Politecnico di Milano - Bovisa
via La Masa 34, 20156 Milano

Graduate Course on "Multibody System Dynamics"
Ph.D. in Aerospace Engineering, Mechanical Systems Engineering,
and Rotary Wing Aircraft

Analysis of Systems of Differential-Algebraic Equations (DAE)

Paolo Mantegazza, Pierangelo Masarati

November 22, 2012

1 Mechanical Problems in Differential-Algebraic Form

The dynamics of unconstrained mechanical systems is described by Ordinary Differential Equations (ODE). Its solution does not present significant problems. The dynamics of constrained systems implies the addition of kinematic constraints in form of algebraic relationships between the kinematic variables that describe the motion of the unconstrained system. Type of problems can be solved in rather different manners; the most appropriate approach needs to be carefully considered, often on a case by case basis. This discussion aims at providing useful indications to determine the most appropriate approach.

1.1 Unconstrained Systems

The dynamics of a generic mechanical system is described by the differential equation

$$\mathbf{M}(\mathbf{q})\,\ddot{\mathbf{q}} = \mathbf{f}(\dot{\mathbf{q}}, \mathbf{q}, t), \tag{1}$$

where $\mathbf{q}$ and its derivatives represents the kinematics of the system, while $\mathbf{M}$ is the mass matrix and $\mathbf{f}$ are the remaining forces. The same equation describes the dynamics of a continuum after discretization.

1.2 Constrained Systems

Kinematic constraints are naturally expressed in form of algebraic relationships between kinematic variables. Two kinds of constraints can be formulated: *holonomic* and *non-holonomic*:

- holonomic constraints express a finite algebraic relationship between kinematic variables, e.g.

  $$\boldsymbol{\phi}(\mathbf{q}, t) = \mathbf{0}; \tag{2}$$

  when an explicit dependence on time t is present, the constraint is called *rheonomic*, otherwise it is called *scleronomic*;

- non-holonomic constraints express an algebraic relationship in differential, non-integrable form, e.g.

  $$\mathbf{A}(\mathbf{q}, t)\,\dot{\mathbf{q}} - \mathbf{b}(\mathbf{q}, t)' = \mathbf{0}. \tag{3}$$

  The constraint is usually expressed, as above, in form of the time derivative of $\mathbf{q}$, but it should be thought as

  $$\mathbf{A}(\mathbf{q}, t)\,\mathrm{d}\mathbf{q} - \mathbf{b}(\mathbf{q}, t)'\,\mathrm{d}t = \mathbf{0}, \tag{4}$$

  i.e. like the differential of a non-existing continuous function of $\mathbf{q}$ and t.

Exercise 1.1 *Formulate the equation that expresses the (non-holonomic) constraint of a disc rolling on a straight line in two-dimensions.*

Exercise 1.2 *Formulate the equations that express the (non-holonomic) constraints of a 'marble' rolling on a flat surface.*

The presence of holonomic or non-holonomic constraints implies the addition of algebraic equations in the form of Eq. (2) or Eq. (3). Constraint reactions arise, in the form of Lagrange multipliers $\boldsymbol{\lambda}$. Consider the case of holonomic constraints: the Lagrange multipliers result from augmenting the approach used to generate the unconstrained dynamics equations, Eq. (1), with the virtual perturbation of a contribution of the form

$$\boldsymbol{\lambda} \cdot \boldsymbol{\phi}\left(\mathbf{q}, t\right), \tag{5}$$

which yields

$$\delta\left(\boldsymbol{\lambda} \cdot \boldsymbol{\phi}\left(\mathbf{q}, t\right)\right) = \delta\boldsymbol{\lambda} \cdot \boldsymbol{\phi}\left(\mathbf{q}, t\right) + \delta\mathbf{q} \cdot \boldsymbol{\phi}_{/\mathbf{q}}^{T}\boldsymbol{\lambda}. \tag{6}$$

The term that multiplies $\delta\boldsymbol{\lambda}$ is the constraint equation, while the one that multiplies $\delta\mathbf{q}$ is added to the equations of motion, yielding

$$\mathbf{M}\left(\mathbf{q}\right)\ddot{\mathbf{q}} = \mathbf{f}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right) + \boldsymbol{\phi}_{/\mathbf{q}}^{T}\boldsymbol{\lambda} \tag{7a}$$

$$\boldsymbol{\phi}\left(\mathbf{q}, t\right) = \mathbf{0}. \tag{7b}$$

Similarly, in case of non-holonomic constraints, one obtains

$$\mathbf{M}\left(\mathbf{q}\right)\ddot{\mathbf{q}} = \mathbf{f}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right) + \mathbf{A}^{T}\boldsymbol{\lambda} \tag{8a}$$

$$\mathbf{A}\left(\mathbf{q}, t\right)\dot{\mathbf{q}} + \mathbf{b}\left(\mathbf{q}, t\right) = \mathbf{0}. \tag{8b}$$

In this case, one might be tempted to write a form

$$\boldsymbol{\lambda} \cdot \left(\mathbf{A}\left(\mathbf{q}, t\right)\dot{\mathbf{q}} + \mathbf{b}\left(\mathbf{q}, t\right)\right) \qquad \boxed{\text{NO!}} \tag{9}$$

and perturb it much like in the case of holonomic constraints; *this would lead to incorrect results!* One should rather consider the form of Eq. (4) and replace $\mathrm{d}\mathbf{q}$ with $\delta\mathbf{q}$ at fixed time, according to the definition of virtual displacement, i.e. with $\mathrm{d}t = 0$, resulting in Eq. (8).

Non-holonomic constraint equations are algebraic, despite they contain the time derivatives of the kinematic variables, because they do not contain the time derivatives of the algebraic variables (the multipliers).

1.3 Multi-Field Problems

Often DAE systems arise from the formulation of multi-field problems, where different fields of analysis are glued together. Consider for example coupled problems where a mechanical system is actuated by a hydraulic subsystem, which in turn is controlled by an electrical network, and the whole system is immersed in an aerodynamic field whose boundary conditions are determined by the deformation of the structure and the motion of hydraulically actuated control surfaces.

The dynamics of electrical and hydraulic networks are often written in terms of current and flow balance equations at the nodes of the network, complemented by the constitutive

properties of each branch of the network (capacitors, resistors, inductors, current and voltage generators, motors, or pipes, orifices, accumulators, pressure and flow generators, actuators, servo-valves).

The interface between different physical models, for generality, may be expressed in terms of compatibility equations between homogeneous quantities using algebraic constraints.

1.4 Solution Approaches

Two main approaches can be followed to solve constrained system dynamics:

- reduction to ODE;
- direct solution of DAE.

The two approaches are rather different, but present significant common aspects. The convenience of each approach needs to be considered with care, although in computer methods the latter is emerging. These notes should help the reader in determining the most appropriate approach for specific problems.

An intermediate approach, substantially analogous to the first one, consists in removing from the external loads the portion that would cause a constraint violation, without formally reducing the number of equations and kinematic variables.

1.5 Direct Elimination of Lagrange Multipliers

This is the case of the above mentioned intermediate approach. The exact solution of the constraint problem implies to reformulate the constraint equations in terms of the second derivative of the kinematic variables. When the constraint is holonomic, this implies

$$\mathbf{0} = \dot{\boldsymbol{\phi}}(\mathbf{q}, t) = \boldsymbol{\phi}_{/\mathbf{q}}\dot{\mathbf{q}} + \boldsymbol{\phi}_{/t} \qquad = \boldsymbol{\phi}_{/\mathbf{q}}\dot{\mathbf{q}} + \mathbf{b}' \tag{10a}$$

$$\mathbf{0} = \ddot{\boldsymbol{\phi}}(\mathbf{q}, t) = \boldsymbol{\phi}_{/\mathbf{q}}\ddot{\mathbf{q}} + \left(\boldsymbol{\phi}_{/\mathbf{q}}\dot{\mathbf{q}} + \boldsymbol{\phi}_{/t}\right)_{/\mathbf{q}}\dot{\mathbf{q}} + \left(\boldsymbol{\phi}_{/\mathbf{q}}\dot{\mathbf{q}} + \boldsymbol{\phi}_{/t}\right)_{/t} \qquad = \boldsymbol{\phi}_{/\mathbf{q}}\ddot{\mathbf{q}} + \mathbf{b}'' \tag{10b}$$

The case of non-holonomic constraints is formally equivalent to that of holonomic constraints starting from their first derivative, namely

$$\mathbf{0} = \mathbf{A}\ddot{\mathbf{q}} + \left(\mathbf{A}\dot{\mathbf{q}} - \mathbf{b}'\right)_{/\mathbf{q}}\dot{\mathbf{q}} + \left(\mathbf{A}\dot{\mathbf{q}} - \mathbf{b}'\right)_{/t} \qquad = \mathbf{A}\ddot{\mathbf{q}} + \mathbf{b}''. \tag{11}$$

At this point, $\ddot{\mathbf{q}}$ can be made explicit from the equations of motion, and substituted in the derivative of the constraint equations, yielding Lagrange's multipliers,

$$\boldsymbol{\lambda} = -\left(\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{q}}^{T}\right)^{-1}\left(\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{M}^{-1}\mathbf{f}(\dot{\mathbf{q}}, \mathbf{q}, t) + \mathbf{b}''\right). \tag{12}$$

After substituting the multipliers in the equations of motion a formally purely differential problem results, that intrinsically complies with the (second derivative of the) constraints (Gauss [1], Udwadia and Kalaba [2]):

$$\mathbf{M}(\mathbf{q})\,\ddot{\mathbf{q}} = \left(\mathbf{I} - \boldsymbol{\phi}_{/\mathbf{q}}^{T}\left(\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{q}}^{T}\right)^{-1}\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{M}^{-1}\right)\mathbf{f}(\dot{\mathbf{q}}, \mathbf{q}, t) - \boldsymbol{\phi}_{/\mathbf{q}}^{T}\left(\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{q}}^{T}\right)^{-1}\mathbf{b}''. \tag{13}$$

The problem is a bit more complicated in case of non-ideal constraints, i.e. when constraint reactions do work for a virtual displacement (e.g. in the case of friction).

This problem is formally purely differential; the resulting equations appear as *pure*, since constraint reactions are not present. However, the problem is not constrained. As Eq. (13) shows, constraints are complied with by removing the portion of the forces that would violate the second derivative of the constraint with the (oblique) projector

$$\mathbf{P} = \mathbf{I} - \boldsymbol{\phi}_{/\mathbf{q}}^T \left(\boldsymbol{\phi}_{/\mathbf{q}} \mathbf{M}^{-1} \boldsymbol{\phi}_{/\mathbf{q}}^T \right)^{-1} \boldsymbol{\phi}_{/\mathbf{q}} \mathbf{M}^{-1}. \tag{14}$$

Since the original constraint is not explicitly enforced, a numerical integration of Eq. (13) suffers from drift and thus is impractical unless measures are taken to control it, as discussed in subsequent sections.

1.6 System Reduction to Pure Equations of Motion

According to the projection illustrated in the previous section, the problem of unconstrained dynamics is replaced by a problem formally of the same dimensions, whose solution is constrained in a subspace of the original space that is tangent to the constraint manifold.

At this point one can ask the question: can we reduce the problem to the dimensions of the subspace that is actually unconstrained? This can be obtained using the constraint equations to express the constrained variables as functions of the unconstrained ones, the actual coordinates in a Lagrangian sense.

For example, the kinematic variables can be partitioned in unconstrained (subscript a) and constrained (subscript o),

$$\mathbf{q} = \left\{ \begin{array}{ll} \mathbf{q}_a & \text{dof used in the analysis} \\ \mathbf{q}_o & \text{dof omitted} \end{array} \right\}, \tag{15}$$

where, in analogy with the notation used in NASTRAN, the subscript o indicates the *omitted* variables, while the subscript a indicates those that are preserved in the *analysis*.

The first derivative of the holonomic constraint equation, or the non-holonomic constraint equation, can be used to express the constrained variables as functions of the unconstrained ones,

$$\dot{\mathbf{q}}_o = -\boldsymbol{\phi}_{/\mathbf{q}_o}^{-1} \left(\boldsymbol{\phi}_{/\mathbf{q}_a} \dot{\mathbf{q}}_a + \boldsymbol{\phi}_{/t} \right). \tag{16}$$

This is possible as soon as matrix $\boldsymbol{\phi}_{/\mathbf{q}_o}$ is square (this is guaranteed by the fact that the constrained variables must be as many as the constraint equations) and non-singular. Singularity can only occur when constraints are redundant, possibly indicating a singular configuration.

1.7 Maggi's Equations

The derivatives of the kinematic variables can be expressed in compact form as

$$\dot{\mathbf{q}} = \left[\begin{array}{c} \mathbf{I} \\ -\boldsymbol{\phi}_{/\mathbf{q}_o}^{-1} \boldsymbol{\phi}_{/\mathbf{q}_a} \end{array} \right] \mathbf{e} + \left\{ \begin{array}{c} \mathbf{0} \\ -\boldsymbol{\phi}_{/\mathbf{q}_o}^{-1} \boldsymbol{\phi}_{/t} \end{array} \right\} = \mathbf{T}\mathbf{e} + \mathbf{t}, \tag{17}$$

where $\mathbf{e} = \dot{\mathbf{q}}_a$.

Exercise 1.3 *Show that* $\dot{\mathbf{q}}$ *from Eq. (17) intrinsically comply with the first derivative of the holonomic constraint, Eq. (10a).*

The second derivative of the kinematic variables is

$$\ddot{\mathbf{q}} = \mathbf{T}\dot{\mathbf{e}} + \mathbf{h}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right), \tag{18}$$

where $\mathbf{h}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right) = \dot{\mathbf{T}}\mathbf{e} + \dot{\mathbf{t}}$ contains everything that does not contain $\dot{\mathbf{e}}$, while $\mathbf{T}$ is the matrix of the linear part of the transformation.

