TU WIEN TECHNISCHE UNIVERSITÄT WIEN

BACHELOR'S THESIS

Constrained rigid body mechanics: A DPG finite element approach

submitted to the

Institute of
Analysis and Scientific Computing
TU Wien

under the supervision of

Univ.Prof. Dipl.-Ing. Dr.techn. Joachim Schöberl

by

Norbert Hammer

Matriculation number: 12225948

Vienna, October 6, 2025

TODO: correct date!, improve

Contents

Appendix A. Theoretical remarks

Introduction

For computing the motion of a rigid body by the laws of physics, the starting point is often a variational principle such as Livens' principle. (This principle shall be introduced in the first chapter.) A usual next step would be to derive an ODE from the variational principle. The ODE can then be solved numerically. Alternatively, one might directly employ numerics of PDEs. The downside of the latter approach is that the time interval of interest usually cannot be broken up into time steps for numerical computation.

Here, a (Discontinuous Petrov-Galerkin) FEM approach shall be taken to remedy this disadvantage. This leads to time stepping algorithms capable of computing the motion of constrained mechanical systems.

The outline of this thesis is as follows: Chapter 1 consists of the derivation of the family of time steppers that is the subject of this thesis. In Chapter 2, we proceed to prove convergence of the time steppers, for simplified assumptions. Subsequently, Chapter 3 illustrates that the time steppers are symmetric, symplectic, and conserve energy; at least in a simple setting. Furthermore, Chapter 4 provides the tools for applying the time steppers to systems of rigid bodies. Finally, in Chapter 5, the results of numerical experiments are given. In a benchmark problem, one of the time steppers showcases fourth order convergence of the global error. All tests exhibit only minor oscillations in the energy.

The aforementioned time stepper with convergence order four[1] is as follows; the symbols that appear in it are explained below.

$$
\begin{aligned}
\operatorname{alg} P_{q(t_k)}\Big[\frac{t_{k+1}-t_k}{6}[F(t_k, q(t_k), v(t_k)) - \nabla V(q(t_k))&\\
-(M\dot{v}(t_k)) + \lambda(t_k)^T G(q(t_k))] - (Mv(t_k) - \hat{p}(t_k))\Big] &= 0\\
\operatorname{alg} P_{q\left(t_{k+1/2}\right)}[F(t_{k+1/2}, q(t_{k+1/2}), v(t_{k+1/2})) - \nabla V\left(q\left(t_{k+1/2}\right)\right)&\\
-\left(M\dot{v}\left(t_{k+1/2}\right)\right) + \lambda\left(t_{k+1/2}\right)^T G\left(q\left(t_{k+1/2}\right)\right)] &= 0\\
\operatorname{alg} P_{q(t_{k+1})}\Big[\frac{t_{k+1}-t_k}{6}[F(t_{k+1}, q(t_{k+1}), v(t_{k+1})) - \nabla V(q(t_{k+1}))&\\
-(M\dot{v}(t_{k+1})) + \mu(t_{k+1})^T G(q(t_{k+1}))] + (Mv(t_{k+1}) - \hat{p}(t_{k+1}))\Big] &= 0\\
\\
g\left(q\left(t_{k+1/2}\right)\right) &= 0\\
g(q(t_{k+1})) &= 0\\
G(q(t_{k+1}))M^{-1}\hat{p}(t_{k+1}) &= 0\\
\\
\operatorname{alg}_q v &= v^a\\
\\
\forall \delta p_1, \delta p_2 :\ (\dot{q}(t_k) - \operatorname{alg}^{-1}_{q(t_k)} v^a(t_k)) \cdot \operatorname{alg}^{-1}_{q(t_k)} \delta p_1&\\
+2\left(\dot{q}(t_{k+1/2}) - \operatorname{alg}^{-1}_{q(t_{k+1/2})} \frac{v^a(t_k) + v^a(t_{k+1})}{2}\right) \cdot \operatorname{alg}^{-1}_{q(t_{k+1/2})}(\delta p_1 + \delta p_2)&\\
+(\dot{q}(t_{k+1}) - \operatorname{alg}^{-1}_{q(t_{k+1})} v^a(t_{k+1})) \cdot \operatorname{alg}^{-1}_{q(t_{k+1})} \delta p_2 &= 0\\
\\
q(t_{k+1/2}) &\in \mathbb{R}^3 \times SO(3)\\
q(t_{k+1}) &\in \mathbb{R}^3 \times SO(3)
\end{aligned}
\tag{0.1}
$$

Here is a short overview of the symbols used in the equations above: t_k and t_{k+1} are beginning and end of the time step, $t_{k+1/2}$ is their mean value; q, v and $\hat{p}$ are the (generalized) position, velocity and momentum, respectively. v^a denotes the tangent space representation of the velocity. q is assumed to be a polynomial of degree 2, whereas v^a is assumed to be a polynomial of degree one. alg maps to the Lie algebra of the configuration manifold while P projects to the tangent space of that manifold. F, V, M, g and G denote the "external force", potential energy, mass matrix, constraint function and derivative of the constraint function, respectively.

Abstract acknowledgement?, reference to Linus?

[1] In benchmark testing.

1. Derivation of a time stepper

Better way to make force time-depende All I have found in analytical mechanics allows only for F(q,v)!

In this section, the derivation of a time-stepping algorithm from Livens' principle will be discussed.

Let

- $\mathcal{Q}$ be a differentiable manifold which is the configuration space of a dynamical system, and assume (without loss of generality[1]) $\mathcal{Q} \in \mathbb{R}^{\mathfrak{M}}$ for some finite set $\mathfrak{M}$;
- $q : [0, T] \to \mathcal{Q}$ be the generalized position of that dynamical system;
- $v(t) \in T_{q(t)}\mathcal{Q}$ be its generalized velocity[2];
- $p(t) \in T^*_{q(t)}\mathcal{Q}$ be the corresponding momentum[3]
- M be a linear function induced by a symmetric, positive-definite, invertible matrix (the mass matrix)
- $V \in C^1(\mathbb{R}^{\mathfrak{M}})$ be the potential energy of the dynamical system;
- $L(q, v) = \frac{1}{2} v \cdot Mv - V(q)$ be the Lagrangian of the system.

Let $q(0)$ and $q(T)$ be given. Then Livens' principle states that the movement of our dynamical system solves

$$0 = \delta \int_0^T L(q, v) + p \cdot (\dot{q} - v) \; dt, \tag{1.1}$$

where δ denotes the first variation; see Definition A.1.1. This variational principle is also called Hamilton-Pontryagin principle.
(See [15, page 213] and [20, page 3].)
Note that it is a boundary value problem. Furthermore, the original version of Livens' principle would have to be defined on a parameter space for $\mathcal{Q}$. However, the use of the results of Section A.1.1 leads to a generalized result, which is Equation (1.1).

In addition, we want the dynamical system to fulfill the constraint(s) $g(q) = 0$[4]. We assume that $g \in C^1(\mathbb{R}^{\mathfrak{M}}, \mathbb{R}^n)$, with derivative G. Such a constraint can be accommodated through a supplement to Equation (1.1). With Lagrange parameters $\lambda : [0, T] \to \mathbb{R}^n$ and the definition

$$S(q, v, p, \lambda) := \int_0^T L(q, v) + p \cdot (\dot{q} - v) + \lambda \cdot g(q) \; dt, \tag{1.2}$$

[1] By the Whitney embedding theorem, see for example [18, Theorem 6.15].

[2] $T_x\mathcal{Q}$ denotes the tangent space on $\mathcal{Q}$ at the point x. [13, Definition 2.2.20]

[3] $T^*_x\mathcal{Q}$ denotes the dual space of $T_x\mathcal{Q}$. [13, Definition 2.2.20]

[4] This kind of constraint is called scleronomic, i.e. it depends only on the generalized position. Therefore, it is also holonomic; that is, it depends on not more than the position and time.
[9, Chapter 2.1.1]

the stationary condition of interest turns into $\delta S(q, v, p, \lambda) = 0$ (for any given $q(0)$ and $q(T)$). [16, Section 2.3]

1.1. Taking the variation

For this section, let any test function from the variation in $\delta S(q, v, p, \lambda) = 0$ be split up component-wise into $(\delta q, \delta p, \delta v, \delta\lambda)$, where δq runs through the tangent space, as in Section A.1.1. Also, assume that $q, \delta q \in C^1([0, T])$ and $p, v, \lambda, \delta p, \delta v \in C([0, T])$.

Proposition 1.1.1
$\delta S(q, v, p, \lambda) = 0$ is equivalent to:

$$\begin{aligned} \delta q(0) = \delta q(T) = 0 \Rightarrow 0 &= \int_0^T -\nabla V(q) \cdot \delta q + (Mv) \cdot \frac{\partial \delta q}{\partial t} + \lambda^T G(q) \cdot \delta q \; dt \\ \dot{q} &= v \\ p &= Mv \\ \forall t \in [0, T] : \; 0 &= g(q(t)) \end{aligned} \tag{1.3}$$

for all δq.

Proof. In the following, exchanging the derivative with respect to ε and the integral can be justified by the Leibniz rule. Regarding the condition $\delta q(0) = \delta q(T) = 0$, it stems from the fact that $q(0)$ and $q(T)$ are given, which is a restriction on the set X from Theorem A.1.1.

$\delta S(q, v, p, \lambda) = 0$ holds if and only if for all $\delta q, \delta p, \delta v$ and $\delta\lambda$ such that $\delta q(0) = \delta q(T) = 0$, the following holds:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \varepsilon}\int_0^T L(q+\varepsilon\delta q, v+\varepsilon\delta v) + (p+\varepsilon\delta p)\cdot(\dot{q}+\varepsilon\frac{d\delta q}{dt} - v - \varepsilon\delta v)\\
&\qquad + (\lambda+\varepsilon\delta\lambda)\cdot g(q+\varepsilon\delta q)\ dt\Big|_{\varepsilon=0}\\
&= \int_0^T \nabla L(q+\varepsilon\delta q, v+\varepsilon\delta v)\cdot(\delta q,\delta v)\\
&\qquad + \delta p\cdot(\dot{q}+\varepsilon\frac{d\delta q}{dt} - v - \varepsilon\delta v) + (p+\varepsilon\delta p)\cdot(\frac{d\delta q}{dt} - \delta v)\\
&\qquad + \delta\lambda\cdot g(q+\varepsilon\delta q) + (\lambda+\varepsilon\delta\lambda)^T G(q+\varepsilon\delta q)\delta q\ dt\Big|_{\varepsilon=0}\\
&= \int_0^T \nabla L(q,v)\cdot(\delta q,\delta v) + \delta p\cdot(\dot{q}-v) + p\cdot(\frac{d\delta q}{dt} - \delta v)\\
&\qquad + \delta\lambda\cdot g(q) + \lambda^T G(q)\delta q\ dt\\
&= \int_0^T -\nabla V(q)\cdot\delta q + (Mv)\cdot\delta v + \delta p\cdot(\dot{q}-v) + p\cdot\frac{d\delta q}{dt} - p\cdot\delta v\\
&\qquad + \delta\lambda\cdot g(q) + \lambda^T G(q)\delta q\ dt\\
&= \int_0^T -\nabla V(q)\cdot\delta q + p\cdot\frac{d\delta q}{dt} + \lambda^T G(q)\delta q\ dt + \int_0^T (\dot{q}-v)\cdot\delta p\ dt\\
&\qquad + \int_0^T (Mv-p)\cdot\delta v\ dt + \int_0^T g(q)\cdot\delta\lambda\ dt
\end{aligned}$$

The last part of the above chain of equalities can be split up with respect to the different test functions. This yields

$$\begin{aligned}
\forall\delta q:\ \left[\delta q(0)=\delta q(T)=0 \Rightarrow 0 = \int_0^T -\nabla V(q)\cdot\delta q + p\cdot\frac{\partial\delta q}{\partial t} + \lambda^T G(q)\cdot\delta q\ dt\right]\\
\forall\delta p:\ \ 0 = \int_0^T (\dot{q}-v)\cdot\delta p\ dt\\
\forall\delta v:\ \ 0 = \int_0^T (Mv-p)\cdot\delta v\ dt\\
\forall\delta\lambda:\ \ 0 = \int_0^T g(q)\cdot\delta\lambda\ dt.
\end{aligned}$$

The application of the fundamental lemma of the calculus of variations to the last three of these equations (and the substitution $p = Mv$ in the first equation) implies the system of equations (1.3). The other half of the equivalence can be obtained by multiplying with test functions and integrating.

□

Proposition 1.1.2

Assume that (q, v, p, λ) solve $\delta S(q, v, p, \lambda) = 0$. Furthermore, define $\hat{p}(0) := p(0)$ and $\hat{p}(T) := p(T)$. Then (q, v, p, λ) also solves

$$\begin{aligned} \forall \delta q: \ 0 &= \int_0^T -\nabla V(q) \cdot \delta q + (Mv) \cdot \frac{\partial \delta q}{\partial t} + \lambda^T G(q) \cdot \delta q \ dt - \hat{p} \cdot \delta q \Big|_0^T & \text{(1.4a)} \\ \dot{q} &= v & \text{(1.4b)} \\ p &= Mv & \text{(1.4c)} \\ \forall t \in [0, T]: \ 0 &= g(q(t)). & \text{(1.4d)} \end{aligned}$$

Proof. Per Theorem 1.1.1, $\delta S(q, v, p, \lambda) = 0$ is equivalent to (1.3). Equations (1.4b) to (1.4d) are the same as the second to fourth lines of (1.3).

In regard to the respective first lines, the restriction $\delta q(0) = \delta q(T) = 0$ leads to the disappearance of the boundary term

$$-\hat{p} \cdot \delta q \Big|_0^T.$$

□

Proposition 1.1.3

If Equation (1.4) has a solution $(q, v, p, \lambda, \hat{p}(0), \hat{p}(T))$, then

$$\begin{aligned} \delta S(q, v, p, \lambda) &= 0 \\ \hat{p}(0) &= p(0) \\ \hat{p}(T) &= p(T) \end{aligned}$$

holds.

Proof. In analogy to the proof of Theorem 1.1.2, only the first line of (1.4) is of interest. For $\delta q(0) = \delta q(T) = 0$, it leads to the first line of (1.3).

Partial integration of the first line of Equation (1.4) yields

$$\forall \delta q: \ 0 = \int_0^T (-\nabla V(q) - (M\ddot{q}) + \lambda^T G(q)) \cdot \delta q \ dt + (M\dot{q} - \hat{p}) \cdot \delta q \Big|_0^T.$$

Here, the choice of $\delta q(0) = \delta q(T) = 0$ results in the disappearance of the boundary terms. The application of the fundamental lemma of the calculus of variations leads to $-\nabla V(q) - (M\ddot{q}) + \lambda^T G(q) = 0$. Thus, the integral above is zero, resulting in $(M\dot{q} - \hat{p}) \cdot \delta q\big|_0^T = 0$ for all δq. Therefore, $p(0) \overset{(1.4)}{=} M\dot{q}(0) = \hat{p}(0)$ and $p(T) \overset{(1.4)}{=} M\dot{q}(T) = \hat{p}(T)$ hold.

□

1.2. An initial value problem

The system of equations (1.4) is clearly equivalent to

$$\begin{aligned}&\forall \delta q, \delta\lambda :\\ &\quad 0 = \int_0^T -\nabla V(q)\cdot\delta q + (Mv)\cdot\frac{\partial\delta q}{\partial t} + \lambda^T G(q)\cdot\delta q\ dt - \hat{p}\cdot\delta q\Big|_0^T + \int_0^T g(q)\cdot\delta\lambda\ dt\\ &\quad \dot{q} = v\\ &\quad p = Mv.\end{aligned}$$

Assume now that $q \in C^2([0,T])$. Then partial integration of $(Mv)\cdot\frac{\partial\delta q}{\partial t}$ yields

$$\begin{aligned}&\forall \delta q, \delta\lambda :\\ &0 = \int_0^T (-\nabla V(q) - (M\dot{v}) + \lambda^T G(q))\cdot\delta q\ dt + (Mv - \hat{p})\cdot\delta q\Big|_0^T + \int_0^T g(q)\cdot\delta\lambda\ dt \qquad (1.5)\\ &\dot{q} = v\\ &p = Mv.\end{aligned}$$

This system of equations shall now be seen as an initial value problem where the initial values are $q(0)$ and $\hat{p}(0)$.

