All's Well That Ends Well: Guaranteed Resolution of Simultaneous Rigid Body Impact

ETIENNE VOUGA, University of Texas at Austin
BREANNAN SMITH, Columbia University
DANNY M. KAUFMAN, Adobe Research
RASMUS TAMSTORF, Walt Disney Animation Studios
EITAN GRINSPUN, Columbia University

Iterative algorithms are frequently used to resolve simultaneous impacts between rigid bodies in physical simulations. However, these algorithms lack formal guarantees of termination, which is sometimes viewed as potentially dangerous, so failsafes are used in practical codes to prevent infinite loops. We show such steps are unnecessary. In particular, we study the broad class of such algorithms that are conservative and satisfy a minimal set of physical correctness properties, and which encompasses recent methods like Generalized Reflections as well as pairwise schemes. We fully characterize finite termination of these algorithms. The only possible failure cases can be detected, and we describe a procedure for modifying the algorithms to provably ensure termination. We also describe modifications necessary to guarantee termination in the presence of numerical error due to the use of floating-point arithmetic. Finally, we discuss the challenges dissipation introduce for finite termination, and describe how dissipation models can be incorporated while retaining the termination guarantee.

CCS Concepts: • **Computing methodologies → Physical simulation**; *Collision detection*;

Additional Key Words and Phrases: Collision response, Elastic impact, Rigid Bodies, Gauss-Seidel, Termination, Physical simulation

ACM Reference format:
Etienne Vouga, Breannan Smith, Danny M. Kaufman, Rasmus Tamstorf, and Eitan Grinspun. 2017. All's Well That Ends Well: Guaranteed Resolution of Simultaneous Rigid Body Impact. *ACM Trans. Graph.* 36, 4, Article 1 (July 2017), 19 pages.
https://doi.org/http://dx.doi.org/10.1145/3072959.3073689

1 INTRODUCTION

The numerical simulation of collisions between multiple objects in *simultaneous impact* is challenging. While conservation of energy and momentum completely determine the motion of a pair of colliding rigid balls, the same conservation laws are not sufficient to determine the behavior when three or more balls collide [Glocker 2004]. Additional assumptions are needed about the material properties of the objects and the way shocks propagate through them in order to make the contact problem well-posed.

Simultaneous impacts can occur in the time-continuous setting, but become even more prevalent when time is discretized, and one attempts to resolve all interferences that occur within the span of a time integration step. The problem is particularly well studied in the case of colliding rigid bodies, and it arises naturally in the simulation of granular media [Nguyen and Brogliato 2014]. This work focuses on responding to these impacts at one frozen instant in time, at the velocity level; responding to impact thus entails applying impulses to the colliding objects so that they are no longer approaching. We will not discuss the many alternative formulations based on position corrections, soft (acceleration-level) responses, etc, in detail; see any of several comprehensive surveys [Bender et al. 2014; Gilardi and Sharf 2002; Khulief 2012] for an overview.

When multiple objects are touching, resolving collisions between some of them often creates new collisions between others. For example, in a perfect head-on pool break, the cue ball first collides (only) with the ball at the tip of the pyramid, but a resolution of this collision in isolation induces new collisions against the two balls on the next level of the pyramid, and so forth. For this reason, the usual approach for solving the multiple impact problem is to adopt an iterative strategy [Han and Gilmore 1993]: a rule R is chosen for how to modify the velocities of objects to fix some subset of the collisions (we will call these rules *impact operators*). Applying the impact operator fixes some collisions while perhaps causing others; the operator is thus applied again, repeatedly, until all collisions are resolved. Algorithm 1 illustrates the general structure of algorithms that adopt this strategy; we call this structure *Gauss-Seidel-like*, by analogy to the iterative splitting method of the same name used to solve linear complementarity problems [Cottle et al. 1992; Erleben 2007]

A natural question that we must ask is ***will the iterative algorithm terminate, producing a collision-free solution in finite time?*** If not, what are the obstructions to termination? Clearly, Algorithm 1 is not guaranteed to terminate if the impact operator R is completely arbitrary. However, the rule R is necessarily restricted by physical considerations, and we can ask whether physical assumptions on R are sufficient to guarantee termination. In this setting termination has been shown for simple geometries (such as a straight line of balls of different masses, or impacts involving a limited number of objects [Jia et al. 2013]) but the general question remains open.

We thank Andy Ruina for discussions early in this project, and Eric Price for advice on proving our inexact arithmetic results. This work was supported in part by the National Science Foundation (DMS-1304211, 1117257, 1319483, 1409286, 1441889), Adobe Systems, The Walt Disney Company, Pixar, Nvidia Corporation, and Google.

0730-0301/2017/7-ART1 $15.00
https://doi.org/http://dx.doi.org/10.1145/3072959.3073689

Algorithm 1 Gauss-Seidel-like impact operator

```
function ResolveImpacts(configuration q, velocity q̇)
    N ← ActiveConstraintGradients(q)
    q̇_0 ← q̇
    for i := 0, ∞ do
        if N^T q̇_i ≥ 0 then
            return q̇_i
        end if
        q̇_{i+1} ← R(q̇_i)
    end for
end function
```

The lack of full answers to these termination questions has led to a perception that Gauss-Seidel-like methods are unprincipled or dangerous, despite them working well in many cases [Uchida et al. 2015]. One common workaround is to abort the loop after a fixed number of iterations, and proceed by either permitting collisions to remain unresolved, or performing a gross approximation (called a *failsafe*) that introduces artificial dissipation and sticking [Provot 1997]. With a better understanding of when methods for resolving the multiple impact problem are guaranteed to converge, the need for these non-physical failsafes decreases.

Overview and Contributions. We will first focus on the subspace of Gauss-Seidel-like algorithms where R is *frictionless*, *elastic* (energy-preserving) and satisfies a minimum set of physical correctness properties (§2). Our analysis encompasses pairwise operators common in computational mechanics, as well as the recent *Generalized Reflection* operator proposed by Smith et al. [2012]. **We fully characterize finite termination of Algorithm 1 for this broad subspace of impact operators.**

We prove that the *only* possible obstruction to termination for such R is the presence of certain geometric degeneracies in the set of contact constraints (*implicit equality constraints*, §4), and we show that, surprisingly, absence of these implicit equality constraints guarantees termination after a finite number of iterations. The key ingredient of our proof will turn out to be conservation of energy, as it rules out the possibility of R applying an infinite sequence of dissipative impulses that never fully resolve the collisions (§6.1). Moreover, we describe an algorithm for detecting and directly removing implicit equality constraints from the multiple impact solver. **We arrive at a modified algorithm that provably resolves collisions in a finite number of iterations**. This result places Gauss-Seidel-like methods on sure theoretical footing, and gives practitioners the option to pay a small additional computational cost to guarantee progress and termination of their simulations, without sacrificing physical correctness at the hands of ad-hoc failsafes (§4.1).

It is often desirable in practice to incorporate dissipation into impact response. Unlike elastic impact, which can be reasoned about from first physical principles, dissipation is always phenomenological: when two objects collide, some of their kinetic energy is lost due to vibration of the objects, sound generation, internal friction and heating of the objects, etc; it is intractable to simulate these small-scale phenomena explicitly, and so dissipation is incorporated instead using coarse-scale empirical laws like kinetic Coulomb friction or coefficients of restitution. There are several ways to modify Algorithm 1 to include such dissipative laws, not all of which are guaranteed to terminate. We discuss the challenges dissipation introduces to termination, and propose a dissipation model that retains the termination guarantee (§6).

Finally, finite termination is most useful in practice if the guarantees continue to hold in the presence of inexact arithmetic and numerical noise. We revisit our analysis of finite termination and **prove that termination is guaranteed when Algorithm 1 is computed using inexact arithmetic** (§7).

2 RESOLVING MULTIPLE IMPACTS

We consider physical systems with d-dimensional configuration spaces Q. As described above, we focus on resolving contact by iteratively applying an impact operator which resolves some collisions, while potential causing others; all of this happens at one instant in time, when the physical simulation has configuration $\mathbf{q} \in Q$. Let $\dot{\mathbf{q}}_0$ denote the starting configurational velocity. Some number m of inequality constraints $g_i : Q \to \mathbb{R}$, with gradients $\mathbf{n}_i = \nabla g_i$, encode the fact that no objects in the system are interpenetrating.[1] We will denote by $\mathcal{N}$ the set of constraints, and by N the matrix with columns $\mathbf{n}_i$. Typically one differentiates between *active* constraints $g_i(\mathbf{q}) \leq 0$ (corresponding to penetrations between objects) and *inactive* constraints $g_i(\mathbf{q}) > 0$ (corresponding to objects far apart whose interactions can be ignored); however since we are concerned only with one instant in time we will ignore inactive constraints entirely and assume all constraints are active. Instead we distinguish between constraints with $\dot{\mathbf{q}} \cdot \mathbf{n}_i \geq 0$, where no response is needed since objects are already separating, and constraints with $\dot{\mathbf{q}} \cdot \mathbf{n}_i < 0$, where active intervention is needed to resolve the impact.

We will call a configurational tangent vector $\mathbf{v}$ *feasible with respect to constraint* i if $\mathbf{v} \cdot \mathbf{n}_i \geq 0$ (and infeasible with respect to that constraint or vector otherwise). We will also call a tangent vector simply *feasible* if it is feasible with respect to all constraints.

Finally, let $M_{d\times d}$ be the system mass matrix, so that $\frac{1}{2}\dot{\mathbf{q}}_0^T M \dot{\mathbf{q}}_0$ is the initial kinetic energy of the system. Since the magnitudes of the contact constraints g_i is arbitrary, we may take the constraint gradients to be normalized with respect to the energy metric, $\mathbf{n}_i^T M^{-1} \mathbf{n}_i = 1$, which simplifies some calculations; we assume this convention through the remainder of the text.

In the special case of a network of b balls[2] in the plane, $Q = \mathbb{R}^{2b}$ and $\mathbf{q}$ is the concatenation of the x, y coordinates of the centers of the b balls. For each pair of balls that are touching, a constraint g_i enforces positive signed distance between the balls. M is a $2b \times 2b$ diagonal matrix containing the masses of each ball.

Systems of b rigid bodies also fit into this formulation. Configuration space now includes both translational and rotational degrees of freedom for each body, but for any $\mathbf{q} \in Q$, we can parameterize

[1] We assume throughout the paper, for clarity of the exposition, that the constraint gradients $\mathbf{n}_i$ arise from holonomic constraints, as described in this paragraph. However since the Gauss-Seidel-like algorithms discussed in this paper run during a single, frozen instant in time, it is not essential that the $\mathbf{n}_i$ arise from true inequality constraints. Our analysis extends without difficulty to settings where the $\mathbf{n}_i$ are chosen using any other scheme (by a collision detection algorithm, for instance).