The equilibrium equation can be rewritten in terms of $\dot{\mathbf{e}}$,

$$\mathbf{M}\mathbf{T}\dot{\mathbf{e}} = \mathbf{f}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right) - \mathbf{M}\mathbf{h}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right) + \boldsymbol{\phi}_{/\mathbf{q}}^{T}\boldsymbol{\lambda}. \tag{19}$$

The problem is reduced to the subset of the unconstrained kinematic variables (*minimal coordinate set*) by premultiplication by the transpose of matrix $\mathbf{T}$,

$$\mathbf{T}^{T}\mathbf{M}\mathbf{T}\dot{\mathbf{e}} = \mathbf{T}^{T}\mathbf{f}^{*}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right) + \mathbf{T}^{T}\boldsymbol{\phi}_{/\mathbf{q}}^{T}\boldsymbol{\lambda}, \tag{20}$$

with $\mathbf{f}^{*}(\dot{\mathbf{q}}, \mathbf{q}, t) = \mathbf{f}(\dot{\mathbf{q}}, \mathbf{q}, t) - \mathbf{M}\mathbf{h}(\dot{\mathbf{q}}, \mathbf{q}, t)$. Matrix $\mathbf{T}$ expresses a subspace that is intrinsically orthogonal to that described by the multipliers' matrix, $\boldsymbol{\phi}_{/\mathbf{q}}^{T}$,

$$\begin{bmatrix} \mathbf{I} \\ -\boldsymbol{\phi}_{/\mathbf{q}_o}^{-1}\boldsymbol{\phi}_{/\mathbf{q}_a} \end{bmatrix}^{T} \begin{bmatrix} \boldsymbol{\phi}_{/\mathbf{q}_a}^{T} \\ \boldsymbol{\phi}_{/\mathbf{q}_o}^{T} \end{bmatrix} = \boldsymbol{\phi}_{/\mathbf{q}_a}^{T} - \left(\boldsymbol{\phi}_{/\mathbf{q}_o}^{-1}\boldsymbol{\phi}_{/\mathbf{q}_a}\right)^{T}\boldsymbol{\phi}_{/\mathbf{q}_o}^{T} = \mathbf{0}. \tag{21}$$

Eqs. (20) thus reduce to the pure equations

$$\mathbf{T}^{T}\mathbf{M}\mathbf{T}\dot{\mathbf{e}} = \mathbf{T}^{T}\mathbf{f}^{*}\left(\dot{\mathbf{q}}, \mathbf{q}, t\right). \tag{22}$$

They are called *Maggi's equations* [3], from the name of the Italian mathematician that first formulated them at the end of the XIXth century. They are also known as *Kane's equations*, from the name of the American mathematician that independently rediscovered them in the second half of the XXth century [4]. They are well illustrated in [5].

The resulting system is second-order differential in the unconstrained variables $\mathbf{q}_a$, and first-order differential in the constrained variables $\mathbf{q}_o$,

$$\hat{\mathbf{M}}\ddot{\mathbf{q}}_a = \hat{\mathbf{f}}\left(\dot{\mathbf{q}}_a, \dot{\mathbf{q}}_o, \mathbf{q}_a, \mathbf{q}_o, t\right) \tag{23a}$$

$$\dot{\mathbf{q}}_o = -\boldsymbol{\phi}_{/\mathbf{q}_o}^{-1}\left(\boldsymbol{\phi}_{/\mathbf{q}_a}\dot{\mathbf{q}}_a + \boldsymbol{\phi}_{/t}\right), \tag{23b}$$

where $\hat{\mathbf{M}} = \mathbf{T}^{T}\mathbf{M}\mathbf{T}$ and $\hat{\mathbf{f}} = \mathbf{T}^{T}\mathbf{f}^{*}$.

The constrained variables actually need to be solved according to the constraint equation, Eq. (2), and not only to its derivative, to avoid numerical drift. As a consequence, the values of $\mathbf{q}_o$ resulting from the integration of Eq. (23b) need to be corrected to comply with Eq. (2).

1.8 Minimal Set and Orthogonality

The main concept at the roots of this formulation consists in the determination of a subspace $\mathbf{T}$ of the space of kinematic variables that is locally orthogonal to the constraint manifold, such that $\mathbf{T}^T \boldsymbol{\phi}_{/\mathbf{q}}^T = \mathbf{0}$. The expression of matrix $\mathbf{T}$ resulting from Eq. (17) is not the only possible one; on the contrary, often the choice of $\mathbf{q}_o$ is not easy and may result in algorithms that are not sufficiently robust.

Exercise 1.4 *Consider a point-mass pendulum, described by the constraint equation* $\phi = (x^2 + y^2)^{1/2} - L = 0$. *Why matrix* $\mathbf{T}$ *cannot be determined according to Eq. (17) in a manner that is equally valid for any position of the point mass?*

An efficient and robust algorithm is based on the QR decomposition [6]. This decomposition factorizes a generic rectangular matrix into the product of a square unit matrix, $\mathbf{Q}$, such that $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, and a rectangular matrix, $\mathbf{R}$, that is upper-triangular. In this case, the generic rectangular matrix is $\boldsymbol{\phi}_{/\mathbf{q}}^T$, the matrix that expresses the effect of Lagrange's multipliers on the equations of motion. The dimensions of matrix $\boldsymbol{\phi}_{/\mathbf{q}}^T$ are $n \times m$, with $n > m$. The resulting unit matrix is $n \times n$, while matrix $\mathbf{R}$ is also $n \times m$. Its upper $m \times m$ portion, $\mathbf{R}_1$, is upper-triangular, while its lower $(n-m) \times n$ portion, $\mathbf{R}_2$, is zero:

$$\boldsymbol{\phi}_{/\mathbf{q}}^T = \mathbf{QR} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_2 \end{bmatrix} \begin{bmatrix} \mathbf{R}_1 \\ \mathbf{0} \end{bmatrix} = \mathbf{Q}_1\mathbf{R}_1. \tag{24}$$

Only submatrix $\mathbf{Q}_1$ of matrix $\mathbf{Q}$, of dimensions $n \times m$, is used, since submatrix $\mathbf{Q}_2$ multiplies a null matrix. However, submatrix $\mathbf{Q}_2$, of dimensions $n \times n - m$, represents a convenient subspace that is guaranteed to be orthogonal to the subspace tangent to the constraint manifold. In fact, if one chooses $\mathbf{T} = \mathbf{Q}_2$, Eq. (21) is intrinsically satisfied, since $\mathbf{T}^T\boldsymbol{\phi}_{/\mathbf{q}}^T = \mathbf{Q}_2^T\mathbf{Q}_1\mathbf{R}_1$, and $\mathbf{Q}_2^T\mathbf{Q}_1 = \mathbf{0}$ by definition.

Using matrix $\mathbf{Q}_2$, $\dot{\mathbf{q}}$ can be expressed as a function of $\mathbf{e}$, according to

$$\dot{\mathbf{q}} = \mathbf{Q}_2\mathbf{e} + \mathbf{Q}_1\mathbf{p}', \tag{25}$$

where $\mathbf{p}'$ is an unknown vector of size $m \times 1$. Vector $\mathbf{p}'$ is determined by substituting $\dot{\mathbf{q}}$ in the first derivative of the constraint equation,

$$\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{Q}_2\mathbf{e} + \boldsymbol{\phi}_{/\mathbf{q}}\mathbf{Q}_1\mathbf{p}' + \mathbf{b}' = \mathbf{0}. \tag{26}$$

This yields

$$\mathbf{R}_1^T \underbrace{\mathbf{Q}_1^T\mathbf{Q}_2}_{\equiv\mathbf{0}}\mathbf{e} + \mathbf{R}_1^T \underbrace{\mathbf{Q}_1^T\mathbf{Q}_1}_{\equiv\mathbf{I}}\mathbf{p}' + \mathbf{b}' = \mathbf{0}, \tag{27}$$

and thus

$$\mathbf{p}' = -\mathbf{R}_1^{-T}\mathbf{b}'. \tag{28}$$

This implies that $\mathbf{p}' = \mathbf{0}$ when the constraint is rheonomic.

Then, since $\ddot{\mathbf{q}}$ is needed,

$$\ddot{\mathbf{q}} = \mathbf{Q}_2\dot{\mathbf{e}} + \mathbf{Q}_1\mathbf{p}'', \tag{29}$$

with $\mathbf{p}''$ unknown as well. Substitute $\ddot{\mathbf{q}}$ in the second derivative of the constraint equation,

$$\boldsymbol{\phi}_{/\mathbf{q}}\mathbf{Q}_2\dot{\mathbf{e}} + \boldsymbol{\phi}_{/\mathbf{q}}\mathbf{Q}_1\mathbf{p}'' + \mathbf{b}'' = \mathbf{0}. \tag{30}$$

This yields

$$\mathbf{R}_1^T \underbrace{\mathbf{Q}_1^T\mathbf{Q}_2}_{\equiv\mathbf{0}}\dot{\mathbf{e}} + \mathbf{R}_1^T \underbrace{\mathbf{Q}_1^T\mathbf{Q}_1}_{\equiv\mathbf{I}}\mathbf{p}'' + \mathbf{b}'' = \mathbf{0} \tag{31}$$

and thus

$$\mathbf{p}'' = -\mathbf{R}_1^{-T}\mathbf{b}''. \tag{32}$$

The acceleration becomes

$$\ddot{\mathbf{q}} = \mathbf{Q}_2\dot{\mathbf{e}} - \underbrace{\mathbf{Q}_1\mathbf{R}_1^{-T}}_{\boldsymbol{\phi}^+_{/\mathbf{x}}}\mathbf{b}'' \tag{33}$$

Exercise 1.5 *Prove that* $\mathbf{Q}_1\mathbf{R}_1^{-T}$ *corresponds to the Moore-Penrose pseudo-inverse of matrix* $\boldsymbol{\phi}_{/\mathbf{x}}$.

The problem is integrated to yield $\mathbf{e}$; from this, $\dot{\mathbf{q}}$ is computed using Eq. (25). Finally, $\dot{\mathbf{q}}$ is integrated to yield $\mathbf{q}^{(0)}$, an estimate of $\mathbf{q}$. The estimate may need to be refined by enforcing the constraint equation, which requires to solve the implicit nonlinear problem $\boldsymbol{\phi}(\mathbf{q}) = \mathbf{0}$.

A minimal norm solution can be obtained considering a solution of the form

$$\mathbf{q} = \mathbf{q}^{(0)} + \boldsymbol{\phi}^T_{/\mathbf{q}}\boldsymbol{\nu}, \tag{34}$$

where $\Delta\boldsymbol{\nu}$ is computed by solving the problem

$$\boldsymbol{\phi}_{/\mathbf{q}}\Delta\mathbf{q} = -\boldsymbol{\phi}. \tag{35}$$

Since, according to Eq. (34), $\Delta\mathbf{q} = \boldsymbol{\phi}^T_{/\mathbf{q}}\Delta\boldsymbol{\nu}$,

$$\Delta\boldsymbol{\nu} = -\left(\boldsymbol{\phi}_{/\mathbf{q}}\boldsymbol{\phi}^T_{/\mathbf{q}}\right)^{-1}\boldsymbol{\phi}, \tag{36}$$

one obtains

$$\mathbf{q}^{(k+1)} = \mathbf{q}^{(k)} - \underbrace{\boldsymbol{\phi}^T_{/\mathbf{q}}\left(\boldsymbol{\phi}_{/\mathbf{q}}\boldsymbol{\phi}^T_{/\mathbf{q}}\right)^{-1}}_{\boldsymbol{\phi}^+_{/\mathbf{q}}\left(\mathbf{q}^{(k)}\right)}\boldsymbol{\phi}\left(\mathbf{q}^{(k)}\right). \tag{37}$$

Exercise 1.6 *Apply the QR decomposition to the Jacobian matrix of the pendulum constraint equation of exercise 1.4 and determine the optimal* $\mathbf{Q}_2$ *(suggestion: consider* $\boldsymbol{\phi}_{/\mathbf{q}}\boldsymbol{\phi}^T_{/\mathbf{q}}$ *with* $\boldsymbol{\phi}_{/\mathbf{q}}$ *in QR form).*

A similar result can be obtained using the Singular Value Decomposition (SVD) [7, 8],

$$\boldsymbol{\phi}_{/\mathbf{q}}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T = \begin{bmatrix} \mathbf{U}_1 & \mathbf{U}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} \\ \mathbf{0} \end{bmatrix} \mathbf{V}^T = \mathbf{U}_1 \boldsymbol{\Sigma} \mathbf{V}, \tag{38}$$

where $\mathbf{U}$ and $\mathbf{V}$ are unit matrices respectively of dimensions $n \times n$ and $m \times m$, while $\mathbf{S}$ is $n \times m$. Matrix $\boldsymbol{\Sigma}$ is diagonal, $m \times m$, and contains the singular values of matrix $\boldsymbol{\phi}_{/\mathbf{q}}$. Singular values are non-negative by definition; they are positive if the constraints are well-posed, and thus the rank of matrix $\boldsymbol{\phi}_{/\mathbf{q}}$ is m. Matrix $\mathbf{U}_2$ has the same role that matrix $\mathbf{Q}_2$ has within the QR decomposition.

Exercise 1.7 *Apply the SVD decomposition to the Jacobian matrix of the pendulum constraint equation of exercise 1.4 and determine the optimal* $\mathbf{U}_2$ *(suggestion: consider* $\boldsymbol{\phi}_{/\mathbf{q}} \boldsymbol{\phi}_{/\mathbf{q}}^T$ *in SVD form).*

Other approaches are related to the zero-eigenvalue theorem [9], the Gramm-Schmidt orthonormalization [10, 11], and more. A detailed discussion of projection algorithms is given in [12] and [13].

Exercise 1.8 *The zero-eigenvalue theorem builds on top of eigenanalysis of* $\boldsymbol{\phi}_{/\mathbf{q}}^T \boldsymbol{\phi}_{/\mathbf{q}}$*. Can you infer the resulting optimal* $\mathbf{T}$*?*

Exercise 1.9 *Compare the zero-eigenvalue theorem and the SVD approach; why is the latter more appealing than the former in practical applications?*

2 Ordinary Differential Equations

A system of *Ordinary Differential Equations* (ODE) in explicit form is given by

$$\dot{\mathbf{y}} = \mathbf{f}\left(\mathbf{y}, t\right). \tag{39}$$

The study of this type of problems, with the exception of Linear, Time-Invariant (LTI) and Linear, Time-Periodic (LTP) problems, is nearly impossible without resorting to numerical methods.

Using numerical methods, usually the system cannot be studied; only specific problems can be solved, consisting of Eq. (39) with a set of initial values, $\mathbf{y}(t_0) = \mathbf{y}_0$, yielding an Initial Value Problem (IVP),

$$\dot{\mathbf{y}} = \mathbf{f}\left(\mathbf{y}, t\right), \qquad \mathbf{y}\left(t_0\right) = \mathbf{y}_0. \tag{40}$$

Its solution is represented by function $\mathbf{y}\left(t\right)$.

2.1 Solution Approximation

The solution of an IVP is called *integration*. When integration in closed form is not possible, numerical integration methods need to be used to approximate the solution. Various approaches are available, with fairly different properties. Criteria need to be defined to be able to choose the most appropriate method for a specific problem. Various classifications are possible; fundamental properties are:

- implicit/explicit
- conditionally/unconditionally stable
- single step/multiple step

Consider a simple mechanical problem, consisting in an undamped system with no excitation (this problem is rather critical) in the form

$$m\ddot{x} + kx = 0. \tag{41}$$

A finite-difference approximation can be used, consisting in

$$\ddot{x} = \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2}. \tag{42}$$

The approximation can be formulated in different manners:

- *centered* differences when the numerical derivative is evaluated in the center of the interval, namely $\ddot{x} = \ddot{x}_k$;
- *forward* differences when the numerical derivative is evaluated at the current time step, namely $\ddot{x} = \ddot{x}_{k+1}$.