1.3. FEM in time

To numerically obtain a solution path for initial values $q(0)$ and $\hat{p}(0)$, a FEM discretization can be employed. That is, consecutive intervals $[t_k, t_{k+1}] \subseteq [0,T]$ are seen as the elements of a FEM mesh. On each of these intervals, Equation (1.5) is solved (for boundary points $[t_k, t_{k+1}]$ instead of $[0,T]$). In this setting, let the trial space for q be the space of polynomials of degree $j > 0$. Then the test space of the δq also has to be $j+1$-dimensional. Furthermore, the exact integral can be approximated using a $j+1$ point (Gauss-)Lobatto rule. These quadratures are characterized by their use of the endpoints of the integration interval as evaluation points (see Section A.5 and [11, Section 8.12]).

Let $\alpha_i, i = 1,\ldots,j+1$ be the weights and $\xi_i, i = 1,\ldots,j+1$ the evaluation points of the $j+1$

point Gauss-Lobatto rule on $[t_k, t_{k+1}]$. Then the discretization leads to

$$\begin{aligned}
0 = &\int_{t_k}^{t_{k+1}} (-\nabla V(q) - (M\dot{v}) + \lambda^T G(q)) \cdot \delta q \; dt + (Mv - \hat{p}) \cdot \delta q \Big|_{t_k}^{t_{k+1}} \\
&\quad + \int_{t_k}^{t_{k+1}} g(q) \cdot \delta\lambda \; dt \\
\approx\, & \alpha_1[-\nabla V(q(t_k)) - (M\dot{v}(t_k)) + \lambda(t_k)^T G(q(t_k))] \cdot \delta q(t_k) \\
&\qquad - (Mv(t_k) - \hat{p}(t_k)) \cdot \delta q(t_k) \\
&\quad + \sum_{i=2}^{j} \alpha_i[-\nabla V(q(\xi_i)) - (M\dot{v}(\xi_i)) + \lambda(\xi_i)^T G(q(\xi_i))] \cdot \delta q(\xi_i) \\
&\quad + \alpha_{j+1}[-\nabla V(q(t_{k+1})) - (M\dot{v}(t_{k+1})) + \lambda(t_{k+1})^T G(q(t_{k+1}))] \cdot \delta q(t_{k+1}) \\
&\qquad + (Mv(t_{k+1}) - \hat{p}(t_{k+1})) \cdot \delta q(t_{k+1}) \\
&\quad + \sum_{i=0}^{j+1} \alpha_i g(q(\xi_i)) \cdot \delta\lambda(\xi_i)
\end{aligned} \tag{1.6}$$

for all δq and $\delta\lambda$. This discretized expression is zero if and only if, for all[5] $\delta q, \delta\lambda$

$$\begin{aligned}
[\alpha_1[-\nabla V(q(t_k)) - (M\dot{v}(t_k)) + \lambda(t_k)^T G(q(t_k))] - (Mv(t_k) - \hat{p}(t_k))] \cdot \delta q(t_k) &= 0 \\
\forall i \in \{2, \dots, j\} : \; [-\nabla V(q(\xi_i)) - (M\dot{v}(\xi_i)) + \lambda(\xi_i)^T G(q(\xi_i))] \cdot \delta q(\xi_i) &= 0 \\
[\alpha_{j+1}[-\nabla V(q(t_{k+1})) - (M\dot{v}(t_{k+1})) + \lambda(t_{k+1})^T G(q(t_{k+1}))] & \\
+(Mv(t_{k+1}) - \hat{p}(t_{k+1}))] \cdot \delta q(t_{k+1}) &= 0 \\
\forall i \in \{1, \dots, j+1\} : \; g(q(\xi_i)) \cdot \delta\lambda(\xi_i) &= 0.
\end{aligned}$$

The above system of equations is equivalent to

$$\forall \delta q_k \in T_{q(t_k)}, \forall i \in \{2, \dots, j\} \; \forall \delta q_{\xi_i} \in T_{q(\xi_i)}, \forall \delta q_{k+1} \in T_{q(t_{k+1})} :$$

$$\begin{aligned}
[\alpha_1[-\nabla V(q(t_k)) - (M\dot{v}(t_k)) + \lambda(t_k)^T G(q(t_k))] - (Mv(t_k) - \hat{p}(t_k))] \cdot \delta q_k &= 0 \\
[-\nabla V(q(\xi_i)) - (M\dot{v}(\xi_i)) + \lambda(\xi_i)^T G(q(\xi_i))] \cdot \delta q_{\xi_i} &= 0 \\
[\alpha_{j+1}[-\nabla V(q(t_{k+1})) - (M\dot{v}(t_{k+1})) + \lambda(t_{k+1})^T G(q(t_{k+1}))] & \\
+(Mv(t_{k+1}) - \hat{p}(t_{k+1}))] \cdot \delta q_{k+1} &= 0 \\
g(q(t_k)) &= 0 \\
g(q(\xi_i)) &= 0 \\
g(q(t_{k+1})) &= 0,
\end{aligned} \tag{1.7}$$

which is independent of the concrete choice of the test spaces for $\delta q(t)$ and $\delta\lambda(t)$. These equations are meant to be solved given $q(0)$ and $\hat{p}(0)$.

Throughout this discretization process, the advantage of the $j+1$ point quadrature rule has come into play: For any higher amount of quadrature points, the constraint $g(q(t_{k+1})) = 0$ would - in full generality - not be enforced.

[5] It is assumed that the spaces of δq and $\delta\lambda$ separate all points ξ_i.

1.4. Projection of $\hat{p}$

Seeing $t = t_k$ and $t = t_{k+1}$ as the start and end of a time step, Equation (1.7) can be used as the description of an implicit time stepper. However, there are three issues with that.

Firstly, $\lambda(t_{k+1})$ is required to compute $\hat{p}(t_{k+1})$. In the next time step, though, $\lambda(t_{k+1})$ would have to be calculated from $q(t_{k+1})$ and $\hat{p}(t_{k+1})$ using a different equation. As one can see, the two results of $\lambda(t_{k+1})$ need not be the same.

Secondly, $\hat{p}(t_{k+1}) \in T^*_{q(t_{k+1})}\mathcal{Q}$ does not have to hold.

In analogy to the derivation of the RATTLE solver (see [1, Chapter II] and [10, Chapter VII.1.4, 1.25]), both of these issues can be resolved simultaneously. This is achieved by approximating - on the time step $[t_k, t_{k+1}]$ - $\lambda(t_{k+1})$ with a Lagrange parameter $\mu(t_{k+1})$ and adding the extra equation

$$G(q(t_{k+1}))M^{-1}\hat{p}(t_{k+1}) = 0.$$

This equation stems from the abstract equations $0 = \dot{g} = G(q)v$ and $p = Mv$.

Along with others, this adjustment can be found in (1.8).

1.5. Final adjustments

The system of time stepper equations still has to be seen in conjunction with the equation $\dot{q} = v$. This is problematic since, in full generality, the discrete q leaves the manifold $\mathcal{Q}$; therefore, the corresponding $\dot{q}$ does not have to lie in the tangent space of $\mathcal{Q}$. In benchmark testing, it has proven fruitful[6] to instead choose v such that it lies in the tangent space of $\mathcal{Q}$ for the points in time $\xi_i, i = 1, \ldots, j+1$. As a result, v becomes a mere approximation of $\dot{q}$. An example for this approach (for rigid body mechanics) can be found in Subsection 5.2.2.

Finally, an "external force" term $F(t, q, v) \in T^*_{q(t)}\mathcal{Q}$ can be added, in analogy to [20, Chapter 4].

[6] See the comparison between the subsections 5.2.1 and 5.2.2.

The resulting system of equations is:

$$\forall \delta q_k \in T_{q(t_k)}, \forall i \in \{2, \dots, j\} \; \forall \delta q_{\xi_i} \in T_{q(\xi_i)}, \forall \delta q_{k+1} \in T_{q(t_{k+1})} :$$

$$\begin{aligned}
[\alpha_1[F(t_k, q(t_k), v(t_k)) - \nabla V(q(t_k)) - (M\ddot{q}(t_k)) + \lambda(t_k)^T G(q(t_k))]& \\
-(M\dot{q}(t_k) - \hat{p}(t_k))] \cdot \delta q_k &= 0 \\
[F(\xi_i, q(\xi_i), v(\xi_i)) - \nabla V(q(\xi_i)) - (M\ddot{q}(\xi_i)) + \lambda(\xi_i)^T G(q(\xi_i))] \cdot \delta q_{\xi_i} &= 0 \\
[\alpha_{j+1}[F(t_{k+1}, q(t_{k+1}), v(t_{k+1})) - \nabla V(q(t_{k+1})) - (M\ddot{q}(t_{k+1}))& \\
+\mu(t_{k+1})^T G(q(t_{k+1}))] + (M\dot{q}(t_{k+1}) - \hat{p}(t_{k+1}))] \cdot \delta q_{k+1} &= 0 \\
\\
g(q(\xi_i)) &= 0 \\
g(q(t_{k+1})) &= 0 \\
G(q(t_{k+1}))M^{-1}\hat{p}(t_{k+1}) &= 0 \\
\\
v &\approx \dot{q} \\
\\
q(\xi_i) &\in \mathcal{Q} \\
q(t_{k+1}) &\in \mathcal{Q}
\end{aligned} \tag{1.8}$$

(The last two lines do not appear out of nowhere, but are meant as a reminder. Before, they were always implicitly assumed.)

2. On convergence and error estimation

The convergence of the time stepping method from the previous chapter is yet to be discussed. Here, a simplified case will be analyzed. To be precise, it is assumed that

- $\mathcal{Q} = \mathbb{R}^m$,
- $g = 0$,
- $\nabla V(q) - F(t, q, v) = Kq - F(t)$, where K is symmetric and positive definite with its smallest eigenvalue being denoted by $\kappa > 0$, and F is a piecewise[1] polynomial of degree $j - 1$, and that
- the smallest eigenvalue of M, denoted by $\nu > 0$, fulfills $\nu > \|K\|_2$.

In addition, the choice $v = \dot{q}$ is made. This already has for result that v lies in the tangent space of $\mathcal{Q}$.

Provided that T is small enough, the assumption $\nu > \|K\|_2$ has for result that a $\beta > 0$ exists such that

$$\nu \geq T^2(2\beta + \|K\|_2) \tag{2.1a}$$

$$\nu \geq 2\beta + \|K\|_2. \tag{2.1b}$$

However, should T be too large, the interval $[0, T]$ can simply be split up. Therefore, we can assume without loss of generality that (2.1) holds true. Another assumption that can be made without loss of generality is that $F(t)$ is a (non-piecewise, continuous) polynomial function: Otherwise, time intervals can be pieced together.

The assumptions result in the following simplified problem: Given $q(0) = q_0, \hat{p}(0) = p_0$, solve

$$\forall \delta q : \ 0 = \int_0^T (F(t) - Kq) \cdot \delta q + (M\dot{q}) \cdot \frac{\partial \delta q}{\partial t} \, dt - \hat{p} \cdot \delta q \Big|_0^T. \tag{2.2}$$

For the analysis, functional analytic results from the Discontinuous Petrov-Galerkin method(ology) (DPG) will be used. First, a boundary value problem will be analyzed for the error of its discrete DPG solution. Subsequently, adjustments will be made to obtain the numerical solution of an initial value problem.

The aforementioned boundary value problem is: Solve

$$\forall \delta q \in H_0^1([0, T])^m : \ 0 = \int_0^T (F(t) - Kq) \cdot \delta q + (M\dot{q}) \cdot \frac{\partial \delta q}{\partial t} \, dt - \hat{p} \cdot \delta q \Big|_0^T \tag{2.3}$$

for $q \in H^1([0, T])^m$ where $q(0) = q_0$ and $q(T) = q_T$ are fixed. This is a Dirichlet boundary value problem. We assume without loss of generality [2] that $q_0 = q_T = 0$, which is equivalent

[1] F can be allowed to have finitely many discontinuities.

[2] For the general case, see Section 2.3.

to q being in H_0^1.

2.1. Inf-sup conditions and the broken test space

This section deals with the application of a theorem:

Theorem 2.1.1
Let $X_0, \hat{X}$ and Y be Hilbert spaces, and Y_0 be a closed subspace of Y. On these spaces, take two bilinear forms

$$\begin{aligned} b_0 :& X_0 \times Y \to \mathbb{R} \\ \hat{b} :& \hat{X} \times Y \to \mathbb{R} \end{aligned}$$

and define $X := X_0 \times \hat{X}$ (which, for the Cartesian product norm, is a Hilbert space) as well as the bilinear form

$$b((x_0, \hat{x}), y) = b_0(x_0, y) + \hat{b}(\hat{x}, y).$$

In this setting, assume that

- b_0 and $\hat{b}$ are continuous,
- $Y_0 = \{y \in Y \mid \forall \hat{x} \in \hat{X} : \hat{b}(\hat{x}, y) = 0\}$,
- and that there exist constants $c_0, \hat{c} > 0$ such that

$$\inf_{x_0 \in X_0 \setminus \{0\}} \sup_{y_0 \in Y_0 \setminus \{0\}} \frac{b_0(x_0, y_0)}{\|x_0\|_{X_0} \|y_0\|_Y} \geq c_0$$
$$\inf_{\hat{x} \in \hat{X} \setminus \{0\}} \sup_{y \in Y \setminus \{0\}} \frac{\hat{b}(\hat{x}, y)}{\|\hat{x}\|_{\hat{X}} \|y\|_Y} \geq \hat{c}.$$

(The inequalities from the last bullet point are called inf-sup conditions.)
Under these assumptions, b is continuous. In addition, there exists a $c > 0$ such that b fulfills the inf-sup condition

$$\inf_{(x_0, \hat{x}) \in X \setminus \{0\}} \sup_{y \in Y \setminus \{0\}} \frac{b((x_0, \hat{x}), y)}{\|(x_0, \hat{x})\|_X \|y\|_Y} \geq c.$$

c can be calculated, using the formula

$$c = \sqrt{\frac{1}{\frac{1}{c_0^2} + \frac{1}{\hat{c}^2}\left(\frac{\|b_0\|}{c_0} + 1\right)^2}}.$$

Finally, the identity

$$\{y_0 \in Y_0 \mid \forall x_0 \in X_0 : b_0(x_0, y_0) = 0\} = \{y \in Y \mid \forall (x_0, \hat{x}) \in X : b((x_0, \hat{x}), y) = 0\} \quad (2.4)$$

holds.

The theorem above is merely a version of a theorem of Carstensen, Demkowicz and Gopalakrishnan [4, Theorem 3.3]. A proof can be found there.

With the last theorem, we have all the required tools to prove sufficient conditions for the existence of (continuous and discrete) solutions. We therefore verify that its assumptions hold in our case. Firstly, b_0 shall be dealt with.

Equation (2.3) can be written as

$$\forall \delta q \in H_0^1([0,T])^m : \; b_0(q, \delta q) = \ell(\delta q),$$

where the bilinear form b_0 and the linear form ℓ are defined as

$$\begin{aligned} b_0(q, \delta q) &:= \int_0^T -(Kq) \cdot \delta q + (M\dot{q}) \cdot \frac{\partial \delta q}{\partial t} \; dt \\ \ell(\delta q) &:= \int_0^T -F(t) \delta q \; dt + \hat{p} \cdot \delta q \Big|_0^T . \end{aligned}$$

b_0 is coercive on $H_0^1([0,T])^m$ (see Proposition A.3.2):

$$b_0(q,q) \geq \beta \|q\|_{H^1}^2$$

Thus, b_0 also fulfills the inf-sup condition

$$\inf_{q \in H_0^1([0,T])^m \setminus \{0\}} \quad \sup_{\delta q \in H_0^1([0,T])^m \setminus \{0\}} \frac{|b_0(q, \delta q)|}{\|q\|_{H^1} \|\delta q\|_{H^1}} \geq \beta,$$

which is easy to see by choosing $\delta q = q$.