[2] Here and throughout the paper, we use "ball" to refer to a point particle of finite extent, which has a position but no orientation or angular velocity.

the space of configurational velocities at $\mathbf{q}$, consisting of the linear and angular velocity of each body, by $\mathbb{R}^{6b}$. The system mass matrix $M_{6b\times 6b}$ is then block-diagonal, with each rigid body contributing a diagonal mass matrix and a 3×3 inertia tensor to M.

If $N^T\dot{\mathbf{q}}_0 \geq 0$ nothing needs to be done as all objects in contact are already separating; typically this is not the case, though, and $\dot{\mathbf{q}}_0$ is infeasible with respect to at least one constraint. Algorithm 1 iteratively applies an *impact operator* R to modify velocity, until the velocity is feasible.

Let $R^i_{\mathcal{N}}$ be a map $R^i_{\mathcal{N}} : TQ \to TQ$ from configuration tangent space to itself, depending both on the gradients of the active constraints $\mathcal{N}$ and the current iteration i of Algorithm 1 (so that $R^i_{\mathcal{N}}$ might act differently on the same velocity during different iterations). For brevity of notation we will elide these implicit parameters and write simply R.

Absent additional assumptions, this notion of R allows all manner of non-physical behavior (for example, simply freezing all objects, $R(\dot{\mathbf{q}}) = 0$, trivially yields a feasible velocity, but conserves neither energy nor momentum). We therefore insist that R satisfy a minimum set of physical correctness criteria:

(**NORM**) **Normal impulses.** R acts only by applying impulses in directions that lie in the span of the constraint normals $\mathbf{n}_i$:

$$R(\dot{\mathbf{q}}) = \dot{\mathbf{q}} + M^{-1}N\lambda,$$

where $\lambda \in \mathbb{R}^m$. (If a ball collides against a wall, contact impulses can push the ball away from the wall, but cannot modify the ball's tangential velocity.)

(**KIN**) **Energy conservation.** Elastic impact conserves kinetic energy:

$$\frac{1}{2}\dot{\mathbf{q}}^T M\dot{\mathbf{q}} = \frac{1}{2}R(\dot{\mathbf{q}})^T MR(\dot{\mathbf{q}}).$$

(**ONE**) **One-sided impulses.** Impulses may push bodies apart but not pull them together. This condition requires $\lambda \geq 0$ in (NORM).

In addition to imposing these physical restrictions on the behavior of R, we will also require R to be well-behaved algorithmically:

(**VIO**) **Only violated constraints exert impulses.** If $\dot{\mathbf{q}}$ is feasible with respect to constraint i, then $\lambda_i = 0$.

(**MOD**) **Infeasible velocities are modified.** An infeasible velocity is not a fixed point, i.e., if $R(\dot{\mathbf{q}}) = \dot{\mathbf{q}}$, then $\dot{\mathbf{q}}$ is feasible.

Notice that (VIO) disallows a single application of R from exerting an impulse between two objects that are at rest relative to each other. However, (VIO) does not prevent iterated application of R from separating initially stationary objects [Smith et al. 2012]. Such separation is required to reproduce Newton's cradle, where a stationary line of balls is hit on one end by a moving ball.

The first three desiderata, sometimes termed kinetic and energetic consistency [Nguyen and Brogliato 2014], bear similarity to and can be viewed as a sharpening of those proposed by Smith et al. [2012] as essential to any physically-principled impact operator; for the sake of generality we have omitted properties like symmetry- or wave-effect-presevation, which are not germane to our study of termination.

The above five requirements are all stated with respect to any single application of R. Our final desideratum is concerned with iterated application of R:

(**FIN**) **Finite termination.** Impact should be fully resolved after a finite number of applications of R.

Our main result is that the first five properties are sufficient to guarantee the sixth (§5), except for an implicit equality failure case which can be fully characterized and detected (§4). We will present a modification of Algorithm 1 that detects and removes these implicit equality constraints, and prove that the entire family of Gauss-Seidel-like impact resolution algorithms, once so modified, satisfy:

(NORM) + (KIN) + (ONE) + (VIO) + (MOD) ⇒ (FIN).

Moreover each of the five properties is essential to termination, in the sense that removing any one property allows the existence of a Gauss-Seidel-like impact operator R that satisfies the others yet never terminates (§5.3).

Relation to Common Operators. The first five properties (NORM)–(MOD) are not very restrictive, and many existing Gauss-Seidel-like algorithms for solving the multi-impact problem satisfy all five.

For instance, consider the most straightforward possible impact operator: select any one constraint g_i for which the current velocity is infeasible, and correct that constraint by applying an impulse. For this single impact, (NORM) and (KIN) uniquely determine the impulse applied,

$$R(\dot{\mathbf{q}}) = \dot{\mathbf{q}} - 2(\dot{\mathbf{q}}^T\mathbf{n}_i)M^{-1}\mathbf{n}_i\ . \tag{1}$$

This formula is the familar reflection of two elastic balls off of each other. Repeatedly applying R in the case of a network of balls in the plane can thus be interpreted as treating one pair of collisions between balls as "happening first," resolving that impact, and then repeating the process, stopping when (or if) the configurational velocity is feasible.

Many possible strategies exist for choosing g_i at each iteration, from simple lexicographic order [d'Alembert 1743; MacLaurin 1742] (i.e., always select the constraint g_j with the least j with respect to which the current velocity is infeasible) to random order [Crassous et al. 2007; Ivanov 1995] to more sophisticated selection [Chatterjee and Ruina 1998; Ivanov 1995; Johnson 1976]. Regardless of the selection process, all variants of this *pairwise Gauss-Seidel* approach satisfy the five criteria above.

Gauss-Seidel-like operators can also take into accout more than one violated constraint at a time, and such global approaches can avoid the artificial symmetry-breaking present in pairwise methods, or better handle the wave-like propagation of shocks through a network of touching bodies. Two such operators are Smith et al. [2012]'s *Generalized Reflection* (GR) operator and Zhang et al. [2015]'s Quadratic Contact Energy approach, both of which satisfy the five criteria. Related operators include the PLUS model of Uchida et al. [2015]; although it is intended primarily for dissipative contact without (KIN), its approach for selecting an impulse in the presence of redundant constraints satisfy the criteria.

To our knowledge, no characterization of which of these operators satisfy (FIN) currently exists. We begin with a simple case where understanding termination is straightforward—a line of balls in 1D—as ideas from this case will guide our general analysis (§3). We then look at one situation where GS-like operators are known to *fail* to

terminate: impact in the presence of *implicit equality constraints* (§4). We will then prove that this is the only possible failure case; in the absence of implicit equality constraints, any method satisfying the five criteria (NORM)–(MOD), including GR and all flavors of pairwise GS, also have (FIN) (§5).

3 TERMINATION IN 1D

The simplest first example for analyzing termination is *Newton's cradle* in one dimension: a line of balls of unit mass and unit radius, with each interior ball touching its two neighbors. Suppose we now assign an arbitrary initial velocity $\dot{\mathbf{q}}_0$ to the balls, and attempt to solve the multiple impact problem using lexicographic Gauss-Seidel. Will the algorithm terminate?

The key insight is that in one dimension, in the case of equal masses, the pairwise reflection (1) amounts to *swapping* the velocities of the two colliding balls. Lexicographic GS is thus a bubble sort of the ball velocities: $\dot{\mathbf{q}}_0$ is infeasible with respect to the constraint coupling the first two balls if and only if the first ball has greater velocity than the second, in which case the first iteration of GS will swap the velocities of these two balls, etc. Since bubble sort terminates after finitely many iterations, so must GS.

But we can say more: index the d balls from left to right, so that the first coordinate of $\dot{\mathbf{q}} \in TQ = \mathbb{R}^d$ is the velocity of the first ball, etc. Then the energy $\dot{\mathbf{q}}^T \begin{bmatrix} 1 & 2 & \cdots & d \end{bmatrix}$ measures the sortedness of $\dot{\mathbf{q}}$: it is maximized when $\dot{\mathbf{q}}$ is sorted, and increases whenever out-of-order consecutive entries of $\dot{\mathbf{q}}$ are swapped. This energy certifies convergence and finite termination (since $\dot{\mathbf{q}}$ can be permuted in only finitely many different ways) of lexicographic GS, as well as all other flavors of GS.

We will prove termination in higher dimensions using a similar argument: we will show that there exist energies that measure how close the configurational velocity is to being feasible, and that increase each time R is applied. Unfortunately, complications arise in higher dimensions that are not evident in 1D: balls colliding in (2+)D at glancing angles do not swap velocities, so that GS can no longer be interpreted as a sorting algorithm with a finite set of possible states. In the next section we examine a situation where the higher-dimensional multiple impact problem can fail to terminate, then we will return in §5 to prove that this is the *only* failure case.

4 LINEALITY SUBSPACES

Consider the simple didactic two-dimensional example illustrated in Fig. 2, left, of a single ball wedged against two parallel vertical walls. If the velocity of the ball is also exactly vertical, then no collision occurs and the ball proceeds unobstructed. If the ball's velocity has any component in the horizontal direction, however, lexicographic Gauss-Seidel fails to terminate: resolving one ball-wall constraint reflects the ball's horizontal velocity while leaving its vertical velocity untouched. Resolving the ensuing collision against the second wall then simply returns the velocity to its starting value — this cycle repeats with no progress towards convergence.

If we replace the single ball with two smaller horizontally-aligned balls (Fig. 2, center), the same failure occurs if any of the velocities have a component in the horizontal direction. Moreover, it is not necessary to start the simulation from such contrived, effectively infeasible initial conditions in order to encounter these "fatal velocity cycles"; Fig. 2, right, shows how a collision can trigger an infinite loop at any arbitrary point in an otherwise-uneventful simulation.

Returning to the single wedged ball example, notice that the root cause of nontermination is the existence of two constraint gradients that point in opposing (configurational) directions. Reflecting velocity along one direction and then the other, not surprisingly, returns velocity to its original state.

Similarly, the two-ball example in Fig. 2, center, has a subset of constraint gradients that are also opposing, although in a more general sense that these gradients form a subset $\mathbb{B}$ of N whose combined application as linear inequality constraints $N_{\mathbb{B}}^T \dot{\mathbf{q}} \geq 0$ is equivalent to applying them as *equality constraints* $N_{\mathbb{B}}^T \dot{\mathbf{q}} = 0$. In particular, numbering constraints in Fig. 2, center, from left to right and writing the configuration as $\mathbf{q} = (x_1, y_1, x_2, y_2)^T$, we have $\mathbf{n}_1 \propto (1,0,0,0)^T$, $\mathbf{n}_2 \propto (-1,0,1,0)^T$, and $\mathbf{n}_3 \propto (0,0,-1,0)^T$. The combined enforcement of $\mathbf{n}_2^T \dot{\mathbf{q}} \geq 0$ and $\mathbf{n}_3^T \dot{\mathbf{q}} \geq 0$ then requires that $(-1,0,0,0)^T \dot{\mathbf{q}} \geq 0$, which is directly opposed to the constraint $\mathbf{n}_1^T \dot{\mathbf{q}} \geq 0$. This direct opposition then reduces (as in the simpler case) to a system that implicitly enforces linear equality constraints, in this case of the form $\mathbf{n}_1^T \dot{\mathbf{q}} = 0$ and $\mathbf{n}_3^T \dot{\mathbf{q}} = 0$.