Exercise 2.1 *Evaluate the accuracy of centered and forward differences (hint: integrate a polynomial function $f = a_0 + a_1 t + a_2 t^2 + \ldots$ over a time step h, and find the lowest coefficient that differs from the expected value).*

2.1.1 Centered Differences

When $\ddot{x} = \ddot{x}_k$ the prediction is straightforward:

$$\ddot{x}_k = \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2}, \tag{43}$$

resulting in

$$m\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + kx_k = 0 \tag{44}$$

and thus

$$x_{k+1} = \left(2 - h^2\frac{k}{m}\right) x_k - x_{k-1}. \tag{45}$$

2.1.2 Forward Differences

When $\ddot{x} = \ddot{x}_{k+1}$ the unknown is simultaneously present in the derivative and in the state, thus the formula is implicit:

$$\ddot{x}_{k+1} = \frac{x_{k+1} - 2x_k + x_{k-1}}{h^2}, \tag{46}$$

which yields

$$m\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + kx_{k+1} = 0 \tag{47}$$

and thus

$$\left(\frac{m}{h^2} + k\right) x_{k+1} = \frac{m}{h^2}\left(2x_k - x_{k-1}\right). \tag{48}$$

One can easily verify that the first formula (centered differences) gives higher accuracy, but its stability is conditioned by the need to select an appropriate time step below a given limit that depends on m and k. On the contrary the second formula (forward differences), although less accurate, yields stability properties that do not depend on the size of the time step.

Proof: The solution of a homogeneous difference LTI equation is $x_k = \rho x_{k-1}$, which yields $x_k = \rho^k x_0$. In order to be asymptotically stable, $|\rho| < 1$ must hold.

Centered differences yield

$$\frac{m}{h^2}\rho^2 = \frac{m}{h^2}\left(2\rho - 1\right) - k\rho, \tag{49}$$

namely

$$\rho = \left(1 - \frac{kh^2}{2m}\right) \pm j\sqrt{1 - \left(1 - \frac{kh^2}{2m}\right)^2}. \tag{50}$$

As soon as the roots are complex conjugates, namely

$$1 - \left(1 - \frac{kh^2}{2m}\right)^2 > 0, \tag{51}$$

$|\rho|$ is unit, and accuracy is maximal: the amplitude of the solution does not reduce, which is correct for an undamped system, while the period of the oscillation presents some error, since ρ is complex and not real.

However, the roots are no longer complex conjugates when $h = 2\sqrt{m/k}$; at this point, one of the real roots makes $\rho > 1$, thus the numerical solution of a stable problem becomes unstable: the stability of the integration method is conditional, depending on the time step.

In case of forward differences the equation becomes

$$\left(\frac{m}{h^2} + k\right)\rho^2 - 2\frac{m}{h^2}\rho + \frac{m}{h^2} = 0, \tag{52}$$

namely

$$\rho = \frac{1 \pm j\sqrt{\frac{k}{m}h^2}}{\left(1 + \frac{k}{m}h^2\right)}. \tag{53}$$

The two roots are now always complex conjugates, except for $h = 0$; the modulus of ρ is always less than 1, except for $h = 0$. The homogeneous solution of an undamped (stable) system will eventually converge to zero, unless $h \equiv 0$.

This characteristic should not be considered absolutely negative. In fact, what is typically of interest when solving dynamic systems are often forced cases. As a consequence, an unconditionally stable integration method that introduces some algorithmic dissipation will cancel the response of the system to perturbations of the initial conditions, while the negative effect of the integration scheme on the forced response is limited, the more the smaller the time step is, and thus ρ is closer to 1.

What makes forward differences not so appealing when mechanical systems need to be integrated is the fact that for $h > 0$ they introduce algorithmic dissipation even in systems with null physical damping, since $|\rho| < 1 \; \forall h > 0$.

2.2 Implicit and Explicit Methods

A numerical integration method is *explicit* when the function to be integrated, $\dot{\mathbf{y}} = \mathbf{f}(\mathbf{y}, t)$, needs to be evaluated only at times earlier than the one the problem is being solved at. Otherwise, a method is *implicit*.

Example: explicit Euler

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\mathbf{y}_k, t_k) \tag{54}$$

Example: implicit Euler

$$\mathbf{y}_{k+1} = \mathbf{y}_k + h\mathbf{f}(\mathbf{y}_{k+1}, t_{k+1}) \tag{55}$$

It is worth noticing that in the case of the implicit Euler the unknown $\mathbf{y}_{k+1}$ appears at the left- and at the right-hand side, and thus cannot be directly evaluated.

Exercise 2.2 *Evaluate the accuracy of the explicit and implicit Euler methods.*

2.3 Conditionally and Unconditionally Stable Methods

A method is *conditionally stable* if its stability is guaranteed only for time steps that belong to a finite time interval (starting from 0, as illustrated in the following). It is *unconditionally stable* if its stability is guaranteed regardless of the integration time step size. Rigorous stability criteria will be formulated in the following. They make the evaluation of the stability properties of the solution approximation methods possible.

Exercise 2.3 *Evaluate the stability of the explicit and implicit Euler methods, highlighting any limitation on the time step.*

2.4 Single- and Multistep Methods

Methods are often classified based on the number of previous steps the function to be integrated needs to be evaluated to compute the solution at the current time step. Single-step methods start from the knowledge of the solution at the last known step (or from the initial conditions when the first step is performed). Multistep methods require the knowledge of the solution more than one step behind. Single-step methods are *self-starting*, i.e. the method can be used to start from the initial time, while multistep methods require some start criterion.

2.5 Linear Multistep Methods

Typical linear multistep methods fall into the general formula

$$\sum_{j=0,k} \alpha_j \mathbf{y}_{n-j} = h \sum_{j=0,k} \beta_j \mathbf{f}\left(\mathbf{y}_{n-j}, t_{n-j}\right) \tag{56}$$

with $\alpha_0 = 1$; explicit methods have $\beta_0 = 0$. When the unknown at the current time step is isolated,

$$\mathbf{y}_n = -\sum_{j=1,k} \alpha_j \mathbf{y}_{n-j} + h \sum_{j=0,k} \beta_j \mathbf{f}\left(\mathbf{y}_{n-j}, t_{n-j}\right). \tag{57}$$

In practice, the unknown is expressed as a linear combination of its value and of the function to be integrated at previous time steps.

Interesting variants of the multistep methods are the so-called *one-leg* methods,

$$\sum_{j=0,k} \alpha_j \mathbf{y}_{n-j} = h\mathbf{f}\left(\sum_{j=0,k} \beta_j \mathbf{y}_{n-j}, \sum_{j=0,k} \beta_j t_{n-j}\right), \tag{58}$$

where the function is evaluated only at the current time step using a linear combination of the unknown at previous time steps; if the unknown value is included, namely $\beta_0 \neq 0$, the method is implicit.

Exercise 2.4 *Formulate explicit and implicit Euler according to the one-leg formula.*

A very interesting implicit method is the one-step *trapezoid rule* (a.k.a. Crank-Nicolson):

$$\mathbf{y}_n = \mathbf{y}_{n-1} + h\frac{\mathbf{f}\left(\mathbf{y}_n, t_n\right) + \mathbf{f}\left(\mathbf{y}_{n-1}, t_{n-1}\right)}{2} \tag{59}$$

which, in the one-leg variant, becomes

$$\mathbf{y}_n = \mathbf{y}_{n-1} + h\mathbf{f}\left(\frac{\mathbf{y}_n + \mathbf{y}_{n-1}}{2}, \frac{t_n + t_{n-1}}{2}\right) \tag{60}$$

with the same stability and accuracy properties. In practice function $\mathbf{y}$ is evaluated at step n as its value at step $n-1$ plus its forward projection with slope equal to either the average of the slopes at both ends of the time step, or that at mid-step.

Exercise 2.5 *Evaluate the accuracy of both forms of the one-step trapezoid rule.*

Exercise 2.6 *Evaluate the stability of both forms of the one-step trapezoid rule.*

2.6 Single-Step Methods

Typical single-step methods (other than multistep ones with $k = 1$, which represent the link between the two categories) can be cast into the Runge-Kutta scheme. They consist in defining a set of intermediate points within the time step, where the function is computed using an interpolation of pseudo-values of the derivative, which in turn are computed interpolating estimated values of the function. The solution approximation is obtained from an interpolation like

$$\mathbf{y}_n = \mathbf{y}_{n-1} + h \sum_{j=1,s} b_j \mathbf{Y}_j \tag{61a}$$

$$\mathbf{Y}_i = \mathbf{f}\left(\mathbf{y}_{n-1} + h \sum_{j=1,s} a_{ij} \mathbf{Y}_j, t_{n-1} + c_i h \right), \tag{61b}$$

subjected to $c_i = \sum_j a_{ij}$. The definition of the pseudo-derivative is recursive; moreover

- when the matrix of coefficients a_{ij} is lower sub-triangular ($a_{ij} = 0$ when $j \geq i$) the method is *explicit*;
- when it is lower triangular ($a_{ij} = 0$ when $j > i$), the method is *diagonally implicit* (Diagonally Implicit Runge-Kutta, DIRK); finally,
- if the diagonal coefficients a_{ii} are all equal, the method is *singly implicit* (Singly Implicit Runge-Kutta, SIRK)[1].

In the explicit case, the pseudo-derivatives $\mathbf{Y}_i$ at some intermediate point i only depend on those evaluated at previous intermediate points, while in the implicit case they also depend on themselves and possibly on subsequent ones.

However, in the Diagonally- and Singly-Implicit cases, there is no dependence between an intermediate point and the subsequent ones. This gives some computational advantages.

Runge-Kutta methods differ by the values assumed by the different coefficients, yielding different stability and accuracy properties. In case they are applied to DAEs, some additional requirements need to be met to guarantee stability.

The coefficients of the Runge-Kutta methods are usually represented using Butcher's matrix:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \tag{62}$$

Single-step methods usually require to evaluate function $\mathbf{f}$ at intermediate steps. When the method is explicit, the evaluations are performed in cascade. When the method is

[1] Nomenclature is not unique; often SIRK is used in the literature to indicate diagonally implicit methods, without any distinction between DIRK and SIRK.

implicit, the solution may require the simultaneous solution of the problem at multiple intermediate times. Only in case of DIRK the solution, although implicit, occur in cascade, and thus the size of each single implicit problem that needs to be solved is equal to the size of $\mathbf{f}$.

A single step algorithm can thus be much more computationally expensive than a multistep one. However, the freedom in the selection of the intermediate times (which need not be equally spaced as in multistep algorithms) and of the weights gives more flexibility in the design of the stability and accuracy properties of the algorithm.

2.7 Synthesis of Both Classes

Both methods can be unified interpreting them in form of discrete integration by collocation of a function approximated by interpolating elementary functions, in the spirit of the finite elements method [14]. The approximate solution can be expressed as

$$\hat{\mathbf{y}}(t) = \sum_{i=0,s} m_i(t)\,\mathbf{y}_{k-i} + h \sum_{i=0,s} n_i(t)\,\dot{\mathbf{y}}_{k-i} \tag{63}$$

The solution at a generic step according to the fundamental theorem of integral calculus applied to $\mathbf{f}$ is

$$\mathbf{y}_k = \mathbf{y}_{k-s} + \int_{k-s}^{k} \mathbf{f}(\mathbf{y}, t)\,\mathrm{d}t. \tag{64}$$

The integral can be approximated by collocation,

$$\mathbf{y}_k \cong \mathbf{y}_{k-s} + \sum_{j=1,n} w_j \mathbf{f}(\mathbf{y}(t_j), t_j). \tag{65}$$

Using the approximation of the solution to evaluate the function, one obtains

$$\mathbf{y}_k \cong \mathbf{y}_{k-s} + \sum_{j=1,n} w_j \mathbf{f}(\hat{\mathbf{y}}(t_j), t_j), \tag{66}$$

a form analogous to Runge-Kutta involving more than one step. Various methods known from the literature and a series of new methods with intermediate characteristics can be obtained by carefully selecting the shape functions and the numerical integration technique.

This approach shows how the coefficients of the Runge-Kutta methods can be interpreted as the result of applying a collocation integration method with a solution interpolation criterion. This is at the roots of the finite elements method.

One could use a Galerkin-like integration instead of the collocation (the latter can be considered a special case of the former). This makes the approach even more general for the design of integration methods, under the umbrella of weighted residuals in an extended interpretation, which includes approaches like *finite elements in time* (FET), even with discontinuities. However, in practical applications, the analytical integration of Galerkin-like formulas is impractical. As a consequence, one would need to resort again to collocation techniques.

2.8 Accuracy

An important property of solution approximation methods is accuracy. An intuitive definition of accuracy of linear methods is represented by the order of the polynomial that they can integrate exactly. A measure of the local error resulting from the integration, in turn, is given by the coefficient of the first non-null term of the remainder polynomial.

A more rigorous definition is represented by the limit, when the time step tends to zero, of the ratio between the errors when the time steps halves, namely

$$n = \lim_{h \to 0} \log_2 \left(\frac{E\left(2h\right)}{E\left(h\right)} \right) - 1 \tag{67}$$

The ratio is that the remainder of a method of order n is $o(h^n)$ and thus $O(h^(n+1))$; then $o((2h)^n) \approx 2^{n+1} o(h^n)$; thus

$$\lim_{h \to 0} \log_2 \left(\frac{E\left(2h\right)}{E\left(h\right)} \right) = \lim_{h \to 0} \log_2 \left(\frac{o((2h)^n)}{o(h^n)} \right) \approx \lim_{h \to 0} \log_2 \left(\frac{2^{n+1} o(h^n)}{o(h^n)} \right) = n + 1. \tag{68}$$

The convergence in damping and phase of a set of methods is reported in Fig. 1, while the error in damping and phase is reported in Fig. 2.

3 Linear Stability

The linear stability of a solution approximation method is studied by applying the method to a simple scalar linear differential problem

$$\dot{y} = \lambda y. \tag{69}$$

It represents a local linearization of a generic nonlinear problem about a fixed point, thus it contains information on the stability of the nonlinear problem in the vicinity of that solution. In this sense linear stability is also called *local stability*, since its results are only valid in the vicinity of an equilibrium point, and only assume a general value for Linear, Time Invariant problems (LTI).

