Elsevier Editorial System(tm) for Mathematics and Computers in Simulation
Manuscript Draft

Manuscript Number:

Title: Spook: a variational time-stepping scheme for rigid multibody systems subject to dry frictional contacts

Article Type: Special Issue: DDS 2010

Keywords: constrained mechanics;
classical ghost variables;
variational mechanics;
non-smooth mechanics;
Coulomb friction;
nonlinear complementarity problem;
constraint stabilization;
regularization;
simulation based training;
multibody dynamics;
heavy machinery simulation;

Corresponding Author: Dr. Claude Lacoursiere, Ph.D.

Corresponding Author's Institution: Umea University

First Author: Claude Lacoursiere, Ph.D.

Order of Authors: Claude Lacoursiere, Ph.D.; Mattias Linde, MSc

Abstract: The present work introduces a stable, semi-implicit, one stage integration scheme for rigid multibody systems subject to mixed holonomic and nonholonomic constraints as well as dry frictional impacts and contacts. A stable direct-iterative splitting scheme to solve the latter is also presented, and is shown to be suitable for real-time simulation of large multibody systems, such as those used for 3D graphics, simulation-based training systems. We use a Lagrangian framework in conjunction with the discrete-time D'Alembert's principle to introduce physically motivated singular perturbations which regularize and stabilize the numerical method. Lower bounds on the perturbations which guarantee numerical stability are provided, and their physical validity is demonstrated at the numerical level. The theoretical formulation uses massless ghost particles in the Lagrangians of mechanical systems. The coordinates and velocities of these converge strongly to Lagrange multipliers for holonomic and nonholonomic constraints, respectively, in the singular limit of zero relaxation. The ghost formulation allows for the systematic treatment of non-ideal, dissipative, constitutive laws such as Coulomb friction. A variational model for the latter is constructed and proven to be solvable in discrete time. Several splitting schemes are investigated mathematically and compared at the numerical level. Convergence and non-convergence properties of these are demonstrated mathematically and numerically.

SPOOK: a variational time-stepping scheme for rigid multibody systems subject to dry frictional contacts

Claude Lacoursière[a,*], Mattias Linde[b]

[a] *HPC2N, UmeåUniversity, SE-901 87 Umeå, Sweden*
[b] *Department of Computing Science, UmeåUniversity, SE-901 87 Umeå, Sweden*

Abstract

The present work introduces a stable, semi-implicit, one stage integration scheme for rigid multibody systems subject to mixed holonomic and nonholonomic constraints as well as dry frictional impacts and contacts. A stable direct-iterative splitting scheme to solve the latter is also presented, and is shown to be suitable for real-time simulation of large multibody systems, such as those used for 3D graphics, simulation-based training systems. We use a Lagrangian framework in conjunction with the discrete-time D'Alembert's principle to introduce physically motivated singular perturbations which regularize and stabilize the numerical method. Lower bounds on the perturbations which guarantee numerical stability are provided, and their physical validity is demonstrated at the numerical level. The theoretical formulation uses massless ghost particles in the Lagrangians of mechanical systems. The coordinates and velocities of these converge strongly to Lagrange multipliers for holonomic and nonholonomic constraints, respectively, in the singular limit of zero relaxation. The ghost formulation allows for the systematic treatment of non-ideal, dissipative, constitutive laws such as Coulomb friction. A variational model for the latter is constructed and proven to be solvable in discrete time. Several splitting schemes are investigated mathematically and compared at the numerical level. Nonvergence and non-convergence properties of these are demonstrated mathematically and numerically.

Keywords: constrained mechanics, classical ghost variables, variational mechanics, non-smooth mechanics, Coulomb friction, nonlinear

*Corresponding author
Email addresses: claude@hpc2n.umu.se (Claude Lacoursière), linde@cs.umu.se (Mattias Linde)

complementarity problem, constraint stabilization, regularization, simulation based training, multibody dynamics, heavy machinery simulation,

1. Introduction

Simulation based training systems have proven useful in application domains ranging from aircraft and ship pilots, surgeons, and ground vehicle operators just to name a few. These systems involve interactive physics simulations coupled with 3D rendering, motion platform, as well as dynamic input controls and output sensors. Each of these components imposes hard constraints on computational budgets and thus, fast and stable numerical integration methods for mechanical systems are required. The soft real-time nature of these systems also demands for fixed-step time integration techniques. In practice, this means that one stage methods of first or second order methods are to be favored when at all possible, justifying the work we present below which focuses on these. The training context also imposes requirements on faithfulness, i.e., the simulated motion must satisfy the laws of physics to a sufficient degree as to avoid false training. This is not identical to accuracy in numerical analysis terms though since local errors have little or no bearing on global ones, or on geometric properties of trajectories, something fundamental in physics. The faithfulness requirement rules out using very stable implicit methods which have too much numerical dissipation such as the implicit Euler method. This is why we focus on variational mechanical integrators (31, 49) which have the desired properties at least in general.

Rigid multibodies are used extensively to model physical systems used in operator training systems. This is a natural choice for ground vehicle, but it can also apply to lumped element models of cables and beams, for instance. This leads to the study of discrete mechanical systems subject to mixed holonomic and nonholonomic constraints, as well as dry frictional contacts and impacts. The latter is particularly important in the simulation of ground vehicles such as cranes, wheel loaders and tree harvesters to name a few, for which there is strong demand from vocational training institutes . Training scenarios call for dynamic reconfiguration meaning that system reduction and analysis prior to simulation is impossible. For instance, geometric collision detection generates variable numbers of contact constraints

between integration steps, a tree harvesting machine cuts trees into logs, a wheel loader trainer can involve attaching and removing equipment from the basic vehicle, etc. One can therefore not assume that constraints are non-degenerate, non-redundant, or even consistent. The nature of multibody systems is such that the inertia opposing a force is configuration dependent and unpredictable. Very high velocity can also develop as in the case of a long chain attached at an anchor and released from a horizontal position. This leads to whiplash effects and are essentially unavoidable. This calls for robust methods which recover gracefully with predictable errors from such cases. This can be achieved by introducing suitable regularization and dissipation in the physics model as we do in this article.

As coordinate reduction techniques are not generally applicable to such systems, we use extended coordinates and compute constraint forces explicitly which requires the solution of systems of nonlinear equations. For the very stiff systems considered, this is a numerical challenge. But given the real-time context and fixed computational budget, it is not possible to enforce strict bounds on constraint violation and so a stabilization scheme is needed to avoid constraint drift.

Constructing stable and faithful time-stepping schemes for constitutive laws, dry friction in particular, and the treatment of impacts is yet another difficulty. One can use discrete-time mechanics for these once such laws are formulated in the variational framework of D'Alembert's principle and though there is recent work in that direction (see 43) there is yet much to do especially with regards to time discretization. There is at least one variational formulation of dry friction (see 50) which has been used for discrete time-stepping, but this does not have a proof that the resulting time-stepping equations have a solution.

Efficient and reliable numerical methods for computing dry frictional forces do not yet exist. Though there is a number of iterative approximation schemes which can deliver reasonable estimates quickly, these have dubious convergence properties which have not received sufficient attention as of yet. The magnitude of global residual errors of simple iterative schemes make them simply unusable for applications involving multibody systems subject to hard constraints.

Variational methods for integrating the DAEs of motion of multibody systems exist, are well-known and widely used in some circles. The first of these, SHAKE, is of first order in velocities and second order in position and requires the solution of one system of nonlinear equations per step. The second is

RATTLE (31) which needs to solve two such systems. Extensions to noholonomic constraints are possible as well (24). Extensions to non-conservative systems are possible as well (49), and even friction can be addressed (50) to some extent. However, the nonlinear equations which needs to be solved in these methods may not have a solution when the velocities are too high (17). This is due to a feature common to all these integrators which is a restriction on the search direction used to locate the constraint surface and is explained below in Fig. 23.

A robust integrator must me able to handle such cases and recover gracefully from such configurations which are essentially outside of the range of what is visible for the user at a fixed frame rate, as is the case for our simulation.

Projection methods offer an interesting alternative as they do not suffer from non-existence. With sufficient computational work, one can always project back to the constraint surface. These have been used and analyzed extensively (33). Recent advances in these include symmetric projection methods (30) which have better energy conservation properties. Though it is possible to use a projection method in combination with a one stage integration scheme such as Verlet's, one still requires the nearly exact solution of linear systems of equations, and again for the case of high velocities, a simplified Newton's method might fail in the same way that SHAKE and RATTLE do and for the same reason.

Previous work on index reduction (33) and constraint stabilization for integrating the differential algebraic equations (DAEs) of motion of rigid multibody systems (9–12, 14) is unsatisfactory as parametrization of existing schemes is never based on physical parameters. Most of these require high order integration as well, in part due to high frequency oscillations introduced by the constraint stabilization itself. Likewise, penalty methods based on spring and damper systems suffer either from strong numerical dissipation, as is the case for the implicit first order Euler or high oscillations for the implicit midpoint method (15). These two integration methods are interesting in the way they mirror each other as Euler's method always dissipate too much, but the implicit midpoint can never dissipate enough, and this gets worse at higher penalties. Worse yet, few of the stabilization methods can be proven to have linear stability and one can easily construct counter examples.

Recent work on penalty methods includes SyLiPN (55) for instance in which a Newmark method is linearized in such a way as to preserve symplecticity, which is not the case for the linearization of the implicit midpoint

method for instance. However, that article provides results where the relaxation parameter is 10^{-6} at time steps of $h = 0.1$. Our method uses and requires such small values as 10^{-12} for time steps of $1/60 \approx 16.7ms$ which is the requirement for real-time rendering which works best at 60Hz. Such small regularizations are required as our systems involve such large mass ratios and they are in fact limited only when there is constraint degeneracy which can cause severe ill-conditioning. Large contact problems are usually strongly degenerate for instance. Other types of penalty methods break down entirely in this regime because the condition numbers of the linear systems grow proportionally with the inverse of the regularization, and because the high oscillations must be damped in some way as discussed in Sec. 3, and the correct damping is usually strongly dependent both on the details of the integration scheme and, much worse, on the system to simulate. Clearly, this cannot work for dynamically reconfigurable systems, and especially so for interactive simulations.

But none of the methods listed above allow for regularization and thus require full rank constraint Jacobians, something that cannot be assumed in general. Indeed, the simple slider-crank mechanism has constraint degeneracy at four different configurations (9) (see also Fig. 1).

Our specific contribution here is the Spook stepper which is a semi-implicit regularized method which requires the solution of a single, well-conditioned linear system per step, and which is proven to have linear stability. The validation of regularization parameters in correspondence to their physical interpretation is briefly demonstrated (this was analyzed more extensively previously (see 54)). A variational formulation of nonholonomic and non-ideal constraints are provided via the ghost theory and these are then discretized using discrete-time D'Alembert's principle. A nonlinear variational formulation of Coulomb friction is also presented along with proof of solvability. Properties of the resulting nonlinear complementarity problem (NCP) which needs to be solved at each time-step are investigated, as well as properties of different splitting schemes. One of these is shown to be bounded and others are shown to have erratic properties in numerical experiments.

The rest of the paper is organized as follows. We demonstrate how penalty and regularization methods differ numerically in Sec. 3. Sections 4 and Sec. 5 introduce classical ghost particles and their relation to Lagrange multipliers of constrained systems for holonomic, nonholonomic, and nonideal systems. The connection to impacts is briefly discussed in Sec. 6. The discrete-time variational principle is covered in Sec. 7 and applied to ghost particles in

Sec. 8. This is used to construct the SPOOK stepper in Sec. 9. The linear stability of SPOOK is demonstrated in Section 10. Discrete-time frictionless impacts Sec. 11. Section 12 presents a non-linear variational formulation of Coulomb friction, the properties of which are investigated in Sec. 13 along with various linearization schemes. Solution methods solving contact forces based on splitting schemes are described and analyzed in Sec. 14, Sec. 15 and Sec. 16. Numerical illustrations and counter-examples are presented in Section 17. Mathematical proofs have been relegated to Appendices in Sec. 19 and Sec. 20.

2. Notation

The notation is not entirely uniform or consistent because we study different aspects of the problem, namely, time stepping, the complementarity problem resulting from dry frictional contacts, and the numerical methods for solution of said. The notation pertinent to each section is clarified but symbols can change meaning from section to section.

Upper case letters refer almost exclusively to matrices, lower case variables are always vectors. Greek letters are used both for scalar parameters and ghost variables, which are vectors.

Symmetric and bisymmetric matrices are represented with symmetric letters.

Letter h is always a time step, and k is the discrete time.

Inner products are written as $x^T y$ or $x \cdot y$, and $(\cdot)^T$ is always a transposition operator.

Indices which are not representing vector or matrix components or discrete time index are written as $x^{(j)}$. Subscripts x_k usually refer to discrete time and x_i or a_{ij} are vectors or matrix elements.

Time derivatives are written as $\dot{x}$.

All binary operations are understood componentwise when vectors are concerned.

3. Penalties and regularizations

Consider a mechanical system with generalized coordinates and velocities $x, v = \dot{x}$ with constant mass M, and subject to a strong force derived from a potential of the form $U_\epsilon = (1/(2\epsilon))\|g(x)\|^2$, where $g : \mathbb{R}^n \mapsto \mathbb{R}^m, m \leq n$ is an indicator function such that $g(q) = 0$ is a smooth manifold, and the

Jacobian $G = \partial g/\partial q$ has full row rank. The potential then generates a force $f = -(1/\epsilon)G^T g$. We first consider the linear case where $g = Gq + b$, choose the implicit midpoint rule (see 31) to integrate the equations of motion

$$M\dot{v} = f(x, v) \tag{1}$$

as per Newton's second law. Writing $x_{k+1/2} = (1/2)(x_k + x_{k+1})$ and similarly for $v_{k+1/2}$, the stepping scheme reads

$$\begin{aligned} x_{k+1} &= x_k + hv_{k+1/2}, \text{ and} \\ v_{k+1} &= v_k + hM^{-1}f(x_{k+1/2}, v_{k+1/2}), \end{aligned} \tag{2}$$

where $f(x, v)$ is the total force applied on the system. The stepping scheme amounts to solving

$$\begin{aligned} \left[M + \frac{h^2}{4\epsilon}G^T G\right] v_{k+1} &= hf(x_k) + \left[M - \frac{h^2}{4\epsilon}G^T G\right] v_k \\ x_{k+1} &= x_k + v_{k+1/2} \end{aligned} \tag{3}$$

This system of equations has very nice properties as the stepping matrix

$$K_\epsilon = \left[M + \frac{h^2}{4\epsilon}G^T G\right]^{-1} \left[M - \frac{h^2}{4\epsilon}G^T G\right] \tag{4}$$

is a Cayley transform and has pure imaginary unit eigenvalues. This means that this scheme produces iterates x_k, v_k which are uniformly bounded. One can show that this stepping scheme preserves linear and angular momentum for general mechanical systems, as well as all invariants which are quadratic functions of the position and velocities (31). This stepping scheme is also symplectic and can be shown to produce the solution of a physical system which differs by $O(h^2)$ from the original one. Not only that but the numerical trajectories shadow those of the physical system, meaning that the former intersect the latter at each time step.

In the limit where $\epsilon \downarrow 0$ however, the trajectory generated satisfies $Gv_{k+1} = -Gv_k$ and similarly for Gx_k. Numerically though, very small values $0 < \epsilon \ll 1$ are unstable, even with the initial conditions $g(x_0) = Gv_0 = 0$. Detailed analysis of this case can be found in the literature (e.g. 15, 16). Linear damping of the form $-\gamma Gv$ changes the stepping matrix to

$$K_{\epsilon,\gamma} = \left[M + \left(\frac{h^2}{4\epsilon} + \frac{h\gamma}{2}\right) G^T G\right]^{-1} \left[M - \left(\frac{h^2}{4\epsilon} + \frac{h\gamma}{2}\right) G^T G\right] \tag{5}$$

which does nothing to remove the oscillatory parts and fails to damp the system but has the same issues with regards to ill-conditioning and fast oscillations in the limit $\epsilon \downarrow 0$, despite damping. This kind of noise is dramatic even when using direct methods since values of $h^2/\epsilon = O(10^{-\alpha})$ after rescaling so that $M = O(1)$ loose α digits of accuracy per step. Assuming double precision, there are no significant digits left after $10^{16-\alpha}$ steps which, for simulations, assuming one hundred steps per second of real-time, occurs after a few seconds or a few minutes at best. At that point, the weak forces, gravity for instance, play no role in the dynamics. This effect can be seen in all implicit methods when using implicit integrators directly on the strong forces though the symplectic methods, such as the implicit midpoint rule, are much better at separating the fast oscillations from the rest. A case in point is the implicit Euler method which can mask the force of gravity even for moderately strong forces as seen in Fig. 5.

None of this should be surprising since there is apparently no limit to U_ϵ as $\epsilon \downarrow 0$ to the stated problem, not in the form given above in any case. One could divide through with ϵ and recover a reasonable limit but that would not remedy the problem of bad conditioning resulting from the ratio M/ϵ.