Characterizing Implicit Equality Constraints. There are several conditions we can check to discover whether or not N contains any such implicit equality constraints. One equivalent condition is that some positive combination of constraint gradients sums to zero: suppose k constraints $M \subset N$ have this property, so that $\mathbf{0} = \sum_{i=0}^{k} \lambda_i \mathbf{m}_i$ with $\lambda_i > 0$. Then for any feasible velocity $\mathbf{v}$,

$$0 \leq \lambda_0 \mathbf{m}_0 \cdot \mathbf{v} = -\sum_{i=1}^{k} \lambda_i \mathbf{m}_i \cdot \mathbf{v} \leq 0,$$

so that the only way the inequalities are satisfied is if $\mathbf{v} \cdot \mathbf{m}_i = 0$. Conversely, if all feasible velocities $\mathbf{v}$ satisfy $\mathbf{w} \cdot \mathbf{v} = 0$, some positive linear combination of constraint gradients must sum to $\mathbf{w}$ and some other combination to $-\mathbf{w}$, and hence $\mathbf{w} + -\mathbf{w} = \mathbf{0}$ is a positive combination of constraint gradients.

Therefore N contains implicit equality constraints if and only if the kernel of N intersects the positive orthant; i.e. if there exists a $\lambda \geq \mathbf{0}$, with $\lambda \neq 0$, and $N\lambda = 0$. Notice that this condition is stronger than simple linear dependence of the constraint gradients: consider for instance the trio of constraint gradients $\mathbf{n}_1 = (1,1)^T$, $\mathbf{n}_2 = (-1,1)^T$, and $\mathbf{n}_3 = (0,1)$. These constraint gradients are linearly dependent and redundant (the third constraints imposes no additional restrictions on feasible velocities) but they do not oppose, and so do not form an implicit equality constraint.

Finally, another equivalent condition to the presence of implicit equality constraints, which will be useful in our proof of termination, is that the set of all feasible velocities lies in a proper linear subspace (of dimension $< d$) of $\mathbb{R}^d$. This condition is exactly the geometric interpretation of an equality constraint on $\mathbf{v}$.

Subspace determination. To help eliminate potential nontermination, we can detect these implicit equality constraints and ensure that they are *explicitly* enforced at every iteration of impact response,

by projecting out from all constraint gradients their component in the direction of the implicit equality constraints.

The problem of finding a maximal set of equality constraints implied by a set of inequalities, or equivalently of finding the *lineality space* – the maximal linear subspace of the normal cone formed by a positive span of vectors – is a well-studied problem in numerical optimization [Caire et al. 2008; López 2011; Telgen 1983; ten Dam 1997; Wets and Witzgall 1967].

One solution involves solving a series of linear programming problems. For instance, we can employ the *Support and Shave* algorithm [López 2011] that finds the lineality space with at most $2m$ linear programming solves of size d. In the best and by far most common case, when no implicit equalities exist, only a single linear programming solve is required.

Algorithm 2 `GS-like operator with anti-locking`

function ResolveImpacts(configuration $\mathbf{q}$, velocity $\dot{\mathbf{q}}$)
  $N \leftarrow$ ActiveConstraintGradients($\mathbf{q}$)
  $E \leftarrow$ ImplicitEqualityGradients(N)
  $\dot{\mathbf{q}}_0 \leftarrow \text{argmin}_{\dot{\mathbf{q}}_0} \|\dot{\mathbf{q}} - \dot{\mathbf{q}}_0\|_M^2 \quad \text{s.t.} \quad E^T\dot{\mathbf{q}}_0 = 0$
  for $i := 1, m$ **do**
    $\mathbf{x}_i \leftarrow \text{argmin}_{\mathbf{x}} \|\mathbf{x} - \mathbf{n}_i\|_M^2 \quad \text{s.t.} \quad E^T\mathbf{x} = 0$
    if $\|\mathbf{x}_i\|_M = 0$ **then**
      remove $\mathbf{n}_i$ from N
    else
      $\mathbf{n}_i \leftarrow \mathbf{x}_i / \|\mathbf{x}_i\|_M$
    end if
  end for
  for $i := 0, \infty$ **do**
    if $N^T\dot{\mathbf{q}}_i \geq 0$ **then**
      return $\dot{\mathbf{q}}_i$
    end if
    $\dot{\mathbf{q}}_{i+1} \leftarrow R(\dot{\mathbf{q}}_i)$
  end for
end function

4.1 Handling implicit equality constraints

When a lineality subspace is detected in the constraint set, a minor modification of Algorithm 1 can guarantee that each iteration of the impact operator respects the induced equality constraints: Q should be replaced by the subspace $\bar{Q}$ of the configuration space that respects the linear equality constraints, and each inequality constraint g_i should be restricted to $\bar{Q}$. This modification amounts to projecting out from $\dot{\mathbf{q}}_0$ and all constraint gradients $\mathbf{n}_i$ their components in the direction of the implicit equality constraints. Algorithm 2 outlines this modification. The function `ImplicitEqualityGradients` analyzes N and extracts a basis for the implicit equality constraints in N. Notice that if N is free of implicit equality constraints, the algorithm reduces to the unmodifed Algorithm 1.

Termination with lineality. Without implicit equality constraint detection and handling, (FIN) is not guaranteed. Systems containing degrees of freedom confined to tight spaces are more likely to experience these termination difficulties. In Fig. 1 we simulate one such example: a ball fired along a shaft collides with a plug of stationary balls. On impact, both Gauss-Seidel and GR *without lineality detection* fail to terminate. Both succeed with lineality detection.

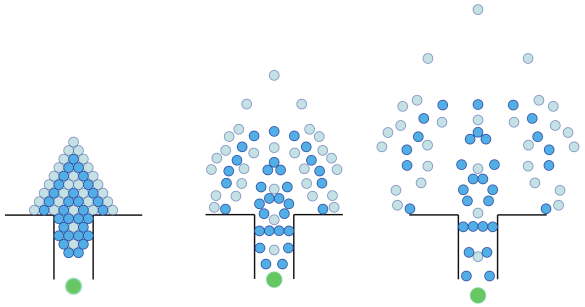

Fig. 1. **Ball Geyser**: A ball fired along a shaft should dislodge the plug made of stationary balls (left). Without Lineality detection, however, Gauss-Seidel and GR algorithms are not able terminate for the initial impact resolution solve. GR with Lineality detection resolves the multi-impact problem and generates a trajectory that retains all physical, desiderata (right).

5 TERMINATION IN ND

We now turn to the problem of termination in arbitrary dimension. We know that the presence of implicit equality constraints can preclude termination. In the remainder of this section, we will show that this is the *only* possible failure: that any impact operator satisfying (NORM)–(MOD), and free of implicit equality constraints, will terminate in finitely many iterations given any set of constraints and initial velocities in any dimension. We will do so in two steps: first we will show *convergence*: that in the limit of applying an impact operator infinitely many times, the velocity approaches some (possibly infeasible) limit velocity (§5.1). We will prove that the number of iterations needed is in fact finite, and that the final velocity is feasible (§5.2).

5.1 Convergence proof

First, let us characterize the geometry of the constraints in configuration tangent space. For any configurational velocity $\mathbf{v}$ and any constraint g_i, one of three things must be true:

- the velocity is *strictly feasible* with respect to that constraint: $\mathbf{v}^T\mathbf{n}_i > 0$. The velocity doesn't violate the constraint, and won't start violating the constraint even under infinitesimal perturbations;
- the velocity is *tangent* to the constraint: $\mathbf{v}^T\mathbf{n}_i = 0$;

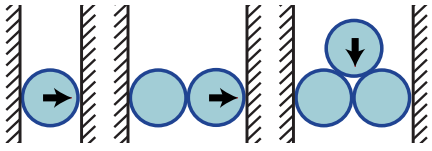

Fig. 2. **Failure of pairwise impulse methods to terminate**: Gauss-Seidel and GR fail to terminate for a single ball *(left)* or two horizontally-aligned balls *(center)* wedged between vertical walls, if the balls' velocities have any component in the horizontal direction. Even if the system's initial conditions are feasible *(right)*, an impact later in the simulation can cause it to enter a nonterminating state.

- the velocity *violates* (is infeasible with respect to) the constraint: $\mathbf{v}^T\mathbf{n}_i < 0$.

We can partition configuration tangent space TQ into up to 3^m sets that differ as to whether $\mathbf{v}$ is strictly feasible, tangent, or infeasible with respect to each constraint. Two other important regions of tangent space are the *feasible set* F containing all feasible tangent vectors and the *strictly feasible set* F_I (containing all tangent vectors $\mathbf{v}$ whose inner product $\langle \mathbf{v}, \mathbf{n}_i \rangle$ is strictly positive for all constraints). The strictly feasible set is the intersection of half-spaces passing through the origin and so is a convex cone. Furthermore F_I is empty if and only if N contains implicit equality constraints. If there are implicit equality constraints, F is a subset of the linear space of dimension less than d that satisfies the implicit equality constraint, so has no interior F_I. Conversely, if F_I is empty, then since F is convex it is contained in a hyperplane of dimension $d-1$. The fact that all feasible velocities lie in this hyperplane is an equality constraint that the feasible velocities satisfy.

As in Algorithm 1, let $\dot{\mathbf{q}}_i$ be the velocity after i application of the impact operator R on $\dot{\mathbf{q}}_0$. We will now prove a higher-dimensional analogue of the bubble-sort result from 1D:

Lemma 5.1. *Let $\mathbf{w}$ be any feasible velocity ($\mathbf{w} \in F$). Then for an impact operator satisfying (NORM), (KIN), (ONE), and (VIO), and any initial velocity $\dot{\mathbf{q}}_0$, $\langle \mathbf{w}, \dot{\mathbf{q}}_i \rangle_M$ converges to a real number $k_{\mathbf{w}}$ as $i \to \infty$.*

Since energy is conserved, the velocity $\dot{\mathbf{q}}_i$ after every iteration can be interpreted as a point on the ellipsoid $\{\mathbf{v} \in TQ \mid \langle \mathbf{v}, \mathbf{v} \rangle_M = \langle \dot{\mathbf{q}}_0, \dot{\mathbf{q}}_0 \rangle_M\}$. Geometrically, the lemma then states that for any feasible velocity $\mathbf{w}$, the velocity $\dot{\mathbf{q}}_i$ after repeated applications of the iteration map, interpreted as a point on the ellipsoid of constant energy, approaches some hyperplane perpendicular to the vector $\mathbf{w}$ (see Fig. 4), with distance to this hyperplane decreasing monotonicaly (see Fig. 3).

