INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN ENGINEERING
Int. J. Numer. Meth. Engng (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/nme.1582

A linearly implicit trapezoidal method for integrating stiff multibody dynamics with contact, joints, and friction

Florian A. Potra[1], Mihai Anitescu[2,*,†], Bogdan Gavrea[1] and Jeff Trinkle[3]

[1]*Department of Mathematics, University of Maryland, Baltimore County, U.S.A.*
[2]*Mathematics and Computer Science Division, Argonne National Laboratory, U.S.A.*
[3]*Department of Computer Science, Rensselaer Polytechnic Institute, U.S.A.*

SUMMARY

We present a hard constraint, linear complementarity based, method for the simulation of stiff multibody dynamics with contact, joints and friction. The approach uses a linearization of the modified trapezoidal method, incorporates a Poisson restitution model at collision, and solves only one linear complementarity problem per time step when no collisions are encountered. We prove that, under certain assumptions, the method has order two, a fact that is also demonstrated by our numerical simulations. For the unconstrained (ODE) case, the method achieves second-order convergence and absolute stability while solving only one linear system per step. When we use a special approximation of the Jacobian matrix for the case where the stiff forces originate in springs and dampers attached to two points in the system, the linear complementarity problem can be solved for any value of the time step and numerical simulation demonstrate that the method is stiffly stable. The method was implemented in UMBRA, an industrial-grade virtual prototyping software. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: multibody dynamics; rigid bodies; Coulomb friction; stiff methods; linear complementarity problems; linearly implicitly methods

1. INTRODUCTION

Mechanical systems composed of nominally rigid bodies interacting through frictional contacts are pervasive, and in fact, are responsible for the high levels of productivity and living standard enjoyed by citizens of modern industrialized countries. The most obvious example of such mechanical systems is the internal combustion engine, but many human-powered devices fall into this category too. Spurred by global economic competition and dwindling resources, companies

*Correspondence to: Mihai Anitescu, Mathematics and Computer Science Division, Argonne National Laboratory, 9700 South Cass Avenue, Argonne, IL 60637, U.S.A.
†E-mail: anitescu@mcs.anl.gov

Contract/grant sponsor: National Science Foundation; contract/grant number: DMS-0139701
Contract/grant sponsor: U.S. Department of Energy; contract/grant number: W-31-109-ENG-38

Received 3 June 2004
Revised 2 August 2004
Accepted 19 September 2005

have begun to use analysis and simulation software to optimize existing systems and to design highly efficient new systems. However, when the proper operation of the device relies on the formation and loss of frictional contacts at possibly unpredictable locations, and the stick–slip behaviour of those contacts, the available software is neither efficient nor robust enough to be useful. Compelling applications that fall into this category range from autonomous robots that explore and clean hazardous environments (see the left-hand picture in Figure 8), to intelligent prosthetic hands that can deftly manipulate objects based on high-level requests, to the simulations of surgical procedures for preoperative planning and risk-free training of surgeons, to the automation of manufacturing processes, including automatic fixturing (see Figure 8) or part-feeding systems [1].

When analysing mechanical systems that experience intermittent contacts with dry friction, the rigid body assumption considerably simplifies the mathematical model. The deformable body assumption naturally leads to a system of partial differential equations (PDEs) which depend on material and geometric properties that are not easy to obtain, but are important for reasonably accurate system response, especially when collisions occur. Even without collisions, the modelling of the contact interfaces (which are distributed when the bodies are deformable) is extremely challenging. In fact, one of the key research areas in the Advanced Scientific Computation Initiative (ASCI) project at Sandia National Laboratories is the modelling and numerical analysis of contact interfaces to extend the capability of their suite of PDE solvers. Moreover, the numerical solution of these PDEs is extremely time consuming, so much so as to make design optimization on widely available computers impractical. Under the rigid body assumption, the dynamic equations and constraints of the multibody system can be modelled by a dynamic complementarity problem (DCP). This model consists of a system of ordinary differential equations (ODE) subject to complementarity constraints (defined in Section 4). When bilateral constraints are also present, then the ODE is replaced by a system of mixed differential algebraic equations (DAE).

The DCP formulation of the mechanical system dynamics is advantageous in two main ways: first, it depends on a relatively small number of physical properties, and second, its form is easily exploited in numerical integration schemes. However, when formulated in terms of body accelerations and forces,[‡] this approach also has the problem that the DCP may not have a solution, as has been pointed out by Painlevé [2] 1895. The example found by Painlevé shows in fact that the Coulomb friction model and the equations of classical rigid body dynamics are incompatible in the sense that models based on them may lead to continuous problems without solution in classical sense. However, Mason and Wang were able to resolve the inconsistency expanding the solution space to allow impulsive forces to act at times other than those when impacts occurred [3]. After a century of debate among scientists, it appears that the continuous problem of rigid multibody dynamics is better understood in terms of measure differential inclusions (see, e.g. Reference [4]) and measure differential equations rather than in terms of ODE. A comprehensive survey of rigid multibody dynamics from this point of view can be found in the book by Brogliato [5] and the review article of Stewart [6].

Over the past decade another way of dealing with the rigid multibody dynamics problem has come to the forefront of research. Instead of developing a continuous-time model that admits

[‡]Formulation in terms of accelerations and forces is the most natural approach, since the Newton–Euler equations are typically expressed in these quantities.

generalized solutions (in the sense of differential inclusions or measure differential equations), several researchers have built discrete-time models based on a time-stepping formulation of the problem [4, 7–12]. Generally speaking, these time-stepping formulations are based on the idea of using time integrals of the forces (i.e. impulses) over each time-step, rather than trying to find the instantaneous forces at each instant. This approach implicitly allows impulsive forces to act at any time during contact, not just at the moment an impact occurs. Formally they can be regarded as some sort of numerical integrator applied to the continuous formulation and a generalization of the method used by Mason and Wang to resolve Painlevé's Paradox. With regard to the former point, the methods of Anitescu and Potra [7] and Stewart and Trinkle [12] are based on a semi-implicit Euler scheme, while the model of Anitescu and Potra [8] is based on a linearly implicit Euler scheme. All three formulations require the solution of one linear complementarity problem at each time step. The key difference between the methods presented in References [7, 12] are that in the latter, the non-penetration constraint is written in terms of positions and therefore includes a constraint stabilization term. The method from Reference [7] was recently modified to include constraint stabilization terms as well [13]. In the latter reference, it is proved that constraint stabilization is achieved and the energy stays bounded as the time step goes to zero, even if only one linear complementarity problem is solved per fixed time step, as opposed to a non-linear complementarity problem that needs to be solved in Reference [12].

A physically correct time-stepping formulation has to be dissipative, since the creation of energy in a system is physically inconsistent. At each step of the time-stepping scheme of Anitescu and Potra [7] the kinetic energy cannot exceed that obtained in the same configuration with no contacts enforced. As a consequence, the resulting velocities are uniformly bounded on any finite time interval. This property is essential in proving the convergence of the numerical solution given by the time-stepping scheme as the step size goes to zero. The strongest convergence result to date was obtained by Stewart [14] who showed that under some restrictive assumptions (which are general enough to be satisfied by Painlevé's example and other problems of interest) the numerical trajectories produced by the time-stepping scheme converge, in some suitable sense, as the step size tends to zero, and that the corresponding limits satisfy all the conditions required in the instantaneous-time problem. More precisely, it is shown that for a subsequence, the positions converge uniformly, while the velocities converge pointwise. The impulses are used to construct measures for which there are subsequences that converge *weak** to measures that are solutions to the corresponding measure differential inclusion.

The time-stepping scheme of Anitescu and Potra [7] proved to be very robust in numerical simulations. However, in the presence of springs and dampers attached to points of the multibody system, a prohibitively small time step has to be chosen to accommodate the related numerical stiffness. The time-stepping formulation of Anitescu and Potra [8] was designed to properly handle such stiffness. Numerical experiments showed that this scheme could use relatively large time steps for systems with very stiff dampers and/or springs. It was shown that as the stiffness parameters increase to infinity the time-stepping formulation of Anitescu and Potra [8] converges to a time-stepping formulation where the corresponding springs and dampers are replaced by rigid elements.

The time-stepping formulations mentioned above are of first order. Since the trajectories of multibody systems with contact and friction are piecewise smooth, the accuracy of the formulation can be improved by using higher-order methods, provided a reliable event (collision,

take-off, stick–slip) detection procedure is available, thereby allowing larger steps to be taken by the algorithm. In the present paper we present a time-stepping formulation based on a linearly implicit trapezoidal method that provides new discrete-time LCP model of second order. Only one LCP has to be solved at each time step, as long as there are no collisions occurring. In this work, we do not address the stabilization of the non-linear geometrical constraints. For this algorithm, constraint stabilization that does not significantly increase the energy, even when stiff forces are present, can be achieved by a projection technique, much as in Reference [8].

Since it involves only one linear complementarity problem per step, the second-order model can be implemented with the same computational effort per time step as the first-order models described in References [7, 8, 12]. We prove that the second model has the same properties concerning the kinetic energy as the first model of Anitescu and Potra [7], that it can accommodate moderately stiff forces, and that it can accommodate arbitrarily stiff forces that originate in springs and dampers, as in Reference [8], at the expense of falling back to an order one scheme. Our numerical experiments show that it can efficiently handle stiff problems. To our knowledge this is the first time-stepping scheme of second order that requires the solution of only one LCP per time step. We mention that Tzitzouris [15] derived a time-stepping scheme based on the fully implicit trapezoidal method that requires the solution of non-linear complementarity problem at each time step. The latter is solved by means of a non-smooth Newton method so that the computational complexity of Tzitzouris' scheme depends on the number of Newton iterations required at each time step. Moreover, there are no results about the kinetic energy corresponding to the time-stepping scheme from Reference [15].

Conventions: We denote by $\mathbb{N}$ the set of all non-negative integers. $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{R}_{++}$ denote the set of real, non-negative real, and positive real numbers, respectively. If $\|.\|$ is a vector norm on $\mathbb{R}^n$ and A is a matrix, then the operator norm induced by $\|.\|$ is defined by $\|A\| = \max\{\|Ax\|; \|x\| = 1\}$. If $x^1, x^2, \ldots, x^m \in \mathbb{R}^n$, then the column vector $z \in \mathbb{R}^{mn}$ obtained by concatenating the vectors x^k will be denoted by $\lceil x^1, x^2, \ldots, x^m \rfloor$, i.e.

$$z = \lceil x^1, x^2, \ldots, x^m \rfloor = [x^{1\mathrm{T}}, x^{2\mathrm{T}}, \ldots, x^{m\mathrm{T}}]^{\mathrm{T}} \tag{1}$$

If $x, y \in \mathbb{R}^n_+$ are two non-negative vectors such that $x^{\mathrm{T}} y = 0$, then we say that x and y are complementary and we write $0 \leqslant x \perp y \geqslant 0$. The symbol e represents the vector of all ones, $e = \lceil 1, 1, \ldots, 1 \rfloor$, with dimension given by the context.

If y is a vector depending on a positive parameter h then we write $y = O(h)$ to indicate that, for sufficiently small h, there is a constant $\bar{\alpha} > 0$ such that $\|y\| \leqslant \bar{\alpha} h$. If γ is a scalar depending on h then we write $\gamma = \Omega(h)$ to indicate that there is a constant $\underline{\alpha} > 0$ such that $\gamma \geqslant \underline{\alpha} h$.

If $x \in \mathbb{R}^m$ and $f(x) \in \mathbb{R}^n$, then the Jacobian of f with respect to x is denoted by f_x. This is an $n \times m$ matrix. We will also use the notations $J_x f = f_x$, $\nabla_x f = f_x^{\mathrm{T}}$. In case $n = 1$, the column vector $\nabla_x f$ is the usual gradient of f.

2. THE LINEARLY IMPLICIT TRAPEZOIDAL METHOD

Take a generic differential equation

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f(t, y) \tag{2}$$

where $y \in \mathbb{R}^n$, $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Given an approximation y^l of the solution at time-step t_l, the classical trapezoidal method [16] defines an approximate solution at time step $t_{l+1} = t_l + h$ as the solution of the following non-linear system:

$$y^{l+1} = y^l + \frac{h}{2}(f(t^l, y^l) + f(t^{l+1}, y^{l+1})) \tag{3}$$

Let us consider the case of the dynamics equations of a system of rigid bodies whose state can be described by a generalized position vector $q \in \mathbb{R}^s$ and a generalized velocity vector $v \in \mathbb{R}^s$. In the absence of any constraints these equations can be written as

$$\frac{\mathrm{d}q}{\mathrm{d}t} = v, \quad M(t, q)\frac{\mathrm{d}v}{\mathrm{d}t} = k(t, q, v) \tag{4}$$

where $M(t, q)$ is the generalized mass matrix, considered to be positive definite, and $k(t, q, v)$ are the external forces acting on the system.

In order to apply the trapezoidal method to (4), we first multiply the second equation with $M(t, q)^{-1}$ in order to obtain an equation of the form (2) and then we apply (3) to obtain an approximation $z_T^{l+1} = \lceil q_T^{l+1}, v_T^{l+1} \rfloor$ of the solution at time step t_{l+1} as the solution of the non-linear system

$$\begin{aligned} q_T^{l+1} &= q^l + \frac{h}{2}(v^l + v_T^{l+1}) \\ v_T^{l+1} &= v^l + \frac{h}{2}(M^{-1}(t_l, q^l)k(t_l, q^l, v^l) + M^{-1}(t_{l+1}, q_T^{l+1})k(t_{l+1}, q_T^{l+1}, v_T^{l+1})) \end{aligned} \tag{5}$$

In what follows we will consider the following more convenient variant of the trapezoidal method

$$\begin{aligned} q_{MT}^{l+1} &= q^l + \frac{h}{2}(v^l + v_{MT}^{l+1}) \\ (M(t_l, q^l) + M(t_{l+1}, q_{MT}^{l+1}))(v_{MT}^{l+1} - v^l) &= h(k(t_l, q^l, v^l) + k(t_{l+1}, q_{MT}^{l+1}, v_{MT}^{l+1})) \end{aligned} \tag{6}$$

If the mass matrix M is constant then (5) and (6) coincide. It is known that the trapezoidal method is a second-order method [16]. This follows from the fact that if $z^l = z(t_l)$, where $z(t) = \lceil q(t), v(t) \rfloor$ is the exact solution of (4), then the truncation error satisfies $z_T^{l+1} = z(t_{l+1}) + O(h^3)$. It can be proved that the modified trapezoidal method (6) is also a second-order method. Equations (5) and (6) are non-linear systems that are usually solved by some variant of Newton's method, which requires computation of the Jacobian of the corresponding system. We note that the Jacobian of the non-linear system (6) is easier to compute than that of the non-linear system (5). In this paper, we consider a linearization of (6) which, for non-constant mass matrix, is slightly different from the one given by Newton's method. We call this method the *linearly implicit trapezoidal method*. Given a point $z^l = \lceil q^l, v^l \rfloor$ which approximates $z(t_l)$, this method computes an approximation z_{LT}^{l+1} of $z(t_{l+1})$ as the solution of the following linear system for (6) produces a point $z_{LT}^{l+1} = \lceil q_{LT}^{l+1}, v_{LT}^{l+1} \rfloor$ as the solution of the following

linear system:

$$\begin{aligned} q_{LT}^{l+1} &= q^l + \frac{h}{2}(v^l + v_{LT}^{l+1}) \\ \hat{M} v_{LT}^{l+1} &= \hat{M} v^l + h\hat{k} \end{aligned} \tag{7}$$

where

$$\hat{M} = \frac{1}{2}(M(t_l, q^l) + M(t_{l+1}, q^l + hv^l)) - \frac{h}{2}k_v(t_{l+1}, q^l, v^l) - \frac{h^2}{4}k_q(t_{l+1}, q^l, v^l) \tag{8}$$

$$\hat{k} = \frac{1}{2}\left(k(t_l, q^l, v^l) + k(t_{l+1}, q^l, v^l) + \frac{h}{2}k_q(t_{l+1}, q^l, v^l)v^l\right) \tag{9}$$

If the computation of the mass matrix at a new position is expensive, we can save a mass matrix evaluation by taking

$$\hat{M} = M\left(t_l + \frac{h}{2}, q^l + \frac{h}{2}v^l\right) - \frac{h}{2}k_v(t_{l+1}, q^l, v^l) - \frac{h^2}{4}k_q(t_{l+1}, q^l, v^l) \tag{10}$$

In the next section we will prove that, with any of the two choices for $\hat{M}$ and $z^l = z(t_l)$, we have $z_{LT}^{l+1} - z(t_{l+1}) = O(h^3)$, so that the linearly implicit trapezoidal is also a second-order method. Since the stiffness is usually caused by $k(t, q, v)$, it follows that the linearly implicit trapezoidal method has the same (linear) stability properties as the classical trapezoidal method.

If constraints are added to the dynamics equations (4) then a linear complementarity problem has to be solved at each time-step, instead of a linear system. There, it will be essential to have a positive definite matrix $\hat{M}$. Since M is positive definite it follows that $\hat{M}$ will be positive definite for sufficiently small h. However it is important that $\hat{M}$ be positive definite for moderate values of h so that larger integration steps can be taken. In order to insure the positive definiteness of $\hat{M}$, we will consider appropriate approximations of $\tilde{k}_q^l$ and $\tilde{k}_v^l$ of $k_q(t_{l+1}, q^l, v^l)$ and $k_v(t_{l+1}, q^l, v^l)$ which will lead to an approximate linearly implicit method of the form

$$q^{l+1} = q^l + \frac{h}{2}(v^l + v^{l+1}) \tag{11}$$

$$\tilde{M} v^{l+1} = \tilde{M} v^l + h\tilde{k} \tag{12}$$

where

$$\tilde{M} = \bar{M} - \frac{h}{2}\tilde{k}_v^l - \frac{h^2}{4}\tilde{k}_q^l \tag{13}$$

$$\bar{M} = \frac{1}{2}(M(t_l, q^l + hv^l) + M(t_{l+1}, q^l)) \quad \text{or} \quad \bar{M} = M\left(t_l + \frac{h}{2}, q^l + \frac{h}{2}v^l\right) \tag{14}$$

$$\tilde{k} = \frac{1}{2}\left(k(t_l, q^l, v^l) + k(t_{l+1}, q^l, v^l) + \frac{h}{2}\tilde{k}_q^l v^l\right) \tag{15}$$

We will prove that if

$$\begin{aligned} &q^l = q(t_l) + O(h^3), \quad v^l = v(t_l) + O(h^3), \quad \tilde{k}_q^l = k_q(t_{l+1}, q^l, v^l) + O(h), \\ &\tilde{k}_v^l = k_v(t_{l+1}, q^l, v^l) + O(h^2) \end{aligned} \tag{16}$$

then the approximate linearly implicit trapezoidal method (11) satisfies $z^{l+1} = z(t_{l+1}) + O(h^3)$. As a consequence we will show that (11) is a second-order time-stepping scheme.

