{x}}$ must be full rank throughout this neighborhood.

(c) The step-size α allows the Newton iteration to converge even if $\mathbf{x}$ is not close to the solution $\mathbf{x}_\star$. It is typically chosen so as to ensure that

$$\left\|\mathbf{r}\left(\mathbf{x} + \alpha\Delta\mathbf{x}\right)\right\| < \left\|\mathbf{r}\left(\mathbf{x}\right)\right\| \tag{60}$$

Finding α is the role of a (small) computer code usually labelled “line-search”.

(d) This formula states that the exact, full-step Newton iteration converges quadratically to a solution. That is, the number of accurate digits in the $\mathbf{x}$ is doubled at every iteration. Achieving the quadratic contraction rate requires basically what is stated in the question, namely:

- Exact Newton steps, i.e. an exact Jacobian $M = \frac{\partial \mathbf{r}(\mathbf{x})}{\partial \mathbf{x}}$ is used and system (58a) is solved to machine precision.
- Full steps are taken, i.e. $\alpha = 1$ throughout the iterations.
- The quadratic convergence rate is local, i.e. it occurs in a neighborhood of the solution $\mathbf{x}_\star$.

5. **Simulation**

(a) (3 points) Describe in your own words the trade-offs in terms of computational cost (amount of computations to run a simulation) in Explicit and Implicit Runge-Kutta schemes. In particular, explain for each why a very high number of stages is not advantageous, and why the "optimum" number of stages is different for explicit and implicit

(b) (2 points) How can one spot from the Butcher tableau if the RK method it describes is implicit? Why is it called that way?

(c) (2 points) Explain in your own words why "collocation ⊂ IRK", i.e. why all collocation methods are IRK methods, but not all IRK methods are collocation methods.

(d) (4 points) Consider a collocation method with a single stage ($s = 1$). Write the Butcher tableau and the resulting collocation equations for the semi-explicit DAE

$$\dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \mathbf{z}, \mathbf{u}\right) \tag{61a}$$

$$0 = \mathbf{g}\left(\mathbf{x}, \mathbf{z}, \mathbf{u}\right) \tag{61b}$$

Note: we will define $\prod_{i=1,j\neq i}^{1} \frac{\tau-\tau_j}{\tau_i-\tau_j} = 1$

Solution:

(a) We need to discuss explicit and implicit methods separately here.

- For **explicit** methods, the computational burden arises from evaluation the stage variables $\mathbf{K}_{1,\ldots,\mathrm{s}}$ in the Runge-Kutta method. Each stage requires an evaluation of the model equation $\mathbf{f}$. The order of the method increases with the number of stages. Up to $\mathrm{s} \leq 4$, we have $o = \mathrm{s}$ but for $\mathrm{s} > 4$, the order "stalls" i.e. $o < \mathrm{s}$ (see Table 1 in the appendix). In order to reach a desired integration accuracy, we then have two choices: decreasing Δt (step-size) in the method of increasing the order o. Both impact the accuracy of the integration according to $\mathcal{O}(\Delta t^o)$. Because of the "stalling" effect in the stage-to-order relationship (i.e. Table 1 in the appendix), a very high number of stages does not yield a correspondingly large order, and the benefit of increasing the order becomes outweighted by the computational cost of the number of stage required to achieve that order. In general, up to $o = 4 - 5$, this "stalling" effect is non-existent or very moderate. As a result, explicit integration methods often use this number of stages.
- For **implicit** methods, the computational burden comes mainly from the factorization of the linear systems involved in the Newton method solving the implicit Runge-Kutta equations. For implicit RK methods (of the collocation familly), we have $o = 2\mathrm{s}$. However, the size of the linear systems is $n \cdot \mathrm{s}$ (where n is the dimension of the state in the ODE), and the computational complexity of solving the linear systems is (at worse) cubic, i.e. $(n \cdot \mathrm{s})^3 = \frac{1}{8} n^3 o^3$. We therefore pay a high price for increasing the order. For that reason, implicit RK methods reach their best efficiency typically at a fairly low number of stages, typically $s = 2 - 3$.

(b) An explicit method has $a_{ij} = 0$ at or above the diagonal, while implicit methods can have $a_{ij} \neq 0$ everywhere. The name "implicit" is coming from the fact that for a generic Butcher tableau, the RK equations describe the stage variables $\mathbf{K}_{1,\ldots,\mathrm{s}}$ implicitly (and hence typically require a Newton method).