There is however an alternative. Introduce the auxiliary variable $\lambda = -(h/\epsilon)g(x_{k+1/2})$ and after rearrangement, the stepping equations read

$$\begin{bmatrix} M & -G^T \\ G & \frac{\epsilon}{4h^2} \end{bmatrix} \begin{bmatrix} v_{k+1} \\ \lambda \end{bmatrix} = \begin{bmatrix} Mv_k + hf_k \\ -\frac{4}{h}g(x_k) \end{bmatrix} \qquad (6)$$
$$x_{k+1} = x_k + hv_{k+1/2},$$

where we have introduced weaker forces $hf(x_k)$ which do not require implicit integration by assumption. The discrete-time variational principle introduced below allows for such mixed discretization, though this can also be understood simply as an approximation one would make when handling nonlinear cases and relying on quasi-Newton methods for solving the systems of equation. The matrix appearing in Eqn. (6) is henceforth denoted with K and its symmetric version, a saddle point matrix, is denoted H, with general forms

$$K = \begin{bmatrix} M & -G^T \\ G & T \end{bmatrix}, \quad \text{and } H = \begin{bmatrix} M & G^T \\ G & -T \end{bmatrix}, \qquad (7)$$

where T is assumed symmetric and positive semidefinite in general, though our numerical strategy and theoretical constructions are all built around the assumption that T is block diagonal and strictly positive definite with

spectral radius $\rho(T) \geq \kappa(H)\text{tol}$ where $\kappa(H)$ is the condition number of H and tol is the machine precision.

The unusual form of the matrix in Eqn. (6) is called bi symmetric (see 23, Ch. 1, p 4), i.e., the sum of a symmetric positive definite matrix with that of an antisymmetric one. This is still positive definite though (see 29, also).

At the numerical level at least, the stepping scheme in Eqn. (6) is stable even at the limit $\epsilon = 0$ and suffers no ill-conditioning. In fact, such bisymmetric problems which are strictly positive definite lend themselves to numerical solution with LDLT factorization which is backward stable for the case where $\epsilon \neq 0$ at the machine precision level (22).

One can also see that the form given in Eqn. (6) can be solved via Schur complements namely

$$\begin{aligned} W_\epsilon &= \left[M + \frac{h^2}{4\epsilon} G^T G \right], \text{ solving for } v_{k+1} \text{ or} \\ A_\epsilon &= \left[G M^{-1} G^T + \frac{\epsilon}{4h^2} \right], \text{ solving for } \lambda. \end{aligned} \tag{8}$$

The strict equivalence between the two different forms does not remove the high oscillation though but provided $g(x_0) = Gv_0 = 0$, these are kept small, of order $O(\sqrt{\epsilon}h^2)$ in fact(40). Note however that the first of the two Schur complement in Eqn. (8) is symmetric positive semidefinite in the limit $\epsilon \downarrow 0$ unless G^T has full row rank, in which case there is simply no dynamics. A Cholesky factorization applied to W_ϵ would breakdown at some point. But the second form is symmetric and positive definite. By virtue of backward stability of the Cholesky factorization with a finite but small value of $\epsilon > 0$ (29, 34) the numerical factors respect the sign of the diagonal elements. This is related to the backward stability of the LDLT factorization of H as well. There is no such guarantee in factorization of W_ϵ which means that numerical computations can inject energy into the system. When working with A_ϵ, it is not even possible to inject energy at $\epsilon = 0$ provided G has full row rank which corresponds to strict positive definiteness of $A_{\epsilon=0}$. Backward stability at the numerical factorization level implies similarly that the numerical trajectories computed are those of a slightly different physical system whose trajectories intersects those of the original, undiscretized one(see 31, 38).

Why bother so much about strong penalty forces? One cannot assume apriori that a multibody system is free of stiff forces or fast oscillations unless only constant forces are considered. This is for the well and good reason

that the effective inertia of a multibody system are configuration dependent, according to $(GM^{-1}G^T)^{-1}$. A simple two link mechanism for instance has infinite inertia in the longitudinal direction when fully extended, but zero inertia in the transversal direction. This relates to rank degeneracy of the constraint Jacobian G in that specific configuration, as well as in the vertical position if the links have exactly the same length. A weak spring applying a transversal force, a good idea for a mechanism that could potentially reach the horizontal configuration and would therefore have to be prevented to jam there, would develop very high but short lived oscillations. The conclusion is that a well designed and robust simulation requires at least semi-implicit treatment of forces for realistic mechanism such as the Andrew squeezer(see 33, Sec VII.7). Stiffness is an essential property of mechanisms designed for grasping, and is even used for such mundane items as folding garden chairs. Forces with finite stiffness correspond to $\epsilon > 0$ in the ongoing argument and clearly, unless $\epsilon \to \infty$, at which point one should use W_ϵ in Eqn. (8) instead of A_ϵ, the linearly implicit form in Eqn. (6) is better suited for numerical work.

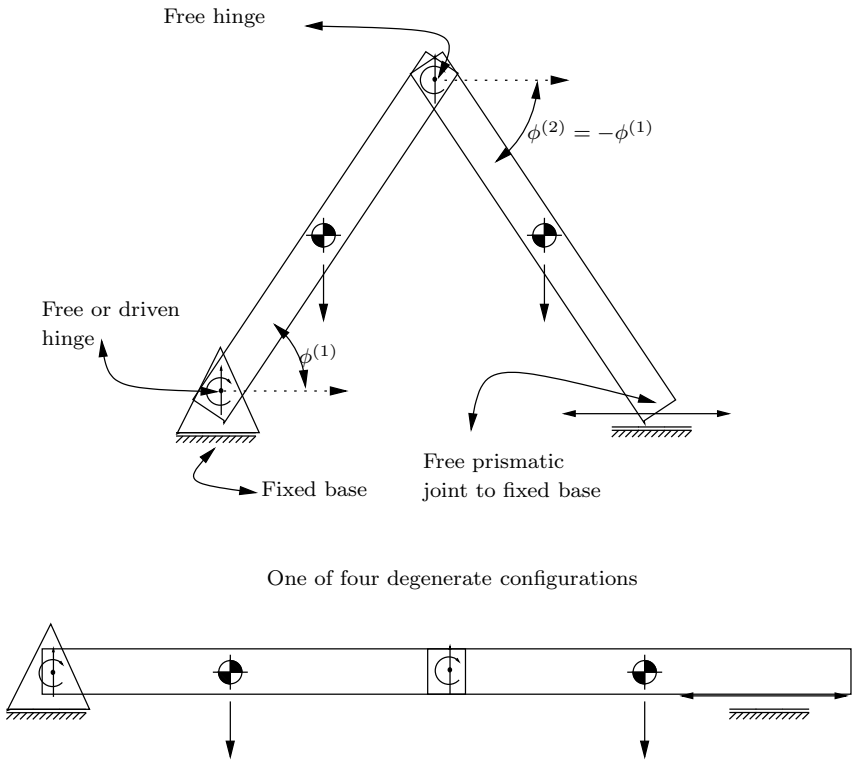

Figure 1: A degenerate situation with a rank deficient Jacobian for the slider crank mechanism.

The stepping scheme of Eqn. (6) is evocative of purely constrained systems. The entire procedure of the limit $\epsilon \downarrow 0$ is in fact related to constraint realization (19). But the equations of motion of multibody systems never include the term $\epsilon > 0$ and the question arises whether one can in fact introduce this regularization in the Lagrangian formulation and the variational principles of mechanics. This is what the ghost particle theory addresses, and this is what we now turn our attention to.

4. Ghost particles

It is customary in physics to use the term "ghost variables" for negative kinetic energy terms. This corresponds to a split in a Lagrangian after suitable change of coordinates, introducing λ for the ghost variables and q for the real ones, to

$$\mathcal{L} = \mathcal{L}_r(q, \dot{q}) - \mathcal{L}_g(\lambda, \dot{\lambda}) + V(q, \lambda), \tag{9}$$

where $V(q,\lambda)$ is a coupling term, and $\mathcal{L}(\lambda,\dot{\lambda})g$ contains non-negative kinetic energy, and potential energy bounded below. For a simple linear system, this splitting corresponds to separating the quadratic forms in the Lagrangian

$$\mathcal{L}(x,\dot{x}) = \frac{1}{2}\dot{x}^T\tilde{M}\dot{x} - \frac{1}{2}x^T\tilde{A}x. \tag{10}$$

assuming that the positive and negative eigenvalues of $\tilde{M}$ and $\tilde{A}$ are matched, transformation to Eqn. (9) can be performed. Ghost variables are truly a plague to be reckoned with though (see 27, Sec. 4), and they introduce real instabilities allowing for trajectories to escape to infinity simultaneously, i.e., $q,\lambda \to \infty$. This does not happen however if the kinetic energy of the ghost variables is strictly zero.

A very relevant question is where do the ghosts come from? Mechanical system subject to constraints $g(q) = 0$ require Lagrange multipliers λ which are the magnitude of forces necessary to keep the variables q on the manifold $g(q) = 0$. Such forces are ideal as they do not produce work and this in turn implies the have the form $f_c = G^T\lambda$. Indeed, $g(q) = 0$ implies $G\dot{q} = 0$ and so the work they perform on the system vanishes, i.e., $f_c^T G\dot{q} = 0$. Geometrically, the constraint forces cancel the components of other forces which are normal to the constraint manifold. But such auxiliary variables can be introduced in the dynamics as demonstrated in Sec. 3.

The physics of potential terms of the form $\frac{1}{2\epsilon}\|g\|^2$ is of importance also, especially in the limit where the amplitude of $\|g\| = O(\sqrt{\epsilon})$. These corresponds to the process of transforming real, physical observations which necessarily contain a variety of time-scales to idealized physical models in which these fast oscillations of small amplitudes are removed. We denote the physical and idealized physical models as $\mathcal{L}_\epsilon$ and $\mathcal{L}_0$ respectively. The idealized model $\mathcal{L}_0$ is usually free of small terms and strong forces are then replaced with constraints which can then be removed entirely by changing coordinate systems.

The idealized Lagrangian may or may not produce physical results however since small terms and fast oscillations can affect the global trajectory in significant ways. In some cases, the idealized model can contain paradoxes and have multiple or no solution in certain cases. The existence and smoothness problem of the Navier-Stokes equations is a case in point demonstrating that one cannot assume that the idealized problem is a sensible representation of a physical problem. The Painlevé paradox (see 21) is an example for which one can show that the equations of motion of the idealized model of a

rigid rod sliding on a plane subject to Coulomb friction fail to exist for some configurations, and allow multiple solutions for others. Especially from the numerical perspective, one should ask whether it is sensible to approximate the idealized model which may have pathological properties, or construct an approximation of a suitably perturbed problem which is more sensible analytically. This boils down to choosing approximations that correspond to physical perturbations or truncation of series expansions.

As analyzed extensively by Bornemann (19), the influence of the fast low amplitude modes on the slow high amplitude ones cannot be neglected in general. A very simple case here is that of contact forces. At the macroscopic level, we can simply express nonpenetration conditions as $g(q) \geq 0$, and the physics encapsulated here is that of repelling forces which become gigantic when $g(q) < 0$. Such forces can only last for short time intervals since a body with finite mass will be quickly accelerated away from the surface. Ultimately, an inconsistent incident velocity $G\dot{q}_- \geq 0$ is changed to a consistent one, $G\dot{q}_+ \geq 0$, and this then preserves the observation that $g(q) \geq 0$, where $g(q)$ is the closest point between the surfaces of solid bodies, for instance. But the previous analysis always assumed the initial conditions $g(q(0)) = 0$ and $G\dot{q}(0) = 0$, but the latter is clearly violated here. A complete analysis of the separation of time and length scales(19) reveals that the influence of some of the fast modes project on the slow mode and one must retain constitutive laws, i.e.,

$$\lim_{\epsilon \downarrow 0} \mathcal{L}_\epsilon \to \mathcal{L}_0 - V_{\text{hom}}, \tag{11}$$

where V_{hom} is the homogenization term and depends directly on the initial values of $G\dot{q}(0)$. This term captures the coupling and residual effect of the fast variables on the slow ones. There are cases where V_{hom} can be computed directly but others where it is not possible. It appears at this time that multibody impact laws fall into the latter category(see 48).

With this in mind, we proceed with the analysis of ghost variables starting with the mathematical equivalence $\equiv$ of two Lagrangians, namely

$$\mathcal{L}(q, \dot{q}) - \frac{1}{2\epsilon}\|g\|^2 \equiv \mathcal{L}(q, \dot{q}) + \frac{\epsilon}{2}\|\lambda\|^2 + \lambda^T g(q). \tag{12}$$

The Lagrangian $\mathcal{L}(q, \dot{q})$ is assumed to contain no high frequency terms, and the equivalence means that the trajectories produced by either formulations are identical at the mathematical level. This can be verified by computing the Euler-Lagrange equations by performing free variations on both q and λ.

This yields

$$q : \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^T}\right) - \frac{\partial \mathcal{L}}{\partial q^T} - G^T\lambda_\epsilon = 0 \tag{13}$$

$$\lambda : \qquad \epsilon\lambda_\epsilon + g(q_\epsilon) = 0, \tag{14}$$

where the subscript ϵ is introduced now to distinguish between trajectories of the regularized system to those of the purely constrained one. As long as ϵ is finite one can eliminate $\lambda_\epsilon = -(1/\epsilon)g(q_\epsilon)$ in the second equation and recover those for the strong potential formulation, which means in fact that the DAEs of motion of mechanical systems are the singular limit of DAEs of index 2, not pure index 3 problems (see 33, for definition of the index of DAEs). The dynamics of λ_ϵ in the limit $\epsilon \downarrow 0$ requires special attention. As it is well-known by now (19, 51), though the trajectories $q_\epsilon, \dot{q}_\epsilon$ and the indicator $g(q_\epsilon)$ converge uniformly in the limit $\epsilon \downarrow 0$, the ghost variables only exhibit weak* convergence, i.e.,

$$\lim_{\epsilon\downarrow 0} \int_0^T \mathrm{d}s\lambda_\epsilon^T\phi \to \int_0^T \mathrm{d}s\lambda^T\phi, \tag{15}$$

for an arbitrary but finite time interval $[0, T]$, and an arbitrary continuous function $\phi : \mathbb{R} \mapsto \mathbb{R}^m$, even though $\lim_{\epsilon\downarrow 0}\lambda_\epsilon$.

The reason is that fast oscillations of finite amplitude persist in λ_ϵ related to the bound $\|g(q_\epsilon)\| = O(\sqrt{\epsilon})$ and the relation $\lambda_\epsilon = -(1/\epsilon)g(q_\epsilon)$. Though the limit of $\lim_{\epsilon\downarrow 0}\lambda_\epsilon$ seems intuitive and is often mentioned in the physics literature (e.g. 42), the resolution of this issue and the realization only weak* convergence is guaranteed is recent (19), as is the connection between weak* convergence and the homogenization potentials arising from inconsistent velocities, $G\dot{q} \neq 0$ which cause impacts in the limit $\epsilon \downarrow 0$.

If there is no impact however, one should observe that D'Alembert's principle applied to $\mathcal{L}(q, \dot{q})$ subject to the constraint $g(q) = 0$ is equivalent the Least action principle applied to the augmented Lagrangian $\mathcal{L}(q, \dot{q}) + \lambda^T g$, performing unrestricted variations on both q, λ. This is in fact an indication of the nature of the ghosts since a stationary point of the Lagrangian is a minimum, but that of the augmented Lagrangian is in fact a saddle point, i.e., a minimum for q but a maximum for λ, a well known fact from primal-dual analysis of constrained optimization.

The ghost reformulation with finite $\epsilon > 0$, which we call "constraint relaxation" is still useful though when considering discretization and the derivation

of semi-implicit integration schemes of the form shown in Eqn. (6). These will be derived by discretizing the ghost variables in time along with the natural q coordinates in Sec. 8.

A natural question arises here namely, the general form of the ghost potential for strong, convex potentials of the form $U_\epsilon = \epsilon^{-1}V(g(q))$. The reason to write $g(q)$ here is that these are the fast, low amplitude variables. The mathematical equivalence

$$\mathcal{L}(q,\dot{q}) - \frac{1}{\epsilon}V(g(q)) \equiv \mathcal{L}(q,\dot{q}) + \epsilon\tilde{V}(\lambda) + \lambda^T g(q). \tag{16}$$

is retained with the definition of $\tilde{V}(\lambda)$ as the Legendre transform

$$\epsilon\tilde{V}(\lambda) = -\min_x \left[\lambda^T g + \epsilon^{-1}V(g)\right], \tag{17}$$

which yields the relation

$$\lambda = -\frac{1}{\epsilon}\frac{\partial V}{\partial g}. \tag{18}$$

By the involutive property of the Legendre transform (see 7, Sec. 14) this implies

$$\epsilon\frac{\partial\tilde{V}}{\partial\lambda} = -g. \tag{19}$$

Clearly, sharp potentials $(1/\epsilon)V$ transform to flat ones $\epsilon\tilde{V}(\lambda)$. But also, for V bounded below, we have $\tilde{V}$ bounded above, another indicator of the ghostly nature of the λ variables. As mentioned in Sec. 3, these high oscillations in λ can cause numerical difficulties and thus some form of damping must be added to filter these away. Since mechanical systems are not ideal we now turn to analytic formulations of nonholonomic constraints and dissipative terms.

This Legendre transformation appeared in the literature already under different guises (25, 50) but without the regularization, thus requiring justification from convex analysis.

5. Non-holonomic constraints and dissipation

The variational formulation of nonholonomic constraints is an old problem in mechanics which requires special attention. Given an indicator $a(q,\dot{q},t)$ which should vanish on the physical trajectory, it is not possible to augment the free Lagrangian with a term of the form $\alpha^T a(q,\dot{q},t)$ as done in the

vakonomic (variational axiomatic) theory (see 8). This would introduce the anomalous term

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\alpha^T\frac{\partial a}{\partial \dot{q}}\right) \tag{20}$$

which is not observed experimentally (see 46). Instead the equations of motion of systems subject to affine nonholonomic constraints of the form $a(q,\dot{q},t) = A(q)\dot{q}$ can be derived via D'Alembert's principle yielding the well-known conclusion that the constraint forces must have the forces of the form $\alpha^T A$ to be ideal, or work less, as in the case of holonomic constraints (42).