In what follows we adapt the time-stepping scheme (11) to the case where the dynamical system (4) is subject to *equality constraints* that usually arise in the presence of joints [17]. Such constraints can be described by equations of the form

$$\Theta^{(i)}(q) = 0, \quad i = 1, 2, \ldots, m \tag{17}$$

where $\Theta^{(i)}$ are sufficiently smooth functions. The force exerted by joint (i) on the system is $c_v^{(i)}(q)v^{(i)}(q)$, where $v^{(i)}(q) = \nabla_q \Theta^{(i)}(q)$ is the gradient of $\Theta^{(i)}(q)$ and $c_v^{(i)}$ is the appropriate Lagrange multiplier [17]. We assume that the equality constraints are not redundant in the sense that the vectors $v^{(i)}(q),\ i = 1, 2, \ldots, m$ are linearly independent.

If we denote $\Theta(q)$ the vector of components $\Theta^{(i)}(q) = 0,\ i = 1, 2, \ldots, m$, then the constraints (17) can be written under the compact form $\Theta(q) = 0$. Also, if we denote by $v(q) = \nabla_q \Theta(q)$ the $s \times m$-matrix with columns $v^{(i)}(q),\ i = 1, 2, \ldots, m$ and by c_v the vector having as components the Lagrange multipliers $c_v^{(i)},\ i = 1, 2, \ldots, m$, then the force exerted by all the joints on the system is represented by the vector $v(q)c_v$. Therefore, the equations of motion of the equality constrained multibody system become

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= v \\ M(t, q)\frac{\mathrm{d}v}{\mathrm{d}t} &= v(q)c_v + k(t, q, v) \\ \Theta(q) &= 0 \end{aligned} \tag{18}$$

This is a system of mixed DAE of index three whose numerical solution is rather difficult [16]. By differentiating the third equation in (18) the following DAE of index 2:

$$\begin{aligned} \frac{\mathrm{d}q}{\mathrm{d}t} &= v \\ M(t, q)\frac{\mathrm{d}v}{\mathrm{d}t} &= v(q)c_v + k(t, q, v) \\ v(q)^{\mathrm{T}} v &= 0 \end{aligned} \tag{19}$$

It is well known that for any initial time t_0 and any initial values q^0, v^0 that are consistent in the sense that

$$\Theta(q) = 0, \quad v(q^0)^{\mathrm{T}} v^0 = 0 \tag{20}$$

there is a unique c_v^0 and unique functions $q(t), v(t), c_n u(t)$ satisfying DAEs (18) and (19) and the initial value condition

$$q(t_0) = q^0, \quad v(t_0) = v^0, \quad c_v(t_0) = c_v^0 \tag{21}$$

The initial vector of Lagrange multipliers c_v^0 can be obtained by solving the following linear system in w^0 and c_v^0:

$$\begin{aligned} M(t_0, q^0)w^0 &= v(q^0)c_v^0 + k(t_0, q^0, v^0) \\ v(q^0)^{\mathrm{T}} w^0 &= -v^{0\mathrm{T}} \nabla_q (v(q)^{\mathrm{T}} v^0) \mid_{q=q^0} \end{aligned} \tag{22}$$

We note that on the right-hand side of the last equation from the linear system above we have the vector of components $-v^{0\mathrm{T}} \nabla_q^2 \Theta_i(q^0) v^0$, $i = 1, 2, \ldots, m$, so that in order to set up this system we need to compute the Hessians of the constraints.

We consider the following time-stepping scheme for the numerical solution of the DAE (19) that are based on the approximate linearly implicit trapezoidal method (11):

$$\begin{aligned} q^{l+1} &= q^l + \frac{h}{2}(v^l + v^{l+1}) \\ \tilde{M} v^{l+1} - hv\left(q^l + \frac{h}{2}v^l\right) c_v^{l+1/2} &= \tilde{M} v^l + h\tilde{k} \\ v\left(q^l + \frac{h}{2}v^l\right)^{\mathrm{T}} v^{l+1} &= -v\left(q^l + \frac{h}{2}v^l\right)^{\mathrm{T}} v^l \end{aligned} \tag{23}$$

where $\tilde{M}$ and $\tilde{k}$ are given by (13)–(15).

We note that both time-stepping schemes need initial values q^0, v^0 satisfying (20). The problem of obtaining consistent initial values q^0, v^0 is well studied and will not be addressed here.

3. THE TRUNCATION ERROR

In this section, we will prove some results about the truncation error of the integrations methods considered in Section 2. In our proofs we will use the following generalization of the classical results concerning the error of the trapezoidal and the midpoint quadrature rules:

Lemma 3.1
If $\psi(t)$ is continuously differentiable and $\phi(t)$ is twice continuously differentiable on the interval $[a, b]$, then there are points $\tau_1, \tau_2\tau_3\tau_4$ in $[a, b]$ such that

$$\int_a^b \left(\phi(t) - \frac{\phi(a) + \phi(b)}{2}\right)\psi(t)\,\mathrm{d}t = \frac{1}{12}(\phi'(c)\psi'(\tau_1) - \phi''(\tau_2)\psi(c))(b-a)^3 + o((b-a)^4)$$

$$\int_a^b \left(\phi(t) - \phi\left(\frac{(a+b)}{2}\right)\right)\psi(t)\,\mathrm{d}t = \frac{1}{12}\left(\phi'(c)\psi'(\tau_3) + \frac{1}{2}\psi(c)\phi''(\tau_4)\right)(b-a)^3 + o((b-a)^4)$$

where $c = (a+b)/2$ is the midpoint of the interval $[a, b]$.

Proof
Recalling the definition of the divided differences

$$[s,t;\phi]=\frac{\phi(t)-\phi(s)}{t-s},\quad [t,t;\phi]=\phi'(t),\quad [r,s,t;\phi]=\frac{[r,t;\phi]-[r,s;\phi]}{t-s}$$

and using their properties (see e.g. Reference [18]) we deduce that

$$\begin{aligned}\phi(t)-(\phi(a)+\phi(b))/2&=\phi(a)+[a,b;\phi](t-a)+[a,t,b;\phi](t-a)(t-b)-(\phi(a)+\phi(b))/2\\&=[a,b;\phi](t-a)-[a,b;\phi](b-a)/2+[a,t,b;\phi](t-a)(t-b)\\&=[a,b;\phi](t-c)+[a,t,b;\phi](t-a)(t-b)\\\phi(t)-\phi(c)&=\phi'(c)(t-c)+[c,c,t;\phi](t-c)^2\\\psi(t)&=\psi(c)+[c,t;\psi](t-c)\end{aligned}$$

Therefore there are points $\tau_i\in[a,b]$, $i=1,\dots,8$ such that

$$\begin{aligned}&\int_a^b\left(\phi(t)-\frac{\phi(a)+\phi(b)}{2}\right)\psi(t)\,\mathrm{d}t\\&=[a,b;\phi]\psi(c)\int_a^b(t-c)\,\mathrm{d}t+[a,b;\phi]\int_a^b[c,t;\psi](t-c)^2\,\mathrm{d}t\\&\quad+\psi(c)\int_a^b[a,t,b;\phi](t-a)(t-b)\,\mathrm{d}t+\int_a^b[a,t,b;\phi][c,t;\psi](t-c)(t-a)(t-b)\,\mathrm{d}t\\&=[a,b;\phi]\psi'(\tau_1)\int_a^b(t-c)^2\,\mathrm{d}t+\frac{1}{2}\psi(c)\phi''(\tau_2)\int_a^b(t-a)(t-b)\,\mathrm{d}t\\&\quad+[a,\tau_5,b;\phi][c,\tau_6;\psi]\int_a^{(a+b)/2}(t-c)(t-a)(t-b)\,\mathrm{d}t\\&\quad+[a,\tau_6,b;\phi][c,\tau_6;\psi]\int_{(a+b)/2}^b(t-c)(t-a)(t-b)\,\mathrm{d}t\\&=\frac{1}{12}([a,b;\phi]\psi'(\tau_1)-\phi''(\tau_2)\psi(c))(b-a)^3+\frac{1}{64}([a,\tau_5,b;\phi][c,\tau_5;\psi]\\&\quad-[a,\tau_6,b;\phi][c,\tau_6;\psi])(b-a)^4\\&=\frac{1}{12}(\phi'(c)\psi'(\tau_1)-\phi''(\tau_2)\psi(c))(b-a)^3-\frac{\psi'(\tau_1)}{24}([a,c,b;\phi]-[a,c,c;\phi])(b-a)^4\end{aligned}$$

$$+\frac{1}{64}([a,\tau_5,b;\phi][c,\tau_5;\psi]-[a,\tau_6,b;\phi][c,\tau_6;\psi])(b-a)^4$$

$$\times\int_a^b\left(\phi(t)-\phi\left(\frac{(a+b)}{2}\right)\right)\psi(t)\,\mathrm{d}t=\phi'(c)\psi(c)\int_a^b(t-c)\,\mathrm{d}t$$

$$+\int_a^b(\phi'(c)[c,t;\psi]+\psi(c)[c,c,t;\phi])(t-c)^2\,\mathrm{d}t+\int_a^b[c,c,t;\phi][c,t;\psi](t-c)^3\,\mathrm{d}t$$

$$=\frac{1}{12}\left(\phi'(c)\psi'(\tau_3)+\frac{1}{2}\psi(c)\phi''(\tau_4)\right)(b-a)^3-\frac{1}{64}([c,c,\tau_7;\phi][c,\tau_7;\psi]$$

$$-[c,c,\tau_8;\phi][c,\tau_8;\psi])(b-a)^4 \qquad \square$$

The following theorem shows that the truncation error of time-stepping scheme (11) is of order three.

Theorem 3.2
The exact solution of the differential equation (4) satisfies the relations

$$q(t_{l+1})=q(t_l)+\frac{h}{2}(v(t_l)+v(t_{l+1}))+O(h^3) \tag{24}$$

$$\tilde{\bar{M}}v(t_{l+1})=\tilde{\bar{M}}v(t_l)+h\tilde{\bar{k}}+O(h^3) \tag{25}$$

where

$$\tilde{\bar{M}}=\bar{\bar{M}}-\frac{h}{2}k_v(t_{l+1},q(t_l),v(t_l))-\frac{h^2}{4}k_q(t_{l+1},q(t_l),v(t_l)) \tag{26}$$

$$\bar{\bar{M}}=\frac{1}{2}(M(t_{l+1},q(t_l)+hv(t_l))+M(t_l,q(t_l)))\quad\text{or}\quad\bar{\bar{M}}=M\left(t_l+\frac{h}{2},q(t_l)+\frac{h}{2}v(t_l)\right) \tag{27}$$

$$\tilde{\bar{k}}=\frac{1}{2}\left(k(t_l,q(t_l),v(t_l))+k(t_{l+1},q(t_l),v(t_l))+\frac{h}{2}k_q(t_{l+1},q(t_l),v(t_l))\right) \tag{28}$$

Proof
By using Lemma 3.1 we deduce that the exact solution of (4) satisfies

$$q(t_{l+1})-q(t_l)=\int_{t_l}^{t_{l+1}}v(t)\,\mathrm{d}t=h\frac{v(t_{l+1})+v(t_l)}{2}+O(h^3) \tag{29}$$

$$\begin{aligned}\int_{t_l}^{t_{l+1}} M(t,q(t))v'(t)\,\mathrm{d}t &= \frac{1}{2}\int_{t_l}^{t_{l+1}} (M(t_l,q(t_l)) + M(t_{l+1},q(t_{l+1})))v'(t)\,\mathrm{d}t \\ &\quad + \int_{t_l}^{t_{l+1}} \left(M(t,q(t)) - \frac{1}{2}(M(t_l,q(t_l)) + M(t_{l+1},q(t_{l+1}))) \right) v'(t)\,\mathrm{d}t \\ &= \frac{1}{2}(M(t_l,q(t_l)) + M(t_{l+1},q(t_{l+1})))(v(t_{l+1}) - v(t_l)) + O(h^3) \\ &= \frac{1}{2}(M(t_l,q(t_l)) + M(t_{l+1},q(t_l) + hv(t_l)))(v(t_{l+1}) - v(t_l)) + O(h^3)\end{aligned}$$

$$\begin{aligned}\int_{t_l}^{t_{l+1}} M(t,q(t))v'(t)\,\mathrm{d}t &= \int_{t_l}^{t_{l+1}} M\left(t_l + \frac{h}{2}, q\left(t_l + \frac{h}{2}\right)\right) v'(t)\,\mathrm{d}t \\ &\quad + \int_{t_l}^{t_{l+1}} \left(M(t,q(t)) - M\left(t_l + \frac{h}{2}, q\left(t_l + \frac{h}{2}\right)\right)\right) v'(t)\,\mathrm{d}t \\ &= M\left(t_l + \frac{h}{2}, q\left(t_l + \frac{h}{2}\right)\right)(v(t_{l+1}) - v(t_l)) + O(h^3) \\ &= M\left(t_l + \frac{h}{2}, q(t_l) + \frac{h}{2}v(t_l)\right)(v(t_{l+1}) - v(t_l)) + O(h^3)\end{aligned}$$

Using the Taylor expansion formula gives

$$\begin{aligned}k(t_{l+1},q(t_{l+1}),v(t_{l+1})) &= k(t_{l+1},q(t_l),v(t_l)) + k_q(t_{l+1},q(t_l),v(t_l))(q(t_{l+1}) - q(t_l)) \\ &\quad + k_v(t_{l+1},q(t_l),v(t_l))(v(t_{l+1}) - v(t_l)) + O(h^2) \\ &= k(t_{l+1},q(t_l),v(t_l)) + \frac{h}{2}k_q(t_{l+1},q(t_l),v(t_l))(v(t_{l+1}) + v(t_l)) \\ &\quad + k_v(t_{l+1},q(t_l),v(t_l))(v(t_{l+1}) - v(t_l)) + O(h^2) \\ &= k(t_{l+1},q(t_l),v(t_l)) + \frac{h}{2}\tilde{k}_q^l(v(t_{l+1}) + v(t_l)) \\ &\quad + \tilde{k}_v^l(v(t_{l+1}) - v(t_l)) + O(h^2)\end{aligned}$$

so that by applying the trapezoidal quadrature scheme we get

$$\int_{t_l}^{t_{l+1}} k(t,q(t),v(t))\,\mathrm{d}t = \frac{h}{2}(k(t_l,q(t_l),v(t_l)) + k(t_{l+1},q(t_{l+1}),v(t_{l+1}))) + O(h^3)$$

$$= hk(t_l, q(t_l), v(t_l)) + \frac{h^2}{4} k_q(t_{l+1}, q(t_l), v(t_l))(v(t_{l+1}) + v(t_l))$$

$$+ \frac{h}{2} k_v(t_{l+1}, q(t_l), v(t_l))(v(t_{l+1}) - v(t_l)) + O(h^3)$$

The statement of our theorem follows from the above relations. □

Corollary 3.3
Assume that the mass matrix $M(t, q)$ is symmetric and uniformly positive definite in the sense that its eigenvalues are bounded below by some positive quantity for all (t, q). Let $q(t), v(t)$ be a solution of the differential equation (4) and consider the time-stepping scheme (11) with $q^l, v^l, \tilde{k}_q^l, \tilde{k}_v^l$ satisfying (16). Then the following relation holds between the approximate solution and the exact solution at time t_{l+1}:

$$q(t_{l+1}) = q^{l+1} + O(h^3), \quad v(t_{l+1}) = v^{l+1} + O(h^3) \tag{30}$$

Proof
Using the notation from (11) and (24) we obtain

$$\tilde{M}(v(t_{t+1}) - v(t_l)) = (\tilde{M} - \tilde{\tilde{M}})(v(t_{t+1}) - v(t_l)) + \tilde{\tilde{M}}(v(t_{t+1}) - v(t_l))$$

$$= \tilde{\tilde{M}}(v(t_{t+1}) - v(t_l)) + O(h^3) = h\tilde{\tilde{k}} + O(h^3) = h\tilde{k} + O(h^3)$$

Since $\tilde{M}(v^{l+1} - v^l) = h\tilde{k}$, we deduce that $\tilde{M}(v(t_{l+1}) - v^{l+1}) = O(h^3)$, and by using the fact that $M(t, q)$ is symmetric and uniformly positive definite we obtain $v(t_{l+1}) = v^{l+1} + O(h^3)$. We also have

$$q(t_{l+1}) - q^{l+1} = \frac{h}{2}(v(t_{l+1}) - v^{l+1}) + O(h^3) = O(h^3)$$

which completes the proof. □

We can obtain similar results for the DAE (18) provided the equality constraints are linear. However, for general non-linear constraints the numerical integrator (23) may experience some order reduction.