Unfortunately, this fact alone is not enough for convergence, since the hyperplane will generally intersect the ellipsoid in an ellipsoid of one lower dimension, and it is conceivable the velocity $\dot{\mathbf{q}}_i$ might bounce infinitely around this intersection ellipsoid without ever settling at a single point; we will address this detail later in the section.

Note that when $d = 1$, we recover the picture of a Newton's cradle with n balls in a row. In that case, $\mathbf{w}$ is an increasing sequence of numbers (for instance, $(1, 2, \ldots, n)$), and the amount that $k_{\mathbf{w}}$ increases during each reflection is bounded away from zero, so that $\dot{\mathbf{q}}_i$ cannot bounce around infinitely on a lower-dimensional ellipsoid.

Proof. Suppose that $\dot{\mathbf{q}}_m$ is feasible for some integer m. Then by (VIO), $R(\dot{\mathbf{q}}_m) = \dot{\mathbf{q}}_m$ and $\dot{\mathbf{q}}_i = \dot{\mathbf{q}}_m$ for all $i \geq m$, so $k_{\mathbf{w}} = \langle \mathbf{w}, \dot{\mathbf{q}}_m \rangle_M$. The other case is where $\dot{\mathbf{q}}_i$ is infeasible for all i. By (KIN),

$$\langle \mathbf{w}, \dot{\mathbf{q}}_i \rangle_M \leq \|\mathbf{w}\|_M \|\dot{\mathbf{q}}_i\|_M = \|\mathbf{w}\|_M \|\dot{\mathbf{q}}_0\|_M$$

and so is bounded above. It therefore suffices to show that $\langle \mathbf{w}, \dot{\mathbf{q}}_i \rangle_M$ increases monotonically at every iteration. And indeed, by (NORM)

$$\begin{aligned}\langle \mathbf{w}, \dot{\mathbf{q}}_{i+1} \rangle_M &= \langle \mathbf{w}, \dot{\mathbf{q}}_i + M^{-1}N\lambda \rangle_M \\ &= \langle \mathbf{w}, \dot{\mathbf{q}}_i \rangle_M + \mathbf{w}^T N\lambda \\ &= \langle \mathbf{w}, \dot{\mathbf{q}}_i \rangle_M + \sum \lambda_i \mathbf{w}^T \mathbf{n}_i,\end{aligned}$$

and since $\mathbf{w} \in F$, $\mathbf{w}^T\mathbf{n}_i \geq 0$. By (ONE), $\lambda_i \geq 0$ as well, so $\sum \lambda_i \mathbf{w}^T \mathbf{n}_i \geq 0$, completing the proof. □

The above lemma guarantees that $\dot{\mathbf{q}}_i$ approaches some hyperplane in tangent space, but as remarked above this is not enough for convergence (and note that the above did *not* assume the lack of implicit equality constraints). The key idea is that we can now pick a *different* $\mathbf{w}_1$ and repeat the above argument: the velocity must now approach two different hyperplanes, and so must approach their intersection. If there are no implicit equality constraints, the strictly feasible region F_I is nonempty, and we can pick a cluster of feasible vectors $\mathbf{w}_j$ from F_I so that their corresponding hyperplanes intersect at a single point. The velocity must then converge to that point. The following proof formalizes this argument.

Lemma 5.2. *Suppose F_I is nonempty. Then it contains d linearly independent vectors $\mathbf{w}_1, \ldots, \mathbf{w}_d$.*

Proof. From the definition of F_I it is clear that it is an open subset of tangent space. Since F_I is nonempty, it contains a ball with center $\mathbf{c}$ and radius r. The vectors $\mathbf{c} + \min(r/2, \|\mathbf{c}\|_1)\mathbf{e}_i$ (where $\mathbf{e}_i$ are the Euclidean coordinate basis functions) are linearly independent, and contained within this ball, and so within F_I. □

Theorem 5.3. *If N is free of implicit equality constraints and R is an impact operator satisfying (NORM), (KIN), and (ONE), then $\lim_{i\to\infty} \dot{\mathbf{q}}_i = \dot{\mathbf{q}}_\infty$ exists.*

Proof. By the above lemmas, we can find d linearly independent vectors $\mathbf{w}_j$ in F_I with $\langle \mathbf{w}_j, \dot{\mathbf{q}}_i \rangle_M$ converging to $k_{\mathbf{w}_j}$. There is therefore a unique solution $\mathbf{r}$ to the linear system

$$\langle \mathbf{w}_j, \mathbf{r} \rangle_M = k_{\mathbf{w}_j} \quad j = 1, \ldots, d,$$

and we will show that $\dot{\mathbf{q}}_i$ converges to $\mathbf{r}$. Let W be the matrix whose rows are $\mathbf{w}_j$, so that

$$WM\mathbf{r} = \mathbf{k}$$

for $\mathbf{k} = \left(k_{\mathbf{w}_1}, k_{\mathbf{w}_2}, \ldots, k_{\mathbf{w}_d}\right)^T$. Since W and M are nonsingular, so is WM, and so WM has a singular value with least magnitude $\sigma > 0$.

We know that $\langle \mathbf{w}_j, \dot{\mathbf{q}}_i \rangle_M$ converges to $k_{\mathbf{w}_j}$, so for any $\epsilon > 0$ there exist integers c_j with

$$\left|\langle \mathbf{w}_j, \dot{\mathbf{q}}_i \rangle_M - k_{\mathbf{w}_j}\right| \leq \frac{\sigma\epsilon}{d}$$

for $i \geq c_j$. Setting $C = \max_j c_j$, for $i \geq C$ we then have that

$$\begin{aligned}\|\dot{\mathbf{q}}_i - \mathbf{r}\| &= \|\dot{\mathbf{q}}_i - (WM)^{-1}\mathbf{k}\| \\ &\leq \sigma^{-1}\|WM\dot{\mathbf{q}}_i - \mathbf{k}\| \\ &\leq \sigma^{-1}\sum_{j=1}^{d} |\langle \mathbf{w}_j, \dot{\mathbf{q}}_i \rangle_M - k_{\mathbf{w}_j}| \\ &\leq \epsilon,\end{aligned}$$

and so $\dot{\mathbf{q}}_i$ converges to $\mathbf{r}$. □

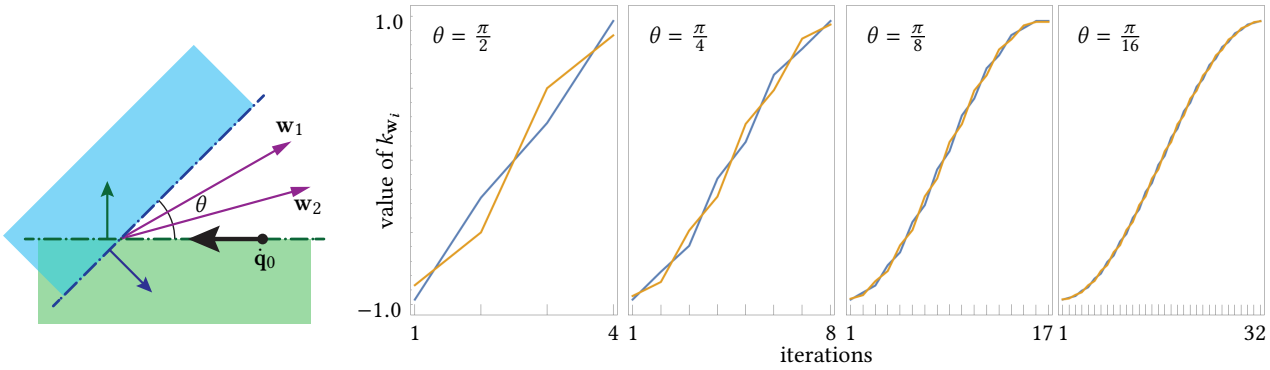

Fig. 3. A ball hits a wedge in the plane made of two constraints with normals $(0, 1)$ and $(\sin\theta, -\cos\theta)$ and the impact is resolved using lexicographic Gauss-Seidel. The smaller θ becomes, the closer the two constraints are to being an implicit equality constraint, and the more iterations are required for termination. For a given value of θ, two feasible velocities are $\mathbf{w}_i = [\cos(i\theta/3), \sin(i\theta/3)]$ for $i \in \{1, 2\}$, and these can be used as certificates of convergence in Lemma 5.1. With each application of the impact operator, their inner product $k_{\mathbf{w}_i}$ with the ball's velocity is guaranteed to increase; once this inner product is sufficiently close to 1.0, the ball is heading in the direction of the opening of the wedge and impact response terminates. The plots show $k_{\mathbf{w}_1}$ (blue) and $k_{\mathbf{w}_2}$ (yellow) as functions of the number of iterations. In each case the algorithm terminates at the last iteration shown.

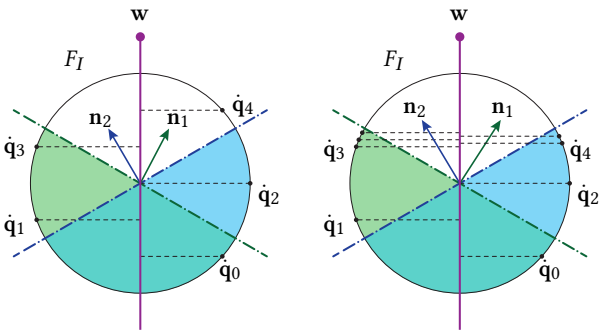

Fig. 4. An iteration map applied in two-dimensional tangent space with two constraints. The initial velocity $\dot{\mathbf{q}}_0$ violates both constraints. The chosen feasible velocity $\mathbf{w}$ can be interpreted as a vector from the origin; under this interpretation, with each application of the iteration map the velocity $\dot{\mathbf{q}}_i$ approaches a hyperplane (line) perpendicular to $\mathbf{w}$. If the velocity ever enters the feasible region, the impact operator terminates (left) but termination, and even convergence, is not guaranteed from looking at $\mathbf{w}$ alone: the velocity could plausibly oscillate near the hyperplane without ever converging to a single point (right).