2.1.1. The broken test space

Given a set system $\mathcal{K}$ (here: over $[0,T]$), the test space can be expanded to form the broken test space

$$\begin{aligned} H^1(\mathcal{K}) &:= \{ f \in L^2([0,T]) \mid \forall \varkappa \in \mathcal{K} : \; f|_\varkappa \in H^1(\varkappa) \} \\ \|f\|_{H^1(\mathcal{K})} &:= \sqrt{\sum_{\varkappa \in \mathcal{K}} \|f\|_{H^1(\varkappa)}^2}. \end{aligned}$$

Here, $\mathcal{K}$ is chosen as the decomposition of $[0,T]$ into $\eta \in \mathbb{N}$ intervals (time steps/mesh elements)

$$\mathcal{K}_\eta := \left\{ \left(\frac{i-1}{\eta} T, \frac{i}{\eta} T \right) \;\middle|\; i = 1, \ldots, \eta \right\}.$$

With this definition, the domain of b_0 can be expanded to $H^1(\mathcal{K}_\eta)$. Then, b_0 is (still) continuous, with a continuity constant that is independent of η. (See Proposition A.3.1.)

We have now verified the conditions on b_0 for Theorem 2.1.1. Hence, we can turn towards doing the same thing for $\hat{b}$.

2.1.2. The interface and broken bilinear forms

For $(\hat{p}_i^*)_{i=0}^{\eta} \in (\mathbb{R}^m)^{\eta+1}$, we define

$$\begin{aligned}\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q) &:= -\hat{p}_\eta^* \cdot \operatorname{tr}\left(\delta q\big|_{\left(\frac{\eta-1}{\eta}T,T\right)}\right)(T) + \hat{p}_0^* \cdot \operatorname{tr}\left(\delta q\big|_{\left(0,\frac{1}{\eta}T\right)}\right)(0) \\ &\quad + \sum_{i=1}^{\eta-1} \hat{p}_i^* \cdot \left[\operatorname{tr}\left(\delta q\big|_{\left(\frac{i}{\eta}T,\frac{i+1}{\eta}T\right)}\right)\left(\frac{i}{\eta}T\right) - \operatorname{tr}\left(\delta q\big|_{\left(\frac{i-1}{\eta}T,\frac{i}{\eta}T\right)}\right)\left(\frac{i}{\eta}T\right)\right] \\ b((q, (\hat{p}_i^*)_{i=0}^{\eta}), \delta q) &:= b_0(q, \delta q) + \hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q),\end{aligned}$$

where tr denotes the trace operator. $\hat{b}$ shall be called the interface bilinear form. (It evaluates the test function on the "interface" of mesh elements.) b shall be called the broken bilinear form.

The interface bilinear form fulfills the following property: $\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q) = 0$ holds for all $(\hat{p}_i^*)_{i=0}^{\eta} \in (\mathbb{R}^m)^{\eta+1}$ if and only if

$$\operatorname{tr}\left(\delta q\big|_{\left(0,\frac{1}{\eta}T\right)}\right)(0) = 0 \tag{2.5a}$$

$$\operatorname{tr}\left(\delta q\big|_{\left(\frac{\eta-1}{\eta}T,T\right)}\right)(T) = 0 \tag{2.5b}$$

$$\operatorname{tr}\left(\delta q\big|_{\left(\frac{i}{\eta}T,\frac{i+1}{\eta}T\right)}\right)\left(\frac{i}{\eta}T\right) = \operatorname{tr}\left(\delta q\big|_{\left(\frac{i-1}{\eta}T,\frac{i}{\eta}T\right)}\right)\left(\frac{i}{\eta}T\right) \qquad \forall i \in \{1, \dots, \eta-1\} \tag{2.5c}$$

holds. Equation (2.5c) is equivalent to δq being in $H^1([0,T])^m$. (For a proof of that, see Section A.2.) Under these circumstances, the combination of Equation (2.5a) and Equation (2.5b) is clearly equivalent to δq being in $H_0^1([0,T])^m$. We conclude that $\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q) = 0$ is equivalent to $\delta q \in H_0^1([0,T])^m$. This equivalence is one of the requirements of Theorem 2.1.1.

2.1.3. The inf-sup condition of the interface bilinear form

Finally, an inf-sup condition for $\hat{b}$ will be required, along with continuity. Both inequalities should be independent of η. We define

$$\|(\hat{p}_i^*)_{i=0}^{\eta}\|_* := \sup_{\delta q \in H^1(\mathcal{K}_\eta)^m} \frac{|\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q)|}{\|\delta q\|_{H^1}}.$$

This expression exists as a result of the continuity of the trace operator. Note that this norm[3] does depend on η.

[3] That it is the norm of a Hilbert space is a result of Carstensen et al. [4]. There, the combination [4, Equation (2b)] and [4, Equation (1a)] is of special importance for this result.

Under that norm, the number one is a continuity constant and inf-sup constant of $\hat{b}$:

$$\begin{aligned}
|\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q)| &= \|\delta q\|_{H^1} \frac{|\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q)|}{\|\delta q\|_{H^1}} \\
&\leq \|\delta q\|_{H^1} \sup_{\delta q' \in H^1(\mathcal{K}_\eta)^m} \frac{|\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q')|}{\|\delta q'\|_{H^1}} \\
&= \|\delta q\|_{H^1} \|(\hat{p}_i^*)_{i=0}^{\eta}\|_*
\end{aligned}$$

$$\inf_{(\hat{p}_i^*)_{i=0}^{\eta} \in (\mathbb{R}^m)^{\eta+1}} \sup_{\delta q \in H^1(\mathcal{K}_\eta)^m} \frac{|\hat{b}((\hat{p}_i^*)_{i=0}^{\eta}, \delta q)|}{\|(\hat{p}_i^*)_{i=0}^{\eta}\|_* \|\delta q\|_{H^1}} = \inf_{(\hat{p}_i^*)_{i=0}^{\eta} \in (\mathbb{R}^m)^{\eta+1}} \frac{\|(\hat{p}_i^*)_{i=0}^{\eta}\|_*}{\|(\hat{p}_i^*)_{i=0}^{\eta}\|_*} = 1$$

2.1.4. The broken form and unique solvability

Now, all requirements of Theorem 2.1.1 are verified. Thus, there exists a $C_1 > 0$ independent of η such that b fulfills the inf-sup condition

$$\inf_{(q,(\hat{p}_i^*)_{i=0}^{\eta}) \in H_0^1([0,T])^m \times (\mathbb{R}^m)^{\eta+1}} \sup_{\delta q \in H^1(\mathcal{K}_\eta)^m} \frac{|b((q,(\hat{p}_i^*)_{i=0}^{\eta}), \delta q)|}{\|(q,(\hat{p}_i^*)_{i=0}^{\eta})\| \|\delta q\|_{H^1}} \geq C_1, \tag{2.6}$$

where the norm for $(q, (\hat{p}_i^*)_{i=0}^{\eta})$ is chosen as the Cartesian product norm

$$\|(q, (\hat{p}_i^*)_{i=0}^{\eta})\|^2 = \|q\|^2_{H^1([0,T])^m} + \|(\hat{p}_i^*)_{i=0}^{\eta}\|^2_*.$$

Apart from that, since b_0 and $\hat{b}$ are continuous with a continuity constant independent of η, b is also continuous with a continuity constant independent of η.

Finally, the coercivity of b_0, in combination with Equation (2.4), leads to

$$\{y_0 \in Y_0 \mid \forall x_0 \in X_0 : \; b_0(x_0, y_0) = 0\} = \{0\}.$$

As a result, all requirements of Lemma 3.6 from [2] are given. That lemma states the existence of a unique (continuous) solution. That solution shall be written as

$$(\tilde{q}, (M\dot{\tilde{q}}(\frac{i}{\eta}T))_{i=0}^{\eta}).$$

2.2. An H^1 a priori error estimator

Our task is now to discretize the problem.

We define $\mathcal{P}_j$ as the space of polynomials of degree at most j and - for a set system $\mathcal{K}$:

$$\begin{aligned} \mathcal{M}_j(\mathcal{K}) &:= \left\{ p \in L_2\left(\bigcup_{\varkappa \in \mathcal{K}} \varkappa\right) \;\middle|\; \forall \varkappa \in \mathcal{K}:\; p|_\varkappa \in \mathcal{P}_j \right\} \\ \mathcal{M}_{j,0}(\mathcal{K}) &:= \mathcal{M}_j(\mathcal{K}) \cap C^0 = \mathcal{M}_j(\mathcal{K}) \cap H^1 \\ \mathcal{M}_{j,0,0}(\mathcal{K}) &:= \mathcal{M}_j(\mathcal{K}) \cap H^1_0. \end{aligned}$$

We choose our discrete trial space as

$$X_d := \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m \times (\mathbb{R}^m)^{\eta+1}.$$

The corresponding test space can be chosen as

$$Y_d := \{ f \mid \forall \varkappa \in \mathcal{K}_\eta:\; f|_\varkappa \in \mathcal{P}_j \}^m.$$

In addition, the integral from b_0 is approximated. To do so, the piecewise $(j+1)$ point Gauss Lobatto rule (on the intervals defined by $\mathcal{K}_\eta$) is used.[4] Let us call the resulting bilinear form $b_{0,\eta}$ and the corresponding broken form b_η. ($\hat{b}$ is not approximated. ℓ is integrated precisely by the quadrature since $F(t)$ has degree $j-1$.)

$b_{0,\eta}$ is coercive (see Section A.4) and has a finite-dimensional domain (namely the discrete trial and test space). Hence, we can derive the existence of a unique discrete solution to our problem

$$(q_d, (\hat{p}_{d,i})_{i=0}^\eta)$$

in full analogy to the continuous case.

2.2.1. Strang's first lemma and its application

When using a quadrature rule, one way of obtaining a finite element error estimate is Strang's first lemma. Ern and Guermond [6, Lemma 2.27] offer a version that can be used for unequal trial and test spaces. With minor simplifications, that version reads:

Lemma 2.2.1 (Strang 1 for inf-sup stable problems)
Let U and V be Hilbert spaces, with respective subspaces U_h and V_h. Assume that $\dim U_h = \dim V_h < \infty$. Furthermore, take a bounded bilinear form $A : U \times V \to \mathbb{R}$ which is inf-sup stable:

$$0 < a < \inf_{u \in U} \sup_{v \in V} \frac{A(u,v)}{\|u\|_U \|v\|_V}$$

Let the (approximating) bilinear form $A_h : U_h \times V_h \to \mathbb{R}$ also be inf-sup stable:

$$0 < a_h < \inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{A_h(u_h,v_h)}{\|u_h\|_U \|v_h\|_V}$$

[4] But only after partial integration of $M\dot{q}$.

We now denote by $w \in U$ and $w_h \in U_h$ the respective solutions of

$$\forall v \in V: \; A(w,v) = f(v)$$
$$\forall v_h \in V_h: \; A_h(w_h, v_h) = f_h(v_h),$$

where f is a continuous functional and f_h is a functional.
In this setting, we have the error bound

$$\begin{aligned} \|w - w_h\|_W \leq \frac{1}{a_h} \sup_{v_h \in V_h} & \frac{|f(v_h) - f_h(v_h)|}{\|v_h\|_V} \\ + \inf_{u_h \in U_h} & \left(\left[1 + \frac{\|A\|_{U_h \times V_h}}{a_h} \right] \|w - u_h\|_{U_h} \right. \\ & \left. + \frac{1}{a_h} \sup_{v_h \in V_h} \frac{|A(u_h, v_h) - A_h(u_h, v_h)|}{\|v_h\|_V} \right). \end{aligned} \tag{2.7}$$

A proof of this lemma has been given by Ern and Guermond [6, Proof of Lemma 2.27].

Theorem 2.2.1 can be applied to our current problem with $A = b$ and $A_h = b_\eta$; the (uniform) inf-sup stability of b_η follows from the uniform coercivity[5] of $b_{0,\eta}$ in analogy to the inf-sup stability of b. This results in:

$$\begin{aligned} & \|(\tilde{q}, (M\dot{\tilde{q}}(\frac{i}{\eta}T))_{i=0}^{\eta}) - (q_d, (\hat{p}_{d,i})_{i=0}^{\eta})\| \\ & \leq \inf_{x_d \in X_d} \left(C_2 \|(\tilde{q}, (M\dot{\tilde{q}}(\frac{i}{\eta}2T))_{i=0}^{\eta}) - x_d\| + C_3 \sup_{\delta q \in Y_d} \frac{|b(x_d, \delta q) - b_\eta(x_d, \delta q)|}{\|\delta q\|_{H^1(\mathcal{K}_\eta)}} \right) \\ & = \inf_{x_d \in X_d} \left(C_2 \|(\tilde{q}, (M\dot{\tilde{q}}(\frac{i}{\eta}2T))_{i=0}^{\eta}) - x_d\| + C_3 \sup_{\delta q \in \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m} \frac{|b_0(x_d, \delta q) - b_{0,\eta}(x_d, \delta q)|}{\|\delta q\|_{H^1}} \right) \\ & = \inf_{s \in \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m} \left(C_2 \|\tilde{q} - s\|_{H^1} + C_3 \sup_{\delta q \in \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m} \frac{|b_0(s, \delta q) - b_{0,\eta}(s, \delta q)|}{\|\delta q\|_{H^1}} \right) \end{aligned}$$

Here, V_d could be replaced by $\mathcal{M}_{j,0,0}$ since the former can be generated from the latter by adding element-wise constant functions. Those functions are however integrated precisely when multiplied with any of the polynomial test functions.

We now write

$$s_\eta := \operatorname*{argmin}_{s \in \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m} \|\tilde{q} - s\|_{H^1}.$$

Together with the fact that the mass term of b_0 is integrated precisely, partial integration of

[5] See Section A.4.

$M\dot{q}$, and a theorem of Duruflé, Grob and Joly [5, Theorem 2.6], we obtain

$$\begin{aligned}
&\|(\tilde{q}, (M\dot{\tilde{q}}(\frac{i}{\eta}T))_{i=0}^{\eta}) - (q_d, (\hat{p}_{d,i})_{i=0}^{\eta})\| \\
&\leq C_2\|\tilde{q} - s_\eta\|_{H^1} + C_3 \sup_{\delta q \in \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m} \frac{|b_0(s_\eta, \delta q) - b_{0,\eta}(s_\eta, \delta q)|}{\|\delta q\|_{H^1}} \\
&\leq C_2\|\tilde{q} - s_\eta\|_{H^1} + C_4 \left(\frac{T}{\eta}\right)^{j+1} \sup_{\delta q \in \mathcal{M}_{j,0,0}(\mathcal{K}_\eta)^m} \frac{\left\|s_\eta^{(j)}\right\|_{L^2} \left\|\frac{d\delta q}{dt}\right\|_{H^1}}{\|\delta q\|_{H^1}} \\
&\leq C_2\|\tilde{q} - s_\eta\|_{H^1} + C_4 \left(\frac{T}{\eta}\right)^{j+1} \left\|s_\eta^{(j)}\right\|_{L^2}
\end{aligned} \tag{2.8}$$

2.2.2. Polynomial approximation in the trial space

The inequality (2.8) can be simplified further.

Since $\tilde{q}$ is the solution of a Dirichlet boundary value problem, it lies in C^∞. [6] Because of this regularity, Theorem 6.4 of [2, Section 6] [7] implies the existence of a $C_5 > 0$ independent of η such that

$$\|\tilde{q} - s_\eta\|_{H^1} \leq C_5 \left(\frac{T}{\eta}\right)^j,$$

and the fact that s_η is bounded in the H^j-norm. Thus, a C_6 exists such that (2.8) can be reformulated to

$$\|(\tilde{q}, (M\dot{\tilde{q}}(\frac{i}{\eta}T))_{i=0}^{\eta}) - (q_d, (\hat{p}_{d,i})_{i=0}^{\eta})\| \leq C_6 \left(\frac{T}{\eta}\right)^j =: C_6 h^j. \tag{2.9}$$

Section on superconvergence can be found here in code.

2.3. Inhomogeneous problems

So far, it was assumed that $q(0) = q(T) = 0$, or in other words: $q \in H_0^1$. Let us now take a look at the general case. Here, we define q^{hom} and q^{inhom} such that

$$q^{\text{inhom}} = \frac{T-t}{T} q_0 + \frac{t}{T} q_T$$
$$q = q^{\text{hom}} + q^{\text{inhom}}.$$

(Recall that the boundary conditions on q were given as $q(0) = q_0$ and $q(T) = q_T$.) Then q^{hom} lies in $H_0^1([0,T])^m$. Now, in the solution theory from above, one can replace q by q^{hom} and the old ℓ by

$$\ell(\delta q) = -b_0(q^{\text{inhom}}, \delta q) + \int_0^T -F(t)\delta q \; dt + \hat{p} \cdot \delta q\Big|_0^T.$$

Using this approach, all previously constructed error estimators are still valid.[8]

[6] See, for example, Theorem 6 on page 343 of [7].