Exercise 3.1 *Show how the generic linear damped oscillator without forcing term, $m\ddot{x} + r\dot{x} + kx = 0$, can be written in the form of Eq. (69).*

The application of a generic solution approximation method yields, in the discrete time domain, a difference problem

$$y_{k+1} = \rho y_k \tag{70}$$

where $\rho = \rho\left(\lambda h\right)$; the *stability region* of the method is given by the set of values of λ and h for which the condition $|\rho| < 1$ holds, i.e. the integration of the linear problem yields a converging, possibly oscillating solution.

Exercise 3.2 *Determine the stability region of explicit/implicit Euler and Crank-Nicolson methods.*

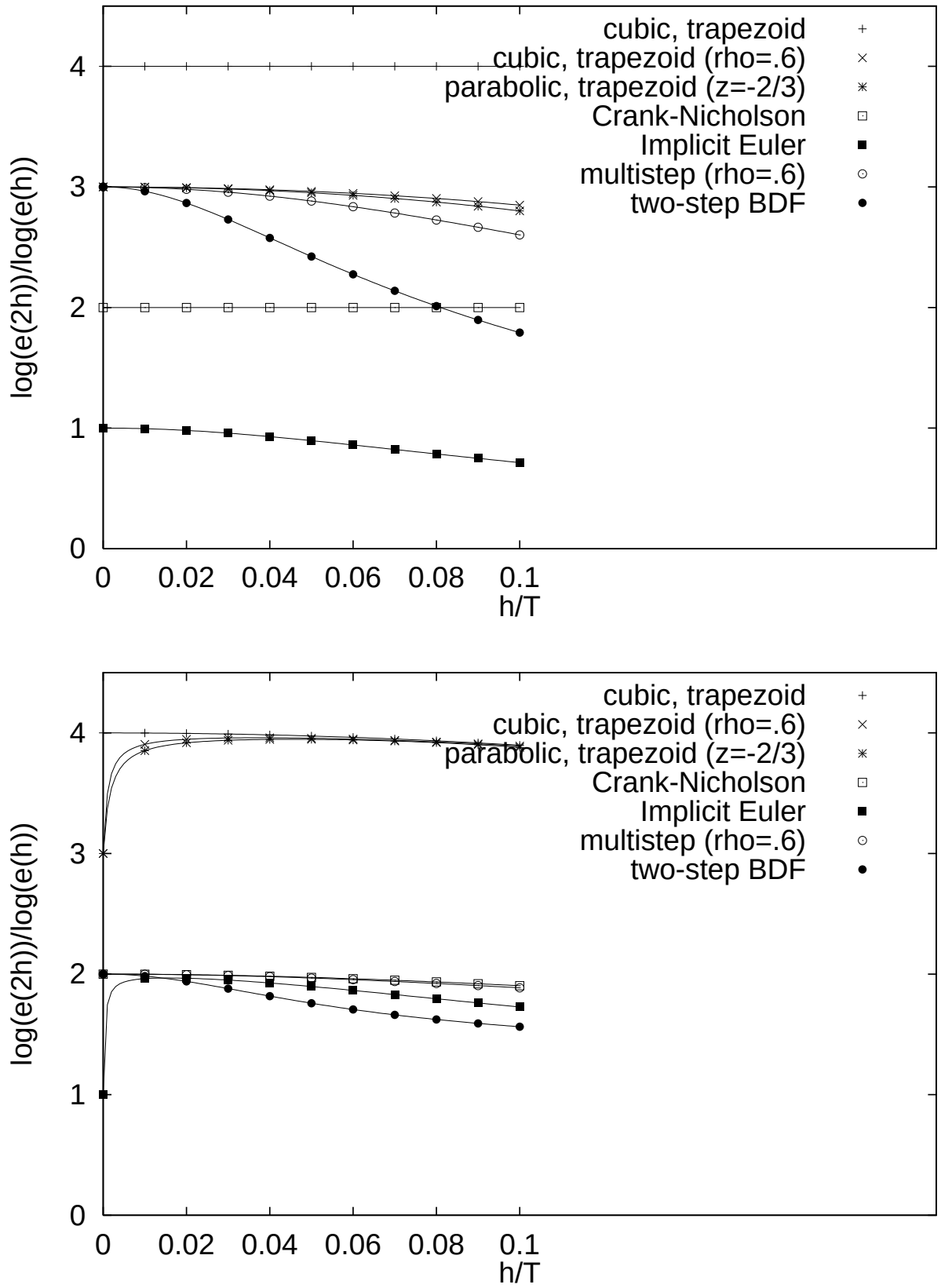

Figure 1: Convergence in damping and phase of a set of methods

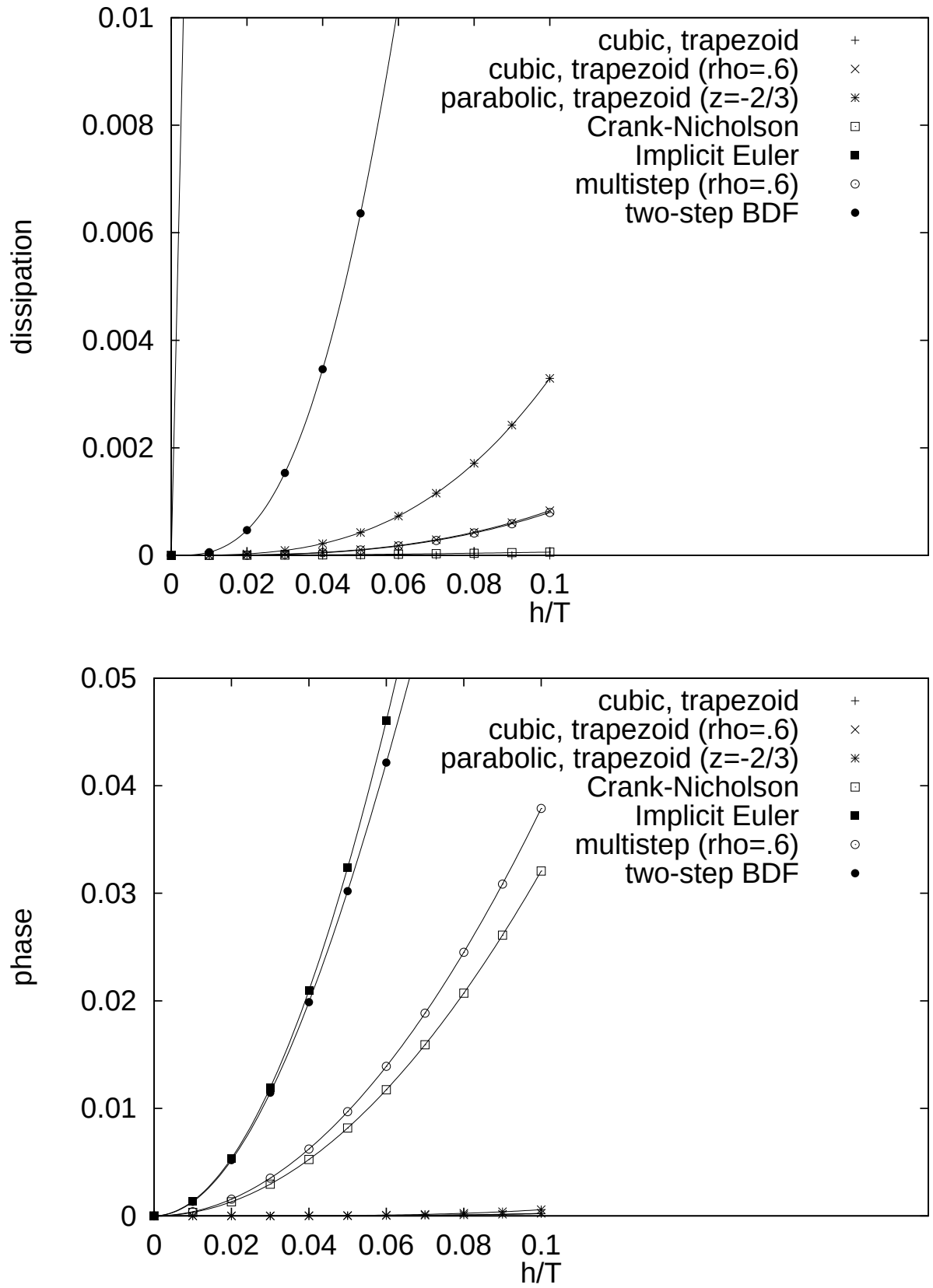

Figure 2: Error in damping and phase of a set of methods

Multistep methods yield a polynomial in ρ of order equal to the number of steps. The polynomial root with largest modulus dominates the linear stability behavior of the method.

The most relevant stability definitions related to ODE and DAE integration are reported in the following.

3.1 Zero-stability

A method is called *zero-stable* when the difference between two approximate solutions is limited as long as the time step is limited (i.e. $h \in [0, h_0]$).

Zero-stability is a necessary requirement for an approximation method to be usable. It rigorously expresses the concept of *conditional stability*. Note that the range of acceptable time steps has zero as lower bound. This excludes all methods that are not stable for time steps below a given threshold.

3.1.1 Dahlquist's First Theorem

Dahlquist's first theorem (1956), also known as *Dahlquist's first barrier*, states that a linear multistep integration method with k steps cannot have accuracy greater than $k+1$ when k is odd, or $k+2$ when k is even, without losing Zero-stability.

Methods of practical use for the solution of structural dynamics problems, where eigenvalues are complex and dominated by the imaginary part (i.e. with very low damping coefficient), usually have two steps ($k = 2$); this means that a conditionally stable scheme cannot have accuracy greater than 4.

Exercise 3.3 *Determine the accuracy of explicit/implicit Euler and Crank-Nicolson methods and check their compliance with Dahlquist's first theorem.*

3.2 A-stability

A method is called A-stable (Dahlquist) when its stability region, for $\mathrm{Re}\,(\lambda) < 0$, does not depend on the size of the time step. This means that the solution converges to zero asymptotically.

Note the definition: $\mathrm{Re}\,(\lambda) < 0$ means that the definition only applies to problems that are asymptotically stable; A-stability requires that asymptotically stable problems result in asymptotically stable approximate solutions, although there is no requirement on how accurate the solution is, i.e. on whether the equivalent damping of the approximate solution is identical to (or larger or smaller than) that of the problem. Moreover, a numerical method that yields an asymptotically stable approximate solution of an unstable problem is A-stable, although this behavior might not be acceptable because a fundamental property of the problem is not captured.

A-stability rigorously expresses the concept of *unconditional stability*.

Table 1: Summary of stability criteria.

Name	Definition
Zero-stability	$\lim_{k\to+\infty} y_k(h_2) - y_k(h_1)$ bounded when $h_1, h_2 \in [0, h_0]$
A-stability	$\lim_{k\to+\infty} y_k(\lambda h)/y_0 = 0 \ \ \forall h > 0$ when $\mathrm{Re}(\lambda) < 0$
L-stability	A-stability and $\lim_{\mathrm{Re}(\lambda h)\to-\infty} y_{k+1}(\lambda h)/y_k(\lambda h) = 0$

3.2.1 Dahlquist's Second Theorem

Another theorem formulated by Dahlquist (1963), also known as *Dahlquist's Second Barrier*, states that an explicit integration scheme cannot be A-stable; the accuracy of an A-stable multistep integration scheme cannot be greater than 2; among the A-stable multistep integration schemes, the one with least error is the trapezoid rule (also known as Simpson or Crank-Nicolson)

There exist methods defined *stiffly stable*, that is A-stable when applied to so-called *stiff* problems (Gear). This notion of stiffness should not be confused with that of stiffness of a spring in a mechanical system, which is usually related to high-frequency, nearly undamped oscillatory behavior. Systems that are stiff in the sense indicated by Gear have very large eigenvalues in the left half-plane, on or very close to the real axis, i.e. with either supercritical or subcritical but nearly critical damping.

Unconditionally stable multistep methods have been specifically formulated for this type of problems that have accuracy higher than second order, in violation of Dahlquist's second barrier. However, these methods fail when applied to problems with complex eigenvalues close to the imaginary axis, typical of mechanical problems. As a consequence, they cannot be considered truly A-stable.

3.3 L-stability

A method is L-stable (Ehle, Axelsson) when it is A-stable (i.e. for $\mathrm{Re}(\lambda) < 0$, $|\rho| < 1$ for all λh) and, for $\mathrm{Re}(\lambda h) \to -\infty$, $y_{k+1} = \rho y_k = 0$, that is the spectral radius contracts into the origin of the unit circle when the time step h increases.

This stability criterion is very important because it suggests a way to eliminate from the approximation of the solution the oscillations related to very large eigenvalues, with $\mathrm{Re}(\lambda h) << -1$, while preserving the accuracy of the method for those terms of the solution related to eigenvalues whose value is up to the order of magnitude of the integration frequency, i.e. $|\measuredangle(\lambda h)| \ll \pi$.

Figure 3 illustrates the spectral radius of a set of methods as a function of the integration step h.

Exercise 3.4 *Is explicit Euler Zero-stable? Is it A-stable? L-stable?*

Exercise 3.5 *Is implicit Euler Zero-stable? Is it A-stable? L-stable?*

Exercise 3.6 *Is Crank-Nicolson Zero-stable? Is it A-stable? L-stable?*

Exercise 3.7 *Are centered differences Zero-stable? Are they A-stable? L-stable?*

Exercise 3.8 *Are forward differences Zero-stable? Are they A-stable? L-stable?*

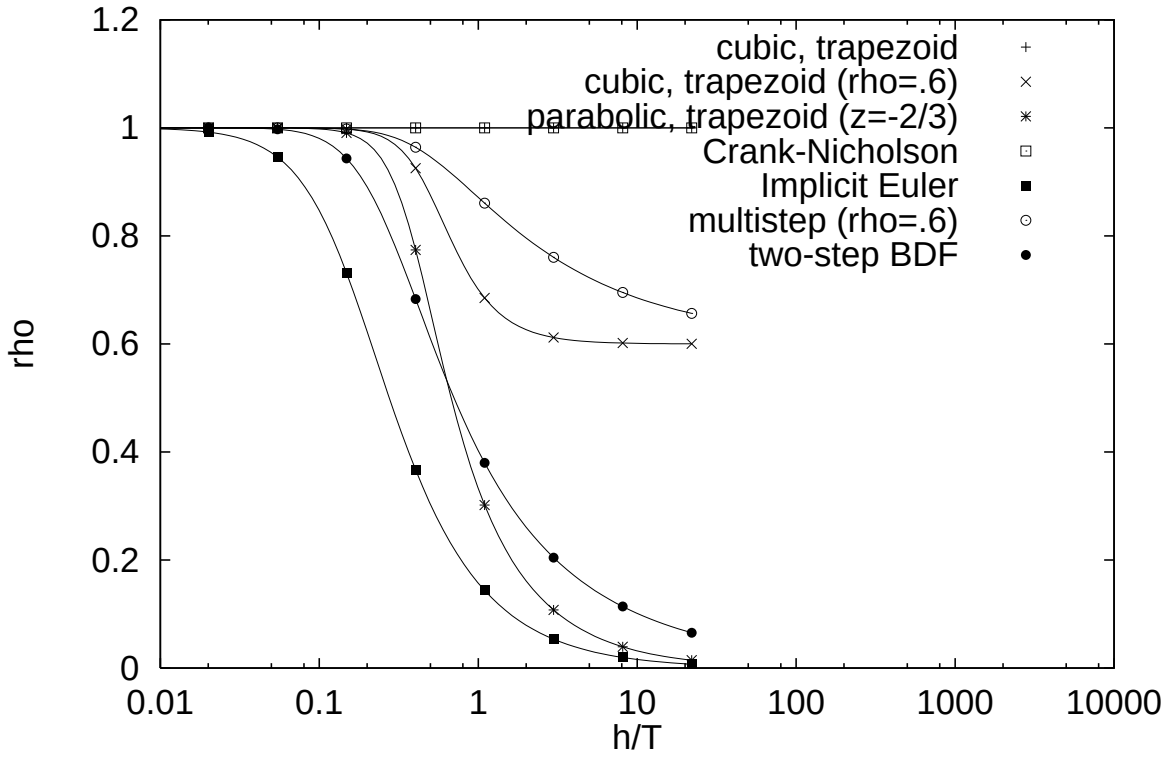

Figure 3: Spectral radius of a set of methods

3.4 Nonlinear Stability

Nonlinear stability, also called *global stability*, is related to the stability of solutions of nonlinear problems. The linear system used to evaluate the local stability of numerical approximation methods can be interpreted as the linearization of a generic nonlinear problem in the vicinity of an equilibrium point.