If holonomic constraints correspond to strong potentials, what do holonomic constraint correspond to physically? The answer is that a very strong force of dissipation of the form

$$f_\delta = -\frac{1}{\delta}A^T a(q,\dot{q},t) = -\frac{1}{2\delta}\frac{\partial\|a(q,\dot{q},t)\|^2}{\partial\dot{q}} \tag{21}$$

does converge to the correct limit $A(q,\dot{q})\dot{q} = 0$ as $\delta \downarrow 0$ (see 20, 39). The problem now is to formulate this with ghost variables.

Consider now polygenic forces derived from potentials of dissipation(42) $\mathfrak{R}(q,\dot{q},t)$, i.e., scalar functions producing forces $-\partial\mathfrak{R}/\partial\dot{q}$. Barring other forms of external forces and assuming an otherwise conservative system, the energy decreases as

$$\frac{\mathrm{d}E}{\mathrm{d}t} = -\dot{q}^T\frac{\partial\mathfrak{R}}{\partial\dot{q}}. \tag{22}$$

This means that such pseudo-potentials are maximally dissipative.

The limiting behavior of strongly damped systems is better behaved than that of highly oscillatory ones as considered in Sec. 4. To relate this to ghost variables, first consider D'Alembert's principle

$$\delta\int_0^T \mathrm{d}s\mathcal{L}(q,\dot{q}) + \int_0^T \mathrm{d}s\delta q\cdot f = 0. \tag{23}$$

If we now introduce $\mathfrak{R} = 1/(2\delta)\|a(q,\dot{q},t)\|^2$ and it's Legendre transform with respect to the indicator a introducing a ghost velocity $\dot{\alpha}$ we have

$$\tilde{\mathfrak{R}}(\dot{\alpha}) = -\min_a\left\{\dot{\alpha}^T a(q,\dot{q},t) + \frac{1}{2\delta}\|a(q,\dot{q},t)\|^2\right\} = -\frac{\delta}{2}\|\dot{\alpha}\|^2. \tag{24}$$

The negative sign confirms that we are dealing with ghost variables. Performing variations on both natural coordinates q and the ghosts α, D'Alembert's principle in Eqn. (23) yields the equations of motion

$$q : \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^T}\right) - \frac{\partial \mathcal{L}}{\partial q^T} - \frac{\partial a(q,\dot{q},t)}{\partial \dot{q}^T}\dot{\alpha} = 0 \tag{25}$$

$$\alpha : \qquad \delta\dot{\alpha} + a(q,\dot{q},t) = 0. \tag{26}$$

Setting $\delta = 0$ recovers the standard equations of motion of nonholonomic systems. By the involutive property of Legendre transform, as observed in Sec. 4, the equations of motion Eqn. (26) including the ghost variables α are mathematically equivalent to the those of the natural variables q subjected to the strong dissipation.

Using ghost velocities $\dot{a}$ here provides for a systematic derivation of non-ideal forces which can include mixing strong potentials with strong dissipation, which does coincide with the physics of fast oscillations since these dissipate energy. Consider the ghost λ with potential $(\epsilon/2)\|\lambda\|^2$ and coupling $\lambda^T g(q)$. Since $\lambda = -(1/\epsilon)g$ then $\dot{\lambda} = -(1/\epsilon)G\dot{q}$ and so a dissipation of the form $(1/(\tau\epsilon))\|\dot{g}\|^2$ Legendre transforms to the ghost dissipator

$$-\frac{\tau\epsilon}{2}\dot{\lambda} - \tau\dot{\lambda}G\dot{q}, \tag{27}$$

and this then yields the equations of motion

$$q : \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^T}\right) - \frac{\partial \mathcal{L}}{\partial q^T} - G^T\lambda - \tau G^T\dot{\lambda} = 0 \tag{28}$$

$$\lambda : \qquad \epsilon(\lambda + \tau\dot{\lambda}) + g(q) + \tau G\dot{q} = 0, \tag{29}$$

showing the consistency of the construction. This will be used below for constraint stabilization at the numerical level in Sec. 910.

The strength of the ghost variables reformulation is that one can construct forcing terms and dissipators directly in the ghost space, without having to perform the Legendre transforms explicitly. This allows then the introduction of constitutive laws and multi-domain coupling in a systematic way, a topic beyond the scope of the present article. The current state of analytic system dynamics (see 43) does not include such a variational formulation. We only consider Coulomb friction in Sec. 12 and leave the multi-domain problems for future work.

6. Inequalities and impact laws

Inequality constraints and impact laws are clear examples of separation of time scales. Surface contact physics involves extremely short time and lent scales and at the macroscopic level, all that really matters is that solids do not interpenetrate, and that normal forces between them are almost exclusively repelling. Exception must be made of Van der Walls forces for instance which are attractive over a very short range, but this only proves the point that perfect geometric constraints do not exist in nature. This provides a guide for choosing acceptable discretizations.

There are difficulties involved in formulating D'Alembert's principle in the presence of inequalities since one must define the variations δq in a consistent way. One could assume first that an impact location is known a priory to occur at t_0 and construct variations $\delta q = \epsilon\eta$ such that $\eta(t_0) = 0$ to respect an inequality condition of the form $c(q) \geq 0$ which vanishes on impact, $c(q(t_0)) = 0$. The alternative is to perform variations consistent with the constraints but without assumption on the time of impact (see 45). From the latter reference, the two scenarios correspond to the strong and weak variational principles, respectively. The weak form fits naturally with constraint regularization since in that case, there is no impact as long as the regularization parameters do not vanish.

Strict inequalities can be regularized with smoothed C-functions (see 26) such as the Fischer-Burmeister (FB) analytic representation of the absolute value operator, namely

$$\min(x) = \lim_{\tau\downarrow 0} \phi^{(\tau)}_{FB}(x) = \lim_{\tau\downarrow 0} \frac{1}{2}\left\{x - \sqrt{|x|^2 + \tau}\right\}. \tag{30}$$

A contact constraint $c(q) \geq 0$ can then be expressed as the strong force

$$\frac{1}{2\epsilon}\|\phi^{(\tau)}_{FB}(c(q))\|^2. \tag{31}$$

This has a residual small attractive force for $c(q) > 0$ with intensity

$$\|f\| = O\left(\frac{\tau^2}{x^3}\right), \text{ when } x > \tau. \tag{32}$$

Using a ghost variable ν, we now have the equation of motion for the ghost

$$\epsilon\nu + \phi^{(\tau)}_{FB}(c(q)) = 0. \tag{33}$$

Given the property of the FB function, this means then that

$$\nu = -\frac{1}{\epsilon}\phi_{FB}^{(\tau)}(c(q)) \geq 0. \tag{34}$$

When $c(q) > 0$ we have $-\sqrt{\tau}/2 \leq \phi_{FB}^{(\tau)}(c(q)) < 0$ which implies that $0 \leq \nu \leq (\sqrt{\tau}/\epsilon)/2$. Otherwise, when $c(q) < 0$, then ν grows without restrictions. In the limit $\tau \downarrow 0$, we recover the complementarity conditions

$$0 \leq \nu \perp \epsilon\nu + c(q) \geq 0, \tag{35}$$

which yield the standard nonsmooth formulation as $\epsilon \downarrow 0$. The complementarity conditions can now produce impacts since there is no restitution force for $c(q) > 0$ so nothing prevents from reaching an impacting configuration $c(q) = 0, N(q)\dot{q} < 0$, where $c = \partial c/\partial q$, the normal Jacobian.

To account for this, we need a model for the momentum change over the time interval it takes to restore a consistent condition with $c(q) \geq 0, N(q)\dot{q} \geq 0$. This will be addressed for the discrete-time case in Sec. 11.

7. Discrete-time mechanics and variational integrators

The theory of variational time integrators (see 49) provides solid grounds for constructing good time-stepping schemes for mechanical systems and this is why ghost variables were introduced. They provide an analytic formulation of all the dynamics involved in multibody systems, including friction as discussed below in Sec. 12.

Consider the action integral segmented over n intervals $I_k = [t_k, t_{k+1}]$, $t_{k+1} > t_k, k = 1, 2, \ldots, n$, so

$$\mathcal{S}[\dot{\gamma}] = \int_0^T \mathrm{d}s\mathcal{L}(q, \dot{q}) = \sum_k \int_{I_k} \mathrm{d}s\mathcal{L}(q, \dot{q}), \tag{36}$$

for a path $\dot{\gamma} : \mathbb{R} \mapsto (q, \dot{q})$. Introduce the discrete Lagrangian as the approximation of the action over the intervals I_k

$$\int_{I_k} \mathrm{d}s\mathcal{L} = \mathbb{L}_d(q_k, q_{k+1}). \tag{37}$$

This quadrature can be approximated with a set of m control points evaluated at times $q(t_j)$ $t_j \in I_k, j = 1, 2, \ldots, m$ (see 6, Appendix 2). The points $t_j \in I_k$

themselves would be provided by any quadrature rule, but that still leaves the issue of computing the states $q(t_j)$. However, these can be computed by any numerical integration method (see 32). In particular, one can choose a symplectic Runge-Kutta method. When this is done, the discrete Lagrangian $\mathbb{L}_d(q_k, q_{k+1}, h)$ still depends only on the endpoints. We now have a discrete-time action

$$\mathbb{S} = \sum_{1}^{n} \mathbb{L}_d(q_k, q_{k+1}, h) \tag{38}$$

which is an approximation of arbitrarily high order but which depends only on the n sample points. This is now simply a multivariate function of the $q_k, k = 1, 2, \ldots, n$.

Any quadrature formula chosen to compute the discrete Lagrangian Eqn. (37) defines an interpolation $q(t) = \eta(q_k, q_{k+1}, t), t \in I_k$, and therefore, we have

$$\delta q(t) = D_1\eta(q_k, q_{k+1}, t)\delta q_k + D_2\eta(q_k, q_{k+1}, t)\delta q_{k+1}. \tag{39}$$

We can then define the discrete forces as

$$\begin{aligned} \int_0^h \mathrm{d}s f \cdot \delta q(s) &= f_d^{(+)}(q_0, q_1)\delta q_0 + f_d^{(-)}(q_0, q_1)\delta q_1 \\ f_d^{(+)}(q_k, q_{k+1}) &= \int_{I_k} \mathrm{d}s f(q, \dot{q}, s) \cdot D_1\eta(q_0, q_1) \\ f_d^{(-)}(q_k, q_{k+1}) &= \int_{I_k} \mathrm{d}s f(q, \dot{q}, s) \cdot D_2\eta(q_0, q_1) \end{aligned} \tag{40}$$

The discrete version of D'Alembert's Fourier principle is the multivariate extremization

$$\left(\frac{\partial \mathbb{S}_d}{\partial q_k} + f_d^{(+)}(q_k, q_{k+1}) + f_d^{(-)}(q_k, q_{k+1})\right)\delta q_k \leq 0, \tag{41}$$

and the inequality holds for cases where the configuration space Q has a closed boundary (see 42). When the boundary is reached, the inequality indicates that the forces must point away from same. Introducing rheonomic constraints $g(q) \geq 0$ with Jacobian $G = \partial g/\partial q$, the conditions produce the discrete time Euler-Lagrange equations of motion

$$\begin{aligned} D_1\mathbb{L}_d(q_k, q_{k+1}) + D_2\mathbb{L}_d(q_{k-1}, q_k) + f_d^{(+)}(q_k, q_{k+1}) + f_d^{(-)}(q_{k-1}, q_k) + G_k^T\lambda = 0 \\ g_{k+1} = 0, \end{aligned} \tag{42}$$

where D_i is the partial derivative with respect to the ith argument. With given initial conditions q_0, q_1, the discrete Euler-Lagrange equations of motion is a nonlinear map $\Phi : (q_k, q_{k-1}) \mapsto (q_k, q_{k+1})$ which can be solved for q_{k+1} given q_{k-1} and q_k for an initial value problem.

Because nonholonomic constraints are treated as strong dissipation pseudo-potentials, there is no need to use the more rigorous theory of variational integration for these (24). Instead, we only use the simpler theory of discrete mechanics related to forcing terms (see 49) and the ghost formulation below in Sec. 8.

Restricting the Lagrangian to a finite dimensional mechanical system with constant mass matrix M and subject only to non-stiff potential $V(q)$

$$\mathcal{L}(q, \dot{q}) = \frac{1}{2}\dot{q}^T M q - V(q). \tag{43}$$

Under the simplest discretization

$$\begin{aligned} \dot{q}(t_{k+1}) &\approx \frac{(q_{k+1} - q_k)}{h} = v_{k+1}, \\ \mathbb{L}_d(q_{k+1}, q_q, h) &= \frac{1}{2h}(q_{k+1} - q_k)^T M (q_{k+1} - q_k) + hV(q_k), \end{aligned} \tag{44}$$

we have

$$D_1\mathbb{L}_d(q_k, q_{k+1}, h) + D_2\mathbb{L}_d(q_{k-1}, q_k, h) = Mv_{k+1} - Mv_k + hf_k = 0, \tag{45}$$

were $f_k = (\partial V/\partial q)_k$ is the force at discrete time k. Adding the constraints to this yields the SHAKE stepping scheme

$$\begin{aligned} Mv_{k+1} - hG_k\lambda &= Mv_k - hf_k \\ g(q_{k+1}) &= 0 \\ q_{k+1} &= q_k + hv_{k+1}. \end{aligned} \tag{46}$$

This is equivalent to solving the following nonlinear equation for λ

$$g(q_k + hv_k + hM^{-1}f_k + hM^{-1}G_k^T\lambda) = 0. \tag{47}$$

Note that the search direction G_k^T is fixed though so there may not be a root λ to that equation (17). This can happen if the velocities are too high or if the solutions of the nonlinear equations are too inaccurate. In both cases, the starting point for the iteration is too far from the constraint manifold

to guarantee a solution. This geometry of the line search for SHAKE and projection methods are presented in Figs. 23, respectively.

The direction of the line search is also independent of the quadrature scheme which means that this is a fundamental problem with these integrators and that fixed-step integration cannot generally be relied upon. That is one of the reason for developing SPOOK presented in Section 9.

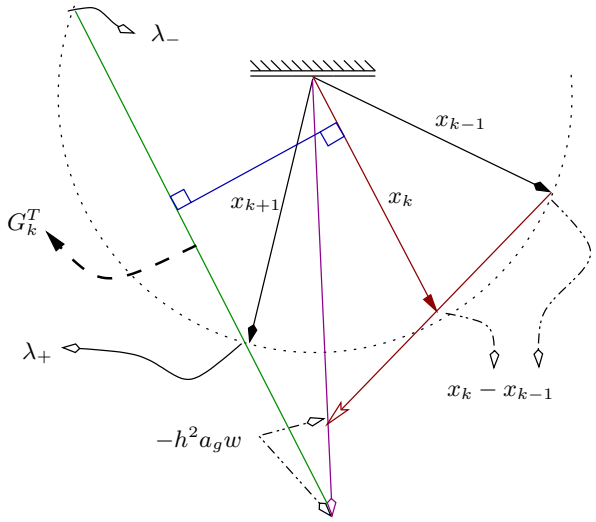

Figure 2: The SHAKE stepping scheme searches for the constraint manifold along a line parallel to the constraint Jacobian computed at step k. There are two possible solutions here labeled $\lambda_\pm$. If the speed was even higher, the line would fail to intersect the circle.

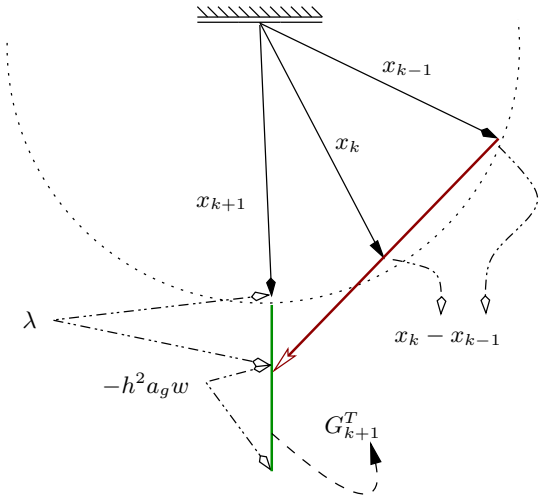

Figure 3: By contrast, projection methods update the search direction G_{k+1}^T and can locate the closest point at the cost of loosing kinetic energy. This can be fixed by using a symmetrized method, however. Projection methods work well provided one updates the Jacobian $G(q)$ as the iterations proceed, which is computationally expensive.

Despite this problem, SHAKE is still a wonder since methods designed for integrating general index 3 DAEs require higher order and, being implicit, require the solutions of much larger systems of nonlinear equations(see 33, Sec. VII).