[7] That theorem extends to subsets of $\mathbb{R}$ as functions on one-dimensional sets can be extended such that they are constant in the added component of the domain. By construction, $\mathcal{M}_{j,0}$ can be replaced by $\mathcal{M}_{j,0,0}$.

[8] The case $j = 1$ requires extra analysis on the first term of the Strang lemma estimate in (2.8). This analysis can be performed in analogy to before, using a theorem of Duruflé et al. [5, Theorem 2.6].

2.4. Towards an initial value problem

So far, the error of a discrete solution has been analyzed. However, that discrete solution was constructed for a boundary value problem. The aim is now to trade the correctness of the position in $t = T$ for the correctness of the momentum in $t = 0$. That way, a different discrete solution shall be constructed, namely the result of the time stepper. Of course, this shall go down without greatly reducing the error. To that end, the stability of the time stepper's numerical flow with respect to the initial values will be used.

2.4.1. The numerical flow

We start by introducing a representation of the numerical flow of the time stepper at hand (Equations (1.8)). It is easy to see that - under the assumptions of this chapter - this time stepper is characterized by the following fact: It generates a solution to the discrete FEM equations from Section 2.2.

Proposition 2.4.1

Since $\nabla V(q) - F(t)$ is affine in q, the numerical flow of the time stepper at hand is also an affine map: There exists a linear function P and a vector F_{flow} such that

$$\left(q_d\left(\frac{i+1}{\eta}T\right), \hat{p}^*_{d,i+1}\right) = P\left(q_d\left(\frac{i}{\eta}T\right), \hat{p}^*_{d,i}\right) + F_{\text{flow}}(t_k),$$

for $i = 0, \dots, \eta - 1$. F_{flow} depends on the starting time of the time step.

Proof. Without loss of generality, there is only a single time step on $[0, T]$.

The system of equations which defines the time stepper (see (1.8)) can be seen as a linear equation where the inhomogeneity is an affine map of the initial values: Linear functions A, B and C exist such that (1.8) can be reformulated to

$$A\begin{pmatrix} (q_d(\xi_i))_{i=2}^{j+1} \\ \hat{p}(T) \end{pmatrix} = B\begin{pmatrix} q_d(0) \\ \hat{p}(0) \end{pmatrix} + CF.$$

It can be seen from (1.8) that A is bijective. If π_T is the projection that fulfills

$$\pi_T : \begin{pmatrix} (q_d(\xi_i))_{i=2}^{j+1} \\ \hat{p}(T) \end{pmatrix} \mapsto \begin{pmatrix} q_d(T) \\ \hat{p}(T) \end{pmatrix},$$

we can now write

$$\begin{pmatrix} q_d(T) \\ \hat{p}(T) \end{pmatrix} = \pi_T A^{-1} B \begin{pmatrix} q_d(0) \\ \hat{p}(0) \end{pmatrix} + \pi_T A^{-1} CF.$$

This leads to the definitions

$$\begin{aligned} P &:= \pi_T A^{-1} B \qquad \text{and} \\ F_{\text{flow}} &:= \pi_T A^{-1} CF. \end{aligned}$$

□

2.4.2. Stability of the numerical flow

Using the representation of the numerical flow from the last subsection, we can write

$$(q_d(T), \hat{p}^*_{d,\eta}) = P^\eta \left(q_d(0), \hat{p}^*_{d,0}\right) + \sum_{i=0}^{\eta-1} P^i F_{\text{flow}}(iT/\eta). \tag{2.10}$$

We want to compare this to the solution with the correct initial values $(q(0), \hat{p}(0))$. Let us call that solution

$$\left(q_{dd}, \left(\hat{p}^*_{dd,i}\right)_{i=0}^{\eta}\right).$$

By definition, $q_{dd}(0) = q(0)$ and $\hat{p}^*_{dd,0} = \hat{p}(0)$ hold. Therefore, this second numerical solution is the actual result of the time stepper. The representation of the numerical flow as an affine map can also be applied here, which results in:

$$(q_{dd}(T), \hat{p}^*_{dd,\eta}) = P^\eta \left(q(0), \hat{p}(0)\right) + \sum_{i=0}^{\eta-1} P^i F_{\text{flow}}(iT/\eta) \tag{2.11}$$

Subtraction of Equation (2.10) from Equation (2.11) then yields

$$(q_{dd}(T) - q(T), \hat{p}^*_{dd,\eta} - \hat{p}^*_{d,\eta}) = P^\eta \left(0, \hat{p}(0) - \hat{p}^*_{d,0}\right). \tag{2.12}$$

We now define the norm

$$\begin{aligned} \|\cdot\|_E : \mathbb{R}^m \times \mathbb{R}^m &\to \mathbb{R} \\ (a, b) &\mapsto \sqrt{\frac{1}{2}(Ka)\cdot a + \frac{1}{2}(M^{-1}b)\cdot b}. \end{aligned}$$

This is a norm since it is a Cartesian product norm. By Proposition 3.3.2, $\|(q,p)\|_E$ represents the square root of the energy in (q,p) if $F = 0$.

Using $\|\cdot\|_E$ in Equation (2.12), we obtain a constant C_E such that

$$\|(q_{dd}(T) - q(T), \hat{p}^*_{dd,\eta} - \hat{p}^*_{d,\eta})\|_2 \leq C_E \|P^\eta \left(0, \hat{p}(0) - \hat{p}^*_{d,0}\right)\|_E.$$

Now note that $(q_{dd} - q_d, (\hat{p}^*_{dd,i})_{i=0}^{\eta} - (\hat{p}^*_{d,i})_{i=0}^{\eta})$ is also a numerical solution, but for a problem where $F = 0$. Therefore, $\|\cdot\|_E$ represents the square root of the energy. However, P conserves the energy (and thus also its square root) as a result of Lemma 3.3.4. Consequently,

$$\|P\left(0, \hat{p}(0) - \hat{p}^*_{d,0}\right)\|_E = \|\left(0, \hat{p}(0) - \hat{p}^*_{d,0}\right)\|_E$$

holds. This in turn implies that

$$\|(q_{dd}(T) - q(T), \hat{p}^*_{dd,\eta} - \hat{p}^*_{d,\eta})\|_2 \leq C_E \|\left(0, \hat{p}(0) - \hat{p}^*_{d,0}\right)\|_E.$$

A consequence of the error bound (2.9) is that $\hat{p}(0) - \hat{p}^*_{d,0}$ has convergence order $\mathcal{O}(h^j)$. Therefore, the last inequality can be turned into

$$\|(q_{dd}(T) - q(T), \hat{p}^*_{dd,\eta} - \hat{p}^*_{d,\eta})\|_2 \leq C_E\ \mathcal{O}(h^j) = \mathcal{O}(h^j). \tag{2.13}$$

It is a direct result that the position q has a global error convergence rate of $\mathcal{O}(h^j)$.

The error bound (2.9) also implies that $\hat{p}(T) - \hat{p}^*_{d,\eta}$ has convergence rate $\mathcal{O}(h^j)$. Together with Equation (2.13), this yields

$$\|\hat{p}(T) - \hat{p}^*_{dd,\eta}\| \leq \|\hat{p}(T) - \hat{p}^*_{d,\eta}\| + \|\hat{p}^*_{d,\eta} - \hat{p}^*_{dd,\eta}\| = \mathcal{O}(h^j) + \mathcal{O}(h^j) = \mathcal{O}(h^j),$$

and we have therefore established that the global error convergence rate in the momentum is also $\mathcal{O}(h^j)$.

2.5. Summary

In conclusion, we have proven that for certain linear ODEs and the choice of $v = \dot{q}$, our time stepper has convergence order at least j in the position and the momentum. It should be noted that since the solver is symmetric (see Section 3.1), the convergence order is always even [10, Chapter II, Theorem 3.2]: If j is odd, the order can be upgraded to $j + 1$.

3. Properties of the time stepper

The introduced time steppers are symmetric. At least in simple cases, they are also symplectic and conserve energy.

3.1. Symmetry

Definition 3.1.1

Let Φ_h be the numerical flow of a one-step method. (The function that maps the numerical solution at time t to the numerical solution at $t+h$.) Then the underlying one-step method is called symmetric if and only if

$$\Phi_h^{-1} = \Phi_{-h}$$

[10, page 144f, Definition 1.4]

Lemma 3.1.2

The solver defined by Equation (1.8) is symmetric if the approximation $v \approx \dot{q}$ is symmetric in time.

Proof. $\lambda(t_k)$ and $\mu(t_{k+1})$ can swap their roles, they are mere auxiliary variables whose values are not necessary for further use.

Also, Equation (1.8) is always assumed in the context of valid input: $q(t_k)$ and $\hat{p}(t_k)$ are assumed to fulfill

$$\begin{aligned} g(q(t_k)) &= 0 \\ G(q(t_k))M^{-1}\hat{p}(t_k) &= 0. \end{aligned}$$

These equations can therefore be appended to Equation (1.8) without changing the time stepping method.

The result is a system of equations that is symmetric in time. Therefore, the resulting time stepper is symmetric. □

3.2. Symplecticity

For some linear ODEs and $v = \dot{q}$, the introduced time steppers are nearly symplectic.

Definition 3.2.1

Let

$$J := \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \in \mathbb{R}^{2m \times 2m},$$

where I denotes the identity matrix of $\mathbb{R}^{m \times m}$. Also, let $A \subseteq \mathbb{R}^{2m}$ be an open set. A differentiable function $f : A \to \mathbb{R}^{2m}$ is called symplectic (with respect to J) if and only if its Jacobian matrix f' fulfills

$$\forall x \in A : \; f'(x)^T J f'(x) = J.$$

[10, page 183, Definition 2.2]

Definition 3.2.2

A one-step method is called symplectic if and only if its numerical flow Φ_h is symplectic.

Lemma 3.2.3

Assume that $-\nabla V(q)$ is a linear function described by a symmetric matrix K, that $F = 0$, $\mathcal{Q} = \mathbb{R}^m$, $v = \dot{q}$, and that there are no constraints. ($g(q) = 0$ for all $q \in \mathcal{Q}$.) Also, assume that the map $q(0) \mapsto q$ has an H^{j+1}-norm that is bounded in $t_{k+1} - t_k$. Then the solver defined by Equation (1.8) is nearly symplectic, with an error of $\mathcal{O}((t_{k+1} - t_k)^{2 \deg q})$.

Proof. Take two numerical solutions from a single time step of length h, obtained by solving Equation (1.8). Let $q^1(t), \hat{p}^1$ and $q^2(t), \hat{p}^2$ denote these solutions and $\delta q : [t_k, t_{k+1}] \to \mathbb{R}^m$ be a polynomial of degree $j = \deg q$. Without loss of generality, the time interval (denoted by $[t_k, t_{k+1}]$) can be assumed to be the same in both cases.

Under these circumstances, reversing the application of the $j+1$ point Gauss-Lobatto rule in Equation (1.6) leads to

$$\int_{t_k}^{t_{k+1}} \left[K q^1 - M \ddot{q}^1\right] \cdot \delta q \; dt + (M \dot{q}^1 - \hat{p}^1) \cdot \delta q \Big|_{t_k}^{t_{k+1}} = \mathcal{O}((t_{k+1} - t_k)^{2j}).$$

Then the choice of $\delta q = q^2$ and partial integration of $\delta q \cdot (M \ddot{q}^1)$ yield

$$\int_{t_k}^{t_{k+1}} q^2 \cdot (K q^1) + \dot{q}^2 \cdot (M \dot{q}^1) \; dt - \hat{p}^1 \cdot q^2 \Big|_{t_k}^{t_{k+1}} = \mathcal{O}((t_{k+1} - t_k)^{2j}). \tag{3.1}$$

The same can be done with the other solution defined at the beginning of this proof and $\delta q = q^1$:

$$\int_{t_k}^{t_{k+1}} q^1 \cdot (K q^2) + \dot{q}^1 \cdot (M \dot{q}^2) \; dt - \hat{p}^2 \cdot q^1 \Big|_{t_k}^{t_{k+1}} = \mathcal{O}((t_{k+1} - t_k)^{2j})$$

Since both K and M are symmetric, the last equation can be reformulated to

$$\int_{t_k}^{t_{k+1}} q^2 \cdot (K q^1) + \dot{q}^2 \cdot (M \dot{q}^1) \; dt - \hat{p}^2 \cdot q^1 \Big|_{t_k}^{t_{k+1}} = \mathcal{O}((t_{k+1} - t_k)^{2j}).$$

This can in turn be subtracted from Equation (3.1) to obtain

$$-\hat{p}^1 \cdot q^2 \Big|_{t_k}^{t_{k+1}} + \hat{p}^2 \cdot q^1 \Big|_{t_k}^{t_{k+1}} = \mathcal{O}((t_{k+1} - t_k)^{2j}),$$

which is equivalent to

$$\hat{p}^2(t_{k+1}) \cdot q^1(t_{k+1}) - \hat{p}^1(t_{k+1}) \cdot q^2(t_{k+1}) = \hat{p}^2(t_k) \cdot q^1(t_k) - \hat{p}^1(t_k) \cdot q^2(t_k) + \mathcal{O}((t_{k+1} - t_k)^{2j}). \tag{3.2}$$

As in Subsection 2.4.1, the numerical flow can be written as

$$(q_d(t_{k+1}), \hat{p}(t_{k+1})) = P(q_d(t_k), \hat{p}(t_k)) + F_{\text{flow}}(t_k).$$

Since we assume that $F = 0$, F_{flow} is (by construction) also zero. Therefore, Equation (3.2) can be rewritten to

$$\begin{aligned} P(q^2(t_k), \hat{p}^2(t_k))_2 \cdot P(q^1(t_k), \hat{p}^1(t_k))_1 - P(q^1(t_k), \hat{p}^1(t_k))_2 \cdot P(q^2(t_k), \hat{p}^2(t_k))_1 \\ = \hat{p}^2(t_k) \cdot q^1(t_k) - \hat{p}^1(t_k) \cdot q^2(t_k) + \mathcal{O}((t_{k+1} - t_k)^{2j}), \end{aligned}$$

which is in turn equivalent to

$$(q^1(t_k), \hat{p}^1(t_k))^T \; P^T J P \; (q^2(t_k), \hat{p}^2(t_k)) = (q^1(t_k), \hat{p}^1(t_k))^T \; J \; (q^2(t_k), \hat{p}^2(t_k)) + \mathcal{O}((t_{k+1} - t_k)^{2j}).$$

Since $(q^1(t_k), \hat{p}^1(t_k))$ and $(q^2(t_k), \hat{p}^2(t_k))$ were left arbitrary, the above results in

$$P^T J P = J + \mathcal{O}((t_{k+1} - t_k)^{2j})$$

which - by definition - would imply symplecticity if there was no error term. The solver is therefore nearly symplectic, with an error of $\mathcal{O}((t_{k+1} - t_k)^{2j})$.

□

3.3. Energy conservation

For $v = \dot{q}$ and $F = 0$, the introduced time steppers nearly conserve energy. For linear ODEs, they conserve the energy precisely.

Definition 3.3.1
The total energy function is given by

$$E(q, p, v) = p \cdot v - L(q, v). \tag{3.3}$$

[16, Equation 4]

Proposition 3.3.2
The total energy can be written as

$$E(q, v) = \frac{1}{2} v \cdot M v + V(q).$$

Proof. We use the result that $p = Mv$ (from Equation (1.4)).

$$E = p \cdot v - L(q, v) = (Mv) \cdot v - \frac{1}{2} v \cdot Mv + V(q) = \frac{1}{2} v \cdot Mv + V(q)$$

□

Lemma 3.3.3

Choose $v = \dot{q}$, and assume that $(-\nabla V(q_d) - M\ddot{q}_d + \theta \cdot G(q)) \cdot \dot{q}_d$ (for the discrete solution q_d) is bounded for varying time step size. Then the time stepper defined by Equation (1.8) nearly conserves the total energy, with a local error of order $2 \deg q$.