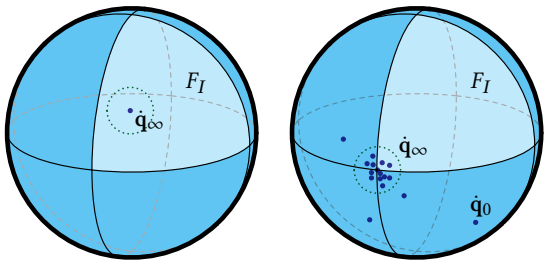

Fig. 5. Left: If the limit velocity $\dot{\mathbf{q}}_\infty$ is strictly feasible, then convergence is guaranteed since the velocity $\dot{\mathbf{q}}_i$ must enter into a ball around $\dot{\mathbf{q}}_\infty$ in finite interations. Right: the tricky case is where $\dot{\mathbf{q}}_\infty$ is feasible but tangent to some of the constraints. Could the velocity remain infeasible for infinitely many iterations? We show this is not possible, given (KIN).

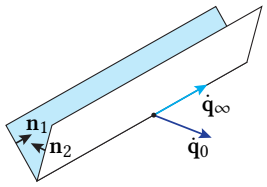

Fig. 6. Two constraints visualized as a trough in three dimensions. If the initial velocity violates both constraints, it is impossible for repeated applications of the impact operator to converge to a final velocity tangent to both constraints ($\dot{\mathbf{q}}_\infty$) by applying impulses only along the $\mathbf{n}_1$ and $\mathbf{n}_2$ directions without dissipating energy.

5.2 Termination proof

We will now show a stronger result: not only does the impact operator converge, it does so in only finitely many iterations, and always to a feasible velocity. The main idea is to look at the limit velocity $\dot{\mathbf{q}}_\infty$, which by the previous theorem must exist. If this velocity is strictly feasible, $\dot{\mathbf{q}}_i$ must enter the feasible region in finitely many iterations since F_I is open. The problematic cases are when $\dot{\mathbf{q}}_\infty$ violates some constraints, or is tangent to constraints: could the velocity hover for infinitely many iterations right outside the feasible region (see figure 5)? We now show that (KIN) disallows this latter possibility: the impact operator cannot take a velocity that violates a constraint to one that is tangent to that constraint (see figure 5, right, and figure 6). Feasibility of the final velocity will then follows from (MOD).

Theorem 5.4. *If N is free of implicit equality constraints and R satisfies the five conditions (NORM)–(MOD), then $\dot{\mathbf{q}}_i$ is feasible after a finite number of applications of the impact operator, i.e., (FIN) is satisfied.*

Proof. We know that $\dot{\mathbf{q}}_\infty$ exists, but is possibly infeasible. We can thus partition the constraints into two sets: a "violated or tangent

set" T, and an "ultimately strictly feasible" set $U = \mathcal{N} \setminus T$,

$$T = \{\mathbf{n} \in \mathcal{N} \mid \dot{\mathbf{q}}_\infty^T \mathbf{n} \leq 0\}$$
$$U = \{\mathbf{n} \in \mathcal{N} \mid \dot{\mathbf{q}}_\infty^T \mathbf{n} > 0\}.$$

We can then define the set F_U of velocities that are strictly feasible with respect to all constraints in U:

$$F_U = \{\mathbf{v} \in \mathbb{R}^d \mid \mathbf{v}^T \mathbf{n} > 0, \mathbf{n} \in U\}.$$

Notice that if U contains all of the constraints, then $F_U = F_I$, and if U is empty, F_U contains the entire tangent space TQ. Moreover by construction of U, $\dot{\mathbf{q}}_\infty \in F_U$. F_U is also clearly open, so there exists some $\epsilon > 0$ with $\mathbf{w} \in F_U$ for all $\mathbf{w}$ with $\|\mathbf{w} - \dot{\mathbf{q}}_\infty\| < \epsilon$. In particular, since $\dot{\mathbf{q}}_i$ converges to $\dot{\mathbf{q}}_\infty$, $\dot{\mathbf{q}}_i$ is in F_U for all $i \geq k$ for some integer k.

Now consider the affine subspace

$$A = \left\{\dot{\mathbf{q}}_k + M^{-1} T\lambda \mid \lambda \in \mathbb{R}^{|T|}\right\}.$$

By (VIO), $\dot{\mathbf{q}}_i \in A$ for all $i \geq k$, and so is $\dot{\mathbf{q}}_\infty$ since A is closed. Therefore

$$\dot{\mathbf{q}}_\infty = \dot{\mathbf{q}}_k + M^{-1} T\lambda$$

for some λ, and by (KIN),

$$\begin{aligned}\|\dot{\mathbf{q}}_\infty\|_M^2 &= \|\dot{\mathbf{q}}_k\|_M^2 \\ &= \|\dot{\mathbf{q}}_\infty - M^{-1}T\lambda\|_M^2 \\ &= \|\dot{\mathbf{q}}_\infty\|_M^2 - 2\langle \dot{\mathbf{q}}_\infty, M^{-1}T\lambda\rangle_M + \|M^{-1}T\lambda\|_M^2.\end{aligned}$$

It follows that

$$\begin{aligned}0 &= -2\langle \dot{\mathbf{q}}_\infty, M^{-1}T\lambda\rangle_M + \|M^{-1}T\lambda\|_M^2 \\ &= \sum_{j=1}^{|T|} -2\lambda_j \dot{\mathbf{q}}_\infty^T \mathbf{n}_j + \|M^{-1}T\lambda\|_M^2,\end{aligned}$$

and since $-\dot{\mathbf{q}}_\infty^T \mathbf{n}_j \geq 0$ and $\lambda_j > 0$ (thanks to (ONE)), both terms are nonnegative and so must each be zero. Therefore $M^{-1}T\lambda = 0$ and $\dot{\mathbf{q}}_k = \dot{\mathbf{q}}_\infty$. Finally, since then $R(\dot{\mathbf{q}}_k) = \dot{\mathbf{q}}_k$, $\dot{\mathbf{q}}_k$ must be feasible by (MOD). □

5.3 Necessity of the Desiderata

It is instructive to trace where the above chain of reasoning breaks down in the presence of implicit equality constraints: Lemma 5.1 still holds, but it is no longer possible to find d different hyperplanes in general position; instead, the implicit equality constraints force all hyperplanes to intersect in a common line, plane, or other higher-dimensional linear space, depending on the degree of degeneracy of the constraint gradients, and the velocity might wander around this linear space instead of converging to a point. Note also that each of the five properties (NORM)–(MOD) is essential to termination, in the sense that removing any one of the desiderata while keeping the others allows for the existence of a impact operator R for which Algorithm 1 does not terminate. For each of the properties, we now show that it is essential, and highlight the portions of the proof that rely on it.

- Without (NORM), two other desiderata ((ONE) and (VIO)) are ill-posed, and both of these are required for guaranteed termination (see below); this dependence also means (NORM) is essential for both convergence and termination proofs.
- If (ONE) is removed and sticking is allowed, the velocity of the system can alternate between two infeasible values even without the presence of implicit equality constraints. Consider the case of a ball flying into a corner, shown in Fig. 7 with normals $\mathbf{n}_1 = (0, 1)$ and $\mathbf{n}_2 = (-1, 0)$. If the ball hits the corner with initial velocity $(2, -1)$, taking $\lambda_1 = \mp 1$ and $\lambda_2 = \pm 1$ each time R is applied will cause the velocity to alternate between $(2, -1)$ and $(1, -2)$, both of which have the same kinetic energy and both of which violate the constraints.

  The convergence proof must fail for this example, and indeed without (ONE) our Lemma 5.1 no longer holds, since an application of R might decrease an energy $k_{\mathbf{w}}$.

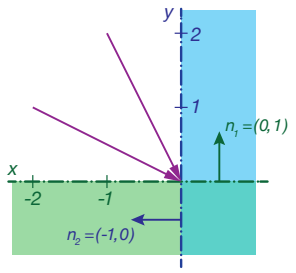

Fig. 7. (ONE) is necessary. A colliding ball flying into a corner without (ONE) can cycle between infeasible velocities as demonstrated with the purple colliding velocities here.

- To see (VIO) is essential, consider a three-ball Newton's cradle, with initial ball velocities 0, 1, 0 from left to right. Only the constraint between the right two balls is violated, but if the left constraint is also allowed to apply a (pushing) impulse, it is possible to infinitesimally perturb the velocities of the three balls (so that the left velocity is now slightly negative, and the right ball's, slighly positive) while keeping the configurational velocity infeasible. This process can be repeated, with the size of the perturbation shrinking during every iteration, so that Algorithm 1 converges, in the limit of infinitely many iterations, to a set of velocities where the middle ball is still moving faster than its rightmost neighbor.

  In terms of our proof, (VIO) is needed in Theorem 5.4 to rule out convergence to an infeasible velocity; without (VIO) there is no guarantee that after sufficiently many iterations of applying R, all intermediate velocities lie in the affine subspace A.
- Obviously, if R violates (MOD) and never modifies some infeasible initial velocity, (FIN) is impossible. In Theorem 5.4 (MOD) was required at the very end to argue that any limit velocity must be feasible.
- Finally (KIN) is an especially interesting case. It is known that if R is allowed to dissipate energy during every application, failure cases exist where Algorithm 1 does not terminate. This phenomenon is known as *inelastic collapse*, and in these cases the velocity does converge, but only in the limit of infinitely many applications of R.

In the proofs, (KIN) is required for convergence, but the proof still holds even if (KIN) is replaced by a weaker energy *non-increase* condition. On the other hand, (KIN) is an essential ingredient in proving (FIN), since it allowed us to argue that once the set of violated constraints stops changing, the velocity cannot change.

We summarize inelastic collapse in the next section, and discuss in more depth how our analysis of termination extends to the inelastic setting.

6 DISSIPATION

We have so far explored termination of Gauss-Seidel-like impact operators in the conservative setting, but for modeling practical physical systems, dissipation is often critical. Many modifications to the Gauss-Seidel-like framework of Algorithm 1 are possible for incorporating dissipation, but it is important to note that *there is no single physically-correct approach* to doing so. Dissipation models are merely phenomenological heuristics that try to capture, at a coarse scale, a wide variety of small-scale effects, such as internal heating of the colliding objects, transient formation and breaking of chemical bonds, etc. (many of them poorly-understood) that in aggregate yield dissipative impact [Brogliato 1999].

One popular model of dissipation during impact is the *coefficient of restitution*

$$c_r = -\frac{\text{relative velocity after}}{\text{relative velocity before}},$$

with $c_r = 1$ corresponding to the perfectly elastic case, and $c_r = 0$ to perfectly inelastic impact (such as when a bean-bag is dropped on the ground). A possible strategy for modeling c_r in the iterative setting is to require each application of R to dissipate energy (so that (KIN) no longer holds). This approach is tempting given its simplicity, but suffers from several drawbacks: first, the amount of energy dissipated during each impact event depends on the number of iterations taken in Algorithm 1, so that the effective c_r of the overall algorithm will depend not only on the geometry of the impact, but also the specific choice of impact operator R. More significantly, this modified algorithm is not guaranteed to terminate, due to the *inelastic collapse* phenomenon.