8. Discretizing the ghosts

As mentioned earlier in Sec. 4 the ghost variables converge weakly* to the Lagrange multipliers which means they must be discretized with care. We use the midpoint rule here so that

$$\mathbb{L}_d(\lambda_0, \lambda_1, h) = \int_0^h \mathrm{d}s \|\lambda\| \approx h \| \frac{\lambda_1 + \lambda_0}{2} \|, \tag{48}$$

and similarly

$$\mathbb{L}_d(q_1, q_0, \lambda_1, \lambda_0) = \int_0^h \mathrm{d}s \mathcal{L} \approx \frac{h}{4} (\lambda_1 + \lambda_0)^T (g_1 + g_0). \tag{49}$$

Dissipative forcing terms on the physical variables q which depend on the ghost λ are then discretized according to

$$\begin{aligned} f_q^{(+)} &= hA_{k+1}^T \frac{(\alpha_{k+1} - \alpha_k)}{h} + h\tau G_{k+1}^T \frac{(\lambda_{k+1} - \lambda_k)}{h}, \\ f_q^{(-)} &= 0, \end{aligned} \tag{50}$$

and the forces acting on the ghosts are then

$$\begin{aligned} f_\lambda^{(+)} &= h\tau\epsilon \frac{\lambda_{k+1} - \lambda_k}{h}, \qquad f_\lambda^{(-)} = 0 \\ f_\alpha^{(+)} &= h\delta \frac{\alpha_{k+1} - \alpha_k}{h}, \qquad f_\lambda^{(-)} = 0. \end{aligned} \tag{51}$$

These correspond to an implicit integration. The complete discretization scheme including constraint stabilization then reads

$$\begin{aligned} D_1\mathbb{L}_d(q_k, q_{k+1}, h) + D_2\mathbb{L}_d(q_{k-1}, q_k, h) + G_k^T\lambda + A_k^T\alpha &= 0 \\ \frac{\epsilon}{h}\lambda + \frac{1}{4}(g_{k+1} + 2g_k + g_{k-1}) + \tau G_{k+1}v_{k+1} &= 0 \\ \frac{\delta}{h}\alpha + A_{k+1}v_{k+1} &= 0, \end{aligned} \tag{52}$$

where $\alpha = (\alpha_{k+1} - \alpha_k)$ and $\lambda = h\lambda_{k+1} + \tau(\lambda_{k+1} - \lambda_k)$. There is no need to integrate the ghosts directly, however. The exact same technique is applied to inequalities since they only differ from equality constraints at impacts as discussed in Sec. 6.

In continuous time, the result of the damping term $\mathfrak{R} = (\tau/2\epsilon)\|G\dot{q}\|^2$ is akin to dissipation terms used in the sequential regularization method of Ascher (9, 13), since it dissipates monotonously toward $g(q) = 0$. For the case where $\epsilon = \delta = 0$, the continuous formulation is precisely the same as for constrained mechanics, and includes nonholonomic constraints systematically.

9. Spook: a semi-implicit stepping scheme for multibodies

The time stepping scheme in Eqn. (52) is nonlinear and suffers from the same problems as SHAKE discussed in Sec. 7. We linearize this here and demonstrate the stability of the resulting scheme in Sec. 10. Of course the linear system will have a solution but of course, this will not yield exact

constraint satisfaction, but constraint stabilization dissipates energy when the velocities are so high that it is difficult to locate the constraints.

Linearizing the second line in Eqn. (52) yields

$$\begin{aligned} G_k v_{k+1} + \frac{\phi 4\epsilon}{h^2}\lambda &= -\frac{4\phi}{h} g_k + \phi G_k v_k - \frac{\phi}{2} v_k^T \frac{\partial^2 g_k}{\partial q^T \partial q} v_k, \text{ where} \\ \phi &= \frac{1}{1 + 4\tau/h}. \end{aligned} \tag{53}$$

Likewise, nonholonomic constraints are approximated by using $A_k v_{k+1}$ instead of $A_{k+1} v_k$ so

$$A_k v_{k+1} + \frac{\delta}{h}\alpha = 0. \tag{54}$$

We also usually neglect the Hessian term, namely, the last term on the first line in Eqn. (53) The linear system of equations to solve then reads

$$\begin{bmatrix} M & -G^T & -A^T \\ G & T & 0 \\ A & 0 & \Delta \end{bmatrix} \begin{bmatrix} v_{k+1} \\ \lambda \\ \alpha \end{bmatrix} = \begin{bmatrix} M v_k + h f_k \\ -\frac{4}{h}\Phi g_k + \Phi G v_k \end{bmatrix} \tag{55}$$

where

$$\begin{aligned} \Phi &= \text{diag}(\phi_1, \phi_2, \ldots, \phi_{m_h}) \\ T &= \text{diag}(\frac{4\epsilon_1\phi_1}{h^2}, \frac{4\epsilon_2\phi_2}{h^2}, \ldots \frac{4\epsilon_{m_h}\phi_{m_h}}{h^2},) \\ \Delta &= \text{diag}(\frac{\delta_1}{h}, \frac{\delta_2}{h}, \ldots, \frac{\delta_{m_{nh}}}{h}), \end{aligned} \tag{56}$$

where m_h is the number of holonomic constraints and m_{nh} is the number of nonholonomic ones.

The main feature in the SPOOK stepping scheme in Eqn. (55) lies in the diagonal perturbations which appear directly from the theory, and the factors multiplying the constraints and constraint velocities on the right hand side of the equations.

The diagonal perturbation is a standard technique in numerical linear algebra (see 34) as it improves conditioning. Such perturbations are usually ad hoc and sometimes performed in software packages using heuristics (see 2, 52). The diagonal perturbation also guarantees backward stability. Indeed, if this is small enough, backward stability of LDLT factorization (22)

guarantees that the numerical factors respect the sign of the perturbation, i.e., if $\tilde{L}$ and $\tilde{D}$ are the numerical factors of a matrix H, then

$$\tilde{L}\tilde{D}\tilde{L}^T = H + E, \tag{57}$$

where E is small and symmetric. The numerical factors then respect the positive definiteness of the perturbation is large enough. This is not necessarily the case for zero perturbation as there is no guarantee on the signs on the diagonal of the error matrix E, only a guarantee that there is a matrix $E = E^T$ which is close to the product of machine precision and condition number.

10. Stability of the Spook stepping scheme

We now investigate the stability of the Spook stepping scheme with respect to constraint satisfaction. It is sufficient for this analysis to use a unit mass matrix $M = I$ and linear, homogeneous constraints of the form $g(q) = Gx = 0$, where $G \in \mathbb{R}^{m\times n}$, and to set external forces to zero here. No assumption is made here on whether matrix G has full row rank or not, and whether $m \leq n$ or otherwise. The discrete dynamics of this simplified system is then given by the following linear recurrence

$$\begin{aligned} q_{k+1} &= q_k + hv_{k+1} \\ v_{k+1} &= v_k + G^T\lambda \\ Gv_{k+1} + T\lambda &= -\frac{4}{h}\Phi G q_k + \Phi G v_k. \end{aligned} \tag{58}$$

After eliminating λ from (58), the resulting stepping formula becomes

$$\begin{aligned} q_{k+1} - hv_{k+1} &= q_k \\ v_{k+1} &= v_k - G^T A_\epsilon^{-1}\left(\frac{4}{h}\Phi G q_k + (I-\Phi)Gv_k\right), \end{aligned} \tag{59}$$

where

$$A_0 = GG^T, \text{ and } A_\epsilon = A_0 + T \tag{60}$$

The dynamics of constraint violation is then given by the new variables $x_k = Gq_k$, and $y_k = hGv_k$. The h factor in the definition of y_k is there

for convenience. After simple rearrangement, the system (59) implies the following dynamics for the constraint violation

$$\begin{aligned} x_{k+1} - y_{k+1} &= x_k \\ y_{k+1} &= y_k - A_0 A_\epsilon^{-1} \left(4\Phi x_k + (I-\Phi) y_k\right) \end{aligned} \tag{61}$$

In block matrix form, the stepping becomes

$$\begin{bmatrix} I & -I \\ 0 & I \end{bmatrix} \begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ -4A_0 A_\epsilon^{-1}\Phi & I - A_0 A_\epsilon^{-1}(I-\Phi) \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \tag{62}$$

The inverse of the matrix on the left hand side is easily computed to yield

$$B^{-1} = \begin{bmatrix} I & I \\ 0 & I \end{bmatrix}, \text{ for } B = \begin{bmatrix} I & -I \\ 0 & I \end{bmatrix}, \tag{63}$$

and therefore, system (62) can be rewritten as the stationary iterative process

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} I & I \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -4A_0 A^{-1}\Phi & I - A_0 A^{-1}(I-\Phi) \end{bmatrix} \begin{bmatrix} x_k \\ y_k \end{bmatrix} \tag{64}$$

which can be written succinctly as

$$z_{k+1} = K z_k + w_k. \tag{65}$$

The system is stable if the spectral radius of matrix K is strictly less than unity. We now proceed to show that this is the case whenever both matrices T and Φ are symmetric and positive definite, and whenever the spectral norm of matrix Φ satisfies $\rho(\Phi) < 1$.

First, write matrix K in factored form

$$\begin{aligned} K &= B^{-1}N, \text{ with} \\ B^{-1} &= \begin{bmatrix} I & I \\ 0 & I \end{bmatrix}, \text{ and} \\ C &= \begin{bmatrix} I & 0 \\ -4A_0 A_\epsilon^{-1}\Phi & I - A_0 A^{-1}(I-\Phi) \end{bmatrix}. \end{aligned} \tag{66}$$

Matrix B^{-1} has only the unit eigenvalue so that $\rho(B^{-1}) = 1$, and so $\rho(K) \leq \rho(B^{-1})\rho(C) = \rho(C)$. Since matrix C is block lower triangular, its spectrum

is contained in the union of the spectra of the diagonal blocks and thus, the spectral radius of C satisfies

$$\rho(C) \leq \max(1, |\lambda_i|), \tag{67}$$

where λ_i are the eigenvalues of the matrix $N = I - S$ where

$$\begin{aligned} S &= A_0 A_\epsilon^{-1} \Theta \\ \Theta &= I - \Phi. \end{aligned} \tag{68}$$

It suffices to show that the spectrum $\sigma(S)$ is positive and $\rho(S) \leq 2$. Consider an eigenvalue $\lambda \in \sigma(S) = \sigma(S^T)$ then

$$\begin{aligned} \Theta A_\epsilon^{-1} A_0 x &= \lambda x \\ \frac{x^\dagger A_0 x}{x^\dagger A_\epsilon \Theta^{-1} x} &= \lambda, \end{aligned} \tag{69}$$

where $x^\dagger$ is the Hermitian conjugate. Clearly, if $\Theta = \theta I$ is a multiple of the identity and $\theta < 1$, the conditions hold. Stability can breakdown however if the variations in θ_i are too large as the real part of $\lambda x^\dagger A_\epsilon \Theta^{-1} x$ can go negative, allowing for $\rho(C) = \rho(N) > 1$. For $\theta = 0$, $\lambda = 0$ and $\rho(N) = 1$.

For the case where all perturbations and damping rates are identical, the southeast corner of the iteration matrix C in Eqn. (66) is diagonalizable. If $\lambda_i^{(a)}$ are the eigenvalues of matrix A, a short computation yields the non-unit eigenvalues of matrix C

$$\begin{aligned} \lambda_i &= \frac{\phi \lambda_i^{(a)} + 4\epsilon\phi/h^2}{\lambda_i^{(a)} + 4\epsilon\phi/h^2} = \phi \frac{\lambda_i^{(a)} + 4\epsilon/h^2}{\lambda_i^{(a)} + 4\epsilon\phi/h^2} \\ &\approx \phi = \frac{1}{1 + 4\tau/h} \approx \frac{h}{4\tau}, \end{aligned} \tag{70}$$

where the last approximation holds provided τ/h is sufficiently large, i.e., when the time step is sufficiently small. Asymptotically, as $h \downarrow 0$, this corresponds to $\|g\| \to \exp(-4t/\tau)$, giving a good physical explanation of the decay rate, and showing also that this is in fact nearly independent of the masses in the limit $\epsilon \downarrow 0$.

This analysis provides a clear cut choice for the parameter τ. If we use a small enough time step h, the physical value of τ can be used. Otherwise, it is safer to choose $\tau \geq 2h$ which makes the error decay as $O(10^{-n})$, i.e.,

one roughly one decade per time step. This is the threshold we use in our simulations.

As far as the perturbations go, they need to be chosen so that the LDLT factorization of matrix in Eqn. (55) is safe.

The effect of perturbations increases both the max and min eigenvalues by amounts proportional to $\tilde{\epsilon} = 4\epsilon\phi/h^2$ which means that is we have a reasonable upper bound on the maximum eigenvalue of a matrix, the worse of which is the trace, and this is dominated by the masses as seen from Eqn. (55). Since the shift on the lowest eigenvalue introduced by the perturbation is $\lambda_{min} \geq \tilde{\epsilon}_{min}$, then the new condition number is of the same order of magnitude as $\rho(A_\epsilon) \leq \lambda_{max}/(\lambda_{min} + \tilde{\epsilon}) \leq \lambda_{max}/\tilde{\epsilon}$, and we can make this moderate by choosing $\tilde{\epsilon} = O(\bar{M}^{-1})$ where $\bar{M}$ is the average inertia, or some measure of that sort. In our simulations, we make sure that $\tilde{\epsilon} \geq 10^{-6}\bar{M}$ and use the biggest mass for $\bar{M}$ giving an estimate of $\rho(A_\epsilon) = O(10^6)$ which is very safe for direct factorization.

11. Discrete time impacts

The case of impacts requires special care and can be treated in two different ways (45) within the variational framework, both for the continuous and discrete-time cases. The first (28) requires the exact location of the impact at some time t_0, and the second (45) requires that the post-impact velocity points away from the constraint surface and does not involve exact even location. This sacrifices exact energy conservation though that is the norm for variational integrators with fixed step. We choose the second of these strategies as it fits with the idea of relaxed constraints as discussed in Sec. 6.

Impacts are detected de post facto and they require a two stage procedure. We diverge here from a formulation based on the positions only (45) and use a velocity construction in terms of the approximation $v_k = (q_k - q_{k-1})/h$ as described above. The first stage imposes the impact law and restores consistency with the constraint at the velocity level, namely $Nv_o \geq -\psi N v_k \geq 0$ where $\psi \in [0, 1]$ is a restitution parameter, and the second stage is a restart which computes v_{k+1} from v_o. This does not guarantee nonpenetration $c(q_{k+1}) \geq 0$. However, due to the strictly dissipative constraint stabilization technique presented above, the trajectory will eventually reach $c(q_{k+m}) \geq -\kappa$ where $\kappa \geq 0$ is small and depends on the relaxation parameter and damping coefficients. This guarantees energy dissipation during impacts. This leaves open the problem of frictional impacts for which the restitution coefficient

can be formulated in a number of different and non-equivalent ways (21), and that of multiple impacts for which there is simply no constitutive law. Presumably, the latter could be constructed from a deeper analysis the action of different contact laws. The case of the infamous Newton's cradle for which different outcomes are observed depending on the exponent of the elastic contact forces between two solids, as derived from Hertz's theory, is enough caution (see 53).

The frictionless non-ideal impact law translates to solving the following LCP

$$\begin{aligned} Mv_+ - G^T{}_k\lambda - N^T{}_k\nu &= Mv_- \\ G_k v_+ + T\lambda &= G_k v_- \\ N_k v_+ + \Psi N v_- + \Phi\nu &= w \\ 0 \leq \nu \quad \perp \quad w &\geq 0, \end{aligned} \tag{71}$$

where Φ is defined as in Eqn. (56), and where the potential forces $-h\nabla V$ were neglected. Indeed, impulsive forces sufficient to revert the incident velocity must be very large in comparison to other forces in the system. Note that this update only affects the velocity variables so the change from this update is restricted to kinetic energy change. This is now evaluated:

$$\begin{aligned} T_+ = v_+^T M v_+ &= v_+^T M v_-^T + v_+^T G^T\lambda + v_+^T N^T\nu \\ &= v_-^T M v_- + \lambda^T G v_- + \nu^T N v_- + \nu^T N v_+ + v_+^T G^T\lambda \\ &= T_- - \lambda^T\lambda + \nu^T(N v_+ + \Psi N v_-) + \nu^T(I-\Psi)N v_- \\ &= T_- - \lambda^T\lambda + \nu^T w - \nu^T\Phi\nu + \nu^T(I-\Psi)N v_- \\ &\leq T_-. \end{aligned} \tag{72}$$

The last inequality is derived from the following facts

$$\begin{aligned} -\lambda^T\lambda \leq 0 &\quad \text{since } T \text{ is symmetric and positive definite,} \\ \nu^T w = 0 &\quad \text{from the complementarity condition in (71),} \\ -\nu^T\Phi\nu \leq 0 &\quad \text{since } \Phi \text{ is symmetric and positive definite,} \\ N v_- \leq 0 &\quad \text{by assumption on the contact conditions,} \\ (I-\Psi)N v_- \leq 0 &\quad \text{from the definition of } \Psi \text{ since } 0 \leq \psi_j \leq 1, j = 1, 2, \ldots, n_c, \\ \nu \geq 0 &\quad \text{in the solution of LCP (71).} \end{aligned} \tag{73}$$

Therefore, this impulsive stage can only decrease the kinetic energy. Since the positions are not changed in this stage, the total energy can only decrease. Once the impulsive stage is computed and velocities updated, the integration proceeds using the computed velocities, v_+, and the previous positions.

This model still lacks friction forces in the direction tangential to the contact plane. The friction model derived in Section 12 can be added to the present formulation without changing the dissipative properties.