What has been discussed so far is essentially based on the analysis of the constant λ, related to the eigenvalues of a LTI problem. In the LTI case this information characterizes the entire evolution of a solution. In a nonlinear problem, the analysis of the eigenvalues of its linearization about an equilibrium point cannot give information about the global evolution of an arbitrary solution. On the contrary, the literature presents examples of problems globally stable where the eigenvalues computed from a linearization about an equilibrium point are structurally in the right half of the complex plane, and vice versa.

Without any ambition to present a detailed discussion, an appropriate measure of the global stability of a system in the neighborhood of a solution is related to the notion of *contractivity*. It expresses the tendency of a perturbed solution to steadily converge towards the corresponding unperturbed solution within a given time interval. This property is clear in the linear case where the solution is $y = y_0 \mathrm{e}^{\lambda t}$, so a perturbed solution $\hat{y} = (y_0 + \delta)\,\mathrm{e}^{\lambda t}$ converges to the unperturbed one as soon as $\mathrm{Re}\,(\lambda) < 0$, regardless of the value of the perturbation δ. In a generic case, the contractivity in general can only be evaluated numerically, after defining an appropriate norm of the perturbation and

contractivity conditions.

The result is a stability definition, when achievable, that directly uses the stability criteria in a Lyapunov sense.

3.4.1 Lyapunov Exponents

Lyapunov Exponents (LE), or Lyapunov characteristic exponents, usually indicated as λ_i, represent a time-average of exponential rates of convergence or divergence of orbits in the state space. They can be effectively interpreted (e.g. [15]) as the exponential evolution of the principal axes of an n-dimensional ellipsoid that grows from an initially infinitesimal n-sphere according to the map $\mathbf{T}$ that describes the time evolution of all phase points, $\mathbf{y}(t) = \mathbf{T}\mathbf{y}(t_0)$. If the i-th axis of the ellipsoid, starting from ${}_i\mathbf{y}(t_0)$, changes into ${}_i\mathbf{y}(t) = \mathbf{T}_i\mathbf{y}(t_0)$, the corresponding LE is

$$\lambda_i = \lim_{t\to+\infty} \frac{1}{t} \ln \frac{\|{}_i\mathbf{y}\,(t)\|}{\|{}_i\mathbf{y}\,(t_0)\|}. \tag{71}$$

For a LTI problem, $\dot{\mathbf{y}} = \mathbf{A}\mathbf{y}$, they correspond to the real part of the eigenvalues of matrix $\mathbf{A}$; for a Linear Time-Periodic (LTP) problem of period T, $\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}$, with $\mathbf{A}(t+T) = \mathbf{A}(t)$, according to Floquet's theory they correspond to the real part, divided by T, of the logarithm of the eigenvalues of the monodromy matrix $\mathbf{H}$ that yields $\mathbf{y}(t+T) = \mathbf{H}\mathbf{y}(t)$ [16].

For ergodic systems, LE are (nearly) independent of the trajectory $\mathbf{y}$ (the fiducial trajectory), as proved by Oseledec [17]. As such, they convey information about the global stability of the problem.

Their definition involves the limit for $t \to +\infty$; when computed from numerical integration, the computation of the solution ${}_i\mathbf{y}(t)$ needs to stop at a finite time. The resulting value λ_i represents an estimation of the actual LE. Whenever the problem evolves towards a stationary solution, the LE estimates also converge to a finite value.

According to the ellipsoid interpretation, a positive LE indicates that the ellipsoid is growing along that direction in the state space; a negative LE indicates contraction. Intuitively, expansion indicates instability, while contraction indicates stability (exponential stability): solutions originating from a perturbation along a direction associated to a positive LE will depart from the original solution, and vice versa.

When the largest LE is negative, the solution is exponentially stable about an equilibrium point in state space, an attractor of the problem. When the largest LE is zero, and all the others are negative, the attractor is a line in state space (e.g. a limit cycle). Higher-order attractors (tori) occur when more than one LE is zero. Positive LE indicate chaotic behavior.

Estimating LE from Eq. (71) can be very tricky, since the function will either contract or expand, soon being dominated by the largest LE estimate. Various methods have been proposed in the literature to preserve information about LE estimates by orthogonalization during the computation of the evolution of ${}_i\mathbf{y}(t)$.

In any case, no information about the stability of the numerical integration can be directly inferred from the application of these methods.

Exercise 3.9 *Compute the LE associated to the problem $\dot{y} = ay$.*

3.4.2 Energy-Based Methods

A completely different approach is used in the so called *energy preserving/dissipating* methods. In this case the integration method is tightly coupled to the formulation of the problem, to exploit intrinsic preservation/dissipation properties (e.g. of mechanical energy, momentum, momenta moment, etc.).

In some sense these methods guarantee global stability with respect to the properties they are based on; however, in many cases they require an *ad hoc* formulation of the problem and thus may not be applicable to generic problems, e.g. multidisciplinary ones.

Moreover, the preservative form of these methods, although quite interesting from a theoretical point of view, seldom finds practical application with some notable exception, because the global preservation of mechanical energy in systems that are stiff in a mechanical sense may not prevent energy transfer from low to high frequency dynamics, the latter often related to the level of spatial discretization rather than on physical properties of the problem, quickly deteriorating the quality of the solution.

4 Solution Approximation Methods

4.1 Multistep

Among the most important multistep methods for the integration of mechanical problems, besides the already mentioned trapezoid rule, a special mention is deserved by the Backward Differentiation Formulas (BDF).

The general formula is

$$\sum_{j=0,s} \alpha_j \mathbf{y}_{k-j} = h\beta_0 \mathbf{f}\left(\mathbf{y}_k, t_k\right), \tag{72}$$

where α coefficients are related by some special relationships. Note that $\mathbf{f}(\mathbf{y}, t)$ only needs to be evaluated at one time value.

BDF are Zero-stable up to sixth order while, according to Dahlquist's second barrier, they are A-stable only up to second order. Significantly, BDF up to second order are actually L-stable (first order BDF correspond to implicit Euler). The second order formula is

$$\mathbf{y}_k = \frac{4}{3}\mathbf{y}_{k-1} - \frac{1}{3}\mathbf{y}_{k-2} + \frac{2}{3}h\mathbf{f}\left(\mathbf{y}_k, t_k\right). \tag{73}$$

Exercise 4.1 *Verify that the second-order BDF of Eq. (73) is second-order accurate.*

Exercise 4.2 *Verify that the second-order BDF of Eq. (73) is L-stable.*

An interesting two-step formula, in the spirit of the formalism presented earlier, results from considering a parabolic interpolation on the value of $\mathbf{y}$ (i.e. with $n_i = 0$). The problem is integrated using the three-point trapezoid rule with accuracy limited to second

order (to comply with Dahlquist's second barrier), thus leaving the integration weights indeterminate by the parameter δ:

$$\begin{aligned}\mathbf{y}_k &= \mathbf{y}_{k-2} \\ &+ h\left(\left(\frac{1}{2}+\delta\right)\mathbf{f}\left(\mathbf{y}_k, t_k\right) + (1-2\delta)\,\mathbf{f}\left(\mathbf{y}_{k-1}, t_{k-1}\right) + \left(\frac{1}{2}+\delta\right)\mathbf{f}\left(\mathbf{y}_{k-2}, t_{k-2}\right)\right). \quad (74)\end{aligned}$$

The resulting form is augmented by adding a Crank-Nicolson integration on the already computed steps $k-1$, $k-2$, weighted by a coefficient $(\beta-1)$,

$$\mathbf{0} = (\beta-1)\left(-\mathbf{y}_{k-1} + \mathbf{y}_{k-2} + \frac{h}{2}\left(\mathbf{f}\left(\mathbf{y}_{k-1}, t_{k-1}\right) + \mathbf{f}\left(\mathbf{y}_{k-2}, t_{k-2}\right)\right)\right). \quad (75)$$

Since the solution at steps $k-1$, $k-2$ is known, this expression must be an identity; as a consequence, the method becomes

$$\begin{aligned}\mathbf{y}_k &= (1-\beta)\,\mathbf{y}_{k-1} + \beta\mathbf{y}_{k-2} \quad (76)\\ &+ h\left(\left(\frac{1}{2}+\delta\right)\mathbf{f}\left(\mathbf{y}_k, t_k\right) + \left(\frac{1}{2}+\frac{\beta}{2}-2\delta\right)\mathbf{f}\left(\mathbf{y}_{k-1}, t_{k-1}\right) + \left(\frac{\beta}{2}+\delta\right)\mathbf{f}\left(\mathbf{y}_{k-2}, t_{k-2}\right)\right).\end{aligned}$$

The resulting method is intrinsically second order accurate; an appropriate choice of the parameters β, δ in the intervals $\beta = 1 \to -1/3$ and $\delta = 0 \to 1/6$ makes the method cover the cases from A- to L-stability (in the latter case, of course, the method reduces to second order BDF).

By imposing the asymptotic coincidence of the two roots of the spectral radius, the latter can be used as the only parameter that characterizes the method,

$$\beta = \frac{4\rho_\infty^2 - (1-\rho_\infty)^2}{4 - (1-\rho_\infty)^2} \quad (77a)$$

$$\delta = \frac{1}{2}\frac{(1-\rho_\infty)^2}{4 - (1-\rho_\infty)^2}. \quad (77b)$$

This makes the tuning of the asymptotic dissipation possible, while preserving at least A-stability and second order accuracy.

4.2 Single Step

There exist many single step methods. Typically, higher-order methods are used; in some case the order can be adaptive. Higher-order methods are typically used to solve very regular problems with very long time steps, without losing accuracy. In case of mechanical problems, fourth order is seldom exceeded.

Consider a method whose Butcher matrix is

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array} \quad (78)$$

This method is known as fourth order Lobatto IIIA (note that Crank-Nicolson can be classified as second order Lobatto IIIA). It can be obtained from the proposed procedure using Hermitian interpolation and integration by collocation with the three-point trapezoid rule. The method is A-stable, fourth order accurate. It provides no algorithmic dissipation: the module of spectral radius for a purely oscillatory problem is unit regardless of the value of the time step h. Note from Butcher's matrix that the method requires to write the problem at mid-step and at the end of the time step.

Exercise 4.3 *Verify that fourth order Lobatto IIIA can be obtained from Eq. (66) using Hermitian shape functions and three point trapezoid rule.*

Starting from this method, by shifting the mid-point evaluation towards the end of the time step, an interesting method is obtained. It presents algorithmic dissipation properties similar to those of the modified BDF method presented earlier, thus showing stability properties intermediate between A- and L-stability.

Consider now a method whose Butcher matrix is

$$\begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array} \tag{79}$$

This method, known as Radau IIA, is third order accurate and L-stable. It results from the proposed procedure using $3/4$ and $1/4$ as weights at $t = t_k - 2/3h$ and $t = t_k$ respectively, and interpolating using quadratic polynomials, with $n_1 = 0$ on $\dot{\mathbf{y}}_{k-1}$.

Exercise 4.4 *Verify that third order Radau IIA can be obtained from Eq. (66) using weights equal to* $3/4$ *and* $1/4$ *at* $t = t_k - 2/3h$ *and* $t = t_k$ *respectively and parabolic polynomials with* $n_1 = 0$.

Exercise 4.5 *List all names Crank-Nicolson's rule is known as.*

5 Differential-Algebraic Equations

Consider a system of purely differential equations (ODE) in fully implicit form

$$\mathbf{s}\left(\dot{\mathbf{y}}, \mathbf{y}, t\right) = \mathbf{0}. \tag{80}$$

It is said to be Differential-Algebraic (DAE) when its partial derivative with respect to the derivative of the state is structurally singular,

$$\det\left(\mathbf{s}_{/\dot{\mathbf{y}}}\right) = 0. \tag{81}$$

Nonetheless the problem is well-posed when the *matrix pencil*

$$\mathcal{P}\left(\lambda\right) = \mathbf{s}_{/\dot{\mathbf{y}}} + \lambda \mathbf{s}_{/\mathbf{y}} \tag{82}$$

is at most singular for a finite number of values of $\lambda \neq 0$, which correspond to the inverse of the eigenvalues of the generalized problem $(\lambda^* \mathbf{E} - \mathbf{A})\delta\mathbf{y} = \mathbf{0}$, with $\mathbf{E} = \mathbf{s}_{/\dot{\mathbf{y}}}$, $\mathbf{A} = -\mathbf{s}_{/\mathbf{y}}$ and $\lambda^* = 1/\lambda$.

In this case an appropriate partitioning of the state $\mathbf{y}$ in differential, $\mathbf{x}$, and algebraic, $\mathbf{z}$, results in rewriting the problem as

$$\mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{z}, t) = \mathbf{0} \tag{83a}$$

$$\mathbf{g}(\mathbf{x}, \mathbf{z}, t) = \mathbf{0} \tag{83b}$$

or, in the special case of explicitable differential part,

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{z}, t) \tag{84a}$$

$$\mathbf{0} = \mathbf{g}(\mathbf{x}, \mathbf{z}, t). \tag{84b}$$

Different types of DAE arise depending on the fact that the differential part $\mathbf{f}(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{z}, t)$ can be made explicit for $\dot{\mathbf{x}}$ and that the algebraic part of the state, $\mathbf{z}$, is actually present in the algebraic part of the problem, $\mathbf{g}(\mathbf{x}, \mathbf{z}, t)$.