Proof. Let $(q_d, \hat{p}(t_{k+1}))$ be the approximate solutions from a single time step $[t_k, t_{k+1}]$. Also, denote by $\theta : [t_k, t_{k+1}] \to \mathbb{R}^n$ the polynomial of degree $j = \deg q$ such that

$$\begin{aligned} \theta(t_{k+1}) &= \mu(t_{k+1}) \\ \forall i \in \{2, \ldots, j\} : \ \theta(\xi_j) &= \lambda(\xi_j) \\ \theta(t_k) &= \lambda(t_k). \end{aligned}$$

That is, θ is the approximation to λ that is used in the time step. Thus, reinsertion into the variational principle from Equation (1.6) results in

$$\int_{t_k}^{t_{k+1}} (-\nabla V(q_d) - M\ddot{q}_d + \theta \cdot G(q)) \cdot \delta q \; dt + (M\dot{q} - \hat{p}) \cdot \delta q \Big|_{t_k}^{t_{k+1}} = \mathcal{O}([t_{k+1} - t_k]^{2j}),$$

since the $j + 1 = \deg q_d + 1$ point Gauss-Lobatto scheme has convergence order $2j$. Note that this convergence order requires the $2j$th derivative of $(-\nabla V(q_d) - M\ddot{q}_d + \theta \cdot G(q)) \cdot \delta q$ to be bounded. By assumption, this is the case for the choice of $\delta q = \dot{q}_d$, which yields

$$\begin{aligned} \mathcal{O}([t_{k+1} - t_k]^{2j}) &= \int_{t_k}^{t_{k+1}} (-\nabla V(q_d) - M\ddot{q}_d + \theta \cdot G(q_d)) \cdot \dot{q}_d \; dt + (M\dot{q}_d - \hat{p}) \cdot \dot{q}_d \Big|_{t_k}^{t_{k+1}} \\ &= \int_{t_k}^{t_{k+1}} (-\nabla V(q_d)) \cdot \dot{q}_d + (M\dot{q}_d) \cdot \ddot{q}_d + (\theta \cdot G(q_d)) \cdot \dot{q}_d \; dt - \hat{p} \cdot \dot{q}_d \Big|_{t_k}^{t_{k+1}} \\ &= \int_{t_k}^{t_{k+1}} -\frac{dV(q_d)}{dt} + \frac{1}{2}\frac{d}{dt}[\dot{q}_d \cdot (M\dot{q}_d)] + \frac{\theta \cdot g(q_d)}{dt} - \dot{\theta} \cdot g(q_d) \; dt - \hat{p} \cdot \dot{q}_d \Big|_{t_k}^{t_{k+1}} \\ &= [-\hat{p} \cdot \dot{q}_d + T(\dot{q}_d) - V(q_d) + \theta \cdot g(q_d)] \Big|_{t_k}^{t_{k+1}} - \int_{t_k}^{t_{k+1}} \dot{\theta} \cdot g(q_d) \; dt \\ &= E(q_d(t_k), \hat{p}(t_k), \dot{q}_d(t_k)) - E(q_d(t_{k+1}), \hat{p}(t_{k+1}), \dot{q}_d(t_{k+1})) \\ &\quad + [\theta \cdot g(q_d)] \Big|_{t_k}^{t_{k+1}} - \int_{t_k}^{t_{k+1}} \dot{\theta} \cdot g(q_d) \; dt. \end{aligned}$$

Since $g(q_d(t_k)) = g(q_d(t_{k+1})) = 0$ and

$$\int_{t_k}^{t_{k+1}} \dot{\theta} \cdot g(q_d) \; dt = \mathcal{O}([t_{k+1} - t_k]^{2j}) + \sum_{i=1}^{j+1} \left[\alpha_i \dot{\theta}(\xi_i) \cdot g(q_d(\xi_i))\right] = \mathcal{O}([t_{k+1} - t_k]^{2j}),$$

this results in

$$E(q_d(t_k), \hat{p}(t_k), \dot{q}_d(t_k)) - E(q_d(t_{k+1}), \hat{p}(t_{k+1}), \dot{q}_d(t_{k+1})) = \mathcal{O}([t_{k+1} - t_k]^{2j})$$

□

Lemma 3.3.4

Assume that the problem at hand is linear, in the sense that: $\mathcal{Q} = \mathbb{R}^m$, $-\nabla V(q) = F + Kq$ for a vector F and matrix K, $g = 0$, and $v = \dot{q}$.
Then the time stepper defined by Equation (1.8) conserves the total energy.

Proof. The imprecision in the last lemma stems from the error of the Gauss-Lobatto quadrature. However, if the integrand is a polynomial of degree at most $2j - 1$, its $2j$th derivative is zero. In that case, the quadrature error representation(s) from Section A.5 can be used to see that the error disappears. This also eliminates the need for the boundedness assumption from the last lemma.

The assumptions made in this lemma are sufficient for all integrands in the proof of the last lemma to be polynomials of degree at most $2j - 1$. If all other aspects of that proof remain the same, this has for result that the solver(s) conserve energy without an error. □

4. Rigid-body dynamics

The time stepper(s) that were constructed so far can be applied to systems of rigid bodies.

4.1. Kinematics: Position and velocity

How can one describe the motion of a rigid body? One way is to split the motion into the translation of the body and the body's rotation around its center of mass.

4.1.1. Position

The set of possible translations of a body can be represented by $\mathbb{R}^3$.

In contrast, the rotational positions around the center of mass may be represented by

$$SO(3) \overset{\text{def}}{=} \{R \in \mathbb{R}^{3\times 3} \ \mid \ R^T R = RR^T = I,\ \det R = 1\}.$$

Thus,

$$\mathbb{R}^3 \times SO(3)$$

can be used to represent the set of all possible (generalized) positions.[1] This set is also a manifold. In the notation from the beginning of Chapter 1, we can therefore choose $\mathcal{Q} = \mathbb{R}^3 \times SO(3)$, or even $\mathcal{Q} = (\mathbb{R}^3 \times SO(3))^d$ for d different bodies.

4.2. Tangent space, momentum and velocity

For the test functions, the momentum, and the velocity, the tangent spaces of the configuration space $\mathcal{Q} = \mathbb{R}^3 \times SO(3)$ will be needed. The tangent space of $SO(3)$ at $R \in SO(3)$ is

$$\{R\widetilde{\Omega} \ \mid \ \widetilde{\Omega} \in \mathbb{R}^{3\times 3} \text{ skew-symmetric}\}$$

By a proposition of Lee [18, Proposition 3.14], the tangent space of a Cartesian product of manifolds is isomorphic to the direct sum of the tangent spaces of the "factors" appearing in

[1] This set is a Lie group with group operation $(x_1, R_1) \circ (x_2, R_2) = (x_1 + x_2, R_1 R_2)$. Another possible group operation on the same set is $(x_1, R_1) \circ (x_2, R_2) = (x_1 + R_1 x_2, R_1 R_2)$. In the latter case, the resulting group is also a Lie group, namely the special Euclidean group in three dimensions, $SE(3)$.

the Cartesian product. For a point $q = (x, R) \in \mathbb{R}^3 \times SO(3)$, the tangent space of $\mathbb{R}^3 \times SO(3)$ is therefore

$$\mathbb{R}^3 \times \{R\widetilde{\Omega} \mid \widetilde{\Omega} \in \mathbb{R}^{3\times 3} \text{ skew-symmetric}\}.$$

The space of all valid velocities is precisely the tangent space [3, page 5]. This can be used to store velocities in computer memory in a convenient way: Using the definition

$$\begin{aligned}\widetilde{\cdot}:\ \mathbb{R}^3 &\to \left\{\widetilde{\Omega} \in \mathbb{R}^{3\times 3} \text{ skew-symmetric}\right\}\\ \Omega &\mapsto \begin{pmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{pmatrix},\end{aligned}$$

the velocity at a point $q = (x, R)$ becomes

$$v = (\dot{x}, R\widetilde{\Omega})$$

for some $\dot{x}, \Omega \in \mathbb{R}^3$. Therefore, the velocity can be stored as $(\dot{x}, \Omega) \in \mathbb{R}^6$.

The set $\mathbb{R}^3 \times \{\widetilde{\Omega} \in \mathbb{R}^{3\times 3}$ skew-symmetric$\}$ is the Lie algebra[2] of $\mathbb{R}^3 \times SO(3)$.

The information presented so far in this chapter (and more) can be found in [3] and [19].

The momentum lies in the dual space of the tangent space. Therefore, the Lax-Milgram lemma implies that the momentum can be identified with an element of the tangent space. As a result, the momentum can be stored in the same way as the velocity. All that is required for this is the choice of a scalar product $\langle\cdot,\cdot\rangle_{\mathbb{R}^{\mathfrak{M}}}$ in order to apply the Lax-Milgram lemma. It is convenient to choose a scalar product that is consistent with the storage format of $\mathbb{R}^3 \times SO(3)$ in computer memory.

4.3. Kinetic energy

Take a body that is represented by the set[3] $B \subseteq \mathbb{R}^3$ and the density function $\rho:\ B \to \mathbb{R}^+$.

We start by analyzing the kinetic energy in $SO(3)$. Under the definition

$$(a, b)_{\mathbb{I}} := \int_B \rho(x)(a \times x) \cdot (b \times x)\ dx$$

for any two vectors $a, b \in \mathbb{R}^3$, the kinetic energy of the body can be written as

$$T(\Omega(t)) = \frac{1}{2}(\Omega(t), \Omega(t))_{\mathbb{I}},$$

where $R(t)\widetilde{\Omega}(t)$ is the body's velocity in the tangent space of $SO(3)$ in $R(t) \in SO(3)$. Let $\mathbb{I}$ be the Gram matrix of $(\cdot,\cdot)_{\mathbb{I}}$. This matrix is called the moment of inertia tensor.
[12, Section 2.1.2]

[2] The Lie algebra of a Lie group is defined as the Lie group's tangent space at the identity of the group operation.

[3] In a sense, one might identify the body and the set.

Furthermore, using the definitions of the mass

$$m := \int_B \rho(x)\ dx \qquad \text{(see [12, Section 2.1.2])}$$

and the mass matrix

$$M_{6\times 6} := \begin{pmatrix} mI_{3\times 3} & \\ & \mathbb{I} \end{pmatrix},$$

the kinetic energy in $\mathbb{R}^3 \times SO(3)$ can be written as

$$T((\dot{x}(t), \Omega(t))) = \frac{1}{2}(\dot{x}(t), \Omega(t)) M_{6\times 6} (\dot{x}(t), \Omega(t))^T.$$

[3]
Therefore, $M_{6\times 6}$ is the equivalent to the operator M from Chapter 1 in the sense that they perform the same linear transformation for different representations of the tangent space and its dual space.

4.4. Adapted time stepper equations

Some software, such as NGSolve[4], is capable of solving the time stepper equations (1.8) almost as they are, merely requiring a suitable representation of ∇V and G. However, some kind of simplification is always required. This is especially true if a simple implementation with only a root finding algorithm and basic linear algebra is desired.

Link to example

Each equation of (1.8) can be written as

$$\forall \delta q \in T_{(x,R)}\left(\mathbb{R}^3 \times SO(3)\right) : \ a \cdot \delta q = 0. \tag{4.1}$$

Here, (x, R) is an element of $\mathbb{R}^3 \times SO(3)$. Strictly speaking, a lies in $(\mathbb{R}^{\mathfrak{M}})^*$. However, it might be more practical to compute a as an element of $\mathbb{R}^{\mathfrak{M}}$. Such a representation exists by the Lax-Milgram lemma, for which we choose the same scalar product $\langle \cdot, \cdot \rangle_{\mathbb{R}^{\mathfrak{M}}}$ as in Section 4.2. Let us denote this representation by $a_{\mathbb{R}^{\mathfrak{M}}}$.

Furthermore, $a_{\mathbb{R}^{\mathfrak{M}}}$ can be decomposed into $a_{\mathbb{R}^{\mathfrak{M}}} = a_T + a_N$, where a_T is the $\langle \cdot, \cdot \rangle_{\mathbb{R}^{\mathfrak{M}}}$-orthogonal projection of $a_{\mathbb{R}^{\mathfrak{M}}}$ onto $T_{(x,R)}\left(\mathbb{R}^3 \times SO(3)\right)$, and a_N is orthogonal to $T_{(x,R)}\left(\mathbb{R}^3 \times SO(3)\right)$. Using this decomposition, Equation (4.1) can be reformulated to

$$\forall \delta q \in T_{(x,R)}\left(\mathbb{R}^3 \times SO(3)\right) : \ \langle a_T + a_N, \delta q \rangle_{\mathbb{R}^{\mathfrak{M}}} = 0,$$

which is in turn equivalent to

$$a_T = 0. \tag{4.2}$$

However, (4.2) is - in a sense - a system of redundant equations. After all, it is already known that a_T lies within the tangent space, which is a genuine subspace of $\mathbb{R}^{\mathfrak{M}}$. Therefore, some of

[4] See https://ngsolve.org/.

its coefficients are automatically equal to 0. A way to circumvent this redundancy is the use of the operator

$$\begin{aligned}\mathrm{alg}_{(x,R)} : \ T_{(x,R)}\left(\mathbb{R}^3 \times SO(3)\right) &\to \mathbb{R}^6 \\ (b, R\,\widetilde{\Omega}) &\mapsto (b, \Omega)\,,\end{aligned}$$

which maps a tangent space element to a corresponding element of the Lie algebra. Together with the operator $P_{(x,R)}$, which we define as the $\langle\cdot,\cdot\rangle_{\mathbb{R}^{\mathfrak{M}}}$-orthogonal projection onto $T_{(x,R)}\left(\mathbb{R}^3 \times SO(3)\right)$, we can write Equation (4.2) as

$$\mathrm{alg}_{(x,R)}\, P_{(x,R)} a = 0.$$

If we identify $\nabla V, M\dot{v}, Mv, \lambda^T G$ and $\hat{p}$ with their $\langle\cdot,\cdot\rangle_{\mathbb{R}^{\mathfrak{M}}}$-representations in $\mathbb{R}^{\mathfrak{M}}$, the system of equations (1.8) is equivalent to

$$\begin{aligned}\mathrm{alg}\, P_{q(t_k)}[\alpha_1[F(t_k, q(t_k), v(t_k)) - \nabla V(q(t_k)) - (M\dot{v}(t_k)) + \lambda(t_k)^T G(q(t_k))]& \\ -(Mv(t_k) - \hat{p}(t_k))] &= 0 \\ \forall i \in \{2,\dots,j\} : \ \mathrm{alg}\, P_{q(\xi_i)}[F(\xi_i, q(\xi_i), v(\xi_i)) - \nabla V(q(\xi_i))& \\ -(M\dot{v}(\xi_i)) + \lambda(\xi_i)^T G(q(\xi_i))] &= 0 \\ \mathrm{alg}\, P_{q(t_{k+1})}[\alpha_{j+1}[F(t_{k+1}, q(t_{k+1}), v(t_{k+1})) - \nabla V(q(t_{k+1})) - (M\dot{v}(t_{k+1}))& \\ +\mu(t_{k+1})^T G(q(t_{k+1}))] + (Mv(t_{k+1}) - \hat{p}(t_{k+1}))] &= 0 \\ \forall i \in \{2,\dots,j\} : \ g(q(\xi_i)) &= 0 \\ g(q(t_{k+1})) &= 0 \\ G(q(t_{k+1}))M^{-1}\hat{p}(t_{k+1}) &= 0 \\ v &\approx \dot{q}.\end{aligned} \tag{4.3}$$

This also implies that in Equation (4.3), M has to map to the tangent space, and not its dual space. Such a representation of M can be constructed using the matrix $M_{6\times 6}$

It should be noted that in practice, (4.3) has to be supplemented by equations that ensure

$$q(\xi_2), \dots, q(\xi_j), q(t_{k+1}) \in \mathbb{R}^3 \times SO(3).$$

To be precise, the 3×3 matrix components of $q(\xi_2), \dots, q(\xi_j), q(t_{k+1})$ have to be elements of $SO(3)$, which can be verified as in Section A.6.

F (as it is passed to the computer) does not have to lie in the tangent space: The part of it which is normal to the manifold will be removed by P in the equations.