12. Solvable nonlinear complementarity model of Coulomb Friction

Consider a contact constraint defined with $c(q) = 0$ and normal vector $n \in \mathbb{R}^3$. Define also tangent vectors $d^{(1)}, d^{(2)}$ so that $d^{(1)}, d^{(2)}, n$ is an orthonormal basis. Define also the Jacobians matrices N and D of dimension $1 \times n$ and $2 \times n$, respectively, where n is the number of degrees of freedom in the system. These matrices define the normal and tangent Jacobians, respectively, so that if $\dot{q}$ is the generalized velocity vector the multibody system, $N\dot{q}$ is the generalized vector whose components are the normal speed at each contact, and the coordinates of the generalized velocities $D\dot{q}$ contain the tangent velocity vectors at each contact.

A contact constraint of the form $c(q) \geq 0$ introduces a ghost $\nu \geq 0$. Consider now the nonholonomic constraint $D(q)\dot{q} = 0$ which enforces stick friction at the contact location introducing the ghost $\dot{\beta}$. Since a ghost is a particle like any other, we now impose the nonholonomic constraint

$$\mu\nu - \|\dot{\beta}\| \geq 0, \tag{74}$$

which leads to the ghost $\dot{\sigma} \geq 0$ and the pseudo-potential

$$\mathfrak{R} = \frac{\delta}{2}\|\dot{\sigma}\|^2 + \dot{\sigma}^T\mu\nu - \|\dot{\beta}\|, \tag{75}$$

where $\delta > 0$ is the regularization. Note that this dissipation potential is again positive as in the case of the physical variables, and this follows directly from the Legendre transformation of the Coulomb condition Eqn. (74) as per the rule defined in Eqn. (17). In fact, $\dot{\sigma}$ corresponds to the sliding speed at the contact. These three ghosts then produce forces according to D'Alembert's

principle and we get the continuous time DAEs

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}^T}\right) - \frac{\partial \mathcal{L}}{\partial q^T} - N^T\nu - D^T\dot{\beta} &= 0 \\ 0 \le N\dot{q} + \epsilon\nu \perp \nu \ge 0 \\ D\dot{q} + \gamma\dot{\beta} + \frac{1}{\|\dot{\beta}\|}\dot{\beta}\dot{\sigma} &= 0 \\ 0 \le \mu\nu - \|\dot{\beta}\| + \delta\dot{\sigma} \perp \dot{\sigma} \ge 0, \end{aligned} \tag{76}$$

where $\epsilon, \gamma \ge 0$ are the regularization. This formulation gives a clear and physical interpretation of the ghost $\dot{\sigma}$ which is the sliding speed according to the third line in Eqn. (76), which also imposes the sliding direction to be parallel to the tangential contact force, producing maximal dissipation. This is to be expected since potentials of dissipations produce forces that maximize energy dissipation according to Eqn. (22).

The regularization parameter $\gamma > 0$ corresponds to viscous damping with magnitude γ^{-1} which becomes infinite as $\gamma \downarrow 0$. Numerically, the perturbation is δ/h as for all nonholonomic constraints (see Eqn. (56)) and this must be kept moderate for numerical stability. Using $\gamma = O(10^{-8})$ is usually good enough for numerical stability and this amount of viscous sliding is entirely negligible in comparison to other discretization and numerical errors coming from the solver.

If we also include dissipation on the ghosts ν with rate τ, the total force along the normal becomes $N^T(\nu + \tau\dot{\nu})$ and therefore, unless we want to integrate ghost velocity directly, we need to replace the friction law with

$$\mathfrak{R} = \frac{\delta}{2}\|\dot{\sigma}\|^2 + \dot{\sigma}(\mu(\nu + \tau\dot{\nu}) - \|\dot{\beta}\|), \tag{77}$$

which produces an additional forcing term on the equation for the ghost ν so that

$$0 \le N\dot{q} + \epsilon\bar{\nu} - \mu\tau\dot{\sigma} \perp \bar{\nu}\dot{\sigma} \ge 0. \tag{78}$$

For the case where $\tau = 1$, we recover the symmetrized model of Anitescu (see 3). Symmetrization is desirable since the complementarity problem to solve then has the P property and linearizations lead to a quadratic program. However, as observed by Anitescu, this introduces an anomaly which prevents steady sliding, the reason being that since $-\mu\dot{\sigma} < 0$, the initiation of sliding at $\nu > 0$ immediately produces $N\dot{q} = \mu\dot{\sigma} > 0$, i.e., contact release. This

said, the reason we introduced damping $\tau > 0$ is to stabilize constraints, meaning that this is useful when we have penetration $c(q) < 0$. When all regularization terms are introduced, the additional forcing term $\tau\mu\dot{\sigma}$ is not necessarily sufficient to break contacts.

It is not necessary to modify the contact law in Eqn. (77) to account for $\tau\dot{\nu}$ however since from the numerical perspective, one can simply ignore the anomalous term, i.e., during sliding,

$$\delta\dot{\sigma} + \mu(\nu + \tau\dot{\nu}) - \|\dot{\beta}\| = -\mu\tau\dot{\nu} \approx 0, \tag{79}$$

which is true for small damping τ and moderate ghost velocity $\dot{\nu}$.

Discretizing this model in time using the techniques described in Sec. 7, Sec. 8 and Sec. 9 yields the nonlinear complementarity problem

$$\begin{bmatrix} M & -G^T & -N^T & -D^T & 0 \\ G & T_G & 0 & 0 & 0 \\ N & 0 & T_N & 0 & 0 \\ D & 0 & 0 & T_D & B \\ 0 & 0 & U & -B^T & T_U \end{bmatrix} \begin{bmatrix} v \\ \lambda \\ \nu \\ \beta \\ \sigma \end{bmatrix} + \begin{bmatrix} q_v \\ q_g \\ q_c \\ q_t \\ q_s \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \rho \\ 0 \\ \eta \end{bmatrix} \tag{80}$$
$$0 \leq \nu \perp \rho$$
$$0 \leq \sigma \perp \eta.$$

We assume here that there is no other inequality constraint beside contacts and Coulomb friction and that all holonomic and nonholonomic equality constraints have been concatenated in one block with Jacobian G for conciseness. The diagonal perturbations T and right hand side vectors b follow as as given in Eqn. (55)(56).

To concentrate on the contact equations in what follows, we transform Eqn. (80) to the reduced problem

$$\begin{bmatrix} A_{NN} & A_{DN}^T & 0 \\ A_{DN} & A_{DD} & B \\ U & -B^T & T_U \end{bmatrix} \begin{bmatrix} \nu \\ \beta \\ \sigma \end{bmatrix} + \begin{bmatrix} q_\nu \\ q_\beta \\ 0 \end{bmatrix} = \begin{bmatrix} \kappa \\ \rho \\ \gamma \end{bmatrix} \tag{81}$$
$$0 \leq \nu \perp \kappa \geq 0$$
$$0 \leq \sigma \perp \gamma \geq 0.$$

which is the Schur complement obtained by eliminating v and λ, which can be recovered after solving Eqn. (81). Linearizations of the friction laws and

polygonization of the cone $\mu\nu - \|\beta\| \leq$ introduce a complementarity condition in Eqn. (81)

$$0 \leq \beta \perp \rho \geq 0. \tag{82}$$

The polygonized models are explained in Sec. 13.

In what follows, we restrict ourselves to poligonized version and abbreviate the corresponding LCP as

$$0 \leq Hz + q \perp z \geq 0, \tag{83}$$

where

$$H = W + \tilde{U} \tag{84}$$

and

$$W = \begin{bmatrix} A_{NN} & A_{DN}^T & 0 \\ A_{DN} & A_{DD} & B \\ 0 & -B^T & T_U \end{bmatrix} \quad \tilde{U} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ U & 0 & 0 \end{bmatrix}. \tag{85}$$

Note that any anomalies in the Coulomb friction law introduced here is of the order of the time step, and these are small when considering large errors resulting from iterative solution methods.

A thorough analysis comparing seemingly and truly different formulations of Coulomb friction formulations belongs to a different article. Suffice to say here that our model can be mapped directly to solvable complementarity formulations as we now show.

13. Linearizations and approximations of the friction model

The perturbations are neglected in this section for brevity, and since they are not related to the various approximations discussed here. We also drop all the time derivatives on the ghosts, relabeling $\sigma \leftarrow \dot{\sigma}$ since we do not need to integrate the ghost velocities.

There is only one nonlinear component in our frictional model and that is the direction of the tangential force which appears in in Eqn. (74). This can be approximated by fixing the presumed sliding direction when sliding occurs based on an educated guess, or by linearization. This can be done either by considering box bounds separately along orthogonal directions in the sliding plane, or by introducing a basis $d^{(i)}, i_1, i_2, \ldots, i_m, m \geq 3$, so that one can write

$$\beta = \sum_i d^{(i)} \xi_i, \quad \text{where } \xi_i \geq 0, \tag{86}$$

and thus approximate the norm

$$\|\beta\| \approx \sum_i \xi_i. \tag{87}$$

This corresponds to the Anitescu-Potra model (4, 5). It can be shown in fact that all linearization reduce to the problem given in Eqn. (81) but with different definitions of matrix U and B, which nevertheless retain the property that $U_{ij} \geq 0$, which is needed for the problem to be solvable, and which explains the negative results of Sec. 14.

Solvability hinges on the fact that matrix H is strictly copositive (see 23) when the perturbations are included. This holds even for the nonlinear version as explained in Appendix II 20

Solvability does not mean "easily solved" however. Most numerical methods for solving complementarity problems work only on P and P_0 problems, with the exception of the Lemke algorithm which can solve all strictly copositive problems (23, 37).

A matrix is P_0 if and only if all principal sub-minors are non-negative (see 23, Thm.3.3.4), and there is a strict inclusion $P_0 \subsetneq P$. Consider the principal sub-minor

$$\begin{vmatrix} \alpha & \beta & 0 \\ \beta & \tau & 1 \\ \mu & -1 & 0 \end{vmatrix} = -1 \begin{vmatrix} \alpha & \beta \\ \mu & -1 \end{vmatrix} = \alpha + \mu\beta \tag{88}$$

and since there is no restriction of sign on β, this can be negative. This also implies that principal submatrices can be degenerate so that applying Newton's method on the Fischer-Burmeister function, smoothed or not, can fail.

14. Solving the frictional contact problem

Lemke's algorithm is far too slow and inefficient for large problems and so one is forced to use various kinds of iterative and splitting methods. A reason of particular relevance here is that pivoting methods proceed by computing rows and columns of the inverse of matrix H, and this destroys sparsity patters which can be exploited by factorization methods. There is recent progress in algorithmic development to improve efficiency (47) but still, the run-times reported recently are in measured in seconds or hundreds of seconds for problem sizes of a few thousand variables. We have budgets of five to 10

milliseconds for problems of this size as described below in Sec. 17 and so this is not a suitable avenue.

Unfortunately, the only mathematical results we can show about splitting methods are negative ones. These hold true for all currently used methods known by these authors. Yet, either the following negative results are original, or they are not well-known yet and so they are included here.

Splitting methods for LCPs are similar to those of linear problems. A matrix H is represented as the sum

$$H = (H - N) + N, \tag{89}$$

and the sequence of problems

$$\begin{aligned} 0 \leq N z^{(k+1)} + q^{(k)} \perp z^{(k+1)} \geq 0 \\ q^{(k+1)} = (H - N) q^{(k)}, \quad k = 1, 2, \ldots \end{aligned} \tag{90}$$

is iterated until, hopefully, a fixed point is reached. This is referred to a q-splitting and written as $(N, H - N)$. Convergence is assured for a variety of splittings provided matrix H is at least symmetric positive semi-definite. There are some results on more exotic and specialized matrices in either P_0 or P classes (see 23, Ch. 5).

For the case of the friction problem, the simplest splitting is that of Eqn. (84). Since W is positive definite and thus a P matrix each problem in Eqn. (90) at step k has a unique solution. In fact, LCP$(W, q^{(k)})$ is equivalent to a QP subject to box bounds on the β variables. There are good solution techniques for that. Our choice is described further below in Sec. 16.

The (W, U) splitting has the property to be generating a bounded sequence of iterates as shown in Lemma 19.1 in the Appendix 19. Unfortunately, though there are converging subsequences for these iterates, there is currently no reliable algorithm for locating them.

15. Projected Gauss-Seidel

Another negative result is derived now regarding convergence of the popular Projected Gauss-Seidel algorithm, in any of its variants, when applied to the frictional contact problem. We list the algorithm for that in Alg 21.1. Well-known convergence results on splitting methods for LCPs (see 23, Sec.5.3) assume that the matrix M in LCP(M, q) is either symmetric, P or at least P_0. Our matrix $W + \tilde{U}$ is neither, however. The projected Gauss-Seidel method

corresponds to the splitting $H = W + \tilde{U} = (L + D + \tilde{U}) + L^T$ where L and D are the strict lower triangular and (block) diagonal parts of W. The PGS iterations on a given contact with scalar normal force ν and normals β_1, β_2 amount to solving the following problems sequentially. First comes the LCP

$$0 \leq a_{ii}\nu + q_i \perp \nu \geq 0 \tag{91}$$

and this is followed by the boxed QP

$$\begin{aligned} &\min_{\beta} \frac{1}{2}\beta^T a_{jj}\beta + \beta^T q_j \\ &\text{subj. to} \\ &-\mu\nu \leq \beta \leq \nu, \end{aligned} \tag{92}$$

where a_{ii}, a_{jj} are diagonal elements of matrix A from Eqn. (81), and q_i, q_j are updated values. The last problem can be decomposed as two one dimensional problems or a single 2×2 one. We choose the latter case here for simplicity. But if β reaches the bounds, the solution of Eqn. (92) is a simple substitution

$$\beta \longleftarrow \pm\mu\nu. \tag{93}$$

Thus if all contacts are in steady sliding, one only solves for the normals and the iterations for these actually become

$$\left[L_{NN} + D_{NN}\right] \nu^{(k+1)} = \left[L_{NN}^T + L_{DN}^T B U\right] \nu^{(k)} - q_N, \tag{94}$$

where $\beta^{(k)} = BU\nu^{(k)}$ is the tangent force of the system for a matrix $B = \text{diag}(\pm, \pm, \ldots, \pm)$. The iterations Eqn. (94) correspond to a Gauss-Seidel splitting of the matrix

$$A_{NN} + A_{DN}^T B \tag{95}$$

which is neither symmetric nor positive definite. Therefore, we can expect sliding configurations to be problematic.

16. Direct-iterative splittings

Since our focus is the simulation of heavy machines and since these have such mass ratios as to require direct methods to guarantee small constraint violations for the joints, we choose a splitting which computes first the constraint forces for the equality constraints and the normals using a direct

method, and follow this with projected Gauss-Seidel iterations on the normals and the tangents. Experience has shown that it is better to iterate both on tangent and normal forces, even though pure tangential iterations would correspond to the splitting discussed in Lemma 19.1 which is non-divergent at least.

This is then a two-level type of splitting in which we first solve the mixed complementarity problem

$$\begin{bmatrix} M & -G^T & -N^T \\ G & T_G & 0 \\ N & 0 & T_N \end{bmatrix} \begin{bmatrix} v^{(k+1/2)} \\ \lambda^{(k+1/2)} \\ \nu^{(k+1/2)} \end{bmatrix} + \begin{bmatrix} q_v - D^T\beta^{(k)} \\ q_g \\ q_c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \rho \end{bmatrix} \qquad (96)$$
$$0 \le \nu \perp \rho \ge 0,$$

using a direct method, and then, continuing with

$$\begin{bmatrix} M & -N^T & -D^T & 0 \\ N & T_N & 0 & 0 \\ D & 0 & T_D & B \\ 0 & U & -B^T & T_U \end{bmatrix} \begin{bmatrix} v^{(k+1)} \\ \nu^{(k+1)} \\ \beta^{(k+1)} \\ \sigma^{(k+1)} \end{bmatrix} + \begin{bmatrix} q_v - G^T\lambda^{(k+1/2)} \\ q_c \\ q_t \\ q_s \end{bmatrix} = \begin{bmatrix} 0 \\ \rho \\ 0 \\ \eta \end{bmatrix} \qquad (97)$$
$$0 \le \nu \perp \rho \ge 0$$
$$0 \le \sigma \perp \eta \ge 0.$$

This last problem is solved using one of the linearizations and discussed in Sec. 13.

Assuming we always terminate such iterations with the direct solver, the only error made is in the approximation of the tangential forces. This introduces anomalous sliding, of course, but minimizes constraint violation for the machinery as well as penetration. The latter two are the most important in our applications.

We use a block pivot method for the direct solver which is is equivalent to Newton-Raphson iterations on the non-smooth Fischer-Burmeister function (see 37), and we give an explicit listing of this in Alg. 21. We prevent cycles by brute force detection and limit the iterations to 10. If we do not find a solution after reaching the maximum number of iterations, we choose the one with the least residual. This has worked well in practice but clearly needs improvements.

In our variant of the projected Gauss-Seidel iterations, we solve for the normal force of a single contact, and use this value to solve for the two

tangential forces. If either reaches its limit, these are projected according to

$$\beta \longleftarrow \mu\nu \frac{1}{\|\beta\|}\beta, \tag{98}$$

which reduces anisotropy.

17. Numerical results

We performed numerical experiments with simple scripts to illustrate the theory on small problems. To be specific, we used OCTAVE (**?**) for all the scripting with some `C++` modules of our own when performance was an issue.