(c) We know that all collocation methods are implicit RK methods, as the collocation equations are identical to the IRK equations. That explains the inclusion "collocation ⊂ IRK". However, not all IRK methods are collocation methods. This fact is easy to argue by observing that a collocation method is eventually entirely described by the "checkpoints" $\tau_{1,\ldots,\mathrm{s}}$ used to build the polynomials, i.e. we have s "degrees of freedom" to select the collocation method (note that these checkpoints are themselves entirely defined if one want to achieve the order $o = ss$). On the other hand, the Butcher tableau has $\mathrm{s}^2 + 2\mathrm{s}$ entries (a, b, c), and therefore as many degrees of freedom (even though many choices of Butcher tableau can be useless for the purpose of integrating the ODE). This mismatch in the number of degrees of freedom explains why "IRK $\nsubseteq$ collocation"

(d) We start with defining the collocation grid points for the method $s = 1$. We observe that we will have a single grid point τ_1, give by the single root of the polynomial:

$$P_1(\tau) = \frac{1}{s!}\frac{\mathrm{d}}{\mathrm{d}\tau}\left(\tau^2 - \tau\right) = 2\tau - 1 = 0 \tag{62}$$

hence we ought to pick $\tau_1 = \frac{1}{2}$. We can build the polynomial supporting the collocation approximation. We use the note in the question to determine that:

$$\ell_1(\tau) = 1 \tag{63}$$

We finally compute the integral of ℓ_1:

$$L_1(\tau) = \int_0^\tau \ell_1(\xi)\mathrm{d}\xi = \tau \tag{64}$$

We are not ready to compute the Butcher coefficients:

$$a_{11} = L_1(\tau_1) = \frac{1}{2}, \quad b_1 = L_1(1) = 1, \quad c_1 = \tau_1 = \frac{1}{2} \tag{65}$$

The Butcher tableau would read as:

$$\begin{array}{c|c} 0.5 & 0.5 \\ \hline & 1 \end{array}$$

The collocation equations resulting from this Butcher tableau for a semi-explicit DAE would read as:

$$\mathbf{K}_1 - \mathbf{f}\left(\mathbf{x}_k + \frac{\Delta t}{2}\mathbf{K}_1, \mathbf{z}_1, \mathbf{u}\left(t_k + \frac{1}{2}\Delta\right)\right) = 0 \tag{66a}$$

$$\mathbf{g}\left(\mathbf{x}_k + \frac{\Delta t}{2}\mathbf{K}_1, \mathbf{z}_1, \mathbf{u}\left(t_k + \frac{1}{2}\Delta\right)\right) = 0 \tag{66b}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta t\mathbf{K}_1 \tag{66c}$$

Appendix: some possibly useful formula

- Lagrange mechanics is built on the equations:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial \dot{\mathbf{q}}} - \frac{\partial \mathcal{L}}{\partial \mathbf{q}} = \mathbf{Q}, \qquad \mathcal{L}\left(\mathbf{q}, \dot{\mathbf{q}}, \mathbf{z}\right) = \mathcal{T} - \mathcal{V} - \mathbf{z}^\top \mathbf{C}, \qquad \mathbf{C} = 0, \qquad \langle \delta\mathbf{q}, \mathbf{Q}\rangle = \delta W, \, \forall\, \delta\mathbf{q} \tag{67}$$

  The kinetic and potential energy of a point mass are given by:

$$\mathcal{T} = \frac{1}{2}m\dot{\mathbf{p}}^\top\dot{\mathbf{p}}, \qquad \mathcal{V} = mg\mathbf{p}_3 \tag{68}$$

  respectively, where $\mathbf{p} \in \mathbb{R}^3$ is the position of the mass in a cartesian reference frame having the third coordinate as the vertical axis pointing up. The generalized forces are identical to the external forces applied to a point mass if the position of that point is expressed in cartesian coordinates in the generalized coordinates $\mathbf{q}$.