Theorem 3.4
The exact solution of the DAE (18) satisfies the relations:

$$q(t_{l+1}) = q(t_l) + \frac{h}{2}(v(t_l) + v(t_{l+1})) + O(h^3) \tag{31}$$

$$\tilde{\tilde{M}} v(t_{l+1}) - hv\left(q(t_l) + \frac{h}{2} v(t_l)\right) c_v\left(t_l + \frac{h}{2}\right) = \tilde{\tilde{M}} v(t_l) + h\tilde{\tilde{k}} + O(h^3) \tag{32}$$

$$\tilde{\tilde{M}} v(t_{l+1}) - \frac{h}{2} v(q(t_l) + hv(t_l)) c_v(t_{l+1}) = \tilde{\tilde{M}} v(t_l) + \frac{h}{2} v(q(t_l)) c_v(t_l) + h\tilde{\tilde{k}} + O(h^3) \tag{33}$$

$$v\left(q(t_l)+\frac{h}{2}v(t_l)\right)^{\mathrm{T}}v(t_{l+1})=-v\left(q(t_l)+\frac{h}{2}v(t_l)\right)^{\mathrm{T}}v(t_l)+\varpi(h) \tag{34}$$

$$v(q(t_l)+hv(t_l))^{\mathrm{T}}v(t_{l+1})=\varpi(h) \tag{35}$$

where $\tilde{\tilde{M}}, \tilde{\tilde{k}}$ are defined in (26)–(28), and

$$\varpi(h)=\begin{cases}0 & \text{if } \Theta(q) \text{ is linear}\\ O(h^2) & \text{if } \Theta(q) \text{ is non-linear}\end{cases} \tag{36}$$

Proof

Relation (31) has already been proved in Theorem 3.2. Relations (32) and (33) follow from the proof of Theorem 3.2 and the fact that the midpoint and the trapezoidal quadrature rules applied to $v(q(t))c_v(t)$ give

$$\begin{aligned}\int_{t_l}^{t_{l+1}} v(q(t))c_v(t)\,\mathrm{d}t &= hv\left(q\left(t_l+\frac{h}{2}\right)\right)c_v\left(t_l+\frac{h}{2}\right)+O(h^3)\\ &= hv\left(q\left(t_l+\frac{h}{2}v(t_l)\right)\right)c_v\left(t_l+\frac{h}{2}\right)+O(h^3)\\ &= \frac{h}{2}(v(q(t_l))c_v(t_l)+v(q(t_{l+1}))c_v(t_{l+1}))+O(h^3)\end{aligned}$$

The exact solution satisfies $v(q(t))v(t)=0$ for any t. If $\Theta(q)$ is linear, then $v(q)$ is a constant matrix so that (34) and (35) are satisfied with $\varpi(h)\equiv 0$. If $\Theta(q)$ is non-linear then we have

$$\begin{aligned}v\left(q(t_l)+\frac{h}{2}v(t_l)\right)\frac{v(t_l)+v(t_{l+1})}{2} &= v\left(q(t_l)+\frac{h}{2}v(t_l)\right)v\left(t_l+\frac{h}{2}\right)+O(h^2)\\ &= v\left(q\left(t_l+\frac{h}{2}\right)\right)v\left(t_l+\frac{h}{2}\right)+O(h^2)=O(h^2)\end{aligned}$$

$$v(q(t_l)+hv(t_l))^{\mathrm{T}}v(t_{l+1})=v(q(t_{l+1}))^{\mathrm{T}}v(t_{l+1})+O(h^2)$$

The proof is complete. □

In Corollary 3.3 we have assumed that the mass matrix $M(q,t)$ is uniformly positive definite. In the presence of equality constraints we assume that $M(q,t)$ is uniformly positive definite on the nullspace of the Jacobian of the constraints, i.e. on the tangent hyper plane to the manifold defined by the equality constraints. As mentioned before, we also have to assume that this Jacobian is full rank. To be more precise we assume that there is a matrix valued mapping $\kappa:\mathbb{R}^s\to\mathbb{R}^{s\times s}$ such that

$$v(q)^{\mathrm{T}}\kappa(q)=0,\quad \kappa(q)^{\mathrm{T}}\kappa(q)=I,\quad \inf_{q}\ \sigma(v(q)^{\mathrm{T}}v(q))>0,\quad \inf_{t,q}\ \sigma(\kappa(q)^{\mathrm{T}}M(t,q)\kappa(q))>0 \tag{37}$$

where $\sigma(m)$ denotes the spectrum of the matrix m. In other words the columns of the matrix $\kappa(q)$ form an orthonormal basis of $v(q)$ (i.e. the tangent hyper plane at q to the position constraint manifold [19, 20]) and the eigenvalues of the symmetric positive definite matrices $v(q)^{\mathrm{T}}M(t,q)v(q)$ and $\kappa(q)^{\mathrm{T}}M(t,q)\kappa(q)$ are bounded below by some positive constant. This implies that the operator norms of their inverses are bounded, i.e.

$$\sup_{q} \|(v(q)^{\mathrm{T}}v(q))^{-1}\| < \infty, \quad \sup_{t,q} \|(\kappa(q)^{\mathrm{T}}M(t,q)\kappa(q))^{-1}\| < \infty \tag{38}$$

Corollary 3.5
Assume that conditions (37) are satisfied and let $q(t), v(t), c_v(t)$ be a solution of DAE (18). Consider the time-stepping scheme (23) with $q^l, v^l, \tilde{k}_q^l, \tilde{k}_v^l$ satisfying (16). Then the following relations hold:

$$q(t_{l+1}) = q^{l+1} + O(h^3), \quad v(t_{l+1}) = v^{l+1} + O(h^{3-\iota}), \quad c_v\left(t_l + \frac{h}{2}\right) = c_v^{l+1/2} + O(h^{2-\iota}) \tag{39}$$

where

$$\iota = \begin{cases} 0 & \text{if } \Theta(q) \text{ is linear} \\ 1 & \text{if } \Theta(q) \text{ is non-linear} \end{cases} \tag{40}$$

Proof
Using the notation (13)–(15) and Equations (31), (32), (34) we obtain

$$\begin{aligned} \tilde{M}(v(t_{l+1}) - v(t_l)) &= (\tilde{M} - \tilde{\tilde{M}})(v(t_{l+1}) - v(t_l)) + \tilde{\tilde{M}}(v(t_{l+1}) - v(t_l)) \\ &= \tilde{\tilde{M}}(v(t_{l+1}) - v(t_l)) + O(h^3) \\ &= hv\left(q(t_l) + \frac{h}{2}v(t_l)\right) c_v\left(t_l + \frac{h}{2}\right) + h\tilde{\tilde{k}} + O(h^3) \\ &= hv\left(q(t_l) + \frac{h}{2}v(t_l)\right) c_v\left(t_l + \frac{h}{2}\right) + h\tilde{k} + O(h^3) \end{aligned}$$

Hence

$$\tilde{M}v(t_{l+1}) = \tilde{M}v(t_l) + hv\left(q(t_l) + \frac{h}{2}v(t_l)\right) c_v\left(t_l + \frac{h}{2}\right) + h\tilde{k} + O(h^3) \tag{41}$$

By subtracting (23) from the system formed by (31), (41) and (34), we get

$$q(t_{l+1}) - q^{l+1} = \frac{h}{2}(v(t_{l+1}) - v^{l+1}) + O(h^3) \tag{42}$$

$$\tilde{M}(v(t_{l+1}) - v^{l+1}) - hv\left(q^l + \frac{h}{2}v^l\right)\left(c_v\left(t_l + \frac{h}{2}\right) - c_v^{l+1/2}\right) = O(h^3) \tag{43}$$

$$v\left(q^l + \frac{h}{2}v^l\right)^{\mathrm{T}}(v(t_{l+1}) - v^{l+1}) = \varpi(h) \tag{44}$$

where $\varpi(h)$ is defined in (36). Let us denote $A = v(q^l + (h/2)v^l)$ and $B = \kappa(q^l + (h/2)v^l)$. Then there are unique vectors a and b such that $v(t_{l+1}) - v^{l+1} = Aa + Bb$ and from (42)–(44) we obtain

$$q(t_{l+1}) - q^{l+1} = \frac{h}{2}(Aa + Bb) + O(h^3)$$

$$A^{\mathrm{T}}\tilde{M}Aa + A^{\mathrm{T}}\tilde{M}Bb - hA^{\mathrm{T}}A\left(c_v\left(t_l + \frac{h}{2}\right) - c_v^{l+1/2}\right) = O(h^3)$$

$$B^{\mathrm{T}}\tilde{M}Aa + B^{\mathrm{T}}\tilde{M}Bb = O(h^3)$$

$$A^{\mathrm{T}}Aa = \varpi(h)$$

If $\Theta(q)$ is linear, then by using the above relations in reverse order, we get successively

$$a = 0, \quad b = O(h^3), \quad c_v\left(t_l + \frac{h}{2}\right) - c_v^{l+1/2} = O(h^2), \quad q(t_{l+1}) - q^{l+1} = O(h^3)$$

If $\Theta(q)$ is non-linear, then we get

$$a = O(h^2), \quad b = O(h^2), \quad c_v\left(t_l + \frac{h}{2}\right) - c_v^{l+1/2} = O(h), \quad q(t_{l+1}) - q^{l+1} = O(h^3)$$

which completes the proof of our corollary. □

4. THE GLOBAL ERROR

Let $q(t), v(t)$ be a solution of the ODE (4) on the interval $[t_0, T]$. For any $h>0$ consider the points

$$t_l = t_0 + lh, \quad l = 0, 1, \ldots, L = \left\lfloor \frac{T - t_0}{h} \right\rfloor \tag{45}$$

The following theorem shows that the distance between the values of the exact solution at these points and the approximate solutions given by the time-stepping scheme (23) is bounded by Ch^2, where C is some constant depending on T, but not on l. This fact will be denoted by the shorthand

$$q(t_l) = q^l + O(h^2), \quad v(t_l) = v^l + O(h^2), \quad l = 0, 1, \ldots, L \tag{46}$$

Theorem 4.1
Assume that the mass matrix $M(t, q)$ is symmetric and uniformly positive definite in the sense that its eigenvalues are bounded below by some positive quantity for all (t, q). Let $q(t), v(t)$ be a solution of the differential equation (4) and consider the time-stepping scheme (11) with starting point

$$q^0 = q(t_0) + O(h^2), \quad v^0 = v(t_0) + O(h^2) \tag{47}$$

and with $\widetilde{k}_q^l, \widetilde{k}_v^l$ satisfying

$$\tilde{k}_q^l = k_q(t_{l+1}, q^l, v^l) + O(h), \quad \tilde{k}_v^l = k_v(t_{l+1}, q^l, v^l) + O(h^2) \tag{48}$$

Then the global estimates (46) hold.

Proof
Let us denote

$$e_q^l = q(t_l) - q^l, \quad e_v^l = v(t_l) - v^l, \quad \varepsilon_q^l = \|e_q^l\|, \quad \varepsilon_v^l = \|e_v^l\|$$

From Theorem 3.2 we have

$$\tilde{\tilde{M}}(v(t_{t+1}) - v(t_l)) - h\tilde{\tilde{k}} = O(h^3)$$

The left-hand side of this equation can be written successively as

$$\begin{aligned}&\tilde{M}(v(t_{t+1}) - v(t_l)) - h\tilde{k} + (\tilde{\tilde{M}} - \tilde{M})(v(t_{t+1}) - v(t_l)) - h(\tilde{\tilde{k}} - \tilde{k})\\ &= \tilde{M}(v(t_{t+1}) - v(t_l)) - h\tilde{k} + O(h\varepsilon_q^l + h\varepsilon_v^l) + O(h^3)\end{aligned}$$

Hence

$$\tilde{M}(v(t_{t+1}) - v(t_l)) - h\tilde{k} = O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3) \tag{49}$$

and by using (11) we deduce that

$$\tilde{M}(e_v^{l+1} - e_v^l) = O(h\varepsilon_q^l + h\varepsilon_v^l) + O(h^3)$$

Since $M(t, q)$ is uniformly positive definite it follows that

$$e_v^{l+1} = e_v^l + O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3)$$

Since

$$q(t_{l+1}) = q(t_l) + h\frac{v(t_l) + v(t_{l+1})}{2} + O(h^3), \quad q^{l+1} = q^l + h\frac{v^l + v^{l+1}}{2}$$

we obtain

$$e_q^{l+1} = e_q^l + h\frac{e_v^l + e_v^{l+1}}{2} + O(h^3) = e_q^l + O(h^2\varepsilon_q^l) + O(h\varepsilon_v^l)$$

By denoting

$$\varepsilon_l = \max\{\varepsilon_q^l, \varepsilon_v^l\}$$

we deduce that there are constants ω and χ such that

$$\varepsilon_{l+1} \leqslant (1 + \omega h)\varepsilon_l + \chi h^3, \quad l = 0, 1, \ldots, L$$

Therefore

$$\varepsilon_l \leqslant (1+\omega h)^l \varepsilon_0 + \chi h^3 \sum_{j=0}^{l-1} (1+\omega h)^i = (1+\omega h)^l \varepsilon_0 + \chi h^3 \sum_{j=0}^{l-1} (1+\omega h)^i$$

$$\leqslant (1+\omega h)^L \varepsilon_0 + \frac{(1+\omega h)^L - 1}{\omega} \chi h^2 \leqslant (\varepsilon_0 + \chi (T - t_0) h^2) e^{\omega (T - t_0)}, \quad l = 0, 1, \ldots, L$$

and the conclusion of our theorem follows in view of (47). □

We end this section by showing that in case the equality constraints are linear then the global error of the time-stepping scheme (23) is of order two both in positions and velocities. If the equality constraints are linear then the matrix valued mappings $v(q)$ and $\kappa(q)$ appearing in (37) are constant. If we denote $A = v(q),\ \kappa(q) = B$, we have $A^{\mathrm{T}} B = 0,\ B^{\mathrm{T}} B = I$, and (37) reduces to the condition

$$\sup_{t,\, q} \| (B^{\mathrm{T}} M(t, q) B)^{-1} \| < \infty \tag{50}$$

which is weaker than the uniform positive definiteness of $M(t, q)$.

Theorem 4.2
Let $q(t), v(t), c_v(t)$ be a solution of DAE (18) and assume that condition (50) is satisfied. Then the approximate solution given by the time-stepping scheme (23) with starting point

$$q^0 = q(t_0) + O(h^2), \quad v^0 = v(t_0) + O(h^2) \tag{51}$$

and with $\widetilde{k}_q^l, \widetilde{k}_v^l$ satisfying (48), admits the following global error estimates:

$$q^l = q(t_l) + O(h^2), \quad v^l = v(t_l) + O(h^2), \quad c_v^{l+1/2} = c_v \left(t_l + \frac{h}{2} \right) + O(h^2)$$

Proof
By denoting

$$e_q^l = q(t_l) - q^l, \quad e_v^l = v(t_l) - v^l, \quad e_c^l = c_v \left(t_l + \frac{h}{2} \right) - c_v^{l+1/2},$$

$$\varepsilon_q^l = \| e_q^l \|, \quad \varepsilon_v^l = \| e_v^l \|, \quad \varepsilon_c^l = \| e_c^l \|$$

and using an argument similar to the one employed in the proof of Theorem 4.1 we deduce that the exact solution of (18) satisfies

$$\tilde{M}(v(t_{l+1}) - v(t_l)) - hAc_v \left(t_l + \frac{h}{2} \right) - h\tilde{k} = O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3)$$

Since for the approximate solution provided by (23) we have

$$\tilde{M}(v^{l+1} - v^l) - hAc_v^{l+1/2} - h\tilde{k} = 0$$

it follows that

$$\tilde{M}(e_v^{l+1} - e_v^l) - hAe_c^l = O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3) \tag{52}$$

Using the fact that the exact solution satisfies $A^{\mathrm{T}}v(t_l) = 0,\ l = 0, \ldots, L$ we deduce that for the approximate solution we have

$$A^{\mathrm{T}}v^l = -A^{\mathrm{T}}v^{l-1} = \cdots = (-1)^l A^{\mathrm{T}}v^0 = (-1)^l A^{\mathrm{T}}v(t_0) + O(h^3) = O(h^3)$$

Therefore the co-ordinate a^l from the decomposition $e_v^l = Aa^l + Bb^l$ satisfies

$$a^l = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}e_v^l = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}(v(t_l) - v^l) = O(h^3), \quad l = 0, \ldots, L$$

and by multiplying (52) with B^{T} we obtain

$$B^{\mathrm{T}}\tilde{M}B(b^{l+1} - b^l) = O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3)$$

According to (50) this implies $b^{l+1} - b^l = O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3)$ and, since $a^l = O(h^3)$, we deduce that

$$e_v^{l+1} - e_v^l = O(h\varepsilon_q^l) + O(h\varepsilon_v^l) + O(h^3)$$

As in the proof of Theorem 4.1 it follows that $\varepsilon_v^l = O(h^2)$ and $\varepsilon_v^l = O(h^2)$. But then from the above equation we have $e_v^{l+1} - e_v^l = O(h^3)$, and by multiplying (52) with A^{T} we deduce that $\varepsilon_c^l = O(h^2)$. □

The result of the preceding theorem can be extended to DAE of the following type:

$$M(q)\frac{\mathrm{d}^2q}{\mathrm{d}t^2} - \sum_{i=1}^{m} v^{(i)}(q)c_v^{(i)} - \sum_{j\in\mathscr{A}}\left(n^{(j)}(q)c_n^{(j)} + \sum_{j_f\in\mathscr{B}_j} d_{j_f}^{(j)}(q)\beta_{j_f}^{(j)}\right) = k\left(t, q, \frac{\mathrm{d}q}{\mathrm{d}t}\right) \tag{53a}$$

$$\Theta^{(i)}(q) = 0, \quad i = 1, 2, \ldots, m \tag{53b}$$

$$\Phi^{(j)}(q) = 0, \quad j \in \mathscr{A} \tag{53c}$$

$$L_1(q)v + L_2(c_n^{(j)})_{j\in\mathscr{A}} + L_3(\beta_{j_f}^{(j)})_{j\in\mathscr{A},\, j_f\in\mathscr{B}_j} + L_4(\lambda^{(j)})_{j\in\mathscr{A}} = 0 \tag{53d}$$

where $L_1(q)$, L_2, L_3, L_4 are matrices of the appropriate size, and $L_1(q)$ depends smoothly on its parameter. Such DAE appear by reducing the differential complementarity problem (59) to a DAE by making an appropriate selection of the active set in its complementarity constraints. For that case, the result of Theorem 4.2 still applies, though the proof is complicated by the additional structure. The corresponding result is the following.

Theorem 4.3
Let $q(t), v(t), c_v(t), c_n(t), \beta(t), \lambda(t)$ be a solution of DAE (53) and assume that the matrix

$$\begin{bmatrix} M & -\nabla\Theta & -\nabla\Phi_{\mathscr{A}} & -\tilde{D}_{\mathscr{A}} & 0 \\ \nabla\Theta^{\mathrm{T}} & 0 & 0 & 0 & 0 \\ \nabla\Phi_{\mathscr{A}}^{\mathrm{T}} & 0 & 0 & 0 & 0 \\ L_1(q) & 0 & L_2 & L_3 & L_4 \end{bmatrix} \tag{54}$$

is constant, except for the M block, that is, all constraints are linear, and has full rank. Here, we denote by $D_{\mathscr{A}} = [d_{j_f}^{(j)}]_{j\in\mathscr{A}, j_f\in\mathscr{B}_j}$. Then the approximate solution given by the linearized trapezoidal method, which is the method (61)–(66) specialized to our choice of active set, with starting point

$$q^0 = q(t_0) + O(h^2), \quad v^0 = v(t_0) + O(h^2)$$

and with $\widetilde{k}_q^l, \widetilde{k}_v^l$ satisfying (48), admits the following global error estimates:

$$q^l = q(t_l) + O(h^2), \quad v^l = v(t_l) + O(h^2)$$

We note that the full rank condition on the matrix (54) reduces to (50) in the joint-only case.