For the larger experiments and real-life examples we used AgX toolkit from Algoryx Simulation (1). This is a commercial code designed for training simulators of the type described in the introduction among other application. We used this in two different ways. The first and simplest was to use the solvers built into the toolkit itself. This uses a direct LDLT factorizer developed by these authors (41), which is tailored specially for multibody systems, and an iterative Projected Gauss-Seidel solver described above. The LCP solver is as described in Sec. 16.

The second configuration uses AgX for collision detection, Jacobian computations, and rendering. We extracted kinematic data into OCTAVE via an HDF5 pipeline to use different prototype solvers. This still allowed for real-time rendering and interaction at frame rates of 20Hz due to communication overhead.

Though the present experiments focus on fidelity, stability, accuracy and convergence rates instead of raw performance, we note that all examples do run in real-time, including the larger ones, including when we use the computationally intensive direct-iterative solvers. This means that numerical work consumes from one to ten milliseconds per frame on the selected examples demonstrating that our methods are indeed usable in situ. In addition, we have not yet introduced parallelism in the solvers though this is in development.

We used a moderately powerful commodity desktop computer with 2x Intel Xeon X5650 CPUs clocked at 2.67GHz with 12MB cache each. The examples need less than 2GB RAM to run properly though the machine had 24GB available. Only the wheel loader problem takes significant amounts of computing time, namely, around 10 ms, more than half the budget for 60 Hz real-time graphics.

With regards to PGS iterations, we looked at a few alternatives for solving each contacts. Extensive testing of these is left for future work but briefly, these are as follows.

All separate: we loop over each equation and apply bounds on the tangential forces separately, performing standard Gauss-Seidel updates on all variables;

Block projection: we solve for normals and tangents with a direct matrix factorization on the 3×3 problems assuming stiction. If sliding is detected, we adjust the magnitude of tangential forces but leave the direction alone;

Block tangents: we solve for the normal first and then after the Gauss-Seidel update, solve for the two tangents at once solving the 2×2 system directly. The length of the resulting vector is then projected within bounds allowed by the normal force;

All these methods have slightly different performance but they can all fail in similar ways.

Our parametrization is nearly uniform with

- ▷ Time step: $h = 1/60$ which is the real-time graphics requirement;
- ▷ Regularization: $\epsilon = 10^{-8}$ to provide perturbations of 10^{-3} according to Eqn. (56) which is a safe regime for the factorizer unless otherwise indicated;
- ▷ Damping: $\tau = 2h$, which gives a relaxation time of two time steps and has proven stable over wide ranges of simulations;
- ▷ Friction: defaults to $\mu = 0.5$ but is set to larger values for some tests as described below;
- ▷ Gauss-Seidel maximum iterations: fifteen by default as a reasonable compromise for simulations, except when indicated otherwise;
- ▷ Direct-iterative coupling: five by default;

17.1. Fidelity and stability

First consider the harmonic oscillator simulated by SPOOK using $g(x) = 0$ as a constraint, and in comparison with the implicit midpoint and implicit Euler methods. The point here is to show that the ghost formulation is faithful to the dynamics, meaning that it reproduces the correct motion, and that it is stable also when the frequencies are unrealistically high at which point the oscillations are quickly filtered out. The results are shown in Fig. 4.

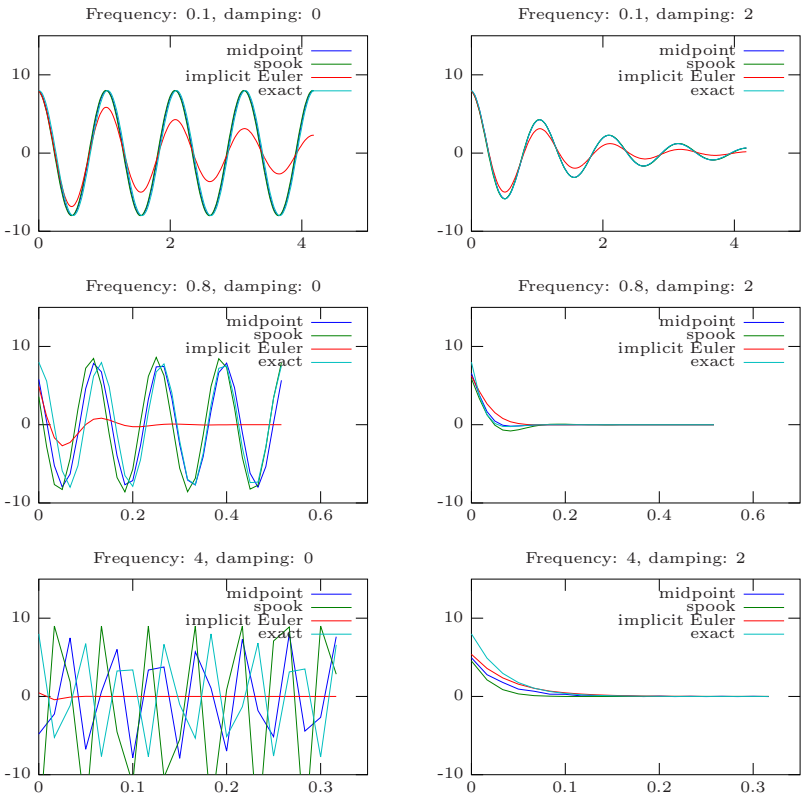

Figure 4: The implicit Euler method dissipates artificially at high rate, and shift the frequency noticeably. Both SPOOK and the implicit midpoint method preserve the energy and frequency adequately though they introduce a phase shift of $O(h^2)$. Under high frequency, SPOOK damps the system in a reliable way and filters out the high frequency oscillations.

Next comes the two dimensional pendulum which illustrates the fidelity

and stability of SPOOK in comparison to penalty and projection methods. As discussed in Sec. 3, the implicit mid-point method can fail, a well known phenomena (see 15). The implicit Euler method reduces the effective strength of gravity, and the projection method nearly removes that force from the motion. Note that such projection methods are widely used in game physics engines today. For this simple case, because the Jacobians always have full row rank, it is possible to set the perturbation $\epsilon = 0$ in SPOOK which is clearly impossible for penalty methods.

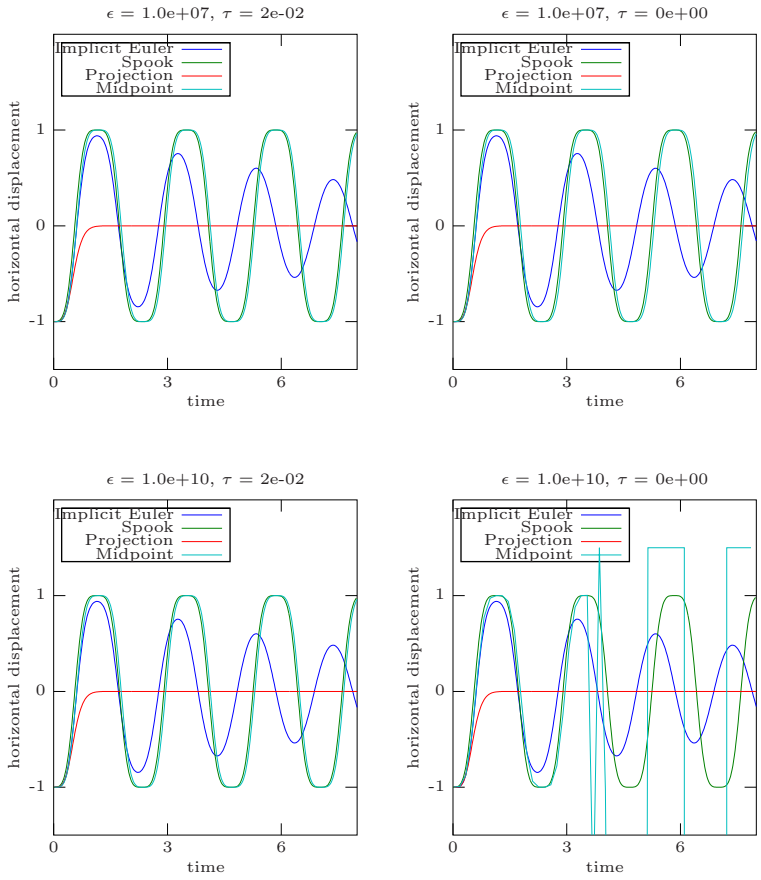

Figure 5: Simulations of the two dimensional pendulum using penalty methods including implicit Euler and implicit midpoint, in comparison with SPOOK. Linearized implicit Euler looses all the physics and the linearized midpoint method can go unstable, but SPOOK is both faithful and stable.

17.2. Dissipation of the impact model

Here comes a numerical illustration of the result of Sec. 11 showing the strict dissipation of the model, something that cannot be guaranteed for all approximate, post facto detection of penetration. Note that without the impact stage, the constraint stabilization used in SPOOK removes all the energy of the system. Because the constraint stabilization works to remove the penetration which exists when the contact is detected, some energy is

lost even for purely elastic impacts, unlike the reference model which requires impact location (28). Impact location is easily done here, but for non-convex geometries in three dimensions, that is nearly impossible except with brute force searches.

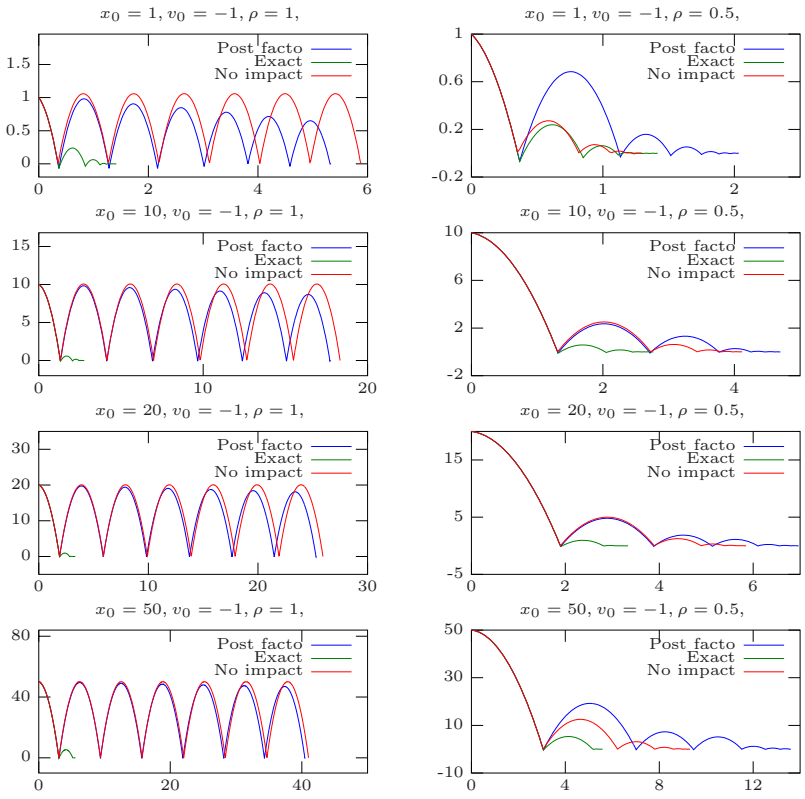

Figure 6: The bouncing ball. This demonstrates long term behavior of the impact resolution model from Sec. 11, in comparison to the reference, energy preserving model (28).

17.3. Different splitting

To demonstrate that the negative results regarding convergence of PGS methods, we tested different techniques for solving contact forces one contact at a time as described in the introduction to this section. These all fail to converge in similar ways. We looked at moderately long time behavior up to

one thousand iterations since that was sufficient to demonstrate the issues, and since the asymptotic relaxation rate is so close to nil in most cases.

For all relaxation plots in Figures 8, 9, and 10, the y-axis is the relative residual. The error relaxation is expected to decay as α^n where n is the iteration count and that should produce a straight line with slope α on the semi-log scale. Gauss-Seidel iterations applied to symmetric positive definite matrices have $\alpha < 1$. Our data clearly shows that this is not the case for PGS applied to the frictional contact problem.

To understand better why the iterations stagnate, we drew vertical lines each time at least one contact equation changed state during a Gauss-Seidel sweep, i.e., when a contact switched from active to inactive, and when a contact switched from sliding to sticking mode. We believe that this is due to the weak conclusions of Lemma 19.1 which does not preclude such high frequency switching of states.

We chose cases that were known to be difficult apriori, because they involve objects with non-homogeneous inertia such as long boxes and long cylinders. We also made sure that there were contacts in sliding configuration since that is predicted to be problematic, something clearly verified by the experiments. These factors impact dramatically on the relaxation rates. But we have not yet identified a configuration in which the Gauss-Seidel iterations explode yet. It appears that contacts eventually separate when errors buildup quickly.

An example that is particularly relevant to our application is the log pile depicted in Fig. 7 which is a catastrophe for simple PGS solvers by combining objects with large ratios of principal inertiae as well as mixed sliding sticking contacts. We let the logs drop on an inclined plane within some guides. This generates non-uniform stacking and mixed set of contacts.

The collapsing brick wall is a popular example, also illustrated in Fig. 7. We added a twist by using rectangular boxes with one dimension larger than the others by a factor of four in one example and compared convergence results with the more commonly seen configuration with cubic or nearly cubic boxes. The convergence profiles shown in Fig. 9 show how plain PGS iterations can stagnate and sometimes jump in mixed contact states, and how this is adversely affected when the boxes dimension differ significantly. For non-uniform boxes, the index set is completely unstable and keeps changing at every Gauss-Seidel sweep.

But the most devastating example from the computational perspective is the "squeezed box" scenario in which we put a tall box between two planes

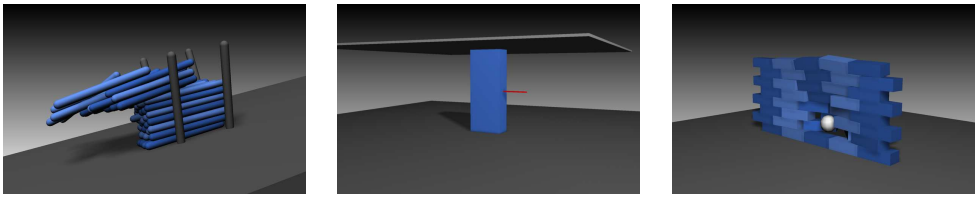

Figure 7: Different scenarios involving only contact constraints.

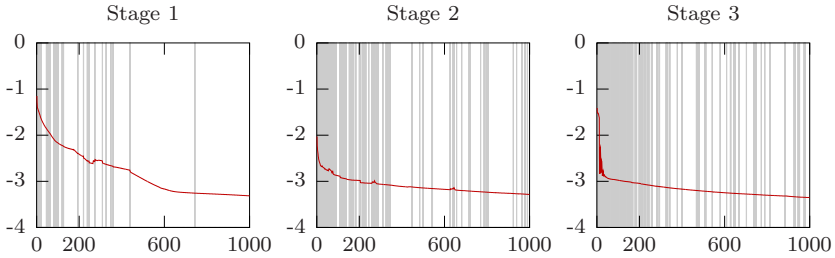

Figure 8: Simulation of a log pile on an inclined plane. The panels correspond to different time steps which contain contacts in different states. The splitting method being used is all separate. Grey vertical lines mark iterations during which at least one constraint changes state from active to inactive or vice versa, corresponding to stick-slip or separation-compression transitions.

and introduced a small compression as shown in Fig. 7. We then applied a large force at the center of mass in the horizontal direction which should force sliding. Mathematically, the regularization of the normals should balance all forces and produce constant and equal contact forces when removing gravity forces, as we did. This is far from what is observed however as shown in Fig. 10.

Relatively small relative errors of $O(10^{-6})$ are reached in some cases but one should note that these are very small examples. The alarming features are the sudden jumps, local exponential increases, fast oscillations, and long plateaus, even after large number of iterations.

17.4. Wheel loader

We provide here a scenario which demonstrates the strength of the direct-iterative splitting with a wheel loader simulation depicted in Fig. 11. The tractor is started at rest and is then driven to the rock pile to scoop a shovel full of rocks. It is then driven over the rock pile. The scenario lasts for ten seconds.

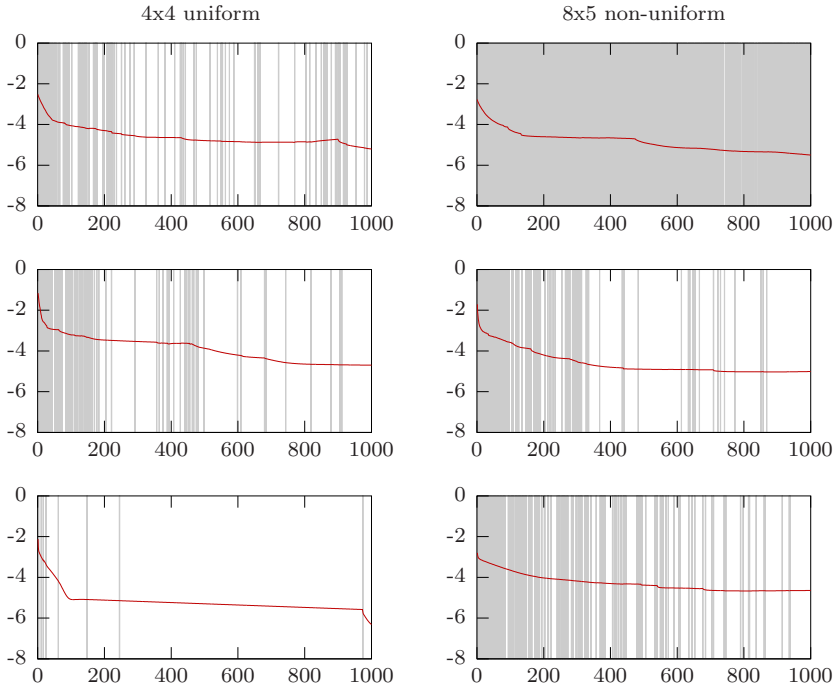

Figure 9: Collapsing brick walls. Regular walls are knocked down by a wrecking ball. The different rows correspond to different snapshots in time. The wall is at rest in the first row, impacted by the wrecking ball on the second, and collapsing on the third.