5.1 Singular Perturbation

Often a DAE system results from considering systems with a clear separation between 'slow' and 'fast' time scales when the dynamics of the 'fast' portion of the problem is neglected. The *singular perturbation* theory expresses this time scale separation in terms of a scalar coefficient ϵ; consider for example the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{z}, \epsilon, t) \tag{85a}$$

$$\epsilon\, \dot{\mathbf{z}} = \mathbf{g}(\mathbf{x}, \mathbf{z}, \epsilon, t) \tag{85b}$$

with $0 < \epsilon \ll 1$. In many cases the identification of a suitable parameter ϵ might not be straightforward; however, for illustrative purposes, it is assumed that all 'fast' dynamics can be identified by a single value of ϵ, so that fast coefficients (e.g. 'small' values of m associated to 'large' values of k) can be replaced by $\epsilon\, w_\epsilon$ (with $w_\epsilon = m/(k\epsilon)$).

The state $\mathbf{x}$ is related to the slow time scale, which typically characterizes the problem, while $\mathbf{z}$ is related to the fast time scale. A problem of this type is usually called *stiff* (in the sense given by Gear, which is not related to the usual stiffness of mechanical systems, but rather to problems whose eigenvalues are close to the real axis in the left half of the complex plane).

Consider for example a problem represented by two masses m_1 e m_2, connected by a spring k_2, while mass m_1 is connected to the ground by a spring k_1,

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \begin{Bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{Bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} f_1 \\ f_2 \end{Bmatrix}. \tag{86}$$

The eigenvalues of the system are

$$s_{1,2}^2 = -\frac{1}{2}\left(\frac{k_2}{m_2} + \frac{k_1 + k_2}{m_1}\right) \pm \frac{1}{2}\sqrt{\left(\frac{k_2}{m_2} + \frac{k_1 + k_2}{m_1}\right)^2 - 4\frac{k_1}{m_1}\frac{k_2}{m_2}}. \tag{87}$$

The first equation can be rewritten as

$$\frac{m_1}{k_1}\ddot{x}_1 + \left(1 + \frac{k_2}{k_1}\right) x_1 - \frac{k_2}{k_1} x_2 = \frac{f_1}{k_1}, \tag{88}$$

where $\epsilon = m_1/k_1$.

In this case, ϵ may go to zero for two reasons:

- $m_1 \to 0$: Eq. (88) degenerates into the algebraic equation

  $$\left(1 + \frac{k_2}{k_1}\right) x_1 - \frac{k_2}{k_1} x_2 = \frac{f_1}{k_1}, \tag{89}$$

  while the eigenvalues become

  $$s_1^2 = \infty \qquad\qquad s_2^2 = \frac{\dfrac{1}{m_2}}{\dfrac{1}{k_1} + \dfrac{1}{k_2}}. \tag{90}$$

  As one would expect, one of the eigenvalues becomes infinite, indicating extremely fast dynamics;

- $k_1 \to \infty$: Eq. (88) degenerates into the algebraic equation

  $$x_1 = 0, \tag{91}$$

  while the eigenvalues become

  $$s_1^2 = \infty \qquad\qquad s_2^2 = \frac{k_2}{m_2}. \tag{92}$$

  As one would expect, one of the eigenvalues becomes infinite, indicating extremely fast dynamics;

This implies that a problem where $\epsilon = 0$ cannot be integrated with a conditionally stable scheme, while when $\epsilon \to 0$, i.e. it is small but finite, the time step required by a conditionally stable scheme is dictated by the fast dynamics, which is essentially irrelevant for the quality of the motion 'at large' of mass m_2.

Singular perturbation theory is used to separately discuss the dynamics of the 'slow' and 'fast' portion of the problem.

5.1.1 Slow Subsystem Dynamics

Consider first the case of $\lim_{\epsilon \to 0}$. If function $\mathbf{g}$ has distinct solutions $\overline{\mathbf{z}}$ for $\epsilon = 0$ in a neighborhood of $\overline{\mathbf{x}} = \mathbf{x}_0$, is differentiable and its partial derivative with respect to $\mathbf{z}$ is not singular, one can locally compute $\mathbf{z} = \hat{\mathbf{g}}(\mathbf{x}, t)$ and replace it in $\mathbf{f}$, obtaining $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \hat{\mathbf{g}}(\mathbf{x}, t), t)$ with initial conditions $\mathbf{x}_0$ and $\overline{\mathbf{z}}$ computed accordingly. This means that the problem is reduced to the minimal set of coordinates that are truly differential and thus need to be independent. As discussed earlier, the reduction may not be easy in many cases; it is only possible locally, so the reduction may need to be recomputed any time the function $\mathbf{f}$ needs to be evaluated; often the original problem is significantly sparse, and sparsity can be exploited efficiently to manipulate it, while the reduced problem may become dense.

5.1.2 Fast Subsystem Dynamics

Consider now a pseudo-time τ related to t by the relationship $t = t_0 + \epsilon\,\tau$, such that its differentiation consists in $\mathrm{d}t = \epsilon\,\mathrm{d}\tau$. The problem of Eqs. (85), written in pseudo-time, becomes

$$\dot{\mathbf{x}} = \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \frac{1}{\epsilon}\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tau} = \mathbf{f}\left(\mathbf{x}, \mathbf{z}, \epsilon, t_0 + \epsilon\,\tau\right) \tag{93a}$$

$$\epsilon\,\dot{\mathbf{z}} = \epsilon\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\tau} = \mathbf{g}\left(\mathbf{x}, \mathbf{z}, \epsilon, t_0 + \epsilon\,\tau\right). \tag{93b}$$

Equation (93a), rewritten as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\tau} = \epsilon\,\mathbf{f}\left(\mathbf{x}, \mathbf{z}, \epsilon, t_0 + \epsilon\,\tau\right), \tag{94}$$

yields $\lim_{\epsilon\to 0} \mathrm{d}\mathbf{x}/\mathrm{d}\tau = 0$, and thus $\mathbf{x} = \mathbf{x}_0$. The limit of Eq. (93b), after writing $\mathbf{z}(t) = \overline{\mathbf{z}}(t) + \Delta\mathbf{z}(\tau)$, yields

$$\frac{\mathrm{d}\Delta\mathbf{z}}{\mathrm{d}\tau} = \mathbf{g}\left(\mathbf{x}_0, \overline{\mathbf{z}} + \Delta\mathbf{z}, 0, t_0\right). \tag{95}$$

As a consequence, the singular perturbation theory provides a useful means to separate the dynamics of the slow subsystem, where a static approximation of the fast variables $\mathbf{z}$ is used to describe the dynamics of the slow subsystem, from those of the fast subsystem, whose dynamics is confined in an infinitely fast *boundary layer* in pseudo-time. The latter results from the solution of Eq. (95) with initial conditions $\Delta\mathbf{z}(0) = \mathbf{z}_0 + \overline{\mathbf{z}}(t_0)$.

5.2 The Concept of Index

Starting from the DAE system in implicit form, one may want to write it in form of ODE by successive differentiation with respect to time. Assuming that $\mathbf{g}$ is locally invertible, consider its time derivative; the system becomes

$$\dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \mathbf{z}, t\right) \tag{96a}$$

$$\mathbf{0} = \mathbf{g}_{/\mathbf{x}}\dot{\mathbf{x}} + \mathbf{g}_{/\mathbf{z}}\dot{\mathbf{z}} + \mathbf{g}_{/t}. \tag{96b}$$

Substitute $\dot{\mathbf{x}}$ into the second equation:

$$\dot{\mathbf{x}} = \mathbf{f}\left(\mathbf{x}, \mathbf{z}, t\right) \tag{97a}$$

$$\dot{\mathbf{z}} = -\mathbf{g}_{/\mathbf{z}}^{-1}\left(\mathbf{g}_{/\mathbf{x}}\mathbf{f}\left(\mathbf{x}, \mathbf{z}, t\right) + \mathbf{g}_{/t}\right). \tag{97b}$$

The problem is now ODE in both sets of variables, $\mathbf{x}$ and $\mathbf{z}$. The *index* of the DAE system is the number of derivatives required to write the system in ODE form.

Note: this does not mean that reducing the system to ODE is a good approach for its solution! On the contrary, as illustrated earlier, this is prone to drift, since only higher order derivatives of the algebraic equations would be actually enforced, and the initial equations would essentially be violated.

5.2.1 Index 1 DAE

Index 1 equations are those described earlier, in the form

$$\mathbf{f}\left(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{z}, t\right) = \mathbf{0} \tag{98a}$$
$$\mathbf{g}\left(\mathbf{x}, \mathbf{z}, t\right) = \mathbf{0}. \tag{98b}$$

A typical example is represented by equations whose algebraic part represents a definition of a term that is used in the differential part; for example

$$\dot{\mathbf{u}} = \mathbf{M}^{-1}\mathbf{v} \tag{99a}$$
$$\dot{\mathbf{v}} = -\mathbf{r} + \mathbf{f}\left(\mathbf{u}, \mathbf{M}^{-1}\mathbf{v}, t\right) \tag{99b}$$
$$\mathbf{r} = \mathbf{k}\left(\mathbf{u}\right), \tag{99c}$$

where the algebraic equation contains the definition of the nonlinear elastic force $\mathbf{R}$.

The solution of these equations usually does not imply special issues, since the theory for DAEs of index 1 is rather complete.

5.2.2 Index 2 DAE

Index 2 equations have the form

$$\mathbf{f}\left(\dot{\mathbf{x}}, \mathbf{x}, \mathbf{z}, t\right) = \mathbf{0} \tag{100a}$$
$$\mathbf{g}\left(\mathbf{x}, t\right) = \mathbf{0}, \tag{100b}$$

i.e. the algebraic equation, $\mathbf{g}$, does not directly depend on the algebraic part of the state, $\mathbf{z}$. In this case two differentiations are required to write the problem in ODE form. A typical case is represented by non-holonomic constrains,

$$\dot{\mathbf{u}} = \mathbf{M}^{-1}\mathbf{v} \tag{101a}$$
$$\dot{\mathbf{v}} = \mathbf{A}\boldsymbol{\lambda} + \mathbf{f}\left(\mathbf{u}, \mathbf{M}^{-1}\mathbf{v}, t\right) \tag{101b}$$
$$\mathbf{0} = \mathbf{A}\mathbf{M}^{-1}\mathbf{v} - \mathbf{b}'. \tag{101c}$$

5.2.3 Index 3 DAE

There exists a type of index 3 equations, called index 3 Hessenberg, that is very important since it describes typical mechanical systems holonomically constrained by kinematic relationships in form of algebraic equations,

$$\dot{\mathbf{x}}_1 = \mathbf{f}_1\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}, t\right) \tag{102a}$$
$$\dot{\mathbf{x}}_2 = \mathbf{f}_2\left(\mathbf{x}_1, \mathbf{x}_2, t\right) \tag{102b}$$
$$\mathbf{0} = \mathbf{g}\left(\mathbf{x}_2, t\right). \tag{102c}$$

As one can easily note, the algebraic equation depends only on that differential portion of the state whose equation does not contain the algebraic portion of the state itself. As a consequence, three differentiations are required to write the problem in ODE form.

The algebraic portion of the state plays the role of constraint reaction forces, usually expressed using Lagrange's multipliers. A typical mechanical example is

$$\dot{\mathbf{u}} = \mathbf{M}^{-1}\mathbf{v} \tag{103a}$$

$$\dot{\mathbf{v}} = \boldsymbol{\phi}_{/\mathbf{u}}^{T}\boldsymbol{\lambda} + \mathbf{f}\left(\mathbf{u}, \mathbf{M}^{-1}\mathbf{v}, t\right) \tag{103b}$$

$$\mathbf{0} = \boldsymbol{\phi}\left(\mathbf{u}, t\right). \tag{103c}$$

The solution of this type of equations implies non-negligible numerical issues. The theory of DAEs of index higher than 1 is far from consolidated.

Exercise 5.1 *What is the index of Eqs. (85) when $\epsilon \to 0$ because $m_1 \to 0$?*

Exercise 5.2 *What is the index of Eqs. (85) when $\epsilon \to 0$ because $k_1 \to \infty$?*

5.3 Solution Strategies

The solution of DAE systems of index higher than 1 represents a challenging task.

5.3.1 Direct Integration of the DAE Problem

If the algebraic part is considered in terms of singular perturbations, the problem presents an extremely fast time scale, compared to that of the differential part. In the limit, when $\epsilon \to 0$, the fast time scale goes to infinity. This implies that an explicit integration scheme cannot be used, because its Zero-stability would only be guaranteed by a null time step. But this, besides being useless, would make the pencil of Eq. (82) structurally singular (or, conversely, as $\epsilon \to 0$, the pencil would be ill-conditioned). A-stability is required; according to Dahlquist's second theorem, this requires to use implicit integration schemes.

However, A-stability may not suffice, because an A-stable scheme would integrate the dynamics of the infinitely fast time scales with a fairly inaccurate time step (recall the phase error plots of Figure 2 for time steps of order comparable to the time scale of the problem).

Algorithmic dissipation is needed to cancel the spurious dynamics of the infinitely large non-physical (and, of course, of the physical but not required because too large) time scales. Otherwise the analysis would quickly diverge with oscillations whose period is strictly related to $2h$, as clearly illustrated in the literature. L-stable methods should be used to exactly cancel in one time step the dynamics related to the infinitely fast time scale of the constraints.

Nowadays the possibility to directly integrate DAEs up to index 2 with A-stable methods, and up to index 3 with L-stable methods is accepted, at the cost of slightly violating the derivatives of the constraints and of losing one order in the accuracy of the constraint reactions.

5.3.2 Spectral Decomposition

One may be tempted to perform a spectral decomposition of the problem, in order to isolate the dynamics of the problem. The theory has been formulated for index 1 DAEs,

and could probably be extended to higher index ones [18]. Consider the (linearized) problem

$$\mathbf{E}(t)\dot{\mathbf{y}} = \mathbf{A}(t)\mathbf{y}; \tag{104}$$

using a Generalized Schur Decomposition, or QZ decomposition, the problem can be rewritten as

$$\tilde{\mathbf{E}}(t)\dot{\tilde{\mathbf{y}}} = \tilde{\mathbf{A}}(t)\tilde{\mathbf{y}} \tag{105}$$

with $\tilde{\mathbf{E}} = \mathbf{QEV}$ and $\tilde{\mathbf{A}} = \mathbf{QAV} - \mathbf{QE}\dot{\mathbf{V}}$.