5. THE CONSTRAINTS

In this section we will use the approximate linearly implicit trapezoidal method (11) to construct a time-stepping method for stiff rigid-multi-body systems with contact and friction. Our construction will be done along the lines of References [7, 8]. We assume that the state of the system of rigid bodies can be described by a generalized position vector $q \in \mathbb{R}^s$ and a generalized velocity vector $v \in \mathbb{R}^s$. We assume that the system is subject to equality, non-interpenetration, contact, and frictional constraints.

The *equality constraints* arise usually in the presence of joints and they have been described in Section 2 (see (17)).

The *non-interpenetration constraints* are generated by the rigid body hypothesis according to which the bodies comprising the system cannot penetrate each other. We assume that for any pair of bodies we can define a continuous signed distance function $\Phi_j(q)$ so that the non-interpenetration constraints can be written as

$$\Phi^{(j)}(q) \geqslant 0, \quad j = 1, 2, \ldots, p \tag{55}$$

where p is the number of pairs of bodies of the system that could get in contact, which in most applications is substantially smaller than the number of all possible pairs of bodies. Although such continuous functions cannot be determined in the most general case, under some weak assumptions it is possible to define them at least in a neighbourhood of all contact configurations [13, 21, 22], which is sufficient for our developments.

The *contact constraints* are complementarity constraints. If two bodies are in contact then $\Phi^{(j)}(q) = 0$ for some index j, and a 'normal' force $c_n^{(j)} n^{(j)}(q)$ (where $n^{(j)}(q) = \nabla_q \Phi^{(j)}(q)$ is

the gradient of $\Phi^{(j)}(q)$) will act at the contact. The force can be only a compression force, which means that $c_n^{(j)} \geqslant 0$. If the two bodies are not in contact, i.e. if $\Phi^{(j)}(q) > 0$, then there is no normal force at contact j, so that $c_n^{(j)} \geqslant 0$. Therefore the complementarity constraints become

$$\Phi^{(j)}(q) \geqslant 0, \quad c_n^{(j)} \geqslant 0, \quad \Phi^{(j)}(q) c_n^{(j)} = 0, \quad j = 1, 2, \ldots, p \tag{56}$$

By ignoring the subscript (j), the above conditions can be written simply as

$$0 \leqslant \Phi(q) \perp c_n \geqslant 0 \tag{57}$$

The *frictional constraints* connect the tangential force, the normal force, and the velocity at contacts. In what follows we adopt the description of frictional constraints from Reference [23]. They are imposed at each contact (j). To simplify notation we omit the superscript (j), although all the quantities refer to the (j)th contact.

The set of possible friction forces, for unitary normal force multiplier, is given by

$$FC_0(q) = \{\bar{D}(q)\bar{\beta} | \bar{\beta} \in \mathbb{R}^d, \psi(\bar{\beta}) \leqslant \mu\}$$

where $\bar{D}(q)$ is a given $d \times s$ matrix, $\psi(\bar{\beta})$ is a convex, positively homogeneous, coercive function, and μ is a non-negative friction coefficient. The total force at the contact belongs to the friction cone

$$FC(q) = c_n(n(q) + FC_0(q)) = \{c_n n(q) + \bar{D}(q)\bar{\beta} | \bar{\beta} \in \mathbb{R}^d, \psi(\bar{\beta}) \leqslant \mu c_n\}$$

If $\bar{D}(q)$ consists of two orthogonal columns that span the friction plane and $\psi(\bar{\beta}) = \|\bar{\beta}\|_2$, then $FC(q)$ becomes the classical circular friction cone [24]. The current representation, however, also covers the representation in global co-ordinates, where $n(q)$ is not necessarily orthogonal to $\bar{D}(q)$ [21].

According to the maximal dissipation principle we choose $\bar{\beta}$ to maximize the dissipation rate $-v^{\mathrm{T}}\bar{D}(q)\bar{\beta}$ over $\bar{D}(q)\bar{\beta} \in c_n FC_0(q)$, which defines $\bar{\beta}$ as the solution of the following optimization problem:

$$\min_{\bar{\beta} \in \mathbb{R}^d} v^{\mathrm{T}}\bar{D}(q)\bar{\beta} \quad \text{subject to } \psi(\bar{\beta}) \leqslant \mu c_n \tag{58}$$

Since our purpose is to construct a time-stepping scheme that requires the solution of a linear complementarity problem at each time step we use a polyhedral approximation of the friction cone [7, 23, 25]. This approximation is generated by the set

$$\{n(q) + d_i(q), i = 1, 2, \ldots, m_C\} \quad \text{with } D(q) = [d_1(q), d_2(q), \ldots, d_{m_C}(q)]$$

where $d_i(q)$ is a collection of direction vectors in $FC_0(q)$. The columns of $D(q)$ are chosen to be balanced in the sense that for any i there is a j such that $d_j(q) = -d_i(q)$ [25], so that one non-negative component β_i is associated with every column $d_i(q)$ of $D(q)$. If we denote the vector of tangential forces by $\beta = (\beta_1, \beta_2, \ldots, \beta_{m_C})^{\mathrm{T}}$, then the total tangential force can be written as $D(q)\beta$ [7, 25]. In terms of these variables, the frictional constraints, including the maximum dissipation principle (58), can be expressed in terms of the following complementarity conditions [7, 23, 25]:

$$0 \leqslant D(q)^{\mathrm{T}} v + \lambda e \perp \beta \geqslant 0, \quad 0 \leqslant \mu c_n - e^{\mathrm{T}}\beta \perp \lambda \geqslant 0$$

If there is relative motion at the contact, $\|D(q)^{\mathrm{T}}v\| \neq 0$, then λ is approximately equal to the norm of the tangential velocity at the contact [25, 26].

When using the maximum dissipation principle in a time stepping scheme, the issue is what is the value of the velocity that should be used in the complementarity constraints. Choosing $v = v^l$, the current value of the velocity, results in an explicit scheme that is unstable in many interesting situations. In the recently developed time stepping scheme that use a linear complementarity subproblem [7, 8, 13, 25], the choice is $v = v^{(l+1)}$ (the future value of the velocity) that makes the treatment of friction implicit. The good stability properties of such a choice has been demonstrated both theoretically and practically. In this work, since we intend to develop a linearized trapezoidal scheme, the appropriate choice is $v = (v^{(l+1)} + v^{(l)})/2$. When matrix $D(q)$ is constant, this can be shown to correspond to a trapezoidal discretization of the integral formulation of the time stepping scheme, which makes our frictional approach well-suited for working with the trapezoidal discretization of the Newton equation.

6. THE TIME-STEPPING SCHEME

If we combine the Newton equation of dynamics with the constraints described in the previous section we obtain the following differential complementarity problem (DCP):

$$
\begin{aligned}
M(q)\frac{\mathrm{d}^2q}{\mathrm{d}t^2} - \sum_{i=1}^{m} v^{(i)}(q)c_v^{(i)} - \sum_{j=1}^{p} (n^{(j)}(q)c_n^{(j)} + D^{(j)}(q)\beta^{(j)}) &= k\left(t, q, \frac{\mathrm{d}q}{\mathrm{d}t}\right) \\
\Theta^{(i)}(q) = 0, \quad i = 1, 2, \ldots, m & \\
0 \leqslant \Phi^{(j)}(q) \perp c_n^{(j)} \geqslant 0, \quad j = 1, 2, \ldots, p & \\
0 \leqslant D^{(j)}(q)^{\mathrm{T}}v + \lambda^{(j)}e^{(j)} \perp \beta^{(j)} \geqslant 0, \quad j = 1, 2, \ldots, p & \\
0 \leqslant \mu^{(j)}c_n^{(j)} - e^{(j)^{\mathrm{T}}}\beta^{(j)} \perp \lambda^{(j)} \geqslant 0, \quad j = 1, 2, \ldots, p &
\end{aligned}
\tag{59}
$$

where $k(t, q, \mathrm{d}q/\mathrm{d}t)$ is the external force. The mass matrix $M(q)$ is considered to be symmetric and uniformly positive definite.

As mentioned in the introduction there are examples for which this DCP does not have a solution in the classical sense. However, by considering integrals of the forces appearing in the DCP over a small time interval of length h it is possible to obtain time-stepping schemes that have solutions under general assumptions. The time-stepping schemes from References [7, 8] are based on the explicit and the linearly implicit Euler methods, respectively. In what follows we will construct a new time-stepping scheme by using the approximate linearly implicit trapezoidal method (11) that requires the numerical solution of a linear complementarity problem at each time step. In order to arrive at a linear complementarity problem, we need to linearize the constraints, a step that is related to the index reduction step in DAE.

From (23) we replace the non-linear equality constraints $\Theta^{(i)}(q^{(l+1)}) = 0,\ i = 1, 2, \ldots, m$ by the linearizations:

$$
v^{(i)}\left(q^{(l)} + \frac{h}{2}v^{(l)}\right)^{\mathrm{T}} \frac{v^{(l+1)} + v^{(l)}}{2} = 0, \quad i = 1, 2, \ldots, m \tag{60}
$$

For practical reasons, it is useful to think about a larger active contact constraint set. During the integration procedure it is possible that while a contact constraint (j) should be theoretically active, the value of $\Phi^{(j)}$ will not be zero because of numerical error. Such a contact will be considered active. Also, some bodies may collide, generating additional active constraints. In this work we will just assume that the active set $\mathscr{A}$ is provided, and we will not discuss the methods for updating the active set unless this has an immediate consequence for the dynamics resolution problem. Using the same strategy as for the non-linear equality constraints, we replace the non-interpenetration constraints $\Phi^{(j)}(q^{(l+1)}) \geqslant 0$ by $n^{(j)}(q^{(l)}+h/2v^{(l)})^{\mathrm{T}}(v^{(l+1)}+v^{(l)})/2 \geqslant 0$, whenever $j \in \mathscr{A}$.

By applying the approximate linearly implicit trapezoidal method (11) to the differential equation (59) and by using the above approximations for the joint constraints and the active non-interpenetration constraints we obtain the following time-stepping scheme:

$$q^{l+1} = q^{(l)} + \frac{h}{2}(v^l + v^{l+1}) \tag{61}$$

$$\tilde{M}v^{l+1} - \sum_{i=1}^{m} v^{(i)} c_v^{(i)} - \sum_{j \in \mathscr{A}} (n^{(j)} c_n^{(j)} + D^{(j)} \beta^{(j)}) = \tilde{M}v^l + \tilde{k} \tag{62}$$

$$v^{(i)^{\mathrm{T}}} \frac{v^{(l+1)} + v^{(l)}}{2} = 0, \quad i = 1, 2, \ldots, m \tag{63}$$

$$0 \leqslant \rho^{(j)} := n^{(j)^{\mathrm{T}}} \frac{v^{(l+1)} + v^{(l)}}{2} \perp c_n^{(j)} \geqslant 0, \quad j \in \mathscr{A} \tag{64}$$

$$0 \leqslant \sigma^{(j)} := \lambda^{(j)} e^{(j)} + D^{(j)^{\mathrm{T}}} \frac{v^{(l+1)} + v^{(l)}}{2} \perp \beta^{(j)} \geqslant 0, \quad j \in \mathscr{A} \tag{65}$$

$$0 \leqslant \zeta^{(j)} := \mu^{(j)} c_n^{(j)} - e^{(j)^{\mathrm{T}}} \beta^{(j)} \perp \lambda^{(j)} \geqslant 0, \quad j \in \mathscr{A} \tag{66}$$

where $v^{(i)} = v^{(i)}(q^l + (h/2)v^l)$, $n^{(j)} = n^{(j)}(q^l + (h/2)v^l)$, and $D^{(j)} = D^{(j)}(q^l + (h/2)v^l)$.

$$\tilde{M} = \left(M(q^l) - \frac{h}{2}\tilde{k}_v^l - \frac{h^2}{4}\tilde{k}_q^l \right), \quad \tilde{k} = \frac{h}{2}\left(k(t_l, q^l, v^l) + k(t_{l+1}, q^l, v^l) + \frac{h}{2}\tilde{k}_q^l v^l \right) \tag{67}$$

and $\tilde{k}_q^l$, $\tilde{k}_v^l$ are some approximations of $k_q(t_{l+1}, q^l, v^l)$, $k_v(t_{l+1}, q^l, v^l)$, respectively.

We notice that equations (62)–(66) represent a mixed linear complementarity problem. If at time-step l the index set of active contact constraints is given by $\mathscr{A} = \{j_1, j_2, \ldots, j_s\}$, and if we denote

$$\tilde{v} = [v^{(1)}, v^{(2)}, \ldots, v^{(m)}], \quad \tilde{c}_v = \lceil c_v^{(1)}, c_v^{(2)}, \ldots, c_v^{(m)} \rfloor, \quad \tilde{D} = [D^{(j_1)}, D^{(j_2)}, \ldots, D^{(j_s)}]$$

$$\tilde{c}_n = \lceil c_n^{(j_1)}, c_n^{(j_2)}, \ldots, c_n^{(j_s)} \rfloor, \quad \tilde{n} = [n^{(j_1)}, n^{(j_2)}, \ldots, n^{(j_s)}], \quad \tilde{\beta} = \lceil \beta^{(j_1)}, \beta^{(j_2)}, \ldots, \beta^{(j_s)} \rfloor^{\mathrm{T}}$$

$$\tilde{\lambda} = [\lambda^{(j_1)}, \lambda^{(j_2)}, \ldots, \lambda^{(j_s)}], \quad \tilde{\zeta} = [\zeta^{(j_1)}, \zeta^{(j_2)}, \ldots, \zeta^{(j_s)}], \quad \tilde{E} = \mathrm{diag}(e^{(j_1)}, e^{(j_2)}, \ldots, e^{(j_s)})$$

$$\tilde{\sigma} = [\sigma^{(j_1)}, \sigma^{(j_2)}, \ldots, \sigma^{(j_s)}], \quad \tilde{\rho} = [\rho^{(j_1)}, \rho^{(j_2)}, \ldots, \rho^{(j_s)}], \quad \tilde{\mu} = \mathrm{diag}(\mu^{(j_1)}, \mu^{(j_2)}, \ldots, \mu^{(j_s)})$$

then the mixed LCP (62)–(66) can be written under the matrix form

$$\begin{bmatrix} \tilde{M} & -\tilde{v} & -\tilde{n} & -\tilde{D} & 0 \\ \tilde{v}^{\mathrm{T}} & 0 & 0 & 0 & 0 \\ \tilde{n}^{\mathrm{T}} & 0 & 0 & 0 & 0 \\ \tilde{D}^{\mathrm{T}} & 0 & 0 & 0 & \tilde{E} \\ 0 & 0 & \tilde{\mu} & -\tilde{E}^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} v^{(l+1)} + v^{(l)} \\ \tilde{c}_v \\ \tilde{c}_n \\ \tilde{\beta} \\ \tilde{\lambda} \end{bmatrix} - \begin{bmatrix} 2\tilde{M}v^l + \tilde{k} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \tilde{\rho} \\ \tilde{\sigma} \\ \tilde{\zeta} \end{bmatrix} \tag{68}$$

$$0 \leqslant \lceil \tilde{c}_n, \tilde{\beta}, \tilde{\lambda} \rfloor \perp \lceil \tilde{\rho}, \tilde{\sigma}, \tilde{\zeta} \rfloor \geqslant 0 \tag{69}$$

The above LCP has the same structure as the ones considered in References [7, 8], so that by using the same analysis we obtain the following result.

Theorem 6.1
If the matrix $\tilde{M}$ given by (67) is positive definite, then the solution set $\mathscr{L}^{\mathrm{T}}(q^l, v^l, h, \tilde{k})$ of LCP (68), (69) is non-empty, and Lemke's algorithm can be used to compute an element of this set.

Since $M(q^l)$ is positive definite, $\tilde{M}$ is guaranteed to be positive definite for sufficiently small h. In order to be able to take large time steps for integration, we have to construct approximations $\tilde{k}_q^l$ and $\tilde{k}_v^l$ of $k_q(t_{l+1}, q^l, v^l)$ and $k_v(t_{l+1}, q^l, v^l)$ such that $\tilde{M}$ is positive definite for the desired time step range. Such approximations are discussed in Reference [7].

Remark
By taking $v = v^{(l+1)}$ instead of $v = (v^{(l)} + v^{(l+1)})/2$ in (63)–(66), we obtain a slightly modified formulation of the integration step. In matrix form this is given by the following MLCP:

$$\begin{bmatrix} \tilde{M} & -\tilde{v} & -\tilde{n} & -\tilde{D} & 0 \\ \tilde{v}^{\mathrm{T}} & 0 & 0 & 0 & 0 \\ \tilde{n}^{\mathrm{T}} & 0 & 0 & 0 & 0 \\ \tilde{D}^{\mathrm{T}} & 0 & 0 & 0 & \tilde{E} \\ 0 & 0 & \tilde{\mu} & -\tilde{E}^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} v^{(l+1)} \\ \tilde{c}_v \\ \tilde{c}_n \\ \tilde{\beta} \\ \tilde{\lambda} \end{bmatrix} - \begin{bmatrix} \tilde{M}v^l + \tilde{k} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \tilde{\rho} \\ \tilde{\sigma} \\ \tilde{\zeta} \end{bmatrix} \tag{70}$$

$$0 \leqslant \lceil \tilde{c}_n, \tilde{\beta}, \tilde{\lambda} \rfloor \perp \lceil \tilde{\rho}, \tilde{\sigma}, \tilde{\zeta} \rfloor \geqslant 0 \tag{71}$$

The representation of the frictional constraints in (70) is more consistent with Coulomb's friction law than the representation in (68). However, the stability properties are stronger for the scheme given by (70). A numerical comparison of the two schemes in terms of friction behaviour is presented in Section 10.

It is very difficult to analyse in full generality the order of time-stepping schemes (68) and (70) in the presence of impacts. Second order can be eventually obtained only with proper event detection and restart, which is detailed in the next section.

7. EVENT DETECTION AND RESOLUTION

Event detection affects two sets of parameters that are essential to preserving the $O(h^2)$ global error of the scheme.