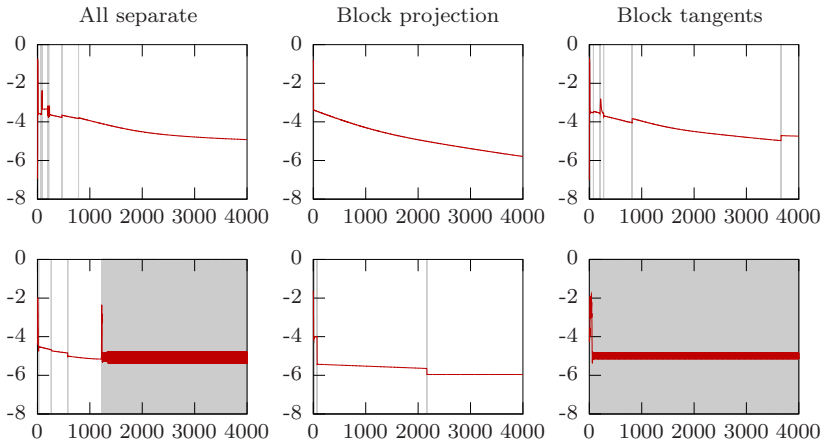

Figure 10: Simulation of a tall box being pulled between two fixed parallel plates. Each column shows a different splitting method and each row a different time step. This illustrates the difficulties linked to sliding as convergence is not guaranteed. This is seen here with jumps in error at index switches but we even have exponential increase with fixed index sets. There are problems with several variants of these splitting as shown here.

As additional parameters, some of the machine constraints were regularized with real physical values such as the constraints which correspond to the wheel suspensions, since these are much bigger than required for numerical stability. The same applies for nonholonomic constraints which represent the motors. These are also implemented as inequalities to account for maximal torques and forces. In this case, the regularization of the nonholonomic constraint, δ in Sec. 5, corresponds to energy loss while reaching the condition $A(q)\dot{q} - \omega(t) = 0$, where $\omega(t)$ is the desired joint velocity input from the user. The friction values for different combinations of ground, wheels, stones and bucket where kept in the range of 0.3–0.4.

The splitting algorithm performs first a direct solve, then performs fifteen PGS sweeps, and finishes with a second direct solve.

The root mean square constraint violation excluding the Coulomb friction conditions is plotted as a function of time in Fig. 12, showing good stability around $\|g\| \approx 10^{-3}$, which is sufficiently small in this context. The residual errors of the direct and iterative solvers are shown in Fig. 13. For the direct solver, we used the norm of the complementarity vector s defined in Eqn. (114). The time history shows that the block pivot method performs well in most cases, though it fails to compute a good solution at five different

instants in this particular run. Such errors are rarely catastrophic, however. The iterative solver can never really decrease the error by more than 10% under such conditions given fifteen iterations. This is a good reason to solve for normal forces in the direct solver along with other constraints since otherwise, penetrations would be far too big.

The convergence of the error as a function of the time step is shown in Fig. 14. This decays initially as $O(h^2)$, something previously reported (40). There is a plateau near $h = 1/60$ since asymptotically, we have $\|g\| = O(\epsilon)$.

In total, there are 164 bodies including the tractor and the rocks which introduces nearly one thousand degrees of freedom. Approximately 350 constraints are active at any time, each constraint including several equations, three for contacts, and five or six for hinge and prismatic joints. This changes over time due to variable number of contacts which are generated dynamically. Pairs of contacting rocks are subject to multiple contact constraints since individual rocks are modeled with non-convex aggregates of spheres of different sizes. This is an approximation of the graphics representation seen in the pictures. This is more detailed for cosmetic reasons.

Overall, there are several times more constraint equations than there are degrees of freedom, nearly four times as many on average. This redundancy cannot really be filtered in advanced since contact selection should in fact be performed by the complementarity algorithm to account for global couplings. Filtering contact constraints between pairs of bodies is possible in principle but there is currently no good method for that. Sphere aggregates offer a reasonable compromise.

As far as performance is concerned, collision detection and contact generation takes 1.5 ms on average for the nearly 800 geometric objects present in the system, the majority of which are spheres. The solver itself takes between 8 to 22ms. The latter applies to the configuration in which the tractor rolls over the rock pile.

18. Conclusion

The theory of ghosts allows the systematic treatment of regularization and stabilization parameters, as well as holonomic, nonholonomic, ideal and nonideal constraints in mechanical systems. Of particular interest is the case of "effort constraints" (see 44), an example of which is Coulomb friction. This alleviates mathematical issues related to nonsmooth analysis and avoids singularities found in idealized problems. The result is a clear understanding

Figure 11: Wheel loader simulation.

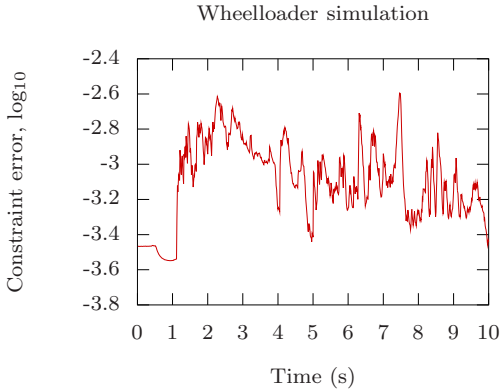

Figure 12: Constraint violations. The y-axis is the log of the root mean square constraint error $\log(\|g\|)$. This is plotted as a function of time for a ten seconds simulation.

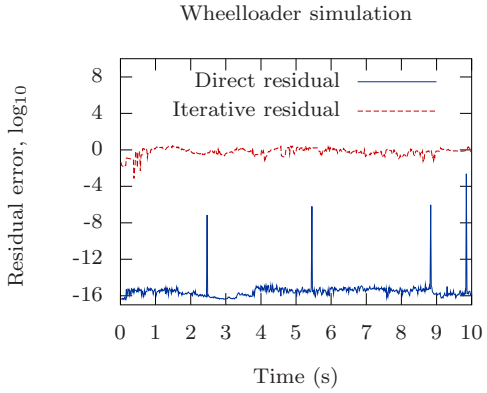

Figure 13: Residual errors from different parts of the split solver. The top line is the logarithm residual error computed by the PGS algorithm after fifteen iterations. The second is the logarithm of the complementarity error, defined in Eqn. (114), reported by the direct solver.

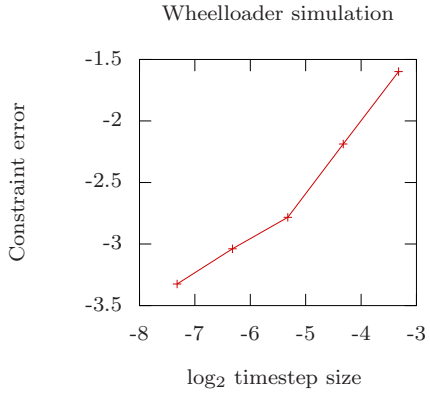

Figure 14: Order of error convergence as a function of the time step for the wheel loader simulation. This is initially $\|g\| = O(h^2)$ but flattens out given that $\|g\| = O(\epsilon)$ asymptotically as $h \downarrow 0$.

of the necessary numerical regularization parameters in terms of the physics they introduce.

The Spook stepper provides constraint stabilization solving a single system of linear equations per step when there are only equality constraints, and one LCP when there are inequalities and friction. There are few if any provably stable single stage schemes for the integration of multibody systems in descriptor form as well as few constraint stabilization schemes which are sufficiently stable for very low order integrators. In addition, the stabilization and regularization parameters are not only physics based but they can be validated in the numerical simulation in contrast with most penalty schemes. The regularization and stabilization parameters need little tuning as the analysis clearly predicts how they should be chosen to guarantee numerical stability, at least for systems which are not too far from being linear. Our experience however shows that stability is preserved over a large range of speeds.

The proof that frictional contact problems do not have the P property in Sec. 13, for either linear or non-linear models in fact, shows that one should expect the existence of multiple solutions, which is one aspect of the Painlevé paradox. The proof that some splittings generate bounded sequences but others don't give strong warnings and warrant extensive experimentation to understand when algorithms fail and how.

The fact that it is the sliding contacts which introduce the biggest errors and can introduce instability in numerical solvers is either a novel result or

a little known one. In any case, this helps understanding which relaxation profiles vary so wildly, and should give perspective regarding published data. One cannot consider arbitrary scenes to be representative of the performance of a solution scheme but one should in fact test for configuration in mixed states.

Our data shows how easy to construct examples which fail to converge, and which even diverge locally. The same data shows clearly how wild index set oscillations relate to stalled convergence, and that slight variations on Gauss-Seidel schemes produce surprisingly different results, though there is no clear winner.

Our direct-iterative scheme also proves sufficiently fast and reliable for real-time simulation of multibody systems subject to hard equality constraints as well as frictional contacts. This also seems to be novel, at least performance-wise. Though there is at least one other solver of a similar type (18), the latter uses a coordinate reduction formulations which cannot handle the type of systems we are looking at, and the performance reported is nowhere near the requirements of our simulations.

There are open issues in solving frictional contact problems efficiently since they do not have the P property and convergent splitting schemes for these do not yet exist. This is part of our present activities and future work.

19. Appendix I

We now prove Lemma 19.1.

Lemma 19.1. *For any matrix $W \in P$, and vector $p \in \mathbb{R}^n, p_i \geq 0$, the set $\{z \in \mathrm{SOL}(\mathrm{LCP}(W, q+p)) \mid p_i \geq 0, i = 1, 2, \ldots, n\}$ is bounded.*

Proof. Introduce index sets $\alpha, \beta \in \{1, 2, \ldots, n\}$ with $\alpha \cup \beta = \{1, 2, \ldots, n\}$ and $\alpha \cap \beta = \emptyset$ so that $z_\alpha \geq 0$ and $z_\beta = 0$. Let $|\alpha| = m = n - |\beta|$. Also introduce the closed convex polytope

$$S_\alpha = \{u \in \mathbb{R}^n_+ \mid u_i \geq 0, \sum u_i = 1, i \in \alpha, u_j = 0, j \in \beta\}. \tag{99}$$

Clearly, if $p \in \mathbb{R}^n_+$, then $p = \lambda u$ for $\lambda \geq 0$, $u \in S_n$. The solution of $\mathrm{LCP}(W, q+p)$ then satisfies

$$W_{\alpha\alpha} z_\alpha = -q_\alpha p_\alpha = -q_\alpha - \lambda u_\alpha \tag{100}$$

where $\lambda > 0, u_\alpha \in S_{|\alpha|}$. Since $W \in P$ then $(W_{\alpha\alpha})^{-1}$ exists and so we can write

$$z_\alpha = (W_{\alpha\alpha})^{-1}(-q_\alpha - \lambda u_\alpha) = r_\alpha - \lambda v_\alpha \geq 0. \tag{101}$$

This implies that $r_i \geq \lambda v_i, i \in \alpha$, and hat we cannot have both $r_i < 0$ and $v_i \leq 0$ simultaneously. If $r_i \geq 0$ and $v_i < 0$, there is no bound on $\lambda > 0$. However, whenever $r_i > 0$ and $v_i > 0$, $\lambda < \infty$. Because W is a P matrix, so is $W_{\alpha\alpha}$ and so is its inverse. This implies the existence of at least one $v_i > 0$, with strict inequality and therefore a bound on $\lambda \leq \bar{\lambda} < \infty$ for any v_α. Now, $f(u_\alpha) = v_\alpha = (W_{\alpha\alpha})^{-1}u_\alpha$ is a continuous function of u_α and so the image $f(S_\alpha)$ is also compact. Since the min and max functions are continuous, so is the function

$$h(v_\alpha) = \max_{v_i > 0} v_i. \tag{102}$$

Since $f(S_\alpha)$ is compact, $h(v_\alpha)$ attains its minimum for a point $v_\alpha^\star \in f(S_\alpha)$. Since $h(v_\alpha) > 0$ given that $(W_{\alpha\alpha})^{-1}$ is a P matrix, $\min(h) = h_{\min} = h(v_\alpha^\star) > 0$. Therefore, since $\lambda \leq \max(-q_\alpha)/\min(h(v\alpha))$, $\lambda < \infty$. The set of vectors $z_\alpha \geq 0$ satisfying Eqn. (101) is the product of two compact sets and is therefore compact, and so is intersection with the polyhedral set $z_\alpha \geq 0$. Therefore, $\|z_\alpha\| < \infty$ for all sets $\alpha \in \{1, 2, \ldots, n\}$ and since the solution SOL(LCP$(W, q + p)$) satisfies Eqn. (101) for some set α by the pigeon hole principle, the conclusion follows. □

Note that Lemma 19.1 does not necessarily hold for P_0 matrices since LCP$(W, q + p), W \in P_0$ can have unbounded or no solutions.

The splitting (W, U) defines a sequence $\{z\}_k, k = 1, 2, \ldots$ in a closed and bounded set and therefore has a converging subsequence. Finding such a subsequence is an open problem, however.

20. Appendix II

Here we present relevant aspects of the theory of solvability for nonlinear complementarity problems and show how this applies to the frictional contact problem.

We introduce the notion of exceptional families of elements (36) to construct the solvability proof. Consider a real $\mathcal{K} \in \mathbb{R}^n$ with dual $\mathcal{K}^\star$, and a continuous function $f : \mathbb{R}^n \mapsto \mathbb{R}^n$. An exceptional family of element is defined as follows

Definition 20.1. *A set of points* $\{z^{(r)}\}_{r>0} \in \mathcal{K}$ *is an exceptional family of elements if the following conditions hold:*

1. $z^{(r)} \in \mathcal{K}$ *for all* $r > 0$*;*
2. $\|z^{(r)}\| \to \infty$ *as* $r \to \infty$*;*
3. *for every* $r > 0$*, there exists a* $\mu^{(r)} > 0$ *such that* $w^{(r)} = \mu^{(r)} z^{(r)} + f(z^{(r)}) \in \mathcal{K}^\star$ *and* ${w^{(r)}}^T {z^{(r)}}^T = 0$*.*

With this definition, a theorem due to Isac (35) establishes existence. This is reproduced here without proof.

Theorem 20.1. *Consider a Hilbert space* H*, a closed convex* $\mathcal{K} \in H$ *and a completely continuous map* $f : H \mapsto H$*. Either there exists a solution to NCP*$(f, \mathcal{K})$ *or* f *has an exceptional family of elements.*

The intuitive picture here is that provided $z^T F(z)$ cannot grow to $-\infty$ too fast, i.e., provided $z^T F(z) \geq -O(\|z\|)^\alpha$ with $\alpha < 2$, for z in the feasible set, the problem is solvable by a fixed point argument since projection iterations cannot grow indefinitely. This is a nonlinear generalization of the copositivity property (see 23, Def. 3.8.1), which guarantees solvability in the linear case.

The nonlinear contact problem can be written in quasi-linear form

$$F(z) + q = H(z)z + q, \tag{103}$$

according to Eqn. (84) and (85), where the matrices $B(z)$ have the form

$$B = \text{diag}(B_{11}, B_{22}, \ldots, B_{n_c n_c}) \tag{104}$$

where n_c is the number of contacts, and the rectangular 1×3 blocks are the normal vectors

$$B_{ii} = \frac{1}{\|\beta^{(i)}\|} \beta^{(i)}, \tag{105}$$

for the ith contact point. Because of the bisymmetry of W in Eqn. (84) and the positivity of U, we have

$$z^T H z = z^T W z + z^T \tilde{U} z \geq 0, \tag{106}$$

where the inequality is strict for the perturbed problems.

Note that only the P property, which generalizes to nonlinear functions, can guarantee uniqueness. The negative result of Eqn. (88) shows that the frictional contact problem is guaranteed to have multiple solutions.

There are already proofs of existence of solutions for linear versions of friction models, one which is equivalent to the present model (5) for rigid bodies and which depends only on the copositive property, and another, longer proof for deformable models (see 26, Sec. 2.7). The latter is also a linear formulation and requires full row rank for the contact Jacobians. This is weak in the sense that contacts cannot be filtered before hand to be linearly independent. The reason being that contact forces are rays in wrench space for rigid bodies. For a point at location $p + x_{cm}$ on a rigid body, the wrench is then

$$w = \begin{bmatrix} n \\ n \times p \end{bmatrix} \tag{107}$$

where $n \times p$ is the usual cross product in $\mathbb{R}^3$. There can be infinitely such rays which are candidates for the contact forces. A point in case is a cube on a plane which has four candidate contacts, though only three are necessary. In general, the normal contact forces lie in the cone corresponding to arbitrarily many such rays and so rank degeneracy of the normal contact Jacobian matrix N defined in Sec. 11 must be assumed. The present result provides solvability with or without full rank, and for arbitrary non-negative friction coefficient.