Matrices $\tilde{\mathbf{E}}$ and $\tilde{\mathbf{A}}$ are real, block-upper triangular; matrix $\tilde{\mathbf{E}}$ can be partitioned such that one diagonal block is non-singular. This identifies the spectrum of the problem that is differential. It can be integrated as an ODE, while the algebraic part can be recovered separately.

This method is clearly impractical, unless $\mathbf{Q}$ and $\mathbf{V}$ can be computed analytically, because it requires repeated very expensive QZ decompositions, and the resulting block matrices destroy any sparsity of the problem.

5.3.3 Baumgarte's Method

The approach proposed in 1972 by Baumgarte [19] is inspired by control theory. The holonomic constraint equation is differentiated twice and linearly combined with its first and second derivative. As a result, a linear combination of the constraint and of its derivatives is enforced instead of the actual constraint,

$$\mathbf{M}\ddot{\mathbf{x}} = \mathbf{f} + \mathbf{g}_{/\mathbf{x}}^T\boldsymbol{\lambda} \qquad \rightarrow \qquad \mathbf{M}\ddot{\mathbf{x}} = \mathbf{f} + \mathbf{g}_{/\mathbf{x}}^T\boldsymbol{\lambda} \tag{106a}$$

$$\mathbf{0} = \mathbf{g}(\mathbf{x}, t) \qquad \rightarrow \qquad \mathbf{0} = \ddot{\mathbf{g}}(\mathbf{x}, t) + 2\alpha\dot{\mathbf{g}}(\mathbf{x}, t) + \beta^2\mathbf{g}(\mathbf{x}, t)\,. \tag{106b}$$

The problem can be conveniently manipulated to expose the second derivative of $\mathbf{x}$ in the modified constraint equation,

$$\begin{bmatrix} \mathbf{M} & -\mathbf{g}_{/\mathbf{x}}^T \\ -\mathbf{g}_{/\mathbf{x}} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \ddot{\mathbf{x}} \\ \boldsymbol{\lambda} \end{Bmatrix} = \begin{Bmatrix} \mathbf{f} \\ -\mathbf{b}'' + 2\alpha\dot{\mathbf{g}}(\mathbf{x}, t) + \beta^2\mathbf{g}(\mathbf{x}, t) \end{Bmatrix} \tag{107}$$

where $\ddot{\mathbf{g}} = \mathbf{g}_{/\mathbf{x}}\ddot{\mathbf{x}} - \mathbf{b}''$ is used.

As already mentioned, a drawback of this approach is that it enforces a combination of the constraint derivatives rather than the constraint itself. The problem that is integrated has a spectrum that differs from that of the original problem: consider the linearization of the original problem

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \delta\ddot{\mathbf{x}} \\ \delta\ddot{\boldsymbol{\lambda}} \end{Bmatrix} + \begin{bmatrix} -\mathbf{f}_{/\dot{\mathbf{x}}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \delta\dot{\mathbf{x}} \\ \delta\dot{\boldsymbol{\lambda}} \end{Bmatrix} + \begin{bmatrix} -\mathbf{f}_{/\mathbf{x}} & -\mathbf{g}_{/\mathbf{x}}^T \\ -\mathbf{g}_{/\mathbf{x}} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \delta\mathbf{x} \\ \delta\boldsymbol{\lambda} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{0} \end{Bmatrix} \tag{108}$$

and that of the modified problem,

$$\begin{bmatrix} \mathbf{M} & \mathbf{0} \\ -\mathbf{g}_{/\mathbf{x}} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \delta\ddot{\mathbf{x}} \\ \delta\ddot{\boldsymbol{\lambda}} \end{Bmatrix} + \begin{bmatrix} -\mathbf{f}_{/\dot{\mathbf{x}}} & \mathbf{0} \\ -2\alpha\mathbf{g}_{/\mathbf{x}} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \delta\dot{\mathbf{x}} \\ \delta\dot{\boldsymbol{\lambda}} \end{Bmatrix} + \begin{bmatrix} -\mathbf{f}_{/\mathbf{x}} & -\mathbf{g}_{/\mathbf{x}}^T \\ -\beta^2\mathbf{g}_{/\mathbf{x}} & \mathbf{0} \end{bmatrix} \begin{Bmatrix} \delta\mathbf{x} \\ \delta\boldsymbol{\lambda} \end{Bmatrix} = \begin{Bmatrix} \mathbf{0} \\ \mathbf{0} \end{Bmatrix}. \tag{109}$$

The two clearly differ, and it is apparent that in the latter case the infinitely fast dynamics related to the constraints become finite and depend on the magnitude of α and β; the larger β the faster the constraint dynamics, but the closer the problem becomes to DAE (in a singular perturbations sense, $\epsilon = 1/\beta^2$). The different spectrum may have a significant impact especially during very fast transients.

The advantage of this method is that it turns a DAE into an ODE by substantially adding a homogeneous differential equation in the constraint violation, $\ddot{\mathbf{g}}+2\alpha\dot{\mathbf{g}}+\beta^2\mathbf{g} = \mathbf{0}$; the stability and the rapidity of asymptotic convergence to zero of the solution of this equation can be controlled by properly choosing the coefficients α and β. Basically, β is related to the rapidity of response (it corresponds to $\sqrt{k/m}$ in a mechanical oscillator), while α is related to the sovraelongation (it corresponds to $r/(2m) = \xi\sqrt{k/m}$ in a mechanical oscillator).

Since the resulting equation is ODE, it could be efficiently integrated with explicit schemes; however, when β is large, the time step of a conditionally stable scheme would be dictated by the modified constraint equation rather than the fundamental dynamics of the problem (in fact, the larger β, the closer the modified problem is to the original DAE).

5.3.4 Constraint Stabilization by Index Reduction

When a higher index DAE problem is written, algebraic constraint equations

$$\mathbf{g}(\mathbf{x}, t) = \mathbf{0} \tag{110}$$

imply compliance with a sufficiently high number of derivatives of the constraint, namely

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\mathbf{g}(\mathbf{x}, t) = \mathbf{0}. \tag{111}$$

Consider the index 3 DAE of Eq. (102); Gear *et al.* in 1985 proposed to 'stabilize' constraints by explicitly enforcing the derivative of the algebraic constraint [20]. Compute the time derivative of Eq. (102c),

$$\mathbf{0} = \dot{\mathbf{g}} = \mathbf{g}_{/\mathbf{x}_2}\dot{\mathbf{x}}_2 + \mathbf{g}_{/t}. \tag{112}$$

To be able to explicitly enforce another constraint equation, another set of multipliers, $\mathbf{w}$, is required. Replacing $\dot{\mathbf{x}}_2$ in Eq. (112) with its expression from Eq. (102b), in order to introduce an explicit dependence on $\mathbf{x}_1$, the augmented DAE becomes

$$\dot{\mathbf{x}}_1 = \mathbf{f}_1(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}, t) \tag{113a}$$

$$\dot{\mathbf{x}}_2 = \mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2, t) + \mathbf{g}_{/\mathbf{x}_2}^T\mathbf{w} \tag{113b}$$

$$\mathbf{0} = \mathbf{g}(\mathbf{x}_2, t) \tag{113c}$$

$$\mathbf{0} = \mathbf{g}_{/\mathbf{x}_2}\mathbf{f}_2(\mathbf{x}_1, \mathbf{x}_2, t) + \mathbf{g}_{/t}. \tag{113d}$$

Note the term $\mathbf{g}_{/\mathbf{x}_2}^T \mathbf{w}$ in Eq. (113b); it states that the original definition of $\dot{\mathbf{x}}_2$ given by Eq. (102b) needs to be relaxed to make the enforcement of the constraint equation possible.

The index of the DAE reduces from 3 to 2. A further reduction to index 1 was proposed by Führer and Leimkuhler in 1991 [21].

Consider the mechanical example of Eqs. (103); when stabilized, it becomes

$$\dot{\mathbf{u}} = \mathbf{M}^{-1}\mathbf{v} + \mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\mu} \tag{114a}$$

$$\dot{\mathbf{v}} = \boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\lambda} + \mathbf{f}\left(\mathbf{u}, \mathbf{M}^{-1}\mathbf{v}, t\right) \tag{114b}$$

$$\mathbf{0} = \boldsymbol{\phi}\left(\mathbf{u}, t\right) \tag{114c}$$

$$\mathbf{0} = \boldsymbol{\phi}_{/\mathbf{u}} \mathbf{M}^{-1}\mathbf{v} + \boldsymbol{\phi}_{/t}. \tag{114d}$$

Note the new multiplier $\boldsymbol{\mu}$ in Eq. (114a); that equation corresponds to the definition of the momentum $\mathbf{p}$ as a function of the derivative of the coordinates $\mathbf{u}$, $\mathbf{p} = \mathbf{M}\dot{\mathbf{u}}$. The definition needs to be relaxed to accommodate the new constraint equation. However, this multiplier has a special meaning: since analytically Eq. (114c) implies Eq. (114d), the violation of the latter can only be minimal, and dictated by numerical integration reasons only. The more Eq. (114d) would be violated by numerical integration, if not explicitly enforced, the larger is $\boldsymbol{\mu}$. So, in a well-behaved analysis, $\boldsymbol{\mu} \cong \mathbf{0}$, and the definition of the momentum is not modified.

Note that Gear *et al.* originally proposed to correct Eq. (114a) using $\boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\mu}$ instead of $\mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\mu}$. Other authors suggest to add the inverse of the mass matrix because this minimizes the correction on the kinetic energy of the problem.

It is important that not only the holonomic constraint, but at least its first derivative are satisfied by the initial conditions of the analysis, otherwise a very stiff transient would dominate the first part of the analysis (it is sometimes called *boundary layer* in the singular perturbations jargon). In most cases, during the analysis, explicit enforcement of the constraint suffices, as soon as the solution is regular enough.

The extension proposed by Führer and Leimkuhler [21] explicitly enforces the second-derivative of the holonomic constraint equations, with an additional set of variables $\mathbf{w}$ that represent the accelerations, and a new set of multipliers $\boldsymbol{\nu}$ to account for the error between the accelerations and the derivative of the velocities,

$$\dot{\mathbf{u}} = \mathbf{M}^{-1}\mathbf{v} + \mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\mu} + \mathbf{M}^{-1}\frac{\mathrm{d}}{\mathrm{d}t}\left(\boldsymbol{\phi}_{/\mathbf{u}}\right)^T \boldsymbol{\nu} \tag{115a}$$

$$\dot{\mathbf{v}} = \mathbf{w} + \boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\nu} \tag{115b}$$

$$\mathbf{w} = \mathbf{f}\left(\mathbf{u}, \mathbf{M}^{-1}\mathbf{v}, t\right) + \boldsymbol{\phi}_{/\mathbf{u}}^T \boldsymbol{\lambda} \tag{115c}$$

$$\mathbf{0} = \boldsymbol{\phi}\left(\mathbf{u}, t\right) \tag{115d}$$

$$\mathbf{0} = \boldsymbol{\phi}_{/\mathbf{u}} \mathbf{M}^{-1}\mathbf{v} + \boldsymbol{\phi}_{/t} \tag{115e}$$

$$\mathbf{0} = \boldsymbol{\phi}_{/\mathbf{u}} \mathbf{M}^{-1}\mathbf{w} + \left(\dot{\boldsymbol{\phi}}\right)_{/\mathbf{u}} \mathbf{M}^{-1}\mathbf{v} + \left(\dot{\boldsymbol{\phi}}\right)_{/t} \tag{115f}$$

5.4 Initialization

A generic explicit ODE problem is defined according to Eq. (40) as

$$\dot{\mathbf{y}} = \mathbf{f}\left(\mathbf{y}, t\right) \qquad\qquad \mathbf{y}\left(t_0\right) = \mathbf{y}_0. \tag{116}$$

The initial conditions consist in the state $\mathbf{y}$ at time t_0, namely $\mathbf{y}(t_0) = \mathbf{y}_0$. Numerical schemes require the knowledge of the derivative of the state at time t_0; this can be explicitly computed evaluating $\dot{\mathbf{y}}(t_0) = \mathbf{f}(\mathbf{y}_0, t_0)$.

When the problem is implicit, namely

$$\mathbf{f}\left(\dot{\mathbf{y}}, \mathbf{y}, t\right) = \mathbf{0} \qquad\qquad \mathbf{y}\left(t_0\right) = \mathbf{y}_0, \tag{117}$$

the derivative of the state results from the solution of the problem

$$\mathbf{f}\left(\dot{\mathbf{y}}_0, \mathbf{y}_0, t_0\right) = \mathbf{0} \tag{118}$$

with respect to $\dot{\mathbf{y}}_0$, which is algebraic in $\dot{\mathbf{y}}_0$. This requires $\mathbf{f}_{/\dot{\mathbf{y}}}$ to be non-singular. However, the latter condition is violated when the problem is DAE.

The initialization of an Index-3 DAE problem in Hessenberg form requires special care. Consider directly Eqs. (103), here reported for completeness,

$$\dot{\mathbf{u}} = \mathbf{M}^{-1}\mathbf{v} \tag{119a}$$

$$\dot{\mathbf{v}} = \boldsymbol{\phi}_{/\mathbf{u}}^T\boldsymbol{\lambda} + \mathbf{f}\left(\mathbf{u}, \mathbf{M}^{-1}\mathbf{v}, t\right) \tag{119b}$$

$$\mathbf{0} = \boldsymbol{\phi}\left(\mathbf{u}, t\right). \tag{119c}$$

The initial conditions can only be defined for a subset of $\mathbf{u}$, $\mathbf{v}$, because of the constraints. However, that subset may not be known in advance, or in any case it may be impractical to restrict the specification of the initial conditions to a subset without prior knowledge of what variables are truly independent when the constraints are expressed in the implicit form of Eq. (103c). On the contrary, it may be practical to specify initial conditions on the entire set of variables $\mathbf{u}$, $\mathbf{v}$, and address any initial constraint violation prior to integration.