- The time step, h, which is *essential* to be chosen such that global second order convergence of the trapezoidal method is preserved. The time step choice is done via interpolation from the current and predicted positions and velocities, based on the principle that if the switching rule is satisfied up to $O(h^2)$ at the time where the simulation is restarted, and if the number of such events is finite, then the global order is preserved. In the simulation of multibody dynamics with contact and friction there are two types of events that generate switching rules: collisions, which typically require discontinuity of the velocity, and stick/slip and take-off transitions, that proceed with continuous (but non-smooth velocity).
- The active set of the non-penetration constraints, $\mathscr{A}$, whose role is to allow for the linearization of the non-penetration constraints to be active only when the gap functions $\Phi^{(j)}(q)$ are no larger than $O(h^3)$, and thus capturing the correct collision sequence.

7.1. Selection of the set of active contact constraints

The list $\mathscr{A}$ of active contact constraints is updated at each time step by setting

$$\mathscr{A} = \{j : \Phi^{(j)}(q) \leqslant \max\{\varepsilon_a, \varepsilon_b h^3\}\} \tag{72}$$

where q, v are the current position and velocity, and $\varepsilon_a > 0$ and ε_b are fixed, user-defined, parameters.

Our choice of an active set selection rule reflects the balance between preserving the order of the method through the $\varepsilon_b h^3$ term and the practical need of advancing the simulation with a time step that is bounded away from 0 which is ensured by the term ε_a.

If the collision events are sufficiently spaced away from each other, then the active set choice is the same as if only either ε_a or $\varepsilon_b h^3$ is considered in the test, and thus the second order of the method is correctly preserved, asymptotically. The same can be said if the time step is chosen sufficiently small, provided that the number of collisions is finite for any finite time interval. Recall, however, that the minimum time interval between two collisions is not known *a priori*. From this perspective, allowing the active set to include only the contacts that penetrate or using only the $\varepsilon_b h^3$ term to define the active set and, in both cases, to expose oneself to the possibility that another collision could occur at very short time interval from the current point, even if only a linear system as opposed to an LCP per step, seems unreasonable. Our choice is also in line with the setup of modern numerical integration methods where a prescribed tolerance level, rather than an order of integration, is satisfied.

In addition, if the number of collision events over a finite time interval is infinite, the second-order approximation of the dynamics may be impossible irrespective of the values of the friction coefficient, and thus unrelated to the Painlevé paradoxes, that are types of singularities that we

will discuss later. An example is the one of a ball that is caught between two rigid walls that have a prescribed motion towards each other, that stop in finite time at a distance equal to the diameter of the ball to each other and that have a restitution coefficient of 1 with respect to the ball. It can be immediately shown that, in any interval that contains the time at which the walls stop, there is no polynomial of finite order, and thus no integration formula, that will provide an approximation that is bounded by ch^2, where c is a constant. In addition, even weaker convergence concepts, such as the ones of measure differential inclusions [14] may not apply for this example since the velocity solution v does not have bounded variation.

7.2. Collision time detection

After updating $\mathscr{A}$ we compute a solution $(\bar{v}, \tilde{c}_v, \tilde{c}_n, \tilde{\beta}, \tilde{\lambda}) \in \mathscr{L}^{\mathrm{T}}(q, v, h, \tilde{k})$ and set $\bar{q} = q + h(v + \bar{v})/2$, where we drop the time step superscript, in order to simplify notation. If all of the non-interpenetration constraints (55) are satisfied then we continue with the next time step. Otherwise, there is at least one j such that $\Phi^{(j)}(\bar{q}) < 0$, which means that a collision has taken place in the time interval $(t, t + h)$. In order to maintain the order of the numerical solution we have to determine the collision time with sufficient accuracy. This can be done by using the unique cubic interpolant $\hat{q}$ satisfying

$$\hat{q}(t) = q, \quad \hat{q}(t + h) = \bar{q}, \quad \hat{q}'(t) = v, \quad \hat{q}'(t + h) = \bar{v}$$

and finding the smallest root t^* of the scalar non-linear equation $\Phi^{(j)}(\hat{q}(\tau)) = 0$ in the interval $(t, t + h)$. If multiple collisions are detected then we consider the earliest collision time

$$t^* = \min_j \left\{ \min_\tau \{\tau \in (t, t + h) : \Phi^{(j)}(\hat{q}(\tau)) = 0\} \ : \ \Phi^{(j)}(\bar{q}) < 0 \right\} \tag{73}$$

and we add the corresponding contact, j^*, to $\mathscr{A}$. The pre-collision position and velocity are defined as

$$q^- = \hat{q}(t^*), \quad \hat{v}^- = q'(t^*) \tag{74}$$

We denote by $q^*(t)$ and $v^*(t)$ the exact solution of the friction problem that uses a fixed active set A and stick–slip choice for each contact. Then $q^*(t)$ and $v^*(t)$ are the solution of a DAE, and are smooth functions. It follows from the theory of Hermite interpolation [18], that $\|\hat{q}(t) - q^*(t)\| \leqslant C_1 h^2$ and $\|\hat{v}(t) - v^*(t)\| \leqslant C_2 h^2$, provided that the numerical integration method used has order 2. Here $v(t) = \hat{q}'(t)$.

We assume now that, for the case of interest, only one collision defined by $\Phi^{(j)}(q)$ reaching 0, without any stick–slip phenomenon, occurs. Denote by $\hat{t}^*$ the *exact* collision time, which satisfies $\Phi^{(j)}(q^*(\hat{t}^*)) = 0$. It then follows that $\Phi^{(j)}(q^*(t^*)) = O(h^2)$, which, in turn, implies that

$$\nabla\Phi^{(j)}(q^*)v^*(\hat{t}^*) \neq 0 \Rightarrow |\hat{t}^* - t^*| = O(h^2) \tag{75}$$

Therefore, if the collision is isolated and non-degenerate (that is, the normal velocity at the collision is not zero), then the collision time and, as a consequence, the position and the velocity, will be computed with a precision of $O(h^2)$.

7.3. The compression phase

Our collision model has two phases: compression and decompression. In the former phase interpenetration is prevented by normal compression contact impulses, while in the latter a fraction of each normal compression contact impulse is restituted to the system (the Poisson hypothesis [27]).

In the compression phase, the dynamic system will respond with constraint impulses generated by joints, contacts, or friction. We denote the impulses by the same symbols as before but with superscript c. Let v^c be the velocity at the end of the compression phase. At the end of this phase, each contact from the list is either maintained $c_n^c \geqslant 0, n^{\mathrm{T}} v^c = 0$, or is breaking, $c_n^c = 0, n^{\mathrm{T}} v^c \geqslant 0$. Therefore, we recover the same complementarity conditions as in Reference [7]. Conservation of impulse requires

$$M(v^c - v^-) - \sum_{i=1}^{m} v^{(i)} c_v^{c(i)} - \sum_{j=1}^{p} (n^{(j)} c_n^{c(j)} + D^{(j)} \beta^{c(j)}) = 0 \tag{76}$$

The conditions on joints and friction with the appropriate complementarity conditions used to solve the compression phase are the same with the ones used in Reference [7]. It follows that

$$(v^c, \tilde{c}_v^c, \tilde{c}_n^c, \tilde{\beta}^c, \tilde{\lambda}^c) \in \mathscr{L}^c(q^-, v^-, 0, 0) \tag{77}$$

where by $\mathscr{L}^c(q, v, h, k)$ we have denoted the solution set of the mixed LCP formulated in Reference [7].

7.4. The decompression phase

During the decompression phase each active contact constraint generates a decompression impulse, dependent on its coefficient of restitution e_j. For instance, contact j generates an impulse $c_n^{d(j)} = e_j c_n^{c(j)} + c_n^{x(j)}$. The additional impulse $c_n^{x(j)} \geqslant 0$ is necessary to prevent interpenetration. Let v^+ be the velocity after the decompression phase, or the post-collision velocity. At the end of the decompression phase contact j either breaks, and then $n^{(j)^{\mathrm{T}}} v^+ \geqslant 0$, $c_n^{x(j)} = 0$, or it is maintained and then $n^{(j)^{\mathrm{T}}} v^+ = 0$, $c_n^{x(j)} \geqslant 0$. Therefore, the following complementarity condition

$$0 \leqslant n^{(j)^{\mathrm{T}}} v^+ \perp c_n^{x(j)} \geqslant 0$$

is generated for each contact. The conditions on joints and friction remain the same as in Reference [7]. The conservation of impulse requires

$$M(v^+ - v^c) - \sum_{i=1}^{m} v^{(i)} c_v^{x(i)} - \sum_{j=1}^{p} (n^{(j)} c_n^{x(j)} + D^{(j)} \beta^{x(j)}) = F^r \tag{78}$$

where F^r is the restitution impulse,

$$F^r = \sum_{j=1}^{p} e_j n^{(j)} c_n^{c(j)} \tag{79}$$

Hence, according to an analogue of Theorem 6.1 from Reference [7], the post-collision velocity v^+ can be found by Lemke's algorithm since

$$(v^+, \tilde{c}_v^x, \tilde{c}_n^x, \tilde{\beta}^x, \tilde{\lambda}^x) \in \mathscr{L}^c(q^-, v^c, 0, F^r) \tag{80}$$

While approaches based on the Poisson hypothesis are widely used in the robotics and mechanics literature [28], it cannot be guaranteed in general that their application results in a decrease of energy after the two collision phases [26]. Although this raises several questions about the appropriateness of the Poisson model, in this work we are concerned with the numerical implications of this observation.

7.5. Stick–slip and take-off (contact deletion) detection

For simplicity in notation, we will ignore the bilateral constraints. A straightforward generalization may be applied to include also bilateral constraints. Suppose we obtain the solution at each time-step by applying some LCP solver, in our case the PATH package. Suppose also that we have detected a change of basis corresponding to the solutions obtained at steps t_{l-1} and $t_{l-1}+h_s$. Let h be any time-step value in the interval $[0, h_s]$. We define,

$$\tilde{N}(h)=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \tilde{E}(h) \\ \tilde{\mu}(h) & -\tilde{E}^t(h) & 0 \end{bmatrix}, \quad H(h)=H(\tilde{q}(t_{l-1}+h))=[\tilde{n}(h)\ \ \tilde{D}(h)\ \ 0]$$

where $\tilde{n}(h)=\tilde{n}(\tilde{q}(t_{l-1}+h))$, $\tilde{D}(h)=\tilde{D}(\tilde{q}(t_{l-1}+h))$, $\tilde{E}(h)=\tilde{E}(\tilde{q}(t_{l-1}+h))$, $\tilde{\mu}(h)=\tilde{\mu}(\tilde{q}(t_{l-1}+h))$ and

$$\tilde{q}(t)=\frac{t_l-t}{h_s}q^{(l-1)}+\frac{t-t_{l-1}}{h_s}q^{(l)} \tag{81}$$

$$\tilde{v}(t)=\frac{t_l-t}{h_s}v^{(l-1)}+\frac{t-t_{l-1}}{h_s}v^{(l)} \tag{82}$$

Using the above notation we can reduce the MLCP solved at step $(l-1)$ to the standard LCP:

$$s=A(0)u+H^t(0)M^{-1}p(0) \tag{83}$$

If (70) is solved, then $A(h)=H^{\mathrm{T}}(h)M^{-1}H(h)+N(h)$, $p(h)=M\tilde{v}(t_{l-1}+h)+\tilde{k}(h)$, and

$$\tilde{k}(h)=\frac{h_s}{2}[k(t_{l-1}+h,\tilde{q}(t_{l-1}+h),\tilde{v}(t_{l-1}+h))+k(t_l+h,\tilde{q}(t_{l-1}+h),\tilde{v}(t_{l-1}+h))$$
$$+h_s k_q(t_{l-1}+h,\tilde{q}(t_{l-1}+h),\tilde{v}(t_{l-1}+h))\tilde{v}(t_{l-1}+h)]$$

A similar construction can be obtained for the formulation given by (68). Now if we consider the index set $J=\{1,2,...,2n\}$, where n is the size of s or u above, at time-step t_{l-1} we have the subset I of J which corresponds to the basic variables of the (t_{l-1})-LCP and I^c the subset of J corresponding to the non-basic variables. Given the index set I and $h\in[0,h_s]$, we construct the following complementarity matrix:

$$C_h(I)=C_{A(h)}(I)(\cdot,i)=\begin{cases} I(\cdot,i) & \text{if } i\in I^c \\ -A(h)(\cdot,i) & \text{otherwise} \end{cases} \tag{84}$$

Now let us denote by $x(0)$ the set of basic variables from step $(l-1)$. Given that I was the index set corresponding to the basis used to solve the $(l-1)$-LCP we have that $x(0)$ is the solution of

$$C_0(I)x(0) = H^{\mathrm{T}}(0)M^{-1}p(0) \tag{85}$$

If this is a non-degenerate solution of the LCP, then $x(0)$ satisfies:

$$\min x(0) > 0$$

A change of basis will correspond to the fact that the solution $x(h_s)$ of the linear system:

$$C_h(I)x(h) = H^{\mathrm{T}}(h)M^{-1}p(h) \tag{86}$$

satisfies:

$$\min x(h_s) < 0$$

Let $F(h) = \min x(h)$. From the constructions above and under the non-degeneracy assumption, we can see that F is continuous on $[0, h_s]$ and

$$F(0)F(h_s) < 0$$

This means that there is a root of F in the interval $(0, h_s)$. Such a root corresponds to a transition, which we denote by h_{TR}. A root of $F(h)$ in $(0, h_s)$ can be found via Brent's method, for example. A numerical experiment related to the approach described here is presented in Section 10. The algorithm is restarted from the switching point after computing the position and velocity by the same formula as (74).

Similarly to the result (75), we can prove that the switching time is determined with a precision of $O(h^2)$ if we have an isolated, non-degenerate, stick–slip or take-off transition (or, in the three-dimensional case with discretization of the friction cone, a change of the active set in the representation of active friction) and if we use the interpolation approach defined by (81) and (82) to restart the algorithm. By non-degeneracy, we mean that the derivative of $F(h)$ at $h = h_{\mathrm{TR}}$ is not zero, uniformly as $h_s \to 0$.

8. THE ALGORITHM

By combining the time-stepping scheme from Section 6 with the collision resolution described in Section 7 we obtain the following algorithm:

Input:

- q^0, v^0—consistent initial position and velocity;
- T—length of simulation time interval;
- h_s—standard step-size;
- ε_a, ε_b—positive parameters.

Set $l = 0$, $t_l = 0$;
while $(t_l < T)$

1. Set $q = q^l$, $v = v^l$, $t = t_l$, $h = h_s$;
2. Determine list of active contacts by (72);

3. Compute $(\bar{v}, \tilde{c}_v, \tilde{c}_n, \tilde{\beta}, \tilde{\lambda}) \in \mathscr{L}^{\mathrm{T}}(q, v, h, \tilde{k})$ (68), (69);
4. Set $\bar{q} = q + h(v + \bar{v})/2$;
if (no collision detected between t and $t + h$)
  if (stick–slip or takeoff detected between t and $t + h$)
    Detect transition time-step h_{TR} (see Section 7);
    Set $h = h_{\mathrm{TR}}$ and **go to** 3;
  else
    if (transition solved at previous step)
      Set $h = h_s - h_{\mathrm{TR}}$;
    else
      Set $h = h_s$;
    end if.
    Set $t_{l+1} = t_l + h$, $q^{l+1} = \bar{q}$, $v^{l+1} = \bar{v}$;
    Set $l \leftarrow l + 1$;
  end if;
else
  Estimate the collision time t^* from (73);
  Add contact j^* to the list of active contacts;
  Determine pre-collision position and velocity q^- and v^- from (74);
  Compute $(v^c, \tilde{c}_v^c, \tilde{c}_n^c, \tilde{\beta}^c, \tilde{\lambda}^c) \in \mathscr{L}^c(q^-, v^-, 0, 0)$;
  Compute $(v^+, \tilde{c}_v^x, \tilde{c}_n^x, \tilde{\beta}^x, \tilde{\lambda}^x) \in \mathscr{L}^c(q^-, v^c, 0, F^r)$;
  Set $t = t^*$, $h = t_l + h_s - t^*$, $q = q^-$, $v = v^+$ and **go to** 3.
end if
end while

9. BOUNDEDNESS OF KINETIC ENERGY AND CONVERGENCE RESULTS

In this section we prove that the kinetic energy of the discrete model described in this paper stays bounded on any given finite interval $[0, T]$ as the step-size h tends to zero. In addition, we give sufficient conditions for the scheme to have second-order convergence.

9.1. Kinetic energy estimates

To simplify analysis we consider the case where the mass matrix M is constant and the force k is of the form

$$k(t, q, v) = f(v) + k_1(t, q, v) \tag{87}$$

where $f(v)$ is the Coriolis force and $k_1(t, q, v)$ are external forces satisfying the following growth condition:

$$\|k_1(t, q, v)\| \leqslant c_1 + c_2\|q\| + c_3\|v\| \tag{88}$$

where the constants c_1, c_2, c_3 depend only on the length T of the given time interval. We also assume that the Coriolis force is given by a bilinear operator

$$[f(v)]_i = \sum_{jk} f_{ijk} v_j v_k$$

This is certainly true in case the system is described by Newton–Euler equations in body co-ordinates [24, Section 2.4] where the matrix $F(v)$ of entries

$$[F(v)]_{ij} = \sum_{k} f_{ijk} v_k$$

is antisymmetric in the sense that

$$u^{\mathrm{T}} F(v) u = 0 \quad \forall u$$

Without loss of generality we may assume that the tensor (f_{ijk}) is symmetric in the sense that $f_{ijk} = f_{ikj}$. In this case it is easily seen that

$$f(v) = F(v)v, \quad f'(v) = 2F(v)$$

where f' is the Fréchet derivative of f. In our time-stepping scheme we take

$$\tilde{k}_q^l = \tilde{k}_{1q}^l, \quad \tilde{k}_v^l = F(v^l) + \tilde{k}_{1v}^l$$

where $\tilde{k}_{1q}^l$ and $\tilde{k}_{1v}^l$ are approximations of $k_{1q}(t_{l+1}, q^l, v^l)$ and $k_{1v}(t_{l+1}, q^l, v^l)$, respectively, that are bounded in the sense that there are constant c_4, c_5 such that

$$\|\tilde{k}_{1q}^l\| \leqslant c_4, \quad \|\tilde{k}_{1v}^l\| \leqslant c_5 \quad \forall l \tag{89}$$

In addition, we will need a regularity condition concerning the total friction cone, *the pointed friction one assumption*, which is an extension of the condition for the contact-only case [29]. We define the total friction cone to be the portion in the velocity space that can be covered by feasible constraint interaction impulses, or

$$FC(q) = \{t = \tilde{\nu} c_v + \tilde{n} c_n + \tilde{D}\tilde{\beta} | c_n \geqslant 0, \tilde{\beta} \geqslant 0, \|\beta^{(j)}\|_1 \leqslant \mu^{(j)} c_n^{(j)}, \forall j \in \mathscr{A}\} \tag{90}$$

Clearly, the cone $FC(q)$ is a convex set.