6.1 Inelastic collapse

Pairwise iterative methods are well- known to suffer from poor convergence whenever $c_r < 1$; in some cases, they do not converge in a finite number of iterations at all [Baraff 1989; McNamara and Young 1994]. A simple example is resolving a five-ball Newton's Cradle using lexicographic GS with $c_r = 0$: each iteration halves the relative velocity of one pair of balls, inducing a Zeno's Paradox ("Achilles and the tortoise") where the velocity remains infeasible for any finite number of iterations of GS. This is illustrated in Fig. 8.

Inelastic collapse can occur even for positive $c_r < 1$. Indeed, the range of c_r for which inelastic collapse can occur increases as the size of the impact problem grows, and quickly approaches unity as the number of colliding bodies becomes sufficiently large [Bernu and Mazighi 1990; McNamara and Young 1994]. Thus inelastic collapse is effectively unavoidable for any large-scale colliding systems whenever $c_r < 1$. As a practical matter, numerical round-off error

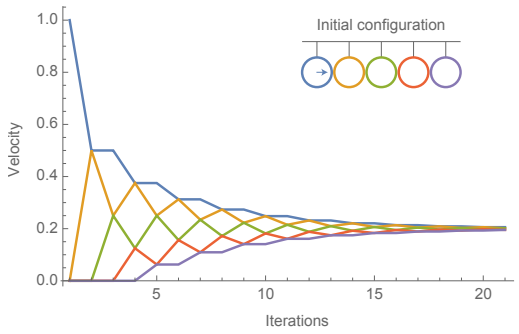

Fig. 8. **Inelastic collapse**: The velocity of each of 5 balls in a 1D Newton's cradle after successive iterations of Gauss-Seidel with $c_r = 0$. Initially all the balls except the leftmost are at rest, while the leftmost ball is given an initial velocity of 1. To reach a feasible solution, the order of the curves would have to be reversed such that the rightmost ball would end up with the highest velocity.

somewhat ameliorates this issue and the iterative process generally terminates [Chatterjee and Ruina 1998], though relying on round-off is not very satisfying, and convergence behavior consistently worsens in proportion to the decrease in c_r.

6.2 Energetic Restitution Revisited

Alternative methods are possible for incorporating c_r into Algorithm 1 without building c_r into R itself, thereby avoiding inelastic collapse. For example, when $c_r = 0$ it is common to replace the entire algorithm by a one-step formulation based on solving a linear complementary problem (LCP) [Anitescu and Potra 1997; Moreau 1985; Stewart 2000]. Dissipative restitution models with $0 < c_r < 1$ remains an open and active area of research [Chatterjee and Ruina 1998; Glocker 2004; Liu et al. 2008; Stewart 2011; Stoianovici and Hurmuzlu 1996]; naive attempts to incorporate $c_r > 0$ into LCP formulations suffer from sticking artifacts and failure to converge to a feasible solution [Smith et al. 2012].

Smith et al. [2012] propose another simple restitution model that avoids inelastic collapse across all c_r values. Here we quickly review the model and show that it inherits (FIN) from the elastic case. First, observe that purely inelastic ($c_r = 0$) multi-impact is well-posed and solvable using the standard inelastic LCP formulation mentioned above, yielding an inelastic, feasible post-impact velocity $\dot{\mathbf{q}}_0$. Similarly, elastic multi-impact is unaffected by collapse: applying Algorithm 1 using any impact operator satisfying (NORM)–(MOD), we obtain $\dot{\mathbf{q}}_1$ in finite iterations. Then c_r can be viewed as the *interpolant* between the two to obtain

$$\dot{\mathbf{q}} = (1 - c_r)\dot{\mathbf{q}}_0 + c_r\dot{\mathbf{q}}_1.$$

Notice that this definition of c_r now allows any amount of dissipation between the maximum (physically) allowable dissipation at $c_r = 0$ and total conservation of energy at $c_r = 1$ while retaining all properties, including (FIN), for all $c_r \in [0, 1]$.

6.3 Friction

While friction plays an obvious key role in persistent contact, it can also be critical in transient collisions [Brogliato 1999]. A standard

way of incorporating friction into iterative contact response algorithms like Algorithm 1 is to add frictional impulses f_k at contacts k, where these impulses are chosen to satisfy a *maximal dissipation* principle [Drumwright and Shell 2011; Goyal et al. 1991] that maximally resists tangential sliding up to a bound given by Coulomb-type constraints, e.g. $\|f_k\| \leq \mu\lambda_k$. In cases of complex contact geometry these frictional impulses might themselves cause additional collisions with their own frictional impulses, etc, so that in principle the frictional impulses must be solved for simultaneously with the contact impulses [Kaufman et al. 2005; Uchida et al. 2015]. However in practice it is often sufficient to assume that frictional impulses depend on the contact impulses, and not vice-versa. This approximation is implemented by solving for and applying friction after computing the frictionless post-impact response. Smith et al. [2012] presented one such friction model whose solution is guaranteed to terminate. Composing this operation with Algorithm 1 (using any impact operator satisfying (NORM)–(MOD)) thus allows simulating friction while retaining (FIN).

7 INEXACT ARITHMETIC

All of the above analysis has assumed that arithmetic is exact; but in a *practical* implementation of Algorithm 1 the question remains whether floating-point errors could lead to nontermination? To begin to study this question, we need to revisit the impact operator desiderata, and rebuild them from the ground up with inexact arithmetic in mind. We will then show that the pairwise Gauss-Seidel methods and Generalized Reflections, with minor modifications, all satisfy all of these desiderata.

The key idea is that inexact arithmetic requires some amount of tolerance when evaluating constraint violation. For example, suppose a tangent vector $\mathbf{v}$ satisfies $\mathbf{n} \cdot \mathbf{v} = -\mu$ for a tiny μ (on the order of the machine epsilon). Although $\mathbf{v}$ is technically infeasible with respect to $\mathbf{n}$, it could well be that no impulse satisfying (NORM) and (KIN) can possibly modify $\mathbf{v}$, stymying termination. We can avoid this situation by declaring a velocity as tangent to a constraint if their inner product is *approximately* zero; more specifically, we relax the definition of feasibility[3] of $\mathbf{v}$ with respect to $\mathbf{n}$ to the condition

$$\langle \mathbf{v}, \mathbf{n} \rangle \geq -\epsilon \|\mathbf{v}\|_M$$

for a chosen unitless tolerance parameter $\epsilon \ll 1$.

We now revisit the desiderata, beginning with (NORM): in the inexact setting, the impulse applied by R will still lie in the span of the constraint gradients, but now we allow some error in both the magnitude and direction of this impulse. For the tolerance threshold $\epsilon < 1$, we approximate (NORM) by:

(ϵ**NORM**) **Normal impulses.**

$$R(\dot{\mathbf{q}}) = \dot{\mathbf{q}} + M^{-1}N\lambda + \mathbf{c},$$
$$\text{with } \|\mathbf{c}\|_M \leq \epsilon\|\lambda\|_1 \text{ and } \|\mathbf{c}\|_M \leq \tfrac{\epsilon}{2}\|\dot{\mathbf{q}}\|_M,$$

which allows some error $\mathbf{c}$ in each application of the impact operator. The first inequality ensures that the amount of noise added to the velocity at each iteration does not overshadow the impulse being applied (otherwise it is not possible to make meaningful progress). Since the allowed error in the constraint satisfaction is scale-invariant, the allowed error in (NORM) must also be scale-invariant, hence the need for $\mathbf{c}$ to be small with respect to $\mathbf{q}$.

Kinetic energy must also be conserved, up to this error:

(ϵ**KIN**) **Energy conservation.** Elastic impact conserves kinetic energy approximately: for the λ corresponding to $R(\dot{\mathbf{q}})$ in (ϵNORM),

$$\frac{1}{2}\|\dot{\mathbf{q}}\|_M = \frac{1}{2}\|\dot{\mathbf{q}} + M^{-1}N\lambda\|_M.$$

(Contrast to exact kinetic energy conservation, which would require a $\mathbf{c}$ on the right-hand side.) We round out the list of requirements with

(ϵ**DRIFT**) **No unbounded energy drift.** In addition to approximate energy conservation, the approximate impact operator R does not cause energy to drift arbitrarily far over multiple iterations. For all $\dot{\mathbf{q}}$ there exists a $C \in \mathbb{R}$ so that for all $k \in \mathbb{Z} > 0$,

$$\left|\frac{1}{2}\|\dot{\mathbf{q}}\|_M^2 - \frac{1}{2}\|R^k(\dot{\mathbf{q}})\|_M^2\right| < C.$$

(ϵ**VIO**) **Only significantly violated constraints** exert any impulses. If $\mathbf{n}_i^T\dot{\mathbf{q}} \geq -\epsilon\|\dot{\mathbf{q}}\|_M$, then $\lambda_i = 0$.

(ϵ**MOD**) **Infeasible velocities are always modifed.** $R(\dot{\mathbf{q}}) = \dot{\mathbf{q}}$ only if $\mathbf{n}_i^T\dot{\mathbf{q}} \geq -\epsilon\|\dot{\mathbf{q}}\|_M$ for all constraints i.

(ϵ**FIN**) **Finite termination.** After a finite number of applications of R, $\mathbf{n}_i^T\dot{\mathbf{q}} \geq -\epsilon\|\dot{\mathbf{q}}\|_M$ for all constraints i.

Algorithm 3 `Inexact GS-like impact operator`

```
1:  function ResolveImpactsApprox(q, q̇, ε)
2:      N ← ActiveConstraintGradients(q)
3:      q̇_0 ← q̇
4:      for i := 0, ∞ do
5:          if N^T q̇_i + ε‖q̇_i‖_M 1 ≥ 0 then
6:              return q̇_i
7:          end if
8:          q̇_{i+1}^tentative ← R(q̇_i)
9:          q̇_{i+1} ← (‖q̇_0‖_M / ‖q̇_{i+1}^tentative‖_M) q̇_{i+1}^tentative
10:     end for
11: end function
```

With some care, pairwise Gauss-Seidel can be modified so that it satisfies these six properties when its computations are performed with floating-point arithmetic; it can be shown (see supplemental material) that Algortihm 3 obeys (ϵNORM)–(ϵMOD), and Generalized Reflections [Smith et al. 2012] can be so modified as well. Notice that the key difference in Algorithm 3, other than the addition of a tolerance when checking for convergence (or choosing a violated constraint, within R) is the renormalization of $\dot{\mathbf{q}}$ at every iteration, in order to ensure (ϵDRIFT).