- In the case $\mathcal{T} = \frac{1}{2}m\dot{\mathbf{q}}^\top W\dot{\mathbf{q}}$ with W constant $\mathcal{V} = \mathcal{V}\left(\mathbf{q}\right)$ and $\mathbf{C} = \mathbf{C}\left(\mathbf{q}\right)$, the Lagrange equations simplify to the dynamics in the semi-explicit index-3 DAE form:

$$\dot{\mathbf{p}} = \mathbf{v} \tag{69a}$$

$$W\dot{\mathbf{v}} + \frac{\partial \mathbf{C}}{\partial \mathbf{q}}^\top \mathbf{z} = \mathbf{Q} - \frac{\partial \mathcal{V}}{\partial \mathbf{q}}^\top \tag{69b}$$

$$0 = \mathbf{C}\left(\mathbf{q}\right) \tag{69c}$$

- The Implicit Function Theorem (IFT) guarantees that a nonlinear set of equations

$$\mathbf{r}\left(\mathbf{y}, \mathbf{z}\right) = 0 \tag{70}$$

  "can be solved" in terms of $\mathbf{z}$ for a given $\mathbf{y}$ iff the Jacobian $\frac{\partial \mathbf{r}(\mathbf{y},\mathbf{z})}{\partial \mathbf{z}}$ is full rank at the solution. More specifically, it guarantees that there is a function $\phi\left(\mathbf{y}\right)$ such that

$$\mathbf{r}\left(\mathbf{y}, \phi\left(\mathbf{y}\right)\right) = 0 \tag{71}$$

  holds in the neighborhood of the point $\mathbf{y}$ where the Jacobian is evaluated. Furthermore, the IFT specifies that:

$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = -\frac{\partial \mathbf{r}}{\partial \mathbf{z}}^{-1}\frac{\partial \mathbf{r}}{\partial \mathbf{y}} \tag{72}$$

- For solving a problem $\mathbf{r}\left(\mathbf{x}\right) = 0$, Newton iterates:

$$\mathbf{x} \leftarrow \mathbf{x} - \alpha\frac{\partial \mathbf{r}}{\partial \mathbf{x}}^{-1}\mathbf{r} \tag{73}$$

  until $\mathbf{r}\left(\mathbf{x}\right) \approx 0$ where $\alpha \in [0, 1]$

- Runge-Kutta methods are described by:

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}$$

$$\mathbf{K}_j = \mathbf{f}\left(\mathbf{x}_k + \Delta t\sum_{i=1}^{s} a_{ji}\mathbf{K}_i, \, \mathbf{u}\left(t_k + c_j\Delta t\right)\right), \quad j = 1, \dots, s \tag{74a}$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta t\sum_{i=1}^{s} b_i\mathbf{K}_i \tag{74b}$$

- For ERK methods, the relationship between the (minimum) number of stages s to the order o is given by:

s	1	2	3	4	6	7	9	11	...
o	1	2	3	4	5	6	7	8	...

Table 1: Stage to order of ERK methods

- Collocation methods use:

$$\dot{\mathbf{x}}(t_k + \Delta t \cdot \tau) \approx \dot{\hat{\mathbf{x}}}(t_k + \Delta t \cdot \tau) = \sum_{i=1}^{s} \mathbf{K}_i \ell_i(\tau), \quad \tau \in [0,\, 1] \tag{75}$$

$$\mathbf{x}(t_k + \Delta t \cdot \tau) \approx \hat{\mathbf{x}}(t_k + \Delta t \cdot \tau) = \mathbf{x}_k + \Delta t \sum_{i=1}^{s} \mathbf{K}_i L_i(\tau) \tag{76}$$

where the Lagrange polynomials are given by:

$$\ell_i(\tau) = \prod_{j=1, j\neq i}^{s} \frac{\tau - \tau_j}{\tau_i - \tau_j}, \quad \text{and} \quad L_i(\tau) = \int_0^{\tau} \ell_i(\xi)\mathrm{d}\xi \tag{77}$$