4.4.1. Multiple bodies

Instead of a single body, a system of multiple, connected bodies can also be simulated. In that case, M becomes a block diagonal matrix of the mass matrices for the single blocks. Apart from that, $\nabla V, g$ and G depend on the positions of all bodies. Among other things, springs can be simulated via ∇V and rigid constraints between bodies can be implemented in g.

5. Numerical experiments

The time stepper(s) described by (4.3) have been tested numerically for $j = 1$ and $j = 2$. The testing consist in running a benchmark problem called the *Heavy Top without kinematic constraints*; the details of this problem can be found in [14, Section 4.1.1]. The implementation of the benchmark has been carried out in NGSolve[1], a FEM library for the programming language Python. The versions used are NGSolve 6.2.2501, Netgen 6.2.2501, and Python 3.13.7.

For the convergence plots, the calculation was performed at the time step sizes $0.004 \cdot 2^{-i}, i = 0, -1, \dots, -7$. The reference solution was calculated at a time step of $0.001 \cdot 2^{-9}$.

Measure for speed?

Update versions of NGSolve and Python

Link to source codes of benchmarks.

5.1. Degree $j = 1$

In this case, we choose $v = \dot{q}$. Note that, for this choice, all evaluations of v that appear in the solver's equations already lie in the tangent space of $\mathbb{R}^3 \times SO(3)$.

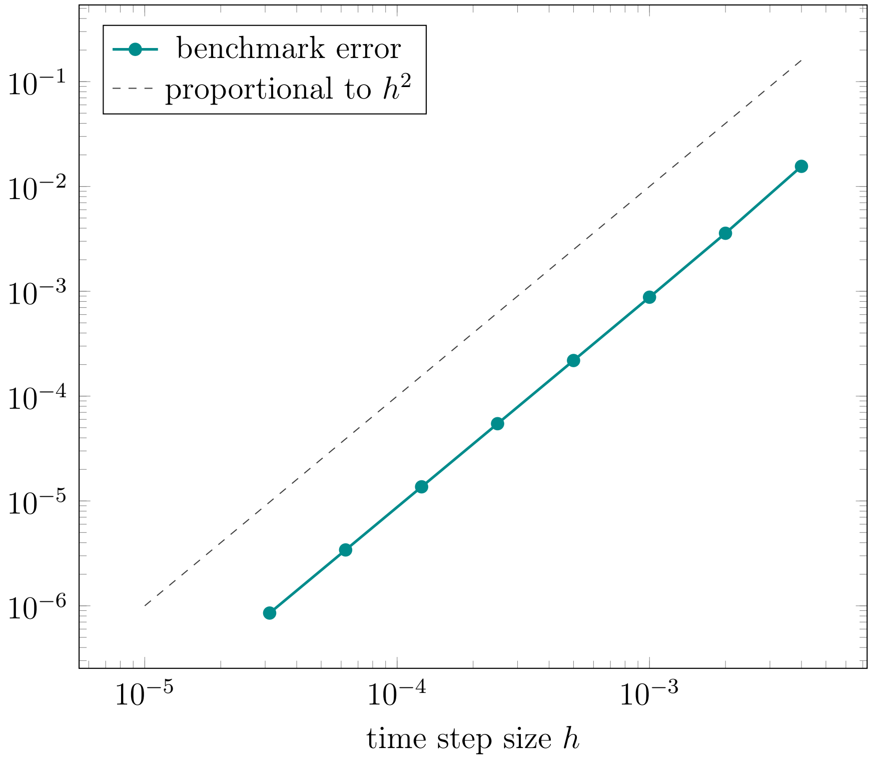
The system of equations (4.3) simplifies to

$$\begin{aligned}
\operatorname{alg} P_{q(t_k)} \Big[\frac{t_{k+1} - t_k}{2} [F(t_k, q(t_k), v(t_k)) - \nabla V(q(t_k)) \\
+ \lambda(t_k)^T G(q(t_k))] - (M\dot{q}(t_k) - \hat{p}(t_k)) \Big] &= 0 \\
\operatorname{alg} P_{q(t_{k+1})} \Big[\frac{t_{k+1} - t_k}{2} [F(t_{k+1}, q(t_{k+1}), v(t_{k+1})) - \nabla V(q(t_{k+1})) \\
+ \mu(t_{k+1})^T G(q(t_{k+1}))] + (M\dot{q}(t_{k+1}) - \hat{p}(t_{k+1})) \Big] &= 0 \\
g(q(t_{k+1})) &= 0 \\
G(q(t_{k+1})) M^{-1} \hat{p}(t_{k+1}) &= 0 \\
q(t_{k+1}) \in \mathbb{R}^3 \times SO(3).
\end{aligned} \tag{5.1}$$

[1] See https://ngsolve.org/.

The numerical experiment yields the following convergence plot of the global error:

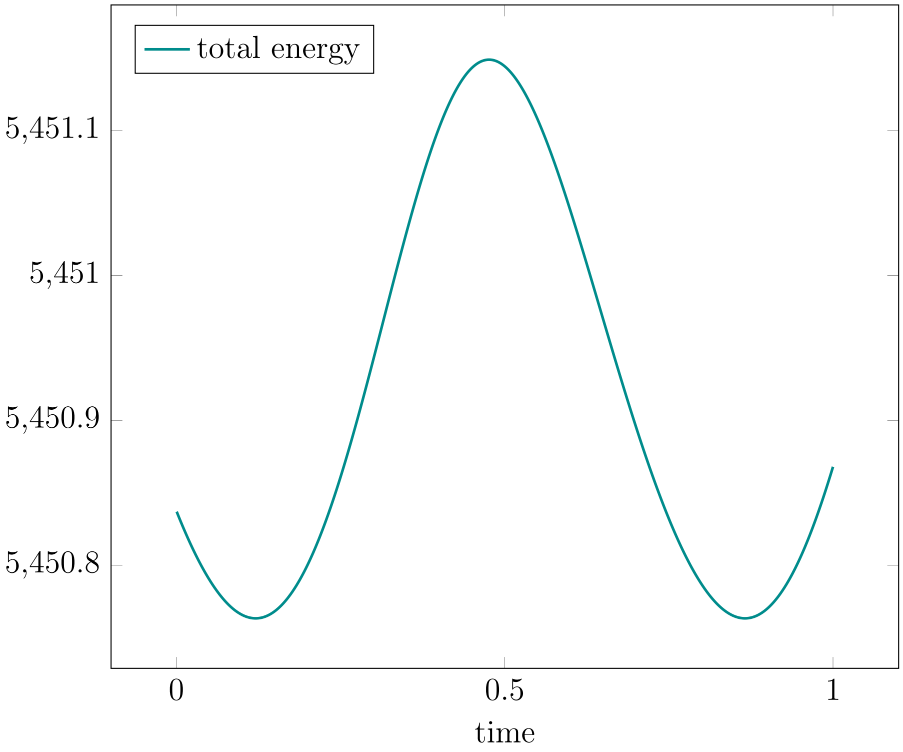
Figure 5.1.: Convergence plot for $j = 1$ and $v = \dot{q}$



This showcases quadratic convergence, which is twice the convergence one would expect given the results of Chapter 2.

Plotting the energy of the system for a fixed time step reveals only minor oscillations:

Figure 5.2.: Total energy for $j = 1$, $v = \dot{q}$, and time step size $5 \cdot 10^{-4}$



5.2. Degree $j = 2$

5.2.1. Version 1: $v = \dot{q}$

Again, we can choose $v = \dot{q}$.

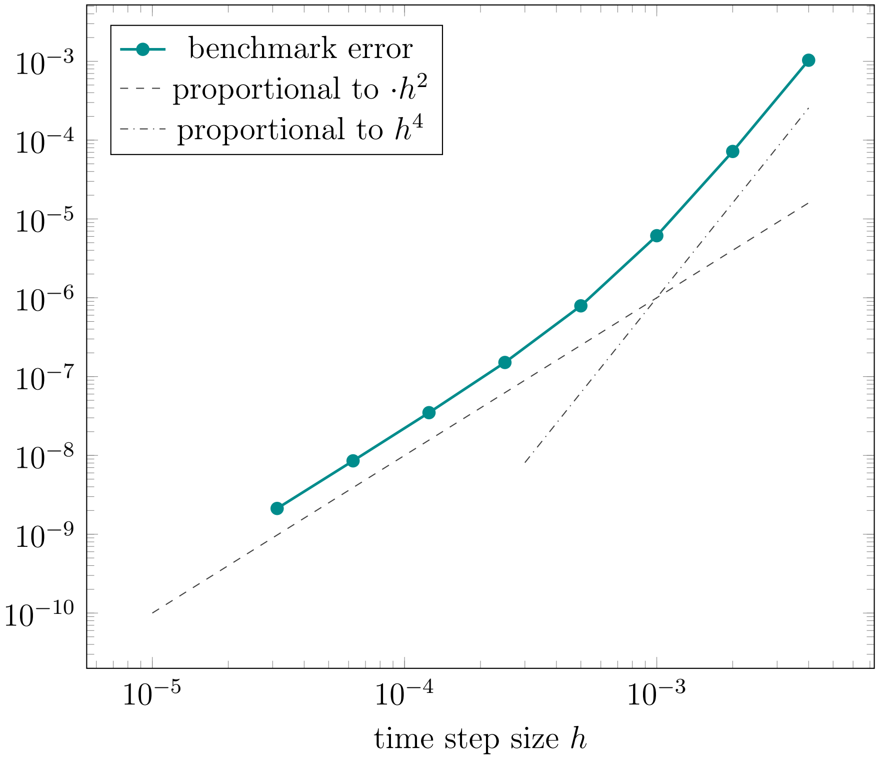
Now, (4.3) simplifies to

$$\begin{aligned}
\operatorname{alg} P_{q(t_k)} \Big[\frac{t_{k+1} - t_k}{6} [F(t_k, q(t_k), v(t_k)) - \nabla V(q(t_k)) - (M\ddot{q}(t_k)) \\
+ \lambda(t_k)^T G(q(t_k))] - (M\dot{q}(t_k) - \hat{p}(t_k)) \Big] &= 0 \\
\operatorname{alg} P_{q\left(t_{k+1/2}\right)} \Bigg[F(t_{k+1/2}, q(t_{k+1/2}), v(t_{k+1/2})) - \nabla V \left(q\left(t_{k+1/2}\right)\right) - \left(M\ddot{q}\left(t_{k+1/2}\right)\right) \\
+ \lambda\left(t_{k+1/2}\right)^T G\left(q\left(t_{k+1/2}\right)\right) \Bigg] &= 0 \\
\operatorname{alg} P_{q(t_{k+1})} \Big[\frac{t_{k+1} - t_k}{6} [F(t_{k+1}, q(t_{k+1}), v(t_{k+1})) - \nabla V(q(t_{k+1})) - (M\ddot{q}(t_{k+1})) \\
+ \mu(t_{k+1})^T G(q(t_{k+1}))] + (M\dot{q}(t_{k+1}) - \hat{p}(t_{k+1})) \Big] &= 0 \\
\\
g\left(q\left(t_{k+1/2}\right)\right) &= 0 \\
g(q(t_{k+1})) &= 0 \\
G(q(t_{k+1}))M^{-1}\hat{p}(t_{k+1}) &= 0 \\
\\
q(t_{k+1/2}) &\in \mathbb{R}^3 \times SO(3) \\
q(t_{k+1}) &\in \mathbb{R}^3 \times SO(3),
\end{aligned} \tag{5.2}$$

where $t_{k+1/2} = (t_{k+1} + t_k)/2$.

The global error convergence plot from the numerical experiment is:

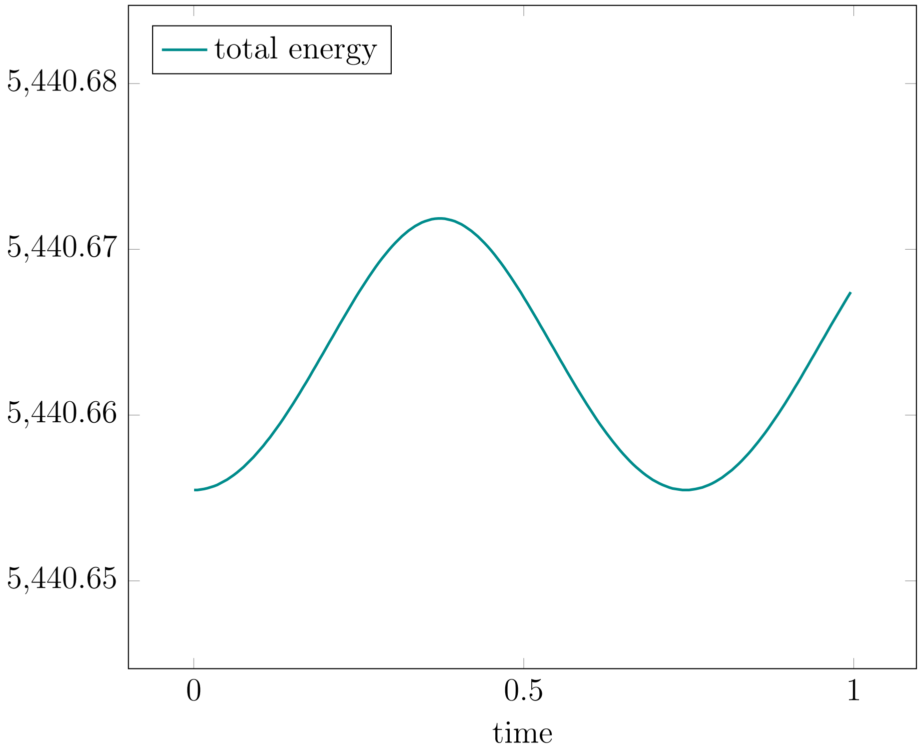
Figure 5.3.: Convergence plot for $j = 2$ and $v = \dot{q}$



Here, the convergence is of order 4 at first, but later reduces to order 2. However, adjustments can be made to improve performance, see Subsection 5.2.2.

The oscillations in energy are smaller than for the previous case of $j = 1$:

Figure 5.4.: Total energy for $j = 2$, $v = \dot{q}$, and time step size $5 \cdot 10^{-4}$



5.2.2. Version 2: Adjustments for fourth order accuracy

The time stepper from the last subsection can be improved. It is a concrete example of an approach where the velocity does not have to lie in the tangent space at the relevant points in time.

However, v can be modified ($v \neq \dot{q}$) in a way that partially mitigates the issue of not lying in the tangent space. Following this rationale, one finds that the resulting method shows better performance in benchmark testing.

The following modification can also be made to time steppers that use higher order polynomial approximation. It consists in approximating $\operatorname{alg}_q \dot{q}$ by a polynomial of degree one. (Thus far, $\dot{q}$ was approximated by a polynomial of degree one.) Let us denote the approximation of $\operatorname{alg}_q \dot{q}$ by v^a. To construct equations that lead to an approximation, the approach followed here is to start with the equation

$$\forall \delta p : \int_{t_k}^{t_{k+1}} (\dot{q} - v) \cdot \delta p \; dt = 0, \tag{5.3}$$

which has already appeared in the proof of Theorem 1.1.1. Depending on which δp are considered, (5.3) is of course equivalent to $\dot{q} = v$. In addition, (5.3) can be made more specific using the tangent space representation of v and δp. This results in the equivalent equation

$$\forall \operatorname{alg}_q^{-1} \delta p^a : \int_{t_k}^{t_{k+1}} (\dot{q} - \operatorname{alg}_q^{-1} v^a) \cdot \operatorname{alg}_q^{-1} \delta p^a \; dt = 0,$$

where v^a and δp^a run through the Lie algebra of $\mathbb{R}^3 \times SO(3)$. Now, a finite element approach can be applied: Assume that v^a and δp^a are polynomials of degree one and define $t_{k+1/2} \coloneqq (t_k + t_{k+1})/2$. Then the application of Simpson's quadrature rule (Gauss-Lobatto) leads to

$$\begin{aligned}
0 = {}& \frac{t_{k+1} - t_k}{6} (\dot{q}(t_k) - \operatorname{alg}_{q(t_k)}^{-1} v^a(t_k)) \cdot \operatorname{alg}_{q(t_k)}^{-1} \delta p^a(t_k) \\
&+ \frac{4(t_{k+1} - t_k)}{6} \left(\dot{q}(t_{k+1/2}) - \operatorname{alg}_{q(t_{k+1/2})}^{-1} \frac{v^a(t_k) + v^a(t_{k+1})}{2} \right) \cdot \operatorname{alg}_{q(t_{k+1/2})}^{-1} \frac{\delta p^a(t_k) + \delta p^a(t_{k+1})}{2} \\
&+ \frac{t_{k+1} - t_k}{6} (\dot{q}(t_{k+1}) - \operatorname{alg}_{q(t_{k+1})}^{-1} v^a(t_{k+1})) \cdot \operatorname{alg}_{q(t_{k+1})}^{-1} \delta p^a(t_k).
\end{aligned}$$

This can be simplified to

$$\begin{aligned}
0 = {}& (\dot{q}(t_k) - \operatorname{alg}_{q(t_k)}^{-1} v^a(t_k)) \cdot \operatorname{alg}_{q(t_k)}^{-1} \delta p_1 && (5.4) \\
&+ 2 \left(\dot{q}(t_{k+1/2}) - \operatorname{alg}_{q(t_{k+1/2})}^{-1} \frac{v^a(t_k) + v^a(t_{k+1})}{2} \right) \cdot \operatorname{alg}_{q(t_{k+1/2})}^{-1} (\delta p_1 + \delta p_2) && (5.5) \\
&+ (\dot{q}(t_{k+1}) - \operatorname{alg}_{q(t_{k+1})}^{-1} v^a(t_{k+1})) \cdot \operatorname{alg}_{q(t_{k+1})}^{-1} \delta p_2 && (5.6)
\end{aligned}$$

for all $\delta p_1, \delta p_2$ in the Lie algebra of $\mathbb{R}^3 \times SO(3)$. To obtain a system of equations for an implementation using a Newton solver, an option would be to insert a basis of the Lie algebra for δp_1 and δp_2.