Definition
We say that

$$FC(q) \text{ is pointed} \Leftrightarrow \forall (c_v, c_n \geqslant 0, \tilde{\beta} \geqslant 0) \neq 0 \text{ such that } \|\beta^{(j)}\|_1 \leqslant \mu^{(j)} c_n^{(j)}, \ \forall j \in \mathscr{A}$$
$$\text{we must have that } \tilde{\nu} c_v + \tilde{n} c_n + \tilde{D}\tilde{\beta} \neq 0 \tag{91}$$

We note that the pointed friction cone assumption is weaker than the linear independence of the active set. Its name originates in the fact that, when there are no joint constraints, the condition is equivalent to the cone not containing any proper linear subspace, and thus being 'pointed'.

For the following, we will assume that the mass matrix $M(q)$ is constant. This is not an exceedingly restrictive assumption and it holds when we use the Newton–Euler formulation in body co-ordinates to describe our problem [24].

Theorem 9.1
Assume that the algorithm from Section 8 solves a number of collisions in the time interval $[0, T]$ that is uniformly upper bounded as $h \to 0$. Suppose that the friction cone is pointed for the possible choices of the position vector q. Suppose that the mass matrix M is a constant symmetric positive definite matrix and that conditions (88), (89) are satisfied. Then there is a constant c such that

$$(v^l)^{\mathrm{T}} M v^l \leqslant \max\{(v^0)^{\mathrm{T}} M v^0, \|q^0\| + 1\} e^{ct_l} \quad l = 0, 1, \ldots, \lfloor T/h \rfloor$$

for all sufficiently small h.

Proof
Since the friction cone is uniformly pointed it follows that the fractional increase in kinetic energy following a collision is uniformly upper bounded. From the properties of the Hermite interpolation and from the definition of our scheme, it follows that the velocity before collision has an uniformly upper bounded fractional increase when compared to the maximum of the velocity before a collision and the velocity after a collision.

Therefore, since the number of collisions is uniformly upper bounded, it is sufficient to prove the conclusion of this theorem for the case where all collisions are totally plastic (that is, the restitution coefficient is zero). Formally, this assumption is the same as assuming that all time steps are computed using just the time stepping LCP (61)–(67), except that some of the time steps may be equal to 0 [26]. Since this does not result in kinetic energy increase, it is sufficient to prove the result for a uniform time step.

Suppose that no collisions are detected in the interval $[t_l, t_{l+1}]$. The new velocity $v^{(l+1)}$ will be determined by solving the LCP (62)–(66).

Left multiplying (62) by $(v^{(l+1)} + v^{(l)})^{\mathrm{T}}$ we get that

$$\begin{aligned}(v^{(l+1)} + v^{(l)})^{\mathrm{T}} \tilde{M}^{(l)} v^{(l+1)} = &\sum_{i=1}^{m} (v^{(i)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}))c_v^{(i)} + \sum_{j \in \mathscr{A}} (n^{(j)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}))c_n^{(j)} \\ &+ \beta^{(j)^{\mathrm{T}}} D^{(j)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}) + \tilde{k}^{(l)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}) \\ &+ (v^{(l+1)} + v^{(l)})^{\mathrm{T}} M v^{(l)} \end{aligned} \tag{92}$$

Using (63), we deduce that $v^{(i)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}) = 0, \quad i = 1, 2, \ldots, m$. Also, using the contact constraints (64), we obtain $c_n^{(j)} n^{(j)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}) = 0, \quad j \in \mathscr{A}$. Finally, from the frictional constraints (65) and (66) we get that

$$\beta^{(j)^{\mathrm{T}}} D^{(j)^{\mathrm{T}}}(v^{(l+1)} + v^{(l)}) = -\lambda^{(j)} \beta^{(j)^{\mathrm{T}}} e^{(j)} = -\mu^{(j)} c_n^{(j)} \lambda^{(j)} \leqslant 0 \quad \forall j \in \mathscr{A}$$

Then (92) implies

$$(v^{(l+1)} + v^{(l)})^{\mathrm{T}} \tilde{M}^{(l)} v^{(l+1)} \leqslant (v^{(l+1)} + v^{(l)})^{\mathrm{T}} \tilde{M}^{(l)} v^{(l)} + (v^{(l+1)} + v^{(l)})^{\mathrm{T}} \tilde{k}^{(l)^{\mathrm{T}}} \tag{93}$$

By expanding the left and right-hand sides of the above inequality and using (88) and (89), it follows that there is a constant c_6 such that for sufficiently small h we have

$$(v^{(l+1)}+v^{(l)})^{\mathrm{T}}\tilde{M}v^{(l+1)} = (v^{l+1})^{\mathrm{T}}Mv^{l+1} - \frac{h}{2}(v^{l+1})^{\mathrm{T}}\left(\tilde{k}_{1v}^{l}+\frac{h}{2}\tilde{k}_{1q}^{l}\right)v^{l+1}$$

$$-\frac{h}{2}(v^{l})^{\mathrm{T}}\left(\tilde{k}_{1v}^{l}+\frac{h}{2}\tilde{k}_{1q}^{l}\right)v^{l+1}$$

$$+v^{(l)^{\mathrm{T}}}Mv^{(l+1)} - \frac{h}{2}v^{(l)^{\mathrm{T}}}F(v^{(l)})v^{(l+1)}$$

which leads to

$$(v^{(l+1)}+v^{(l)})^{\mathrm{T}}\tilde{M}v^{(l+1)} \geqslant (1-c_6h)\|M^{1/2}v^{(l+1)}\|^2 - c_6h\|M^{1/2}v^{(l)}\|\|M^{1/2}v^{(l+1)}\|$$

$$+v^{(l)^{\mathrm{T}}}Mv^{(l+1)} - \frac{h}{2}v^{(l)^{\mathrm{T}}}F(v^{(l)})v^{(l+1)}$$

For the right-hand side we use the same procedure:

$$(v^{(l+1)}+v^{(l)})^{\mathrm{T}}(\tilde{M}v^{(l)}+\tilde{k}) = v^{(l)^{\mathrm{T}}}Mv^{(l)} - \frac{h}{2}v^{(l)^{\mathrm{T}}}\left(\tilde{k}_{1v}^{l}+\frac{h}{2}\tilde{k}_{1q}^{l}\right)v^{(l)}$$

$$-\frac{h}{2}(v^{(l+1)})^{\mathrm{T}}\left(\tilde{k}_{1v}^{l}+\frac{h}{2}\tilde{k}_{1q}^{l}\right)v^{(l)} + \frac{h}{2}(v^{(l+1)}+v^{(l)})^{\mathrm{T}}(k_1(t_l,q^l,v^l)$$

$$+k_1(t_{l+1},q^l,v^l)) + \frac{h^2}{2}v^{(l)^{\mathrm{T}}}\tilde{k}_{1q}v^{(l)} + \frac{h^2}{2}v^{(l+1)^{\mathrm{T}}}\tilde{k}_{1q}v^{(l)}$$

$$+v^{(l+1)^{\mathrm{T}}}Mv^{(l)} + \frac{h}{2}v^{(l+1)^{\mathrm{T}}}F(v^{(l)})v^{(l)}$$

leading to

$$(v^{(l+1)}+v^{(l)})^{\mathrm{T}}(\tilde{M}v^{(l)}+\tilde{k}) \leqslant (1+c_6h)\|M^{1/2}v^{(l)}\|^2 + c_6h\|M^{1/2}v^{(l)}\|\|M^{1/2}v^{(l+1)}\|$$

$$+\frac{h}{2}v^{(l+1)^{\mathrm{T}}}F(v^{(l)})v^{(l)} + v^{(l+1)^{\mathrm{T}}}Mv^{(l)} + c_6h(\|M^{1/2}v^{(l+1)}\|$$

$$+\|M^{1/2}v^{(l)}\|)(1+\|q^{(l)}\|+\|M^{1/2}v^{(l)}\|)$$

Let us denote

$$\rho_l = \|M^{1/2}v^l\|, \quad \sigma_l = \|q^l\|+1$$

Using the symmetry of the mass matrix M, the antisymmetry of $F(v^{(l)})$ as well as the above estimates and notation, it turns out that there exists a constant c_7 such that

$$(1 - c_7 h)\rho_{l+1}^2 \leqslant (1 + c_7 h)\rho_l^2 + c_7 h \rho_l \rho_{l+1} + c_7 h \sigma_l (\rho_l + \rho_{l+1}) \tag{94}$$

for all sufficiently small h. In what follows we analyse the two possibilities $\rho_l < \rho_{l+1}$ and $\rho_l \geqslant \rho_{l+1}$.

If $\rho_l < \rho_{l+1}$, inequality (94) implies that there exists a constant c_8 such that:

$$\rho_{l+1} \leqslant (1 + c_8 h)\rho_l + c_8 h \sigma_l \tag{95}$$

for all sufficiently small h. On the other hand from (61) we have

$$\|q^{l+1}\| \leqslant \|q^l\| + \frac{h}{2}\|M^{-1/2}\|(\|M^{1/2}v^l\| + \|M^{1/2}v^{l+1}\|) \tag{96}$$

Substituting the over-estimate for ρ_{l+1} into (96) gives

$$\sigma_{l+1} \leqslant \frac{h}{2}\|M^{-1/2}\|(2 + c_8 h)\rho_l + \left(1 + c_8 \frac{h^2}{2}\|M^{-1/2}\|\right)\sigma_l$$

It follows that there is a constant c_9 such that

$$\begin{aligned} \rho_{l+1} &\leqslant (1 + c_9 h)\rho_l + c_9 h \sigma_l \\ \sigma_{l+1} &\leqslant c_9 h \rho_l + (1 + c_9 h)\sigma_l \end{aligned} \tag{97}$$

for all sufficiently small h.

In the case when $\rho_{l+1} \leqslant \rho_l$, it is straightforward to see that the equivalent of (97) is

$$\begin{aligned} \rho_{l+1} &\leqslant \rho_l \\ \sigma_{l+1} &\leqslant \sigma_l + h\|M^{-1/2}\|\rho_l \end{aligned}$$

If the pair (ρ_{l+1}, ρ_l) satisfies the above set of inequalities, then it also satisfies (97) for sufficiently small values of h (i.e. one can obviously choose $c_9 > \|M^{-1/2}\|$). Therefore, in both cases the inequalities (97) are satisfied for all sufficiently small h and by taking $c = 2c_9$, we have

$$\begin{aligned} \left\| \begin{bmatrix} \rho_l \\ \sigma_l \end{bmatrix} \right\|_\infty &\leqslant \left\| \begin{bmatrix} 1 + c_9 h & c_9 h \\ c_9 h & 1 + c_9 h \end{bmatrix} \right\|_\infty^l \left\| \begin{bmatrix} \rho_0 \\ \sigma_0 \end{bmatrix} \right\|_\infty = (1 + ch)^l \left\| \begin{bmatrix} \rho_0 \\ \sigma_0 \end{bmatrix} \right\|_\infty \\ &\leqslant e^{clh} \left\| \begin{bmatrix} \rho_0 \\ \sigma_0 \end{bmatrix} \right\|_\infty = e^{ct_l} \left\| \begin{bmatrix} \rho_0 \\ \sigma_0 \end{bmatrix} \right\|_\infty \end{aligned}$$

which proves our theorem. □

9.2. Convergence results

Using our preceding result concerning the boundedness of the kinetic energy, we obtain the following convergence theorem.

Theorem 9.2
Assume that the algorithm from Section 8 solves a number of collisions in the time interval $[0, T]$ that is uniformly upper bounded as $h \to 0$, and that the restitution coefficient at any collision is 0. Suppose that the friction cone is uniformly pointed for the possible choices of the position vector q. Suppose that the mass matrix M is a constant symmetric positive definite matrix and that conditions (88), (89) are satisfied. Assume that the mappings $n^{(j)}(q)$, $v^{(i)}(q)$, $D(q)$ are Lipschitz continuous.

Let $q^{h(t)}$ and $v^h(t)$ be the pointwise position and velocities sequence that can be defined from the numerical sequence produced by our algorithm There exists a subsequence $h_\alpha \to 0$ where

- $q^{h_\alpha}(\cdot) \to q(\cdot)$ uniformly.
- $v^{h_\alpha}(\cdot) \to v(\cdot)$ pointwise a.e.
- $\mathrm{d}v^{h_\alpha}(\cdot) \to \mathrm{d}v(\cdot)$ weak * as Borel measures in $[0, T]$, and every such subsequence converges to a solution $(q(\cdot), v(\cdot))$ of a measure differential inclusion.

Since the formulation (59) cannot have a strong solution in general, one must look for a weak solution, an example of which is the measure differential inclusion concept defined in References [6, 14]. Our algorithm therefore recovers a weak solution of the formulation, in the sense defined in Reference [14]. The proof of this theorem, though straightforward once the kinetic energy bound in Theorem 9.1 is proved, is exceedingly technical, and we omit it from this paper. A blueprint for the proof can be found in References [14, 30], where it is applied to other time stepping schemes.

Theorem 9.2 shows that our algorithm converges in a certain sense as the time step $h \to 0$, which was also shown for other time stepping schemes [14, 30] that are based on Euler's method, and cannot therefore be expected to have a global error estimate better than $O(h)$. However, given that we used the trapezoidal method, we expect that the order of convergence is $O(h^2)$ under certain assumptions, which are spelled out in the following result. In order to state that result, we denote by $\mathscr{A}^{\mathrm{tot}}(t^{(l)}, q^{(l)}, v^{(l)})$ the active set of the MLCP (61)–(66).

Theorem 9.3
Consider the numerical sequence $(t^{h,(l)}, q^{h,(l)}, v^{h,(l)})$ produced by our algorithm, when ε_a is turned to zero in (72), for the time interval $[t_0, T]$. Assume that there exist $t_0 = t_0^h < t_1^h < t_2^h \leqslant \cdots < t_K^h < T = t_{f+1}^h$ such that, for all sufficiently small h we have that

- A1 $\mathscr{A}^{\mathrm{tot}}(t^{h,(l)}, q^{h,(l)}, v^{h,(l)}) = \mathscr{A}_k^{\mathrm{tot}}$, for $t_k^h \leqslant t^{(l)} \leqslant t_{k+1}^h$, $0 \leqslant k \leqslant K$, that is, the sequence of active sets of the MLCP (61)–(66) is constant for sufficiently small time step, and K is bounded uniformly as $h \downarrow 0$.
- A2 The submatrix of the matrix in (68), that corresponds to the active set $\mathscr{A}^{\mathrm{tot}}(t^{h,(l)}, q^{h,(l)}, v^{h,(l)})$ is full rank, uniformly as $h \to 0$.
- A3 The solution of the compression–decompression LCP that is used in solving the collision is unique and stable.

A4 For a sequence $h_\alpha \to 0$ we have that $t_k^{h_\alpha}$ is convergent to t_k^*, $k = 1, 2, \ldots, K^*$ and $t_i^* \neq t_j^*$, for $i \neq j$, as well as $q^{h,(l)} \to q(\cdot)$ and $v^{h,(l)} \to v(\cdot)$.
A5 The velocity sequence $v(\cdot)$ is continuous except at collisional points.
A6 $v^{(i)}(q)$, $i = 1, 2, \ldots, m$ and $D^{(j)}(q), n^{(j)}(q)$, $j = 1, 2, \ldots, p$ are constant.

Then we have that $|v^h(t) - v(t)| = O(h^2)$ and $|q^h(t) - q(t)| = O(h^2)$ as $h \to 0$, and the method has global order 2.

Proof
Using assumptions A1, A2 and A4, we have that, on $[0, t_1^*)$, $q(\cdot)$ and $v(\cdot)$ satisfy the differential algebraic equation that corresponds to dividing with h and taking the limit as $h \to 0$ in (61)–(66). The limit must exist since the linear system that corresponds to the active constraints in the MLCP (61)–(66) has uniform full rank and has a right-hand side that is $O(h)$, once we write Newton's law in terms of the difference in velocities $v^{h,(l+1)} - v^{h,(l)}$. Using assumption A6 and Theorem 4.3, it follows that, on $[0, t_1^*)$, we have that $|v^h(t) - v(t)| = O(h^2)$ and $|q^h(t) - q(t)| = O(h^2)$ as $h \to 0$.

At the switching point, we are now in one of two situations. Either the point t_1^* is not a collision point, and then, using assumption A5, we will have that $v(t_1^{*,+}) = v(t_1^{*,-})$, and, using A2, as well as the fact that our event detection is $O(h^2)$, we can apply the argument from the preceding paragraph to conclude that on $[t_1^*, t_2^*)$, we have that $|v^h(t) - v(t)| = O(h^2)$ and $|q^h(t) - q(t)| = O(h^2)$ as $h \to 0$. Or, the point t_1^* is collisional (the change in basis occurs after a compression–decompression phase), and then, using the fact that our event detection is $O(h^2)$ and assumption A3, it follows that $|v(t_1^{*,+}) - v(t_1^h)| = O(h^2)$. Here we use the convention that at a collision point the numerical velocity is the post-collision velocity. At this point, we apply again the argument from the preceding paragraph to obtain that on $[t_1^*, t_2^*]$, we have that $|v^h(t) - v(t)| = O(h^2)$ and $|q^h(t) - q(t)| = O(h^2)$ as $h \to 0$.

We repeat the argument by induction for all the K^* pieces of the solution, and we reach the stated conclusion. □

Among the assumptions that we required in order to prove the previous result, we view assumptions A2 and A3 as natural technical assumptions. Assumption A4 is a consequence of Theorem 9.2 for the case where the restitution coefficient is equal to zero, including the convergence of a subsequence of the switching points, except for the part where we assume that the switching points stay distinct in the limit.

Assumption A1 is not generally satisfied for systems where the restitution coefficient is different from zero, since we could have an infinite number of bounces in a finite amount of time. This can be remediated by using a restitution coefficient that switches to zero for a small incident velocity, which is frequently done in practice (other causes for accumulation of switching points can remain, though they are rarely encountered in practice). Assumption A5 does not always hold, a distinguished example to that effect being the case where we encounter a Painlevé paradox, where a velocity discontinuity can occur in absence of a collision. The use of assumption A6 is perhaps not fortunate, but we did not succeed in proving theoretically that Theorem 4.2 for the case with non-linear constraints. Nonetheless, as we will show in the numerical results section, we have observed second-order convergence of the positions in the example, which has non-linear constraints, in Section 10.1.