In order to enforce compliance of the initial conditions with the constraint equations one needs to solve the problems

$$\mathbf{0} = \boldsymbol{\phi}\left(\mathbf{u}_0, t_0\right) \tag{120a}$$

$$\mathbf{0} = \boldsymbol{\phi}_{/\mathbf{u}}\mathbf{M}^{-1}\mathbf{v}_0 + \boldsymbol{\phi}_{/t} \tag{120b}$$

to find $\mathbf{u}_0$, $\mathbf{v}_0$ starting from an initial guess $\overline{\mathbf{u}}_0$, $\overline{\mathbf{v}}_0$ represented by the possibly non-compliant initial values. Methods analogous to coordinate partitioning, projection or similar, as discussed in Section 1, can be used to ensure the determination of the most appropriate initial values. For example, one can prove that writing a problem in the form

$$\mathbf{K}\mathbf{u}_0 + \boldsymbol{\phi}_{/\mathbf{u}}^T\boldsymbol{\lambda}^* = \mathbf{K}\overline{\mathbf{u}}_0 \tag{121a}$$

$$\mathbf{D}\mathbf{M}^{-1}\mathbf{v}_0 + \boldsymbol{\phi}_{/\mathbf{u}}^T\boldsymbol{\mu}^* = \mathbf{D}\mathbf{M}^{-1}\overline{\mathbf{v}}_0 \tag{121b}$$

$$\mathbf{0} = \boldsymbol{\phi}\left(\mathbf{u}_0, t_0\right) \tag{121c}$$

$$\mathbf{0} = \boldsymbol{\phi}_{/\mathbf{u}}\mathbf{M}^{-1}\mathbf{v}_0 + \boldsymbol{\phi}_{/t} \tag{121d}$$

yields a minimal norm correction of the initial values, weighted by matrices $\mathbf{K}$, $\mathbf{D}$, starting from the possibly incompatible initial guess $\overline{\mathbf{u}}_0$, $\overline{\mathbf{v}}_0$. The corresponding pseudo-multipliers $\boldsymbol{\lambda}^*$, $\boldsymbol{\mu}^*$ have no special meaning, and are discarded.

The initial value of the derivative of the differential variables $\mathbf{u}$, namely $\dot{\mathbf{u}}$, can be computed from Eq. (103a) as soon as $\mathbf{v}_0$ is known.

As soon as compliant initial values of the differential variables are available, initial values of the algebraic ones are needed. This requires to consider the second derivative of the constraint equation,

$$\mathbf{0} = \boldsymbol{\phi}_{/\mathbf{u}}\mathbf{M}^{-1}\left(\dot{\mathbf{v}} - \mathbf{v}_{/\mathbf{u}}\mathbf{M}^{-1}\mathbf{v}\right) + \left(\dot{\boldsymbol{\phi}}\right)_{/\mathbf{u}}\mathbf{M}^{-1}\mathbf{v} + \left(\dot{\boldsymbol{\phi}}\right)_{/t}. \tag{122}$$

After substitution of $\dot{\mathbf{v}}$ from Eq. (103b), Eq. (122) yields the multipliers

$$\boldsymbol{\lambda} = -\left(\boldsymbol{\phi}_{/\mathbf{u}}\mathbf{M}^{-1}\boldsymbol{\phi}_{/\mathbf{u}}^T\right)^{-1}\left(\boldsymbol{\phi}_{/\mathbf{u}}\mathbf{M}^{-1}\left(\mathbf{f} - \mathbf{v}_{/\mathbf{u}}\mathbf{M}^{-1}\mathbf{v}\right) + \left(\dot{\boldsymbol{\phi}}\right)_{/\mathbf{u}}\mathbf{M}^{-1}\mathbf{v} + \left(\dot{\boldsymbol{\phi}}\right)_{/t}\right). \tag{123}$$

At this point, the initial value of the derivative of the differential variables $\mathbf{v}$, namely $\dot{\mathbf{v}}$, can be directly computed from Eq. (103b).

6 Implementation Aspects

6.1 Nonlinear Algebraic Problem

Implicit ODE and DAE problems at some point need to be turned into an algebraic problem. Consider a generic implicit equation of the form

$$\mathbf{f}\left(\dot{\mathbf{y}}, \mathbf{y}, t\right) = \mathbf{0}. \tag{124}$$

When a multistep scheme is used, the equation needs to be solved at a specific time t. Consider the linearization of Eq. (124),

$$\mathbf{f} + \mathbf{f}_{/\dot{\mathbf{y}}}\Delta\dot{\mathbf{y}} + \mathbf{f}_{/\mathbf{y}}\Delta\mathbf{y} = \mathbf{0}. \tag{125}$$

From the formula of the generic multistep method of Eq. (72) one obtains

$$\Delta\mathbf{y} = hb_0\Delta\dot{\mathbf{y}} = c\Delta\dot{\mathbf{y}} \tag{126}$$

The problem is solved by iteratively computing

$$\left(\mathbf{f}_{/\dot{\mathbf{y}}} + c\mathbf{f}_{/\mathbf{y}}\right)\Delta\dot{\mathbf{y}} = -\mathbf{f}. \tag{127}$$

The matrix of Eq. (127) is structurally non-singular for $c > 0$ except in case of physical singularities of the problem (e.g. labilities). The state and its derivative are consistently updated according to

$$\dot{\mathbf{y}}^{(i+1)} = \dot{\mathbf{y}}^{(i)} + \Delta\dot{\mathbf{y}} \tag{128a}$$

$$\mathbf{y}^{(i+1)} = \mathbf{y}^{(i)} + c\Delta\dot{\mathbf{y}} \tag{128b}$$

6.2 Application of Single-Step Methods to Implicit Problems

Consider the application of a Runge-Kutta-like method to the integration of the implicit problem of Eq. (124). The approximation of the state and of its derivative is

$$\dot{\mathbf{y}}(t_i) = \mathbf{Y}_i \tag{129a}$$

$$\mathbf{y}(t_i) = \mathbf{y}_{n-1} + h \sum_{j=1,s} a_{ij} \mathbf{Y}_j \tag{129b}$$

Their perturbation yields

$$\Delta\dot{\mathbf{y}}(t_i) = \Delta\mathbf{Y}_i \tag{130a}$$

$$\Delta\mathbf{y}(t_i) = h \sum_{j=1,s} a_{ij} \Delta\mathbf{Y}_j \tag{130b}$$

Function $\mathbf{f}$ needs to be evaluated at all collocation points; for this purpose, directly consider its linearization in the i-th point,

$$\mathbf{f} + \mathbf{f}_{/\dot{\mathbf{y}}} \Delta\mathbf{Y}_i + \mathbf{f}_{/\mathbf{y}} h \sum_{j=1,s} a_{ij} \Delta\mathbf{Y}_j = \mathbf{0}. \tag{131}$$

The assembly of all linearizations leads to the iterative solution of an implicit algebraic problem of order equal to the order of the original problem, $\mathbf{f}$, times the number of collocation points s,

$$\begin{bmatrix} \mathbf{f}_{/\dot{\mathbf{y}}} + h a_{11} \mathbf{f}_{/\mathbf{y}} & \cdots & h a_{1s} \mathbf{f}_{/\mathbf{y}} \\ \vdots & \ddots & \vdots \\ h a_{s1} \mathbf{f}_{/\mathbf{y}} & \cdots & \mathbf{f}_{/\dot{\mathbf{y}}} + h a_{ss} \mathbf{f}_{/\mathbf{y}} \end{bmatrix} \begin{Bmatrix} \Delta\mathbf{Y}_1 \\ \vdots \\ \Delta\mathbf{Y}_s \end{Bmatrix} = - \begin{Bmatrix} \mathbf{f}(t_{n-1} + c_1 h) \\ \vdots \\ \mathbf{f}(t_{n-1} + c_s h) \end{Bmatrix}. \tag{132}$$

DAE problems cannot be solved with explicit methods, because $\mathbf{f}_{/\dot{\mathbf{y}}}$ is singular and thus the problem cannot be locally inverted.

An interesting case occurs when a DIRK method is used:

$$\begin{bmatrix} \mathbf{f}_{/\dot{\mathbf{y}}} + h a_{11} \mathbf{f}_{/\mathbf{y}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ h a_{s1} \mathbf{f}_{/\mathbf{y}} & \cdots & \mathbf{f}_{/\dot{\mathbf{y}}} + h a_{ss} \mathbf{f}_{/\mathbf{y}} \end{bmatrix} \begin{Bmatrix} \Delta\mathbf{Y}_1 \\ \vdots \\ \Delta\mathbf{Y}_s \end{Bmatrix} = - \begin{Bmatrix} \mathbf{f}(t_{n-1} + c_1 h) \\ \vdots \\ \mathbf{f}(t_{n-1} + c_s h) \end{Bmatrix} \tag{133}$$

In this case only s subproblems of the same order of the original problem need to be solved, related to the block-diagonal elements of the above matrix; subsequent problems use the solution of earlier ones.

The submatrices of DIRK problems are different, and thus need to be independently solved. When $a_{ii} = d \ \forall i$, the method is SIRK,

$$\begin{bmatrix} \mathbf{f}_{/\dot{\mathbf{y}}} + h d \mathbf{f}_{/\mathbf{y}} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ h a_{s1} \mathbf{f}_{/\mathbf{y}} & \cdots & \mathbf{f}_{/\dot{\mathbf{y}}} + h d \mathbf{f}_{/\mathbf{y}} \end{bmatrix} \begin{Bmatrix} \Delta\mathbf{Y}_1 \\ \vdots \\ \Delta\mathbf{Y}_s \end{Bmatrix} = - \begin{Bmatrix} \mathbf{f}(t_{n-1} + c_1 h) \\ \vdots \\ \mathbf{f}(t_{n-1} + c_s h) \end{Bmatrix}. \tag{134}$$

In principle the subproblems are identical and thus only need to be solved once. However, matrices $\mathbf{f}_{/\dot{\mathbf{y}}}$ and $\mathbf{f}_{/\mathbf{y}}$ may differ at each time collocation point and, in any case, may depend on $\mathbf{y}$.

Moreover, SIRK methods usually result in reduced accuracy, compared to generic implicit RK and DIRK methods of similar size.

References

[1] C. F. Gauss, "Ueber ein neues algemeines Grundgesetz der Mechanik," *J. fuer die reine und angewandte Mathematik*, vol. 4, pp. 232–235, 1829. In German.

[2] F. E. Udwadia and R. E. Kalaba, *Analytical Dynamics*. New York: Cambridge University Press, 1996.

[3] G. A. Maggi, *Principii di stereodinamica: Corso sulla formazione, l'interpretazione e l'integrazione delle equazioni del movimento dei solidi*. Milano: Hoepli, 1903. In Italian.

[4] T. R. Kane and C. F. Wang, "On the derivation of equations of motion," *J. Soc. Ind. Appl. Math.*, vol. 13, no. 2, pp. 487–492, 1965.

[5] T. Levi-Civita and U. Amaldi, *Lezioni di Meccanica Razionale — Vol. II: Dinamica dei Sistemi con un numero finito di gradi di libertà, parte II*. Bologna: Zanichelli, 1974. In Italian.

[6] S. S. Kim and M. J. Vanderploeg, "QR decomposition for state space representation of constrained mechanical dynamic systems," *J. of Mech. Trans.*, vol. 108, no. 2, pp. 183–188, 1986. doi:10.1115/1.3260800.

[7] R. P. Singh and P. W. Likins, "Singular value decomposition for constrained dynamical systems," *J. Appl. Mech.*, vol. 52, pp. 943–948, December 1985. doi:10.1115/1.3169173.

[8] N. K. Mani, E. J. Haug, and K. E. Atkinson, "Application of singular value decomposition for analysis of mechanical system dynamics," *J. Mech. Trans. Auto. Des.*, vol. 107, no. 1, pp. 82–87, 1985. doi:10.1115/1.3258699.

[9] E. C. Steeves and J. Walton, W. C., "A new matrix theorem and its application for establishing independent coordinates for complex dynamical systems with constraints," TR R-326, NASA, 1969.

[10] C. G. Liang and G. M. Lance, "A differentiable null space method for constrained dynamic analysis," *J. of Mech. Trans.*, vol. 109, no. 3, pp. 405–411, 1987. doi:10.1115/1.3258810.

[11] O. P. Agrawal and S. Saigal, "Dynamic analysis of multi-body systems using tangent coordinates," *Computers & Structures*, vol. 31, no. 3, pp. 349–355, 1989. doi:10.1016/0045-7949(89)90382-9.

[12] E. Pennestrì and L. Vita, "Strategies for the numerical integration of DAE systems in multibody dynamics," *Computer Applications in Engineering Education*, vol. 12, no. 2, pp. 106–116, 2004. doi:10.1002/cae.20005.

[13] L. Mariti, N. P. Belfiore, E. Pennestrì, and P. P. Valentini, "Comparison of solution strategies for multibody dynamics equations," *Intl. J. Num. Meth. Engng.*, 2011. doi:10.1002/nme.3190.

[14] P. Masarati, M. Lanz, and P. Mantegazza, "Multistep integration of ordinary, stiff and differential-algebraic problems for multibody dynamics applications," in *XVI Congresso Nazionale AIDAA*, (Palermo), pp. 71.1–10, 24–28 September 2001.

[15] A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, "Determining Lyapunov exponents from a time series," *Physica D: Nonlinear Phenomena*, vol. 16, pp. 285–317, July 1985. doi:10.1016/0167-2789(85)90011-9.

[16] L. Dieci, R. D. Russell, and E. S. Van Vleck, "On the computation of Lyapunov exponents for continuous dynamical systems," *SIAM Journal on Numerical Analysis*, vol. 34, no. 1, pp. 402–423, 1997. doi:10.1137/S0036142993247311.

[17] V. I. Oseledec, "A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems," *Trans. Moscow Math. Soc.*, vol. 19, pp. 197–231, 1968.

[18] V. H. Linh and V. Mehrmann, "Lyapunov, Bohl and Sacker-Sell spectral intervals for differential-algebraic equations," *J. Dyn. Diff. Equat.*, vol. 21, pp. 153–194, 2009. doi:10.1007/s10884-009-9128-7.

[19] J. Baumgarte, "Stabilization of constraints and integrals of motion in dynamical systems," *Comput. Meth. Appl. Mech. Engng.*, vol. 1, pp. 1–36, 1972. doi:10.1016/0045-7825(72)90018-7.

[20] C. W. Gear, B. Leimkuhler, and G. K. Gupta, "Automatic integration of Euler-Lagrange equations with constraints," *J. Comp. Appl. Math.*, vol. 12&13, pp. 77–90, 1985. doi:10.1016/0377-0427(85)90008-1.

[21] C. Führer and B. J. Leimkuhler, "Numerical solution of differential-algebraic equations for constrained mechanical motion," *Numerische Mathematik*, vol. 59, pp. 55–69, December 1991. doi:10.1007/BF01385770.

[22] K. E. Brenan, S. L. V. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. New York: North-Holland, 1989.

[23] E. Hairer, C. Lubich, and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Lecture Notes in Mathematics, Berlin Heidelberg, Germany: Springer-Verlag, 1989.

[24] J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*. Chichester, England: John Wiley & Sons Ltd., 1991.

[25] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations*, vol. II. Berlin Heidelberg, Germany: Springer-Verlag, 1996. 2nd rev. ed.