Note that v^a is (still only) close to the tangent space at t_k, t_{k+1} and $t_{k+1/2}$.

Now that v is defined using $v \coloneqq \operatorname{alg}_q^{-1} v^a$, $\dot{v}$ can be calculated using the product rule:

$$\dot{v} = \frac{\partial \operatorname{alg}_q^{-1}}{\partial t} v^a + \operatorname{alg}_q^{-1} \frac{v^a(t_{k+1}) - v^a(t_k)}{t_{k+1} - t_k}.$$

The outcome is the time stepper

$$
\begin{aligned}
&\operatorname{alg} P_{q(t_k)}\Big[\frac{t_{k+1}-t_k}{6}[F(t_k, q(t_k), v(t_k)) - \nabla V(q(t_k)) \\
&\qquad -(M\dot{v}(t_k)) + \lambda(t_k)^T G(q(t_k))] - (Mv(t_k) - \hat{p}(t_k))\Big] = 0 \\
&\operatorname{alg} P_{q\left(t_{k+1/2}\right)}[F(t_{k+1/2}, q(t_{k+1/2}), v(t_{k+1/2})) - \nabla V\left(q\left(t_{k+1/2}\right)\right) \\
&\qquad - \left(M\dot{v}\left(t_{k+1/2}\right)\right) + \lambda\left(t_{k+1/2}\right)^T G\left(q\left(t_{k+1/2}\right)\right)] = 0 \\
&\operatorname{alg} P_{q(t_{k+1})}\Big[\frac{t_{k+1}-t_k}{6}[F(t_{k+1}, q(t_{k+1}), v(t_{k+1})) - \nabla V(q(t_{k+1})) \\
&\qquad -(M\dot{v}(t_{k+1})) + \mu(t_{k+1})^T G(q(t_{k+1}))] + (Mv(t_{k+1}) - \hat{p}(t_{k+1}))\Big] = 0
\end{aligned}
$$

$$
\begin{aligned}
g\left(q\left(t_{k+1/2}\right)\right) &= 0 \\
g(q(t_{k+1})) &= 0 \\
G(q(t_{k+1}))M^{-1}\hat{p}(t_{k+1}) &= 0
\end{aligned}
\tag{5.7}
$$

$$
\operatorname{alg}_q v = v^a
$$

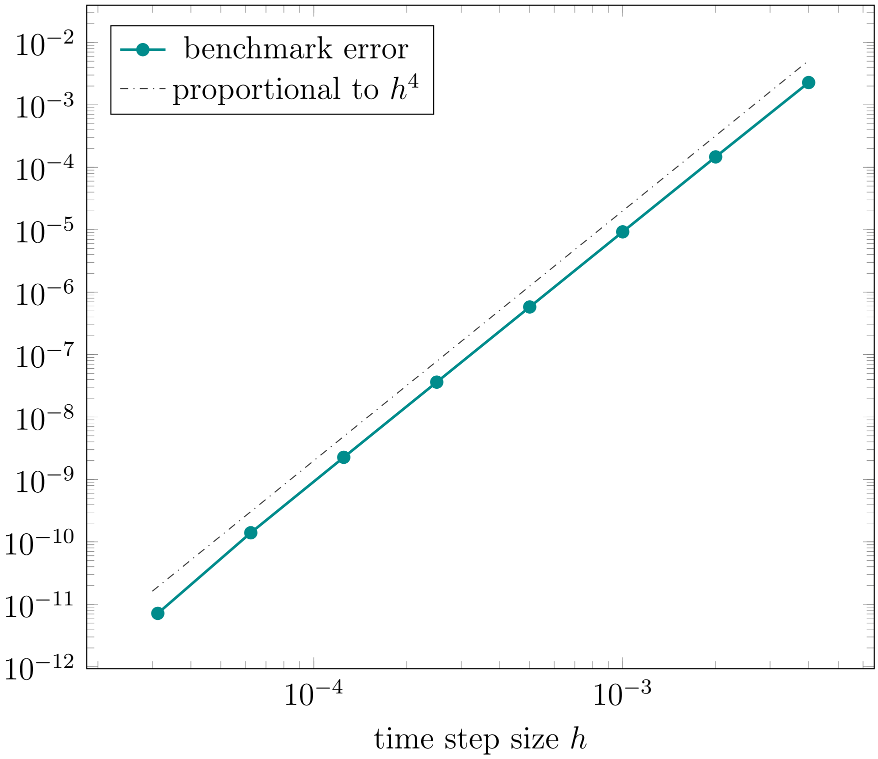
$$
\begin{aligned}
&\forall \delta p_1, \delta p_2 : \ (\dot{q}(t_k) - \operatorname{alg}^{-1}_{q(t_k)} v^a(t_k)) \cdot \operatorname{alg}^{-1}_{q(t_k)} \delta p_1 \\
&+2\left(\dot{q}(t_{k+1/2}) - \operatorname{alg}^{-1}_{q(t_{k+1/2})} \frac{v^a(t_k) + v^a(t_{k+1})}{2}\right) \cdot \operatorname{alg}^{-1}_{q(t_{k+1/2})}(\delta p_1 + \delta p_2) \\
&+(\dot{q}(t_{k+1}) - \operatorname{alg}^{-1}_{q(t_{k+1})} v^a(t_{k+1})) \cdot \operatorname{alg}^{-1}_{q(t_{k+1})} \delta p_2 = 0
\end{aligned}
$$

$$
\begin{aligned}
q(t_{k+1/2}) &\in \mathbb{R}^3 \times SO(3) \\
q(t_{k+1}) &\in \mathbb{R}^3 \times SO(3)
\end{aligned}
$$

where $t_{k+1/2} = (t_k + t_{k+1})/2$.

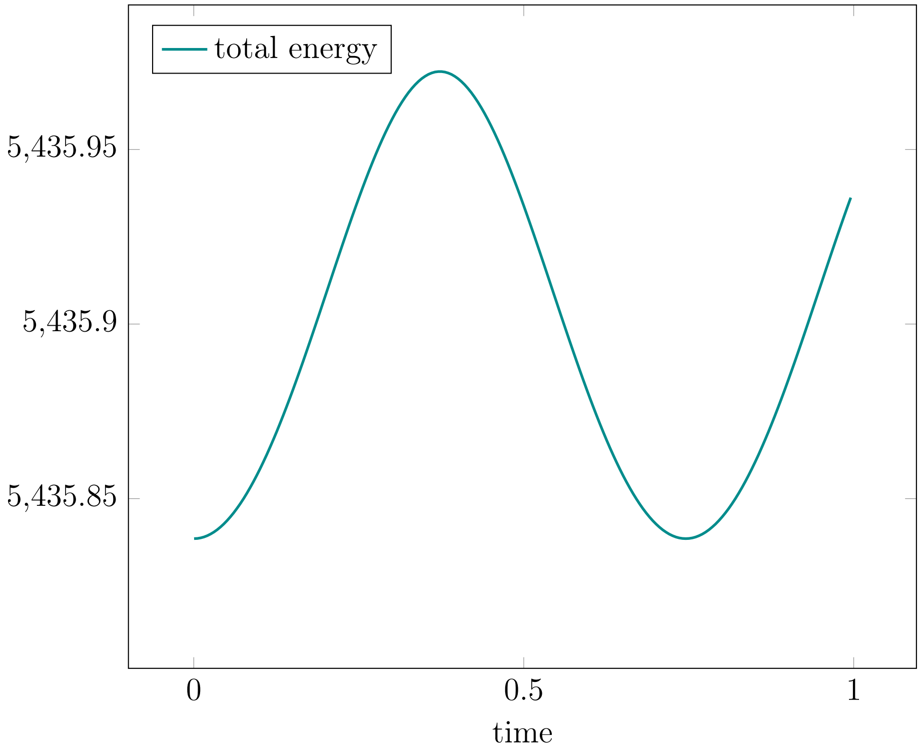
This time stepper showcases a better convergence rate, namely order 4 throughout the entire benchmark:

Figure 5.5.: Convergence plot for $j = 2$ and $v = \dot{q}$



However, the oscillations in the energy can become larger than for the choice of $v = \dot{q}$:

Figure 5.6.: Total energy for $j = 2$, $v = \dot{q}$, and time step size $5 \cdot 10^{-4}$



For an implementation that runs faster, it is advisable to simplify (5.7). Via substitution, the amount of equations in (5.7) can be reduced.

A. Theoretical remarks

A.1. The first variation

Definition A.1.1
Let X be a subset of a real vector space V, $x_0 \in X$, and $\delta x \in V$ such that $x_0 + \varepsilon \cdot \delta x \in X$ for small enough $\varepsilon > 0$. For a map $J : X \to \mathbb{R}$, the first variation is defined as

$$\delta J(x_0, \delta x) := \frac{d}{d\varepsilon} J(x_0 + \varepsilon \cdot \delta x)\bigg|_{\varepsilon=0}$$

if the expression on the right-hand side of the above equation exists.
x_0 is called a stationary point of J if $\delta J(x_0, \delta x) = 0$ for all δx for which $\delta J(x_0, \delta x)$ exists.
The notation $\delta J(x_0) = 0$ will also be used for this.
[8, Chapter 1]

A.1.1. Test functions on manifolds

Under certain circumstances, a stationary condition on the parameter space of a differentiable manifold can be simplified by choosing the test functions to lie in the tangent space of that manifold:

Lemma A.1.2
Let $\mathcal{Q}$ be an n-dimensional differentiable manifold that is a subset of an m-dimensional real vector space V.[a] For the sake of taking derivatives, we assume - without loss of generality - that $V = \mathbb{R}^m$. Also, let φ be any chart of $\mathcal{Q}$. We define the curves $\gamma \in C^1(\mathbb{R}, \operatorname{ran} \varphi)$ and $q := \varphi^{-1} \circ \gamma$ as well as the Fréchet differentiable maps $J : \mathbb{R}^n \to \mathbb{R}$ and $K := J \circ \varphi$.
In this setting, $\delta J(\gamma) = 0$ holds if and only if

$$\frac{\partial}{\partial \varepsilon} K(q + \varepsilon \delta q)\bigg|_{\varepsilon=0} = 0 \tag{A.1}$$

for all $\delta q \in C^1$ such that $\delta q(t) \in T_{q(t)}$ for all t.

[a] By the Whitney embedding theorem (see for example [18, Theorem 6.15]), this can be assumed without loss of generality.

Proof. Let $d\varphi$ denote the differential of φ. $\delta J(\gamma) = 0$ holds if and only if, for all test functions $\delta\gamma$:

$$0 = \frac{\partial}{\partial \varepsilon} J(\gamma + \varepsilon\delta\gamma)\Big|_{\varepsilon=0} = \nabla J(\gamma) \cdot \delta\gamma = \nabla J(\gamma) \cdot d\varphi(q) \cdot d\varphi(q)^{-1} \cdot \delta\gamma \tag{A.2}$$

As for Equation (A.1):

$$\begin{aligned} 0 &= \frac{\partial}{\partial \varepsilon} K(q + \varepsilon\delta q)\Big|_{\varepsilon=0} \\ &= \frac{\partial}{\partial \varepsilon} J(\varphi(\varphi^{-1}(\gamma) + \varepsilon\delta q))\Big|_{\varepsilon=0} \\ &= [\nabla J(\varphi(\varphi^{-1}(\gamma) + \varepsilon\delta q))^T d\varphi(\varphi^{-1}(\gamma) + \varepsilon\delta q))] \cdot \delta q\Big|_{\varepsilon=0} \\ &= [\nabla J(\gamma)^T d\varphi(q)] \cdot \delta q \end{aligned} \tag{A.3}$$

However, the tangent space of $\mathcal{Q}$ in x is $d\varphi(x)^{-1}(\mathbb{R}^m)$ [18, Proposition 3.15, Equation 3.8]. Therefore, the terms $d\varphi(q)^{-1} \cdot \delta\gamma$ from Equation (A.2) and δq from Equation (A.3) correspond. □

It should be noted that there is another consequence of the proof of Theorem A.1.2: The correspondence stated at the end of the proof has for result that certain properties of the per-definition test functions $\delta\gamma$ can simply be carried over to the tangent space test functions δq. Such properties comprise smoothness or the test function being zero at a certain point t. Therefore, the result of Theorem A.1.2 does not require the restriction to a single chart.

A.2. Continuous representatives of broken H^1

Proposition A.2.1
Let $f \in H^1(\{(a,b),(b,c)\})$. Then

$$f \in H^1((a,c)) \Leftrightarrow tr(f|_{(a,b)})(b) = tr(f|_{(b,c)})(b).$$

Proof. By the Sobolev embedding theorem, functions in $H^1((a,c))$ have continuous representatives. It is thus a necessary condition for $f \in H^1((a,c))$ that the values of the traces coincide.

It is also a sufficient one: $f \in L^2((a,c))$ clearly holds. Recall that the distributional derivative of f is said to be in $L^2((a,c))$ if and only if a function Df exists such that

$$\forall \varphi \in C_c^\infty((a,c)) : \int_a^c Df \; \varphi \; dx = -\int_a^c f\varphi' \; dx.$$

In this case, partial integration on $H^1((a,b))$ and $H^1((b,c))$ can be used to obtain

$$-\int_a^c f\varphi' \; dx = \int_a^b Df_- \varphi \; dx + \int_b^c Df_+ \varphi \; dx - \varphi(b)[tr(f|_{(a,b)})(b) - tr(f|_{(b,c)})(b)],$$

where Df_- and Df_+ denote the weak derivatives of f on (a,b) and (b,c), respectively. By assumption, this has for result that one can choose:

$$Df(x) := \begin{cases} Df_-(x) & \text{for } a < x < b \\ Df_+(x) & \text{for } b < x < c \end{cases}$$

□

A.3. Continuity and coercivity of b_0

This section analyses the bilinear form b_0 from Section 2.1.

Proposition A.3.1 (Continuity of b_0)
b_0 is continuous on $H^1(\mathcal{K}_\eta)^m \times H^1(\mathcal{K}_\eta)^m$. Moreover, it has a continuity constant that is independent of η.

Proof. For any matrix A and tuples a_1, a_2, a theorem by Lang [17, Chapter V, Theorem 2.2] states that the 2-norm of the bilinear form $(a_1, a_2) \mapsto (Aa_1) \cdot a_2$ is equal to $\|A\|_2$. Therefore,

$$|(Aa_1) \cdot a_2| \leq \|A\|_2 \; \|a_1\|_2 \; \|a_2\|_2$$

holds true.