Note, however, that even in the absence of the system and the resulting numerical sequence satisfying the assumptions A2–A6, we still can make use of Theorem 9.2 to prove that our scheme is well defined and converges in the weak sense defined in that result. If assumption A1 does not hold, then our scheme cannot terminate finitely, which is a problem with all event detection schemes. We can modify the algorithm in Section 8 in such a way that if a sequence of steps becomes smaller than a prescribed value, event detection other than the choice of the active set in (72) is turned off and, in particular, $\varepsilon_a > 0$ in (72). In the latter case, if that situation is reached, the method loses the second order. Nonetheless, it is perhaps fair to note that a lower bound on the time step is a fairly common practical occurrence in all time-stepping oriented numerical software, including commercial grade ODE solvers.

9.3. Comparison with other methods

The acceleration–force approaches [27, 31] are very valuable tools for simulation, but cannot offer a consistent solution for the case where a Painlevé paradox is encountered, as opposed to our approach. In addition, they need an extra set of derivatives of the data when compared to our method. Previous velocity-impulse time stepping methods, such as the ones based on an LCP [7, 12, 13] as well as the contact dynamics method [32–34] are generally based on an Euler approach, and cannot reach second order even under the best of circumstances.

The fact that our method has a more sophisticated event detection procedure compared to the original LCP-based time stepping scheme makes it perhaps less robust. On the other hand, it solves only one LCP per time step, of a comparable complexity to the one for the original LCP time-stepping method [7, 12, 13], whereas the constraints are satisfied to higher accuracy. In addition, under the conditions defined in Theorem 9.3 we can prove second-order convergence of the method. For the unconstrained case, the method provably achieves second-order converges while solving only one linear system per step and continuing to enjoy the stability properties of the trapezoidal methods; to our knowledge this is also a new result. Finally, we mention that, even in the absence of assumptions A1–A6, the geometrical constraints (the joint and non-penetration constraints) are satisfied globally with one order better, $O(h^2)$, than the methods that are based on an Euler approach, and if constraint stabilization is introduced [13], even $O(h^3)$; all this while still solving only one linear problem per step. Since practitioners tend to consider the geometrical requirements more stringent than the dynamical requirements, we believe that this makes the scheme that we have described very attractive for practical applications.

10. NUMERICAL SIMULATIONS

To verify the theoretical results described above we have implemented a Matlab version of the algorithm. The appropriate LCPs were solved using the PATH package [35]. The numerical experiments performed are concerned with the accuracy of the method and its stability in the presence of stiff forces originated in springs and dampers. To maintain the accuracy of the scheme one has to use an appropriate model for estimating the collision data. Assume that following the strategy presented in Section 7 we have detected that the earliest collision is given by $\phi^{(\bar{j})}(q)$. For simplicity we will omit the superscript $\bar{j}$. The next step will be to construct a cubic interpolation polynomial using the data $\phi(q(t))$, $\phi_q(q(t))^{\mathrm{T}} v(t)$, $\phi(q(t+h))$, and $\phi_q(q(t+h))^{\mathrm{T}} v(t+h)$. The collision time t^* will be found as the solution of the interpolating

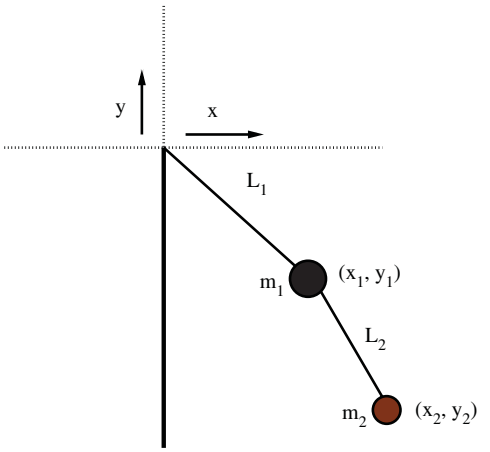

Figure 1. A double pendulum with a long wall.

polynomial in the considered time interval. The position as well as the pre-collision velocity can be computed now by evaluating $\tilde{q}$ and $\tilde{q}'$ at t^* as explained in Section 7.

10.1. A frictionless example

The rigid-body system used in the error analysis consists of a double pendulum in contact with a vertical wall, [15] as illustrated in Figure 1. The wall is situated at the position $x = 0$. The masses of the two bobs are $m_1 = m_2 = 1$, and the lengths of the massless rods are taken to be $L_1 = L_2 = 1$.

In modelling this system we have used the Cartesian co-ordinates of the point masses m_1 and m_2, denoted by $(x_1(t), y_1(t))$ and $(x_2(t), y_2(t))$, respectively. The initial position is given by

$$q(0) = (x_1(0), y_1(0), x_2(0), y_2(0)) = \left(\sin\frac{\pi}{3}, -\cos\frac{\pi}{3}, \sin\frac{\pi}{3} + \sin\frac{\pi}{5}, -\cos\frac{\pi}{3} - \cos\frac{\pi}{5}\right)$$

and the initial velocity is taken to be zero. The restitution coefficients were chosen both to be 0.1. The system was simulated for an interval of 2.5 (s), for different values of the step-size h. The x-components of m_1 and m_2 together with the total energy of the system are plotted in Figure 2.

To measure the error we have computed a reference solution $\tilde{q}^*$ obtained by running the code with $h = 2^{-20}$. The error E_h has been calculated as the 2-*norm* of the difference in positions at the final time T, or more precisely $E_h = \|q_h(T) - \tilde{q}^*(T)\|_2$. The results are presented in Table I.

The first column of Table I lists the values of the time-step used to obtained the numerical solution. The second column gives the error E_h as explained above while the fourth and the sixth columns represent the error in the joints. More precisely, if we denote by $\Theta^{(1)}(q(t)) := x_1(t)^2 + x_2(t)^2 - 1$, and $\Theta^{(2)}(q(t)) := (x_1(t) - x_2(t))^2 + (y_1(t) - y_2(t))^2 - 1$ the bilateral constraints,

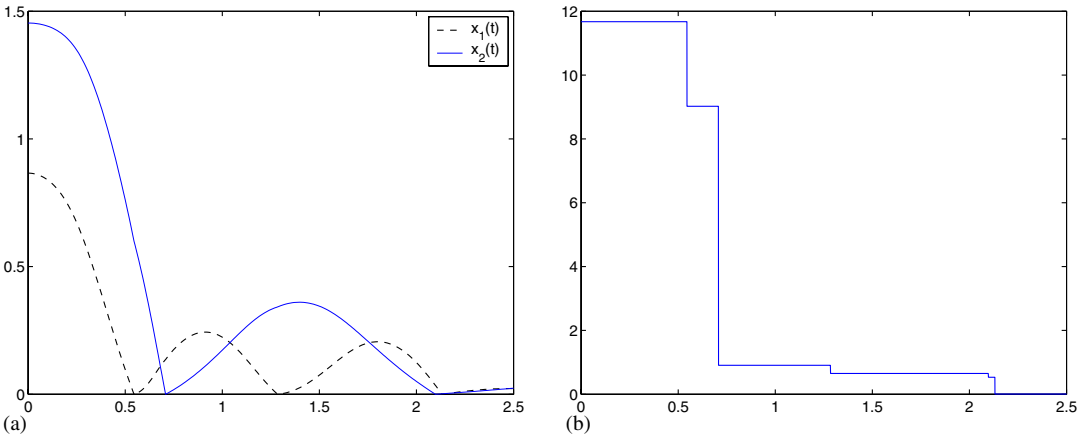

Figure 2. (a) The x-components of the positions of m_1 and m_2; and (b) total energy of the system.

Table I. The error results for the double pendulum.

h	E_h	$R_{0,h}$	$E_{1,h}$	$R_{1,h}$	$E_{2,h}$	$R_{2,h}$
2^{-5}	2.38e − 003	N/A	−1.27e − 003	N/A	−3.18e − 003	N/A
2^{-6}	6.14e − 004	3.88e + 000	−3.14e − 004	4.04e + 000	−8.29e − 004	3.84e + 000
2^{-7}	1.54e − 004	3.97e + 000	−7.82e − 005	4.01e + 000	−2.14e − 004	3.88e + 000
2^{-8}	3.95e − 005	3.91e + 000	−1.95e − 005	4.00e + 000	−5.48e − 005	3.90e + 000
2^{-9}	1.01e − 005	3.92e + 000	−4.88e − 006	4.00e + 000	−1.37e − 005	3.99e + 000
2^{-10}	2.42e − 006	4.17e + 000	−1.22e − 006	4.00e + 000	−3.44e − 006	3.99e + 000
2^{-11}	6.15e − 007	3.93e + 000	−3.05e − 007	4.01e + 000	−8.60e − 007	4.00e + 000

column 4 and 6 represent $\Theta^{(1)}(q_h(t))$ and $\Theta^{(2)}(q_h(t))$, respectively. The ratios $R_{j,h}$, $j = 1, 2$, will then measure the factor of decrease in these errors, i.e.

$$R_{j,h} = \left| \frac{\Theta^{(j)}(q_h)}{\Theta^{(j)}(q_{h/2})} \right|, \quad j = 1, 2$$

These ratios are represented as columns 5 and 7 in Table I. In the same way one defines the ratio represented by the second column, which gives the decrease in the calculated error. As we can see from Table I, by halving the stepsize we achieve an error reduced by approximately a factor of 4, which shows the second-order convergence of the method applied to this system. The same type of behaviour is observed in the error accumulated in the joints.

10.2. Stick–slip detection

We want to analyse the behaviour of the two LCP formulations (68) and (70) for a very simple mechanical system, which is depicted in Figure 3. The block in Figure 3 has been taken to have

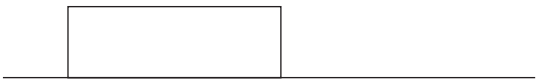

Figure 3. A block sliding on a frictional table.

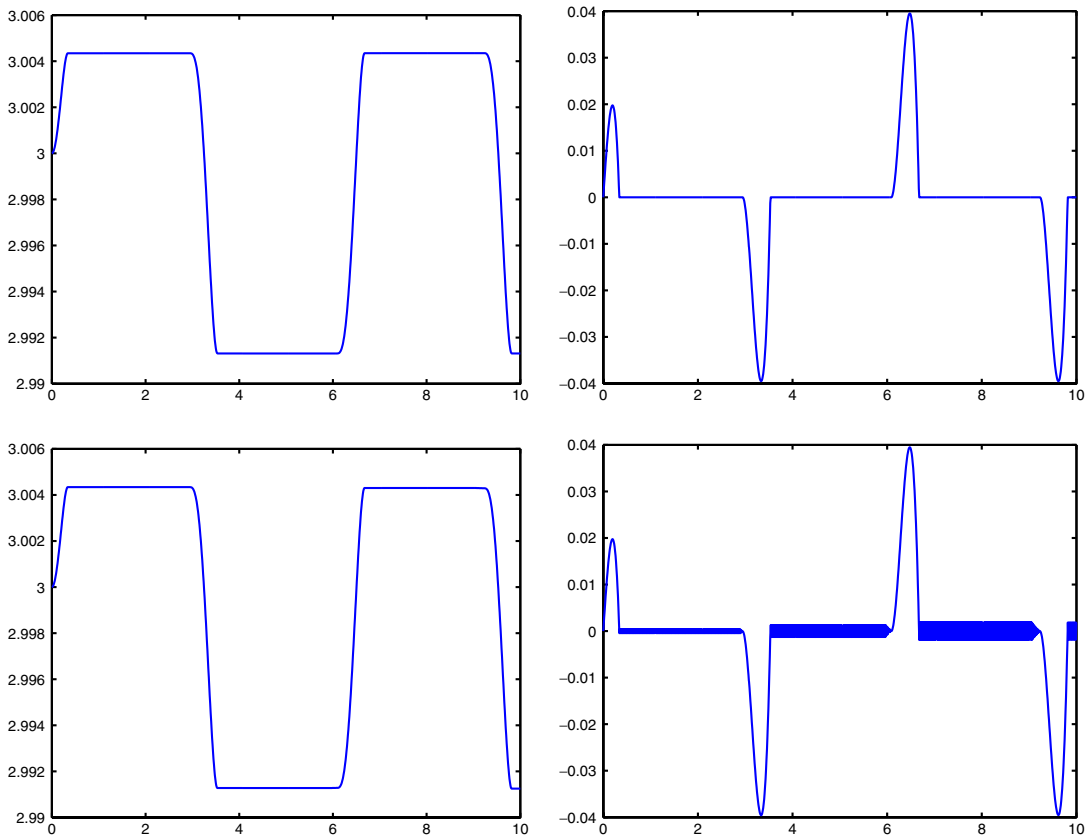

Figure 4. Numerical positions and velocities for $T_f = 10$ (s), and $h_s = 0.01$. The upper level plots were obtained with formulation (70) (position on the left, velocity on the right), while the lower level plots were obtained with formulation (68).

mass $m = 1$. An exterior force $k(t) = 8\cos(t)$ is acting on the block that is sliding on a flat table with friction coefficient $\mu = 0.8$. The initial position of the block is $q_0 = (3, 0)^{\mathrm{T}}$ and the initial velocity is $v_0 = (0, 0)^{\mathrm{T}}$. The gravity $G = (0, -mg)^{\mathrm{T}}$ is calculated with $g = 9.81$. The solutions obtained using the two formulations are shown in Figure 4. The numerical positions obtained by the two formulations cannot be distinguished one from another. However the numerical velocities tell a different story. Let us refer to formulation (68) which we call scheme II, and let us denote by $w^{(l+1)}$ the quantity given by $v^{(l+1)} + v^{(l)}$. Then the matrices corresponding to

Table II. Order analysis for scheme I.

h_s	E_h	Rate
2^{-5}	5.197e − 005	N/A
2^{-6}	1.025e − 005	5.068
2^{-7}	2.555e − 006	4.014
2^{-8}	6.551e − 007	3.899
2^{-9}	1.549e − 007	4.229
2^{-10}	3.959e − 008	3.912

Table III. Order analysis for scheme II.

h_s	E_h	Rate
2^{-5}	8.023e − 005	N/A
2^{-6}	2.310e − 005	3.473
2^{-7}	5.619e − 006	4.111
2^{-8}	1.395e − 006	4.027
2^{-9}	3.652e − 007	3.820
2^{-10}	8.246e − 008	4.429

the two formulations, i.e. Scheme II and (70), which we refer to as scheme I, are identical. The only thing that changes is $r^{(l)}$, which is defined as $r^{(l)} = [Mv^{(l)} + \tilde{k}^{(l)}; 0; 0; 0]$ for Scheme I, and $r^{(l)} = [2Mv^{(l)} + \tilde{k}^{(l)}; 0; 0; 0]$ for scheme II. Using the initial data described above, it is easy to see that both schemes will start in a 'forward-sliding' mode. When stiction occurs, scheme I will give $v_I^{(l+1)} = 0$, while scheme II gives $w^{(l+1)} = 0$, which corresponds to $v_{\mathrm{II}}^{(l+1)} = -v_{\mathrm{II}}^{(l)}$, which explains the oscillations observed when scheme II is used.

We have analysed the order of convergence for both schemes. The stick–slip detection is needed in order to ensure the predicted order of convergence. As done in the previous section, the reference solution $\tilde{q}^*$ was obtained by running the code (given by scheme I) with a small time step, i.e. $h_s = 2^{-18}$, for a period of 2 (s). The results showing a rate of convergence close to quadratic for both schemes are presented in Tables II and III.

10.3. Numerical validation of the stability of the scheme

The second numerical experiment is concerned with the stability analysis of the proposed scheme when the system is subjected to stiff forces that originate in springs and/or dampers between pairs of points of the system, in addition to a non-stiff force. In this case the force $k_1(t, q, v)$ from (87) can be written as (see Reference [8]):

$$k_1(t, q, v) = -\sum_{i=1}^{n_\gamma} \gamma_i \phi^{(i)}(q) \nabla_q \phi^{(i)}(q) - \sum_{j=1}^{n_\delta} \delta_j \nabla_q \psi^{(j)}(q) (\nabla_q \psi^{(j)}(q)^{\mathrm{T}} v) \tag{98}$$

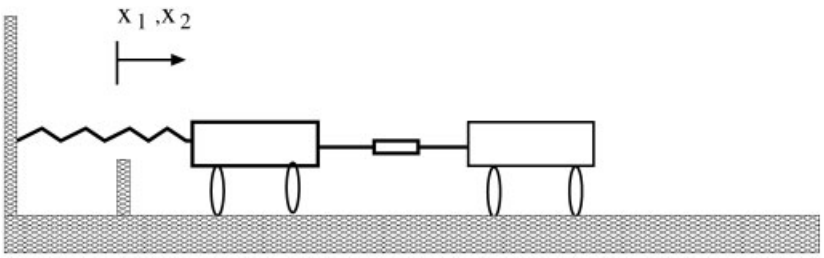

Figure 5. Two carts and a stopper.

Here γ_i, $i = 1, 2, \ldots, n_\gamma$ are spring parameters, and δ_j, $j = 1, 2, \ldots, n_\delta$, are damping parameters. The functions $\phi^{(i)}(q)$ and $\psi^{(j)}(q)$ are related to the distances between the points where the springs and the dampers are attached. They are normalized in such a way that they vanish if the springs and/or dampers are at equilibrium. We assume that the co-ordinates of the system vary in a region where $\phi^{(i)}(q)$ and $\psi^{(j)}(q)$ are differentiable. Here, the term $k_2(t, q, v)$ denotes the non-stiff forces.

We analyse the time-stepping scheme (62)–(66) with

$$\tilde{k}_q^{(l)} = -\sum_{i=1}^{n_\gamma} \gamma_i \nabla_q \phi^{(i)}(q^{(l)}) \nabla_q \phi^{(i)^{\mathrm{T}}}(q^{(l)}) \tag{99}$$

$$\tilde{k}_v^{(l)} = F(v^{(l)}) - \sum_{i=1}^{n_\delta} \delta_j \nabla_q \psi^{(j)}(q^{(l)}) \nabla_q \psi^{(j)^{\mathrm{T}}}(q^{(l)}) \tag{100}$$

We note that $\tilde{M}$ is positive definite for any value of the time step. We can use Theorem 6.1 to *guarantee that Lemke's method will return a solution* of the linear complementarity problem (68), (69) that needs to be solved to obtain the next value of the velocity.