[3] This relaxation might raise concerns about *constraint drift*. In the exact case, advancing the configuration along a finite-size step in the direction of the post-response velocity $\dot{\mathbf{q}}_\infty$ does not necessarily stay within the feasible region of configuration space; this is because $\dot{\mathbf{q}}_\infty$ was computed using only first-order information about the constraints, and the constraint manifold can *curve into* $\dot{\mathbf{q}}_\infty$. In the inexact setting, the drift can be linear, instead of second-order, thanks to the tolerance; however this first-order drift is on the order of ϵ, and so for ϵ not too large relative to the time step size, does not introduce significant additional error.

In the remainder of this section we will show the following generalized termination result: in the absence of implicit equality constraints and for every $\epsilon > 0$

$$(\epsilon\text{NORM}) + (\text{ONE}) + (\epsilon\text{KIN}) + (\epsilon\text{DRIFT}) + (\epsilon\text{VIO}) + (\epsilon\text{MOD}) \Rightarrow (\epsilon\text{FIN}).$$

We now reproduce the lemmas and theorems of §5 and §5.2 in the inexact case, starting with convergence.

Lemma 7.1. *Let* $\mathbf{w}$ *be any velocity with* $\mathbf{w}^T\mathbf{n}_i \geq \epsilon\|\mathbf{w}\|_M$ *for every constraint gradient* $\mathbf{n}_i$*. For an impact operator satisfying* (ϵ*NORM*)*,* (ϵ*KIN*)*, and* (*ONE*)*, and any initial velocity* $\dot{\mathbf{q}}_0$*,* $\langle \mathbf{w}, \dot{\mathbf{q}}_i\rangle_M$ *converges to a real number* $k_{\mathbf{w}}$ *as* $i \to \infty$*.*

Proof. By (ϵDRIFT),

$$\|\dot{\mathbf{q}}_i\|_M < \sqrt{\|\dot{\mathbf{q}}_0\|_M^2 + 2C},$$

so

$$\langle \mathbf{w}, \dot{\mathbf{q}}_i\rangle_M \leq \|\mathbf{w}\|_M\|\dot{\mathbf{q}}_i\|_M \leq \|\mathbf{w}\|_M\sqrt{\|\dot{\mathbf{q}}_0\|_M^2 + 2C}$$

and so is bounded above. It therefore suffices to show that $\langle \mathbf{w}, \dot{\mathbf{q}}_i\rangle_M$ increases monotonically at every iteration. And indeed, by (ϵNORM),

$$\begin{aligned}\langle \mathbf{w}, \dot{\mathbf{q}}_{i+1}\rangle_M &= \langle \mathbf{w}, \dot{\mathbf{q}}_i + M^{-1}N\lambda + \mathbf{c}\rangle_M \\ &= \langle \mathbf{w}, \dot{\mathbf{q}}_i\rangle_M + \mathbf{w}^T N\lambda + \langle \mathbf{w}, \mathbf{c}\rangle_M \\ &= \langle \mathbf{w}, \dot{\mathbf{q}}_i\rangle_M + \sum \lambda_i \mathbf{w}^T\mathbf{n}_i + \langle \mathbf{w}, \mathbf{c}\rangle_M,\end{aligned}$$

and since $\mathbf{w}^T\mathbf{n}_i \geq \epsilon\|\mathbf{w}\|_M$,

$$\begin{aligned}\sum \lambda_i \mathbf{w}^T\mathbf{n}_i + \langle \mathbf{w}, \mathbf{c}\rangle_M &\geq \sum \lambda_i \mathbf{w}^T\mathbf{n}_i - \|\mathbf{w}\|_M\|\mathbf{c}\|_M \\ &\geq \epsilon\|\mathbf{w}\|_M\|\lambda\|_1 - \epsilon\|\mathbf{w}\|_M\|\lambda\|_1 \\ &= 0,\end{aligned}$$

where the inequality in the first line is by Cauchy-Schwarz, completing the proof. □

Lemma 7.1 suggests that the notion of implicit equality constraint must be modified in the inexact setting, and indeed, a set of constraints that *nearly* form an implicit equality constraint must now be treated as an equality constraint. Define the approximate lineality subspace L_ϵ of N to be the largest linear subspace of TQ with $\mathbf{n}_i^T\mathbf{v} \leq \epsilon\|\mathbf{v}\|_M$ for all $\mathbf{n}_i \in \mathcal{N}$ and $\mathbf{v} \in L_\epsilon$; the set of approximate implicit equality constraints in N is then any basis E of L_ϵ.

Theorem 7.2. *If* N *is free of approximate implicit equality constraints and* R *is an impact operator satisfying* (ϵ*NORM*)*,* (ϵ*KIN*)*, and* (*ONE*)*, then a limit configurational velocity* $\lim_{i\to\infty}\dot{\mathbf{q}}_i = R^i(\dot{\mathbf{q}}_0)$ *exists (but may be infeasible).*

Proof. The proof from the exact arithmetic case applies essentially unmodified. □

Theorem 7.3. *If* N *is free of approximate implicit equality constraints and* R *satisfies the six conditions* (ϵ*NORM*)*–*(ϵ*MOD*)*, then* $\dot{\mathbf{q}}_i^T\mathbf{n}_j \geq -\epsilon\|\dot{\mathbf{q}}_i\|_M$ *for all constraint gradients* $\mathbf{n}_j$ *after a finite number of applications of the impact operator* (ϵ*FIN*)*.*

Proof. We follow the same general line of argument as in the exact case, except here the error allowed in the conservation of energy is accounted for by allowing slight violation of the constraints in the final configurational velocity. As in the exact case, we partition the constraints into two sets: a "violated or approximately tangent" set T, and an "ultimately almost-feasible" set $U = \mathcal{N} \setminus T$,

$$\begin{aligned}T &= \{\mathbf{n} \in \mathcal{N} \mid \dot{\mathbf{q}}_\infty^T\mathbf{n} \leq -\epsilon\|\dot{\mathbf{q}}_\infty\|_M\} \\ U &= \{\mathbf{n} \in \mathcal{N} \mid \dot{\mathbf{q}}_\infty^T\mathbf{n} > -\epsilon\|\dot{\mathbf{q}}_\infty\|_M\}.\end{aligned}$$

We can then define the set A of velocities that are almost-feasible with respect to all constraints in U, are are close to violating all of the constraints in T:

$$A = \left\{\mathbf{v} \in \mathbb{R}^n : \begin{array}{l}\mathbf{v}^T\mathbf{n}_i > -\epsilon\|\mathbf{v}\|_M, \mathbf{n}_i \in U \\ \mathbf{v}^T\mathbf{n}_i < -\frac{\epsilon}{2}\|\mathbf{v}\|_M, \mathbf{n}_i \in T\end{array}\right\}.$$

A is clearly open, and by construction contains the limit velocity $\dot{\mathbf{q}}_\infty$ as an element (since $-\frac{\epsilon}{2} > -\epsilon$). Since $\dot{\mathbf{q}}_i$ converges to $\dot{\mathbf{q}}_\infty$, $\dot{\mathbf{q}}_i$ is in A for all $i \geq k$ for some integer k and

$$\dot{\mathbf{q}}_{k+1} = \dot{\mathbf{q}}_k + M^{-1}T\lambda + \mathbf{c}$$

for some λ, which we can rewrite in two ways:

$$\begin{aligned}\dot{\mathbf{q}}_{k+1} - \mathbf{c} &= \dot{\mathbf{q}}_k + M^{-1}T\lambda \\ \dot{\mathbf{q}}_{k+1} - M^{-1}T\lambda - \mathbf{c} &= \dot{\mathbf{q}}_k.\end{aligned}$$

By (ϵKIN), the right-hand sides have equal norms, and so

$$\begin{aligned}\|\dot{\mathbf{q}}_{k+1} - \mathbf{c}\|_M^2 &= \|\dot{\mathbf{q}}_{k+1} - M^{-1}T\lambda - \mathbf{c}\|_M^2 \\ &= \|\dot{\mathbf{q}}_{k+1} - \mathbf{c}\|_M^2 - 2\langle \dot{\mathbf{q}}_{k+1} - \mathbf{c}, M^{-1}T\lambda\rangle_M \\ &\quad + \|M^{-1}T\lambda\|_M^2.\end{aligned}$$

Since $\dot{\mathbf{q}}_{k+1} \in A$,

$$-2\langle \dot{\mathbf{q}}_{k+1}, M^{-1}T\lambda\rangle_M \geq \epsilon\|\dot{\mathbf{q}}_{k+1}\|_M\|\lambda\|_1,$$

and the above bound can be written as

$$0 \geq \epsilon\|\dot{\mathbf{q}}_{k+1}\|_M\|\lambda\|_1 + 2\langle \mathbf{c}, M^{-1}T\lambda\rangle_M + \|M^{-1}T\lambda\|_M^2. \tag{2}$$

By the triangle inequality and (ϵKIN),

$$\begin{aligned}\|\dot{\mathbf{q}}_{k+1}\|_M &\geq \|\dot{\mathbf{q}}_k + M^{-1}T\lambda\|_M - \|\mathbf{c}\|_M \\ &= \|\dot{\mathbf{q}}_k\|_M - \|\mathbf{c}\|_M,\end{aligned}$$

and now since $\|\mathbf{c}\|_M \leq \frac{\epsilon}{2}\|\dot{\mathbf{q}}_k\|_M$, when $\epsilon < 1$ we have that $\|\mathbf{c}\|_M \leq \frac{1}{2}\|\dot{\mathbf{q}}_k\|_M$ and

$$\begin{aligned}\epsilon\|\dot{\mathbf{q}}_{k+1}\|_M\|\lambda\|_1 &\geq \epsilon(\|\dot{\mathbf{q}}_k\|_M - \|\mathbf{c}\|_M)\|\lambda\|_1 \\ &\geq \frac{\epsilon}{2}\|\dot{\mathbf{q}}_k\|_M\|\lambda\|_1 \\ &\geq \epsilon\|\lambda\|_1\|\mathbf{c}\|_M.\end{aligned}$$

Finally since $\|\mathbf{c}\|_M \leq \epsilon\|\lambda\|_1$,

$$\epsilon\|\dot{\mathbf{q}}_{k+1}\|_M\|\lambda\|_1 \geq \|\mathbf{c}\|_M^2. \tag{3}$$

Combining (2) and (3) yields

$$0 \geq \|M^{-1}T\lambda + \mathbf{c}\|_M^2.$$

This inequality is satisfied only when $\|M^{-1}T\lambda + \mathbf{c}\|_M = 0$. But then $\dot{\mathbf{q}}_{k+1} = \dot{\mathbf{q}}_k$ and so $\dot{\mathbf{q}}_k^T\mathbf{n}_j \geq -\epsilon\|\dot{\mathbf{q}}_k\|_M$ for all constraint gradients $\mathbf{n}_j$ by (ϵMOD). □

Notice that as in the proof for exact arithmetic, conservation of energy plays a crucial role here.