Another result that will be needed is that for any function $f \in H^1(\mathcal{K}_\eta)$, the Cauchy-Schwartz inequality implies that

$$\|f\|_{L^2} + \|\dot{f}\|_{L^2} \leq \|(1,1)\|_2 \; \|(\|f\|_{L^2}, \|\dot{f}\|_{L^2})\|_2 = \sqrt{2}\|f\|_{H^1(\mathcal{K}_{2\eta})}.$$

Together with the Hölder inequality, this results in the continuity of b_0 on $H^1(\mathcal{K}_\eta)^m \times H^1(\mathcal{K}_\eta)^m$:

$$\begin{aligned} |b_0(q, \delta q)| &= \left| \int_0^T -(Kq) \cdot \delta q + (M\dot{q}) \cdot \frac{d\delta q}{dt} \, dt \right| \leq \int_0^{2T} |(Kq) \cdot \delta q| + \left| (M\dot{q}) \cdot \frac{d\delta q}{dt} \right| \, dt \\ &\leq \int_0^T \|K\|_2 \|q\|_2 \|\delta q\|_2 + \|M\|_2 \|\dot{q}\|_2 \left\| \frac{d\delta q}{dt} \right\|_2 \, dt \\ &\overset{\text{Hölder}}{\leq} \|K\|_2 \|q\|_{L^2} \|\delta q\|_{L^2} + \|M\|_2 \|\dot{q}\|_{L^2} \left\| \frac{d\delta q}{dt} \right\|_{L^2} \\ &\leq \max\{\|K\|_2, \|M\|_2\} \Bigg(\|q\|_{L^2} \|\delta q\|_{L^2} + \|q\|_{L^2} \left\| \frac{d\delta q}{dt} \right\|_{L^2} \\ &\qquad\qquad + \|\dot{q}\|_{L^2} \|\delta q\|_{L^2} + \|\dot{q}\|_{L^2} \left\| \frac{d\delta q}{dt} \right\|_{L^2} \Bigg) \\ &= \max\{\|K\|_2, \|M\|_2\} (\|q\|_{L^2} + \|\dot{q}\|_{L^2}) \left(\|\delta q\|_{L^2} + \left\| \frac{d\delta q}{dt} \right\|_{L^2} \right) \\ &\leq \max\{\|K\|_2, \|M\|_2\} \sqrt{2}^2 \|q\|_{H^1(\mathcal{K}_\eta)} \|\delta q\|_{H^1(\mathcal{K}_\eta)} \end{aligned}$$

□

Proposition A.3.2 (Coercivity of b_0)
Assuming (2.1), b_0 is coercive on $H_0^1([0,T])^m \times H_0^1([0,T])^m$. β is a valid coercivity constant.

Proof. Let A be any symmetric, positive definite, real matrix. Then a Schur decomposition $A = U^T \Lambda U$ exists, where U is an orthogonal matrix. Since A is symmetric, Λ is a diagonal matrix of eigenvalues of A. Let α denote the smallest eigenvalue of A. This yields

$$(Aa) \cdot a = a^T A a = a^T U^T \Lambda U a \geq \alpha a^T U^T U a = \alpha a \cdot a.$$

Apart from that, the first paragraph of the proof of Theorem A.3.1 leads to the result that $(Aa) \cdot a = |(Aa) \cdot a| \leq \|A\|_2 \|a\|_2^2$. Altogether, we have:

$$(Aa) \cdot a \geq \alpha a \cdot a \tag{A.4a}$$

$$(Aa) \cdot a \leq \|A\|_2 \|a\|_2^2 \tag{A.4b}$$

Coercivity is clearly equivalent to the existence of a constant β such that

$$\forall q : \ [\|q\|_{H^1} = 1 \Rightarrow b_0(q,q) \geq \beta] \,.$$

We will show that this inequality is fulfilled for the constant β from (2.1). To that end, we assume that q lies in $H_0^1([0,T])^m$ and that $\|q\|_{H^1} = 1$. Also, we reformulate $b_0(q,q) \geq \beta$ into

$$\int_0^T (M\dot{q}) \cdot \dot{q} - (Kq) \cdot q \; dt \geq \beta.$$

Here, we can apply Inequality (A.4a) to M and Inequality (A.4b) to K and obtain the sufficient condition

$$\int_0^T \nu \dot{q} \cdot \dot{q} - \|K\|_2 q \cdot q \; dt \geq \beta.$$

This can in turn be written as

$$\nu \|\dot{q}\|_{L^2}^2 \geq \|K\|_2 \|q\|_{L^2}^2 + \beta. \tag{A.5}$$

Now, two cases have to be dealt with.

On the one hand, $\|q\|_{L^2}^2 < 1/2$ might hold. Since we assume $1 = \|q\|_{H^1}^2 = \|q\|_{L^2}^2 + \|\dot{q}\|_{L^2}^2$, this automatically implies that $\|\dot{q}\|_{L^2}^2 \geq 1/2$. Therefore, it is a sufficient condition for Inequality (A.5) that

$$\nu \frac{1}{2} \geq \|K\|_2 \frac{1}{2} + \beta. \tag{A.6}$$

This last inequality is equivalent to Inequality (2.1b), which holds by assumption.

On the other hand, $\|q\|_{L^2}^2 \geq 1/2$ might hold. In that case, we transform Inequality (A.5) into

$$\frac{\nu}{\|K\|_2 + \frac{\beta}{\|q\|_{L^2}^2}} \|\dot{q}\|_{L^2}^2 \geq \|q\|_{L^2}^2.$$

By the Poincaré-Friedrichs inequality (see [2, Theorem 1.5]), it is a sufficient condition that

$$\frac{\nu}{\|K\|_2 + \frac{\beta}{\|q\|_{L^2}^2}} \geq T^2.$$

This inequality is in turn implied by

$$\frac{\nu}{\|K\|_2 + 2\beta} \geq T^2, \tag{A.7}$$

for which we used the assumption of this case, namely $\|q\|_{L^2}^2 \geq 1/2$. However, Inequality (A.7) is equivalent to Inequality (2.1a). The latter holds by assumption. □

A.4. Uniform coercivity of $b_{0,\eta}$

The bilinear form $b_{0,\eta}$ is continuous and coercive, with a continuity constant and a coercivity constant independent of η.

Let us denote the application of the $(j+1)$ point Gauss-Lobatto rule on an interval $[a,b]$ as

$$\int_{[a,b]}^{\mathrm{GL}(j+1)} \cdot \, dt.$$

Proposition A.4.1
Given $C > 0$ such that

$$\int_{[0,T]}^{\mathrm{GL}(j+1)} f^2 \, dt > C \, \|f\|_{L^2([0,T])}^2 \,,$$

the inequality

$$\int_{[0,h]}^{\mathrm{GL}(j+1)} f\left(t\frac{T}{h}\right)^2 \, dt > C \left\| f\left(t\frac{T}{h}\right)\right\|_{L^2([0,h])}^2 \,,$$

holds for any h.

Proof.

$$\begin{aligned}\int_{[0,h]}^{\mathrm{GL}(j+1)} f\left(t\frac{T}{h}\right)^2 \, dt &= \frac{h}{T}\int_{[0,T]}^{\mathrm{GL}(j+1)} f^2 \, dt > \frac{h}{T} \, C \, \|f\|_{L^2([0,T])}^2 = \frac{h}{T} \, C \int_0^T f(s)^2 \, ds \\ &= C \int_0^h f\left(t\frac{T}{h}\right)^2 \, dt = C \left\| f\left(t\frac{T}{h}\right)\right\|_{L^2([0,h])}^2\end{aligned}$$

□

Proposition A.4.2 (Uniform coercivity of $b_{0,\eta}$)
$b_{0,\eta}$ is coercive on the space of piecewise polynomials over $\mathcal{K}_\eta$, with a coercivity constant that is independent of η.

Proof. We start by verifying that

$$(f,g)_{\mathrm{GL},L^2} := \int_{[0,T]}^{\mathrm{GL}(j+1)} fg\ dt$$

is positive definite for polynomials of degree j. Assume that f fulfills $\deg f = j$ and

$$0 = (f,f)_{\mathrm{GL},L^2} = \int_{[0,T]}^{\mathrm{GL}(j+1)} f^2\ dt.$$

This implies that f^2 has $j+1$ roots. Thus, f also has $j+1$ roots. However, f is of degree j, and thus $f = 0$.

Since $(\cdot,\cdot)_{\mathrm{GL},L^2}$ is positive definite, a C exists such that

$$(f,f)_{\mathrm{GL},L^2} > C\,\|f\|_{L^2}^2$$

for all f with $\deg f = j$. Hence, Theorem A.4.1 can be used.

A piecewise polynomial u can now be decomposed into element-wise polynomial functions $u_i, i = 1,\dots,\eta$. That way, we can write

$$\begin{aligned}
b_{0,\eta}(u,u) &= b_{0,\eta}\left(\sum_i u_i, \sum_i u_i\right) \\
&= \sum_i b_{0,\eta}(u_i,u_i) && \text{(orthogonality)} \\
&\geq \min\{\kappa,\nu\} \sum_i \int_{\operatorname{dom} u_i}^{\mathrm{GL}(j+1)} u_i^2 + \dot{u}_i^2\ dt && \text{(smallest eigenv. of } K, M\text{)} \\
&> \min\{\kappa,\nu\} \sum_i C\left(\|u_i\|_{L^2(\operatorname{dom} u_i)}^2 + \|\dot{u}_i\|_{L^2(\operatorname{dom} u_i)}^2\right) && \text{(Theorem A.4.1, shifted)} \\
&= C\min\{\kappa,\nu\} \sum_i \|u_i\|_{H^1(\operatorname{dom} u_i)}^2 \\
&= C\min\{\kappa,\nu\}\,\|u\|_{H^1}^2 && \text{(Pythagoras)}
\end{aligned}$$

□

A.5. (Gauss-)Lobatto quadrature

The (Gauss-)Lobatto quadrature rules are characterized by the fact that they evaluate the integrand at the endpoints of the integration interval. For the interval $[-1,1]$ and the application of the n point rule to a function f, the error can be written as

$$C(n)f^{(2n-2)}(\zeta), \qquad -1 < \zeta < 1.$$

An example for this type of quadrature is Simpson's rule. [11, Section 8.12]

Another example is the trapezoidal rule.

For an integration interval of length h, the error is

$$D(n)f^{(2n-2)}(\zeta)h^{2n-2}$$

for a ζ in the integration interval. This result can be obtained via integral substitution.

A.6. $SO(3)$ verification

To check whether a matrix B lies in $SO(3)$, one might use the equations by which $SO(3)$ is defined, namely

$$B^T B = I_{3\times 3} \tag{A.8a}$$

$$BB^T = I_{3\times 3} \tag{A.8b}$$

$$\det B = 1. \tag{A.8c}$$

However, some of these equations are redundant. In the context of time steppers on $SO(3)$, this redundancy becomes a problem for the numerical equation solver (Newton's method or similar) which is used to solve the equations of the time stepper but also assure that B lies in $SO(3)$.

As a result, these redundant equations have to be removed. Three propositions help with that:

Proposition A.6.1
Let $B \in \mathbb{R}^{i\times i}$. Then $B^T B$ and BB^T are symmetric.

Proof. A matrix is symmetric if and only if it is equal to its transpose. This equality can be verified for $B^T B$:

$$\left(B^T B\right)^T = B^T \left(B^T\right)^T = B^T B$$

The symmetry of BB^T can be derived from this by exchanging B and B^T in the last chain of equalities. □

Proposition A.6.2
Let $B \in \mathbb{R}^{i\times i}$. Then $B^T B = I_{i\times i}$ and $BB^T = I_{i\times i}$ are equivalent.

Proof. $B^T B = I_{i\times i}$ is equivalent to B^T being a left inverse of B. This is in turn equivalent to B being surjective, which is equivalent to B being bijective, making B^T a full inverse of B. Thus, $BB^T = I_{i\times i}$.

The same argumentation can be applied to $BB^T = I_{i\times i}$, replacing left inverses with right inverses. □

Proposition A.6.3
An orthogonal matrix has determinant 1 or -1.

Proof. Let $B \in \mathbb{R}^{i\times i}$ be any orthogonal matrix. Then

$$1 = \det I_{i\times i} = \det B^T B = (\det B)^2.$$

□

From (A.8), Equation (A.8b) can be removed completely as a result of Proposition A.6.2. What is more, a time stepper does not have to check Equation (A.8c): Since the determinant is a continuous function and the matrix in the beginning of the time step has determinant 1, it only takes a small enough time step and Proposition A.6.3 for Equation (A.8c) to hold automatically.

Furthermore, Proposition A.6.1 has for result that $B^T B = I_{3\times 3}$ only has to be verified for the upper (or lower) triangular parts of $B^T B$ and $I_{3\times 3}$. All in all, this is the only verification a time stepper needs to perform.

Bibliography

[1] Hans C. Andersen. Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations. *Journal of computational physics*, 52(1):24–34, 1983.

[2] Dietrich Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*. Cambridge University Press, 3 edition, 2007.

[3] Olivier Brüls, Martin Arnold, and Alberto Cardona. Two lie group formulations for dynamic multibody systems with large rotations. *Proceedings of the ASME Design Engineering Technical Conference*, 4:85–94, 01 2011.

[4] C. Carstensen, L. Demkowicz, and J. Gopalakrishnan. Breaking spaces and forms for the DPG method and applications including Maxwell equations. *Comput. Math. Appl.*, 72(3):494–522, 2016.

[5] M. Durufle, P. Grob, and P. Joly. Influence of gauss and gauss-lobatto quadrature rules on the accuracy of a quadrilateral finite element method in the time domain. *Numerical Methods for Partial Differential Equations*, 25(3):526–551, 2009.

[6] Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*. Applied mathematical sciences. Springer, New York, NY, 2004.

[7] Lawrence C. Evans. *Partial differential equations*. Graduate studies in mathematics. American Math. Soc., Providence, RI, 2. ed. edition, 2010.

[8] Mariano Giaquinta and Stefan Hildebrandt. *Calculus of variations : 1. The Lagrangian formalism*. Die Grundlehren der mathematischen Wissenschaften. Springer, Berlin [u.a.], 1996.

[9] Helmut Haberzettl. *Classical mechanics : lecture notes*. World Scientific, New Jersey London Singapore Beijing Shanghai Hong Kong Taipei Chennai Tokyo, 2021.

[10] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration : structure-preserving algorithms for ordinary differential equations*. Springer series in computational mathematics. Springer, Berlin [u.a.], 2. ed. edition, 2006.

[11] Francis B. Hildebrand. *Introduction to numerical analysis*. International series in pure and applied mathematics. McGraw-Hill, New York, NY [u.a.], 1956.

[12] Darryl D. Holm. *Geometric mechanics : 2. Rotating, translating and rolling*. World Scientific Publ. Imperial College Press, Singapore [u.a.] London, 2008.

[13] Darryl D. Holm. *Geometric mechanics. Part I*. Imperial College Press, London, second edition, 2011. Dynamics and symmetry.

[14] Stefan Holzinger, Martin Arnold, and Johannes Gerstmayr. Evaluation and implementation of lie group integration methods for rigid multibody systems. *Multibody System Dynamics*, 62:273–306, 03 2024.

[15] P. L. Kinon, P. Betsch, and S. Schneider. The GGL variational principle for constrained mechanical systems. *Multibody system dynamics*, 57(3-4):211–236, 2023.

[16] Philipp L. Kinon, Peter Betsch, and Simeon Schneider. Structure-preserving integrators based on a new variational principle for constrained mechanical systems. *Nonlinear dynamics*, 111(15):14231–14261, 2023.

[17] Serge Lang and Nicolas Bourbaki. *Real and functional analysis*. Graduate texts in mathematics. Springer, New York, NY [u.a.], 3. ed. edition, 1993.

[18] John M. Lee. *Introduction to smooth manifolds*. Graduate texts in mathematics. Springer, New York, NY [u.a.], 2. ed. edition, 2013.

[19] Ignacio Romero and Martin Arnold. Computing with rotations: Algorithms and applications. In *Encyclopedia of Computational Mechanics Second Edition*, volume 3, pages 1–27. John Wiley & Sons, Ltd, 2017.

[20] Hiroaki Yoshimura and Jerrold E. Marsden. Dirac structures in lagrangian mechanics part ii: Variational structures. *Journal of geometry and physics*, 57(1):209–250, 2006.

List of Figures

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Bachelorarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Wien, am ____________________ ________________________