We are interested in the behaviour of our discrete model when the stiffness parameters tend to infinity. In particular we investigate the case whether the kinetic energy at the next time step remains bounded when the stiffness parameters tend to infinity. If the latter does occur, we conclude that our scheme is stable for the spring/damper type of stiffness. While we do not present in this paper any theory concerning the stiff stability of the method, such results can be obtained by natural extensions of the results in Reference [8].

The mechanical system considered here consists of two carts of unit mass on a flat surface, [5, 15]. The left-most cart is connected by a spring with elasticity γ to an immovable wall. A stopper acting on the left-most cart is placed at $x = 0$. The spring is at equilibrium when the position of the left cart is $x = 0.2$. The two carts are connected by a damper of stiffness δ. In all the examples presented here, the original position of the second cart will be within 5 units from the left-most cart, and the friction coefficient between the carts and the floor is 0.05. A graphical representation of this system is given in Figure 5. All the simulations are done with a time-step $h = 0.01$ and the restitution coefficient between the left-most cart and the stopper is taken to be $e = 0.3$.

The generalized position of the system is given by the x-co-ordinates of the carts: $x_1(t)$ represents the position of the left-most cart and $x_2(t)$ the position of the right cart. We are

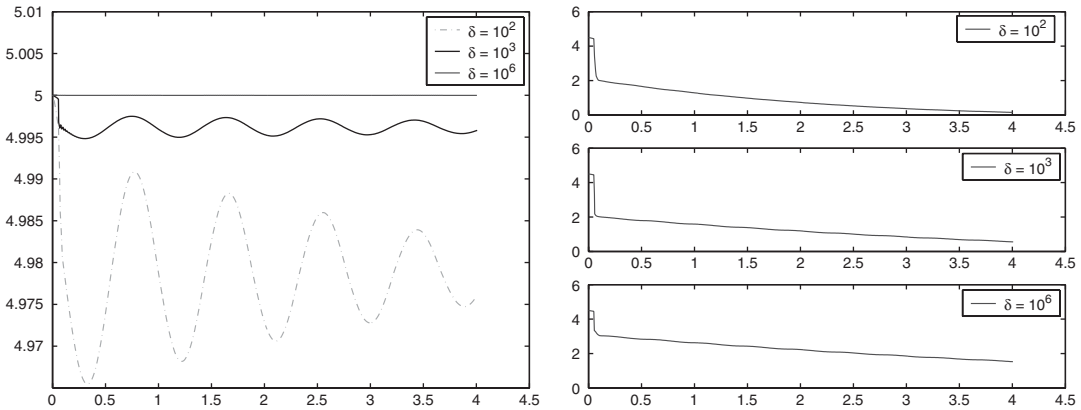

Figure 6. Results for increasing values of δ. On the left we have plotted the variation of the distance between the carts, $x_2(t) - x_1(t)$, and on the right the total energies.

interested in the behaviour of the system when the stiffness parameters γ or δ tend to infinity. We conduct the following experiments:

- *Increasing the damping coefficient* ($\delta \to \infty$). The initial position is $q(0) = (0.1, 5.1)$ and the initial velocity $v(0) = (-2, -2)$. The system is simulated for a total of 4 (s). The elasticity parameter γ is fixed to 100. Figure 6 depicts on the left, the distance between the two carts and on the right, the total energy of the system, for $\delta = 10^2, 10^3$ and 10^6. We can see that as the stiffness increases to infinity the variation in the distance between the carts approaches zero, which means that the method is indeed stable. We also notice that the total energy of the system is bounded even if the stiffness goes to infinity.
- *Increasing the elasticity parameter* ($\gamma \to \infty$). We start with the spring in equilibrium, so the initial position is $q(0) = (0.2, 5.2)$, while the initial velocity is $v(0) = (-3, -3)$. The system is simulated for a total of 2 (s), with the stiffness of the damper fixed to $\delta = 10$ and the elasticity constant of the spring taking the values $\gamma = 10^2, 10^4$ and 10^6. Figure 7 depicts on the left, the distance between the left-most cart and the equilibrium position of the spring, and on the right the total energy of the system for the values of γ discussed above. As in the previous case the total energy of the system is bounded with respect to the stiff parameter, which shows the stability of our scheme.

10.4. The use of the linearized trapezoidal method in a virtual prototying environment

The time-stepping scheme from this work was implemented into an industrial-grade virtual prototyping simulation package, UMBRA [36], by a group at Sandia National Laboratories that included two of the authors of this paper (Trinkle and Potra). Constraint stabilization is achieved by including in the right-hand side of our LCP a term that is connected to the linearization of the constraints [13].

Two of the applications simulated by this package are presented in Figure 8. The image on the left depicts a four-wheeled rover with a robot arm traversing an unknown landscape. With a terrain map and a traction model, one could plan optimal exploration sorties. The image on the right show a pawl (2.3 mm in length) from a micro-mechanical machine being pushed into

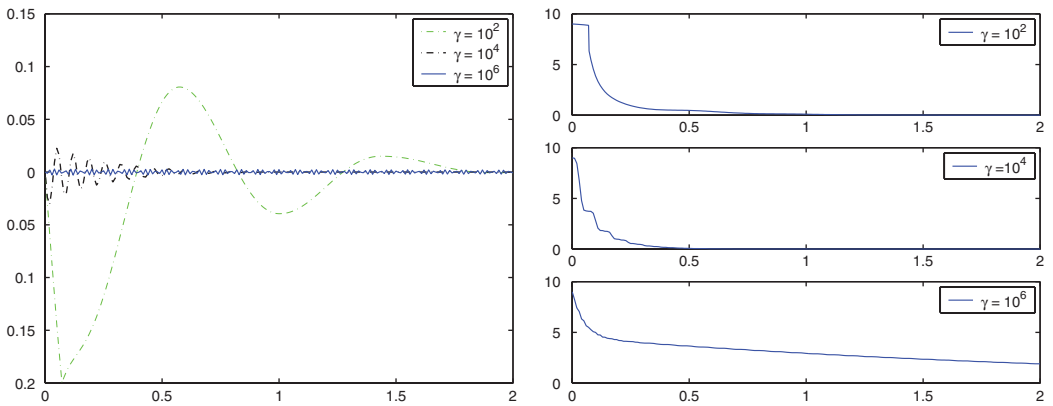

Figure 7. Results for increasing values of γ. On the left we have plotted the distance between the left-most cart and the equilibrium position of the spring and on the right the total energies.

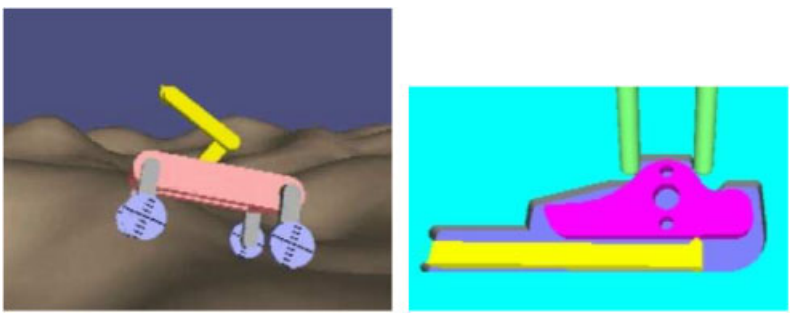

Figure 8. Examples of UMBRA simulations, that uses the time-stepping scheme developed in this work.

an unusually shaped hole as a step in an assembly operation. Rigid body dynamics is used to design the best shape of the pawl, hole, and cantilever beam (below pawl) to make insertion easy. Movies of these systems can be viewed at http://www.cs.rpi.edu/~trink/.

Two main problems arose soon after our initial implementation was complete. First, the size of the active set of contacts was prohibitively large, often on the order of 1000. This problem was traced back to the geometric representations of the bodies and the collision detection algorithm. Our geometric models were obtained from CAD systems that allowed the designer to create smooth curved features, such as drilled holes. However, the best available collision detection methods operate on polyhedral bodies. Thus, CAD tools were used to approximate the bodies as polyhedra, with more polygons needed to represent the curves with greater accuracy. The planar faces of the laminar pawl and the fixture disk (the part with the hole) shown in Figure 8 were each decomposed into about 50 coplanar triangles.

State-of-the-art collision detection tools are 'uninformed' with respect to the dynamics of the system. The best they do is return the distance between a pair of bodies and an estimate of

the penetration depth when they overlap. For stable simulation, the active set should contain not just the closest geometric feature pairs (vertex–face and edge–edge pairs), but others that could come into contact in the next time step. For example, if a cube is dropped onto a plane, it will eventually come to rest on a face. Even before the cube comes to rest, each time step should include four potential contacts. If only the closest vertex is in the active set, then the cube will oscillate for ever. To combat this problem, we modified Sandia's proprietary collision detection algorithm to return all feature pairs within ε of each other.

This algorithm worked well for the land rover shown in Figure 8, but found hundreds of potential contacts for the sliding pawl because many of the interior edges of the pawl were very near many of those of the fixture disk. The same was true for face–vertex pairs. It was not uncommon to have 100–300 or more contacts in the active set. Each contact gives rise to n_f+2 variables in the MLCP, yielding extremely time consuming time-stepping problems. For similar problems when there are two complex planar faces in contact, it is easy to construct filters to identify the feature pairs within the convex hull (these are redundant) of the contact patch, but in general settings, numerical errors in the triangulation makes it difficult to determine redundant contacts. A general method was developed and implemented in UMBRA. When applied to the pawl problem, the size of the MLCP was reduced by an order of magnitude and simulation times reduced three order of magnitude.

The other problem confronted was related to our representation of non-penetration constraints. In particular, these constraints are linearized and then written as a system of (conjunctive) linear inequalities. When viewing the system in configuration space (C-space), the configuration of the system is a point and the collision-free portion of C-space near the point is polytope—typically non-compact. However, when a vertex approaches another vertex or an edge, the local collision-free region is non-convex and thus cannot be properly represented by a system of linear inequalities. The pile of jacks shown at: `http://www.cs.rpi.edu/~trink/rigid_body_dynamics.html` is the last frame of a simulation that failed for these reasons. The solution to this problem is still under investigation. A first solution attempt in which the disjunctivity of the constraints is represented by adding more complementarity relationships to the MLCP is given in Reference [37].

11. CONCLUSIONS AND FUTURE WORK

We have presented a linearized trapezoidal scheme for the simulation of rigid multibody dynamics with contact and friction. The scheme solves only one linear complementarity problems per time step while achieving second-order convergence, a fact that is both proven analytically, for the case where the constraints are linear, and demonstrated by numerical simulations. Stiffness can be accommodated by our scheme, by using a suitable approximation of the Jacobian of the external forces. The scheme needs only first-order derivatives of the problem data. To our knowledge, this is the first presentation of a scheme with these properties. In addition, the scheme has been successfully implemented in and industrial grade rigid multibody dynamics simulating system, UMBRA [36].

Some of the issues that still need to be resolved are achieving constraint stabilization without using any additional projection [13], maintaining second-order convergence while working with a convenient approximate Jacobian, eventually by switching to an exact Jacobian when the time step is sufficiently small, as establishing analytically the order of convergence of the scheme

for the case where the constraints are non-linear. Although none of these issues are completely addressed here, this work contains much of the key results which have allowed to prove or demonstrate the similar conclusions for the linearized backward Euler scheme [8, 13, 14, 25] which leads us to believe that these issues will likely be successfully resolved.

ACKNOWLEDGEMENTS

The authors are very grateful to the editor and the anonymous referees whose comments have substantially improved the quality of the paper. The authors are grateful to Todd Munson and Michael Ferris for providing PATH [35], a solver for the general linear complementarity problem. Florian Potra, Bogdan Gavrea and Jeff Trinkle have been supported by the National Science Foundation through Grant DMS-0139701. Mihai Anitescu was supported by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38.

REFERENCES

1. Song P, Trinkle JC, Kumar V, Pang J-S. Design of part feeding and assembly processes with dynamics. *Proceedings*, *IEEE International Conference on Robotics and Automation*, April 2004.
2. Painlevé P. Sur le lois du frottement de glissemment. *Comptes Rendus de l' Académie des Sciences Paris* 1895; **121**:112–115; 1905; **141**:401–405, 546–552.
3. Mason MT, Wang Y. On the inconsistency of rigid-body frictional planar mechanics. *Proceedings*, *IEEE International Conference on Robotics and Automation*, April 1988; 524–528.
4. Moreau J-J. Bounded variation in time. *Topics in Nonsmooth Mechanics*. Birkhäuser: Basel, Boston, MA, 1988; 1–74.
5. Brogliato B. *Nonsmooth Impact Mechanics*. Springer-Verlag London Ltd.: London, 1996. (Models, dynamics and control)
6. Stewart DE. Rigid-body dynamics with friction and impact. *SIAM Review* 2000; **42**(1):3–39 (electronic).
7. Anitescu M, Potra FA. Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. *Nonlinear Dynamics* 1997; **14**:231–247.
8. Anitescu M, Potra FA. A time-stepping method for stiff multibody dynamics with contact and friction. *International Journal for Numerical Methods in Engineering* 2002; **55**(7):753–784.
9. Monteiro Marques MDP. *Differential Inclusions in Nonsmooth Mechanical Problems*: *Shocks and Dry Friction*. Progress in Nonlinear Differential Equations and Their Applications, vol. 9. Birkhäuser: Basel, Boston, MA, Berlin, 1993.
10. Moreau J-J. Standard inelastic shocks and the dynamics of unilateral constraints. In *Unilateral Problems in Structural Mechanics*, del Piero G, Maceri F (eds). C.I.S.M. Courses and Lectures, vol. 288. Springer: New York, Vienna, 1985; 173–221.
11. Stewart DE, Trinkle JC. Dynamics, friction, and complementarity problems. *Complementarity and Variational Problems* (*Baltimore*, *MD*, *1995*). SIAM: Philadelphia, PA, 1997; 425–439.
12. Stewart DE, Trinkle JC. An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction. *International Journal for Numerical Methods in Engineering* 1996; **39**(15):281–287.
13. Anitescu M, Hart GD. A constraint-stabilized time-stepping approach for rigid multibody dynamics with joints, contact and friction. *International Journal for Numerical Methods in Engineering* 2004; **60**(14):2335–2371.
14. Stewart DE. Convergence of a time-stepping scheme for rigid body dynamics and resolution of Painlevé's problems. *Archive for Rational Mechanics and Analysis* 1998; **145**(3):215–260.
15. Tzitzouris JA. Numerical resolution of frictional multi-rigid-body systems via fully implicit time-stepping and nonlinear complementarity. *Ph.D. Thesis*, The Johns Hopkins University, 2001.
16. Hairer E, Wanner G. *Solving Ordinary Differential Equations. II* (2nd edn). Springer: Berlin, 1996. (Stiff and differential-algebraic problems)
17. Haug EJ. *Computer Aided Kinematics and Dynamics of Mechanical Systems*. Allyn and Bacon: Boston, MA, 1989.

18. Atkinson KE. The numerical-solution of a nonlinear boundary integral-equation on smooth surfaces. *IMA Journal of Numerical Analysis* 1994; **14**:461–483.
19. Potra FA, Rheinboldt WC. Differential geometric techniques for solving differential algebraic equations. In *Real-Time Integration Methods for Mechanical Systems Simulation*, Haug E, Deyo R (eds). Springer: Berlin, 1991; 155–192.
20. Potra FA, Rheinboldt WC. On the numerical solution of Euler–Lagrange equations. *Mechanics of Structures and Machines* 1991; **19**(1):1–18.
21. Anitescu M, Cremer JF, Potra FA. Formulating 3d contact dynamics problems. *Mechanics of Structures and Machines* 1996; **24**(4):405–437.
22. JC, Vaněček G. Isaac: building simulations for virtual environments. *Workshop IFIP IC 5 WG 5.10 on Virtual Environments*, Coimbra, Portugal, 1994.
23. Stewart DE. Rigid-body dynamics with friction and impact. *SIAM Review* 2000; **42**(1):3–39.
24. Murray RM, Li Z, Shankar Sastry S. *A Mathematical Introduction to Robotic Manipulation*. CRC Press: Boca Raton, FL, 1993.
25. Stewart DE, Trinkle JC. An implicit time-stepping scheme for rigid-body dynamics with inelastic collisions and Coulomb friction. *International Journal for Numerical Methods in Engineering* 1996; **39**:2673–2691.
26. Anitescu M, Potra FA. Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. *Nonlinear Dynamics* 1997; **14**:231–247.
27. Glocker C, Pfeiffer F. Multiple impacts with friction in rigid multi-body systems. *Nonlinear Dynamics* 1995; **7**:471–497.
28. Pfeiffer F, Glocker Ch. *Multibody Dynamics with Unilateral Contacts*. Wiley Series in Nonlinear Science. Wiley: New York, 1996. (A Wiley-Interscience Publication)
29. Stewart DE. Convergence of a time-stepping scheme for rigid body dynamics and resolution of Painleve's problems. *Archive for Rational Mechanics and Analysis* 1998; **145**(3):215–260.
30. Anitescu M. Optimization-based simulation of nonsmooth dynamics. *Mathematical Programming* 2004, in press.
31. Trinkle J, Pang J-S, Sudarsky S, Lo G. On dynamic multi-rigid-body contact problems with Coulomb friction. *Zeitschrift fur Angewandte Mathematik und Mechanik* 1997; **77**:267–279.
32. Moreau JJ, Jean M. Numerical treatment of contact and friction: the contact dynamics method. *Engineering Systems Design and Analysis*, 1996; **4**:201–208.
33. Moreau JJ. Numerical aspects of the sweeping process. *Computer Methods in Applied Mechanics and Engineering* 1999; **177**:329–349.
34. Jean M. The nonsmooth contact dynamics method. *Computer Methods in Applied Mechanics and Engineering* 1999; **177**:235–257.
35. Ferris M, Munson T. Interfaces to PATH 3.0: design, implementation and usage. *Computational Optimization and Applications* 1999; **12**:207–227.
36. Gottlieb EJ, McDonald MJ, Oppel FJ, Rigdon JB, Xavier PG. The umbra simulation framework as applied to building hla federates. *Proceedings of the 2002 Winter Simulation Conference*, San Diego, California, 2002; 981–989.
37. Egan KT, Berard S, Trinkle JC. Sensorless acquisition of multiple contact points between planar parts. *Multipoint Interaction Workshop*, April 2004, in press.