The Lagrange polynomials satisfy the conditions of

$$\text{Orthogonality:} \quad \int_0^1 \ell_i(\tau)\ell_j(\tau)\,\mathrm{d}\tau = 0 \quad \text{for} \quad i \neq j \tag{78a}$$

$$\text{Punctuality:} \quad \ell_i(\tau_j) = \begin{cases} 1 & \text{if} \quad j = i \\ 0 & \text{if} \quad j \neq i \end{cases} \tag{78b}$$

and enforce the collocation equations (for $j = 1, \dots, s$):

$$\dot{\hat{\mathbf{x}}}(t_k + \Delta t \cdot \tau_j) = \mathbf{f}\left(\hat{\mathbf{x}}(t_k + \Delta t \cdot \tau_j),\, \mathbf{u}\left(t_k + \Delta t \cdot \tau_j\right)\right), \quad \text{in the explicit ODE case} \tag{79a}$$

$$\mathbf{F}\left(\dot{\hat{\mathbf{x}}}(t_k + \Delta t \cdot \tau_j), \hat{\mathbf{x}}(t_k + \Delta t \cdot \tau_j),\, \mathbf{u}\left(t_k + \Delta t \cdot \tau_j\right)\right) = 0, \quad \text{in the implicit ODE case} \tag{79b}$$

$$\mathbf{F}\left(\dot{\hat{\mathbf{x}}}(t_k + \Delta t \cdot \tau_j), \hat{\mathbf{z}}_j, \hat{\mathbf{x}}(t_k + \Delta t \cdot \tau_j),\, \mathbf{u}\left(t_k + \Delta t \cdot \tau_j\right)\right) = 0, \quad \text{in the fully-implicit DAE case} \tag{79c}$$

- Gauss-Legendre collocation methods select the set of points $\tau_{1,\dots,s}$ as the zeros of the (shifted) Legrendre polynomial:

$$P_s\left(\tau\right) = \frac{1}{s!}\frac{\mathrm{d}^s}{\mathrm{d}\tau^s}\left[\left(\tau^2 - \tau\right)^s\right] \tag{80}$$

They achieve the order $\|\mathbf{x}_N - \mathbf{x}\left(t_{\mathrm{f}}\right)\| = \mathcal{O}\left(\Delta t^{2s}\right)$.

- Maximum-likelihood estimation is based on

$$\max_{\boldsymbol{\theta}} \quad \mathbb{P}\left[e_k = y_k - \hat{y}_k \quad \text{for} \quad k = 1, \dots, N \mid \boldsymbol{\theta}\right] \tag{81}$$

If the noise sequence is uncorrelated, then

$$\mathbb{P}\left[e_k = y_k - \hat{y}_k \quad \text{for} \quad k = 0, \dots, N \mid \boldsymbol{\theta}\right] = \prod_{k=1}^{N} \mathbb{P}\left[e_k = y_k - \hat{y}_k \mid \boldsymbol{\theta}\,\right] \tag{82}$$

- The solution of a linear least-squares problem

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left\|A\boldsymbol{\theta} - \mathbf{y}\right\|^2_{\Sigma_e^{-1}} \tag{83}$$

reads as:

$$\hat{\boldsymbol{\theta}} = \left(A^\top \Sigma_e^{-1} A\right)^{-1} A^\top \Sigma_e^{-1} \mathbf{y} \tag{84}$$

and the covariance of the parameter estimation based is given by the formula:

$$\Sigma_{\hat{\boldsymbol{\theta}}} = \left(A^\top \Sigma_e^{-1} A\right)^{-1} \tag{85}$$

- In system identification, given the a plant $G(z)$ and a noise $H(z)$ model description, the one-step-ahead predictor $\hat{y}(k|k-1)$ can be retrieved with

$$H(z)\hat{y}(z) = G(z)u(z) + (H(z) - 1)y(z) \tag{86}$$

- The Gauss-Newton approximation in an optimization problem

$$\min_{\mathbf{x}} \quad J\left(\mathbf{x}\right) = \frac{1}{2}\left\|\mathbf{R}\left(\mathbf{x}\right)\right\|^2 \tag{87}$$

uses the approximation:

$$\frac{\partial^2 J}{\partial \mathbf{x}^2} \approx \frac{\partial R}{\partial \mathbf{x}}^\top \frac{\partial R}{\partial \mathbf{x}} \tag{88}$$

- The solution to an LTI system $\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}$ is given by:

$$\mathbf{x}(t) = e^{At}\mathbf{x}(0) + \int_0^t e^{A(t-\tau)}B\mathbf{u}(\tau)\mathrm{d}\tau \tag{89}$$

and the transformation state-space to transfer function is given by:

$$G(s) = C\left(sI - A\right)^{-1} B + D \tag{90}$$