21. Appendix III

We list the most important algorithms here. To simplify notation, we agglomerate all Jacobians into a matrix G. Using M for the mass matrix and T for diagonal perturbations, we solve the mixed linear complementarity problem

$$\begin{bmatrix} M & -G^T \\ G & T \end{bmatrix} \begin{bmatrix} v \\ \lambda \end{bmatrix} + \begin{bmatrix} p \\ q \end{bmatrix} = \begin{bmatrix} 0 \\ \rho_+ - \rho_- \end{bmatrix}$$
$$l \leq \lambda \perp \rho_+ \geq 0 \tag{108}$$
$$u \geq \lambda \perp \rho_- \geq 0,$$

where $-\infty \leq l < u \leq \infty$ are lower and upper bound vectors, and $\rho_\pm$ are positive and negative components of the slack vector. Applying Gauss-Seidel iterations directly on this linear system cannot work directly so we operate on the Schur complement

$$A = GM^{-1}G^T + T \tag{109}$$

which is symmetric and positive definite. This leads to the equivalent mixed LCP

$$\begin{aligned} A\lambda + (q + GM^{-1}p) &= \rho_+ - \rho_- \\ l \leq \lambda \perp \rho_+ &\geq 0 \\ u \geq \lambda \perp \rho_- &\geq 0. \end{aligned} \tag{110}$$

There is no need to compute matrix A explicitly however for either Gauss-Seidel or even Conjugate Gradient operations. Instead, we work on the original block matrices M, G and T to perform the necessary operations. In standard form, solving the system $A\lambda = b$ for a linear system consists of performing

$$\begin{aligned} r_i^{(\nu)} &\longleftarrow A_{i\bullet}\lambda - b_i \\ \Delta\lambda_i^{(\nu+1)} &\longleftarrow -A_{ii}^{-1} r_i^{(\nu)} \\ \\ \lambda_i^{(\nu+1)} &\longleftarrow \lambda_i^{(\nu)} + \Delta\lambda_i^{(\nu+1)}, \end{aligned} \tag{111}$$

where $A_{i\bullet}$ is the ith row of the matrix A. The last of these two updates can be performed efficiently for sparse matrices. For the current specialization, the residuals amount to $r = Gv + T\lambda + q$, and the updates need to include

$$v_b^{(\nu+1)} \longleftarrow v_b^{(\nu)} + M_b^{-1} G_{ib}^T \Delta\lambda_i^{(\nu+1)}, \tag{112}$$

for each body b such that $G_{ib} \neq 0$. Summarizing this, we get Alg. 21.1, in which we use n_c for the total number of constraints which may be organized in blocks, and n_{b_k} for the number of bodies connected via the (block) constraint k, and we use b for body labels. Note that body velocities are all updated simultaneously so that even when using a single equation at a time, G_{kb} is at least a $1 \times n_{dof}$ matrix, where n_{dof} is the number of degrees of freedoms for body b. This is six for rigid bodies. To simplify notation, we use $\mathrm{MLCP}(W, b, l, u)$ to denote

$$\begin{aligned} Wz + b &= w_+ - w_- \\ l \leq z \perp w_+ &\geq 0 \\ u \geq z \perp w_- &\geq 0 \end{aligned} \tag{113}$$

and write $z = \mathrm{SOL}(\mathrm{MLCP}(W, b, l, u))$.

Algorithm 21.1 Gauss-Seidel iterations to solve the MLCP in Eqn. (110)

Given p, q, M, G, T, l, u.
initialize $v = p, \quad \lambda = \lambda^{(0)}$.
Compute blocks $A_{kk} = \sum_b G_{kb} M_{bb}^{-1} G_{kb}^T + T_{kk}$, for $k = 1, 2, \ldots, n_c$
repeat
    for $k = 1, 2, \ldots, n_c$ **do**
        $r = q_k + T_{kk}\lambda_k$ ▷ Compute local residual from scratch
        for $b = b_{k_1}, b_{k_2}, \ldots, b_{n_{b_k}}$ **do**
            $r = r + G_{kb} v_b$
        end for
        $z \longleftarrow \text{SOL}(\text{MLCP}(A_{kk}, r - A_{kk}\lambda_k, l_k, u_k))$
        Update the bounds l_k, u_k if desired
        $\Delta\lambda_k = z - \lambda_k$
        $\lambda_k = z$
        for $b = b_{k_1}, b_{k_2}, \ldots, b_{n_{b_k}}$ **do** ▷ Loop over bodies connected via constraint k
            $v_b = v_b + M_{bb}^{-1} G_{kb}^T \Delta\lambda_k$ ▷ Update velocities
        end for
    end for
until Error is small, or iteration time is exceeded

Another relevant algorithm is the block pivot method for mixed LCPs. Here we introduce index sets α and β for the active and slack variables, and split $\beta = \sigma_l \cup \sigma_u$ for the positive and negative slacks, i.e., variables clamped at their upper or lower bounds with $\sigma_l \cap \sigma_u \emptyset$. We have $\alpha \cup \beta = \{1, 2, \ldots, n\}$, $\alpha \cap \beta = \emptyset$. The residual error in this case is computed from the complementarity vector s defined as

$$\begin{aligned} w &\longleftarrow Hz + q \\ s_i &= \begin{cases} |z_i - l_i + u_i - z_i| & \text{if } i \in \alpha \\ -\min(0, w_i) & \text{if } i \in \sigma_l \\ \max(0, w_i) & \text{if } i \in \sigma_u \end{cases} \end{aligned} \tag{114}$$

Algorithm 21.2 Block pivot algorithm for MLCP based on Newton-Raphson iterations applied to nonsmooth formulation.

Given an $n \times n$ real matrix H, an n-dimensional real vectors q and n-dimensional vectors of bounds $-\infty \leq l < u \leq \infty$, integer $\nu_{max} > 1$, tolerance $\tau > 0$ and sets $\alpha, \sigma_l, \sigma_u, \beta = \sigma_l \cup \sigma_u$

repeat

Solve : $H_{\alpha\alpha} z_\alpha = -q_\alpha - H_{\alpha\sigma_l} l_{\sigma_l} - H_{\alpha\sigma_u} u_{\sigma_u}$

Compute : $w_\beta \leftarrow H_{\beta\alpha} z_\alpha + q_\beta$

$\delta_l \leftarrow \{i \in \alpha \mid z_i < l_i\}$ ▷ Variable active but too small

$\delta_u \leftarrow \{i \in \alpha \mid z_i > u_i\}$ ▷ Variable active but too large

$\gamma_l \leftarrow \{i \in \alpha \mid w_i < 0\}$ ▷ Variable at bound but negative residual

$\gamma_u \leftarrow \{i \in \alpha \mid w_i > 0\}$ ▷ Variable at bound but positive residual

$\sigma_l \leftarrow (\sigma_l \setminus \gamma_l) \cup \delta_l$

$\sigma_u \leftarrow (\sigma_u \setminus \gamma_u) \cup \delta_u$

$\beta \leftarrow \sigma_l \cup \sigma_u$

$\alpha \leftarrow \{1, 2, \ldots, n\} \setminus \beta$

Compute s as per Eqn. (114)

until $\|s\| < \tau$ or $\nu > \nu_{max}$

Acknowledgments

Algoryx Simulation AB has supported this research actively. Both authors work for this company as consultant.

This research was supported by High Performance Computing Center North (HPC2N), Swedish Research Council under grant VR7062571 and the eSSENCE-project, EU Mal 2 Structural Funds (UMIT-project), and VINNOVA/ ProcessIT Innovations.

Bibliography

[1] Algoryx Simulation AB (2011).

[2] P. R. Amestoy, I. S. Duff, J. Y. L'Excellent, Multifrontal parallel distributed symmetric and unsymmetric solvers, Computer Methods in Applied Mechanics and Engineering 184 (2–4) (2000) 501–520.

[3] M. Anitescu, Optimization-based simulation of nonsmooth rigid multibody dynamics, Math. Program. 105 (1, Ser. A) (2006) 113–143.

[4] M. Anitescu, F. A. Potra, Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems, Nonlinear Dynamics 14 (1997) 231–247.

[5] M. Anitescu, F. A. Potra, D. E. Stewart, Time-stepping for three-dimensional rigid body dynamics, Computer Methods in Applied Mechanics and Engineering 177 (1999) 183–197.

[6] G. Arfken, Mathematical Methods for Physicists, 3rd ed., Academic Press, New York, 1985.

[7] V. I. Arnol′d, Mathematical Methods of Classical Mechanics, vol. 60 of Graduate Texts in Mathematics, 2nd ed., Springer-Verlag, New York, 1989, translated from the Russian by K. Vogtmann and A. Weinstein.

[8] V. I. Arnold, V. V. Kozlov, A. I. Neishtadt, Mathematical aspects of classical and celestial mechanics, Springer-Verlag, Berlin, 1997.

[9] U. Ascher, H. Chin, L. Petzold, S. Reich, Stabilization of constrained mechanical systems with DAEs and invariant manifolds, M. Mech. Struct. & Mach. 23 (1995) 135–158.

[10] U. Ascher, P. Lin, Sequential regularization methods for higher index DAEs with constraint singularities: The linear index-2 case, SIAM J. Num. Anal. 33 (1996) 1921–1940.

[11] U. Ascher, P. Lin, Sequential regularization methods for nonlinear higher index DAEs, SIAM J. Sci. Comp. 18 (1997) 160–181.

[12] U. Ascher, P. Lin, Sequential regularization methods for simulating mechanical systems with many closed loops., SIAM J. Sci. Comp. 21 (4) (1999) 1244–1262.

[13] U. M. Ascher, H. S. Chin, S. Reich, Stabilization of DAEs and invariant-manifolds, Numerische Mathematik 67 (2) (1994) 131–149.

[14] U. M. Ascher, L. R. Petzold, Stability of computational methods for constrained dynamics systems, SIAM J. Sci. Computing 14 (1) (1993) 95–120.

[15] U. M. Ascher, S. Reich, The midpoint scheme and variants for Hamiltonian systems: Advantages and pitfalls, SIAM J. Sci. Comp. 21 (3) (1999) 1045–1065.

[16] U. M. Ascher, S. Reich, On some difficulties in integrating highly oscillatory Hamiltonian systems, in: Computational molecular dynamics: challenges, methods, ideas (Berlin, 1997), vol. 4 of Lect. Notes Comput. Sci. Eng., Springer, Berlin, 1999, pp. 281–296.

[17] E. Barth, K. Kuczera, B. Leimkuhler, R. D. Skeel, Algorithms for constrained molecular dynamics, Journal of Computational Chemistry 16 (10) (1995) 1192–1209.

[18] F. Bertails-Descoubes, F. Cadoux, G. Daviet, V. Acary, A nonsmooth Newton solver for capturing exact Coulomb friction in fiber assemblies, ACM Trans. Graph. 30 (2011) 6:1–6:14.

[19] F. A. Bornemann, Homogenization in Time of Singularly Perturbed Mechanical Systems, vol. 1687 of Lecture notes in mathematics, Springer, Berlin, 1998.

[20] V. N. Brendelev, On the realization of constraints in nonholonomic mechanics, J. Appl. Math. Mech. 45 (3) (1981) 481–487.

[21] B. Brogliato, Nonsmooth Mechanics: Models, Dynamics and Control, Communication and Control Engineering, 2nd ed., Springer-Verlag, Berlin, 1999.

[22] J. R. Bunch, W. J. Demmel, C. F. V. Loan, The strong stability of algorithms for solving symmetric linear systems, SIAM J. on Mat. Anal.and Appl. 10 (4) (1989) 494–499.

[23] R. W. Cottle, J.-S. Pang, R. E. Stone, The Linear Complementarity Problem, Computer Science and Scientific Computing, Academic Press, New York, 1992.

[24] M. de León, D. M. de Diego, A. S. Merino, Geometric integrators and nonholonomic mechanics, Journal of Mathematical Physics 45 (3) (2004) 1042–1064.

[25] G. De Saxce, Z. Q. Feng, The bipotential method: A constructive approach to design the complete contact law with friction and improved numerical algorithms, Mathematical and Computer Modelling 28 (4–8) (1998) 225–245.

[26] F. Facchinei, J.-S. Pang, Finite-dimensional variational inequalities and complementarity problems. Vol. I, Springer Series in Operations Research, Springer-Verlag, New York, 2003.

[27] A. D. Felice, Are modified gravity models free of ghosts?, Journal of Physics A: Mathematical and Theoretical 40 (25) (2007) 7061.

[28] R. C. Fetecau, J. E. Marsden, M. Ortiz, M. West, Nonsmooth Lagrangian mechanics and variational collision integrators, SIAM J. Appl. Dyn. Syst. 2 (3) (2003) 381–416.

[29] G. H. Golub, C. F. Van Loan, Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences, 3rd ed., Johns Hopkins Press, Baltimore, 1996.

[30] E. Hairer, Symmetric projection methods for differential equations on manifolds, BIT Numerical Mathematics 40 (2000) 726–734.

[31] E. Hairer, C. Lubich, G. Wanner, Geometric Numerical Integration, vol. 31 of Spring Series in Computational Mathematics, Springer-Verlag, Berlin, 2001.

[32] E. Hairer, S. P. Nørsett, G. Wanner, Solving Ordinary Differential Equations I: Nonstiff Problems, vol. 8 of Springer Series in Computational Mathematics, second revised edition ed., Springer-Verlag, Berlin, 1991.

[33] E. Hairer, G. Wanner, Solving Ordinary Differential Equations II: Stiff and Differential Algebraic Problems, vol. 14 of Springer Series in Computational Mathematics, second revised edition ed., Springer-Verlag, Berlin, 1996.

[34] N. J. Higham, Accuracy and Stability of Numerical Algorithms, 2nd ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002.

[35] G. Isac, Leray-Schauder type alternatives and the solvability of complementarity problems, Topol. Methods Nonlinear Anal. 18 (1) (2001) 191–204.

[36] G. Isac, V. Bulavski, V. Kalashnikov, Exceptional families, topological degree and complementarity problems, J. Global Optim. 10 (2) (1997) 207–225.

[37] J. J. Júdice, Algorithms for linear complementarity problems, in: E. Spedicato (ed.), Algorithms for Continuous Optimization, vol. 434 of NATO ASI Series C, Mathematical and Physical Sciences, Advanced Study Institute, NATO, Kluwer Academic Publishers, 1994.

[38] C. Kane, J. E. Marsden, M. Ortiz, M. West, Variational integrators and the Newmark algorithm for conservative and dissipative mechanical systems, International Journal for Numerical Methods in Engineering 49 (2000) 1295–1325.

[39] A. V. Karapetian, On realizing nonholonomic constraints by viscous friction forces and Celtic stones stability, J. Appl. Math. Mech. 45 (1) (1981) 42–51.

[40] C. Lacoursière, Regularized, stabilized, variational methods for multibodies, in: D. F. P. Bunus, C. Führer (eds.), The 48th Scandinavian Conference on Simulation and Modeling (SIMS 2007), Linköping Electronic Conference Proceedings, Linköping University Electronic Press, Linköping, Sweden, 2007.

[41] C. Lacoursière, M. Linde, O. Sabelström, Direct sparse factorization of blocked saddle point matrices, in: Proceedings of Para 2010: State of the Art in Scientific and Parallel Computing, Reykjavík, June 2010, Lecture Notes in Computer Science, Springer Verlag, to appear.

[42] C. Lanczos, The Variational Pinciples of Mechanics, 4th ed., Dover Publications, New York, 1986.

[43] R. A. Layton, Principles of Analytical System Dynamics, Mechanical Engineering Series, Springer-Verlag, Berlin, 1998.

[44] R. A. Layton, B. C. Fabien, Systematic modelling using lagrangian daes, Mathematical and Computer Modelling of Dynamical Systems 7 (3) (2001) 273–304.

[45] R. I. Leine, U. Aeberhard, C. Glocker, Hamilton's principle as variational inequality for mechanical systems with impact, Journal of Nonlinear Science 19 (6) (2009) 633–664.

[46] A. D. Lewis, R. M. Murray, Variational principles for constrained systems: theory and experiment, Internat. J. Non-Linear Mech. 30 (6) (1995) 793–815.

[47] Q. Li, Large-scale computing for complementarity and variational inequalities, Ph.D. thesis, University of Wisconsin (2010).

[48] C. Liu, Z. Zhao, B. Brogliato, Frictionless multiple impacts in multibody systems I. Theoretical framework, Proc. Roy. Soc. A 464 (2100) (2008) 3193–3211.

[49] J. E. Marsden, M. West, Discrete mechanics and variational integrators, Acta Numer. 10 (2001) 357–514.

[50] A. Pandolfi, C. Kane, J. E. Marsden, M. Ortiz, Time-discretized variational formulation of non-smooth frictional contact, Internat. J. Numer. Methods Engrg. 53 (8) (2002) 1801–1829.

[51] H. Rubin, P. Ungar, Motion under a strong constraining force, Communications on Pure and Applied Mathematics X (1957) 65–87.

[52] O. Schenk, K. Gaertner, On fast factorization pivoting methods for sparse symmetric indefinite systems, Elec. Trans. Numer. Anal. 23 (2006) 158–179.

[53] K. Sekimoto, Newton's cradle versus nonbinary collisions, Physical Review Letters 104 (12) (2010) 124302.

[54] M. Servin, C. Lacoursière, N. Melin, Interactive simulation of elastic deformable materials, in: Proceedings of SIGRAD Conference 2006 in Skövde, Sweden, Linköping University Electronic Press, Linköping, 2006.

[55] M. Tao, H. Owhadi, J. E. Marsden, Symplectic, linearly-implicit and stable integrators with pplications to fast symplectic simulations of constrained dynamics, Arxiv.org arXiv:1103.4645v1.