8 DISCUSSION

We have shown that Gauss-Seidel-like algorithms for resolving multiple elastic impacts are guaranteed to terminate, without the need for failsafes, provided that they satisfy a minimum set of physical and algorithmic properties (NORM)–(MOD). In particular, the many different flavors of pairwise Gauss-Seidel, as well as Smith et al.'s Generalized Reflections [2012], all terminate. The only exceptions occur due to geometric degeneracies in the constraints—implicit equality constraints—which can be detected and removed to preserve termination. These algorithms can be modified to include dissipation, without giving up termination, and we have shown that termination continues to hold even when the algorithms are computed using floating-point arithmetic. Several avenues remain for further investigation:

Iterations Needed for Termination. Although we have proven that Algorithm 1 terminates for well-posed impact operators, in practice it would be useful to be able to *predict* how many iterations will be required to resolve the impact event. This would open the door to *appoximate* high-performance impact response schemes, which could choose to ignore collisions, or artifically rigidify parts of a physical system to prevent collisions, in cases where a fully-correct response is estimated to be very expensive.

Our experiments with a wedge in figure 3 suggest that the difficulty should depend on the"diameter" of the feasible cone in configuration tangent space, and some analytic results exist for simple contact geometries [de Felicio and Redondo 1981; Jia et al. 2013; Nguyen and Brogliato 2014]. Unfortunately, it is not immediately obvious from the current proofs how to estimate in general the number of iterations required to terminate: our *convergence* proof (Theorem 5.3) is constructive, but the *termination* proof is not. The proof of Theorem 5.4 suggests that the number of iterations required should depend on how far away the final, post-impact velocity is from the constraint boundary; estimating this distance, without explicitly computing the limit velocity, is a chicken-and-egg problem.

Coupled Friction. While Smith et al. [2012] show that one-way coupled frictional response predictively captures a wide range of impact behaviors, it is desirable and important to allow frictional response to be two-way coupled with iterated impact operators so that the interactions between normal and tangential responses can be balanced. Friction impulses are applied in *tangential* directions and so directly violate our assumption of (NORM). Can we maintain a guarantee of (FIN) with dissipative tangential impulses? Can we appropriately generalize (NORM)? Moreover, we do not know if directly incorporating a coupled frictional response into an impact operator guarantees that a fixed-point solution exists, much less whether iterating with such an operator will converge. Constructing a meaningful, iterated physical impact model with proof of termination, that includes fully coupled, dissipative, tangential response remains a challenging and important open problem.

Alternative Desiderata. In addition to reformulating (NORM) to allow friction, other modifications to the desiderata might be possible while still maintaining a termination guarantee, and might permit use of more sophisticated methods for modeling coefficients of restitution other than 1.0, and other dissipative phenomena. For instance, Mosterman [2001] analyzed dissipative impact of multiple bodies and proposed a two-phase scheme for resolving them; the method as described does not obey (KIN) and (VIO) and so our proof does not currently apply to it. We showed that removing any one desideratum allows for impact operators that do not terminate, but that argument does not rule out termination if the removed conditions are replaced with new ones.

Need for Anti-Locking. In §4 we established that implicit equality constraints can prevent the termination of Algorithm 1. We therefore described an algorithm for detecting and removing such implicit equality constraints from N. Critically, we proved that implicit equality constraints are the *only* obstruction to termination of Algorithm 1. We now interpret this proof from two different perspectives, corresponding to two mutually exclusive conjectures—that implicit equality constraints *do*, or *do not*, arise in practice.

In the absence of a proof to the contrary, the safest conjecture is that implicit equality constraints arise in practice. In this case, Algorithm 1 may fail to terminate (§4), whereas the improved Algorithm 2 guarantees termination. Certainly it is possible to create examples that contain implicit equality constraints (see Fig. 1, for instance), supporting this viewpoint.

On the other hand, all such examples we have found so far are *special* in the following sense: an infinitesimal perturbation of the geometry of the system removes the possibility of encountering implicit equality constraints completely (for the geyser, notice that slightly shrinking the radius of all of the balls accomplishes this). Does there exist a "generic" physical system that will encounter implicit inequality constraints, even if perturbed? More simply, we can look at a box filled with multiple slightly-deformed instances of a single type of rigid body O: is it possible to choose an O complicated enough that the bodies will *self-jam*, forming clusters that lose degrees of freedom due to implicit equality constraints, just by shaking the box?

This question is closely related to that of computing the *contact number* of a rigid body: in a random packing of multiple instances of O, how many other rigid bodies, on average, does each rigid body touch? A naive counting of degrees of freedom and constraints suggests that the higher the contact number, the more likely implicit equality constraints are to occur (with 12 being the threshold where constraints begin to outnumber DOFs). The contact number has been studied for simple geometries, such as spheres, rods [Wouterse et al. 2009], ellipsoids [Donev et al. 2004], etc, but tends to be small. It is possible to construct an O so that arbitrarily many copies of O can be placed in contact with each other [Erickson and Kim 2003], but like the geyser, this geometry is special: the construction fails if the copies are perturbed. A full investigation into the possibility of self-jamming is an interesting topic for future research.

Given these observations, we conjecture instead that implicit equality constraints do *not* arise in practice. In that case, we have proved that Algorithms 1 and 2 are *equivalent* and *both terminate*. Our proof of *convergence* (§5.1) informs us that any artifacts introduced by non-physical failsafes can be mitigated simply by carrying out additional Gauss-Seidel iterations before applying the failsafes.

Our proof of *termination* (§5.2) further empowers us to discard fail-safes altogether, trading additional computational cost for simplicity, elegance, guaranteed progress, and physical correctness.

REFERENCES

Mihai Anitescu and Florian R. Potra. 1997. Formulating Dynamic Multi-Rigid-Body Contact Problems with Friction as Solvable Linear Complementarity Problems. *Nonlinear Dynamics* 14, 3 (1997), 231–247. https://doi.org/10.1023/A:1008292328909

David Baraff. 1989. Analytical Methods for Dynamic Simulation of Non-penetrating Rigid Bodies. In *Proceedings of the 16th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '89)*. ACM, New York, NY, USA, 223–232. https://doi.org/10.1145/74333.74356

Jan Bender, Kenny Erleben, and Jeff Trinkle. 2014. Interactive Simulation of Rigid Body Dynamics in Computer Graphics. *Computer Graphics Forum* 33 (2014), 246–270. Issue 1.

B. Bernu and R. Mazighi. 1990. One-Dimensional Bounce of Inelastically Colliding Marbles on a Wall. *Journal of Physics A: Mathematical and General* 23, 24 (1990), 5745–5754. https://doi.org/10.1088/0305-4470/23/24/016

Bernard Brogliato. 1999. *Nonsmooth Mechanics: models, dynamics, and control* (2nd ed.). Springer-Verlag.

Mario E. Caire, Francisco J. López, and David H. Williams. 2008. Distributed identification of the lineality space of a cone. *The Journal of Supercomputing* 48, 2 (2008), 163–182. https://doi.org/10.1007/s11227-008-0222-0

A. Chatterjee and A. L. Ruina. 1998. A New Algebraic Rigid-Body Collision Law Based on Impulse Space Considerations. *Journal of Applied Mechanics* 65, 4 (1998), 939–951. https://doi.org/10.1115/1.2791938

Richard W. Cottle, Jong Shi Pang, and Richard E. Stone. 1992. *The Linear Complementarity Problem.* Academic Press, New York.

J. Crassous, D. Beladjine, and A.Valance. 2007. Impact of a Projectile on a Granular Medium Described by a Collision Model. *Physical Review Letters* 99 (2007), 248001.

J. d'Alembert. 1743. *Traité de Dynamique.* Paris.

J. R. de Felicio and D. M. Redondo. 1981. Linear collisions revisited. *Am. J. Phys* 49, 147 (1981).

A. Donev, I. Cisse, D. Sachs, E.A. Variano, F.H. Stillinger, R. Connelly, S. Torquato, and P.M. Chaikin. 2004. Improving the density of jammed disordered packings using ellipsoids. *Science* 303 (2004), 990–993.

E. Drumwright and D. Shell. 2011. *Modeling contact friction and joint friction in dynamic robotic simulation using the principle of maximum dissipation.* Springer Berlin Heidelberg, 249–266.

Jeff Erickson and Scott Kim. 2003. *Arbitrarily Large Neighborly Families of Congruent Symmetric Convex 3-Polytopes.* CRC Press.

Kenny Erleben. 2007. Velocity-based Shock Propagation for Multibody Dynamics Animation. *ACM Trans. Graph.* 26, 2, Article 12 (June 2007), 12:1–12:20 pages. https://doi.org/10.1145/1243980.1243986

G. Gilardi and I. Sharf. 2002. Literature survey of contact dynamics modelling. *Mechanism and Machine Theory* 37 (2002), 1213–1239. Issue 10.

Christof Glocker. 2004. Concepts for Modeling Impacts without Friction. *Acta Mechanica* 168 (2004), 1–19.

Suresh Goyal, Andy Ruina, and Jim Papadopoulos. 1991. Planar sliding with dry friction, Part 1. Limit surface and moment function. *Wear* 143 (1991), 307–330.

I. Han and B. J. Gilmore. 1993. Multi-Body Impact Motion with Friction—Analysis, Simulation, and Experimental Validation. *J. Mech. Des* 115, 3 (1993), 412–422.

A. P. Ivanov. 1995. On multiple impact. *Journal of Applied Mathematics and Mechanics* 59, 6 (1995), 887–902. https://doi.org/10.1016/0021-8928(95)00122-0

Y.-B. Jia, M. Mason, and M. Erdmann. 2013. Multiple Impacts: A State Transition Diagram Approach. *International Journal of Robotics Research* 32, 1 (2013), 84–114.

W. Johnson. 1976. Simple Linear Impact. *International Journal of Mechanical Engineering Education* 4, 2 (1976), 167–181.

Danny M. Kaufman, Timothy Edmunds, and Dinesh K. Pai. 2005. Fast frictional dynamics for rigid bodies. *ACM TOG (SIGGRAPH 05)* 24, 3 (2005), 946–956.

Y. A Khulief. 2012. Modeling of Impact in Multibody Systems: An Overview. *J. Comput. Nonlinear Dynam.* 8, 2 (2012), 021012.

Caishan Liu, Zhen Zhao, and Bernard Brogliato. 2008. Frictionless Multiple Impacts in Multibody Systems. I. Theoretical Framework. *Proceedings of the Royal Society A* 464 (2008), 3193–3211.

Francisco López. 2011. An Algorithm to Find the Lineality Space of the Positive Hull of a Set of Vectors. *Journal of Mathematical Modelling and Algorithms* 10 (2011), 1–30. Issue 1. https://doi.org/10.1007/s10852-010-9133-1

Colin MacLaurin. 1742. *A Treatise on Fluxions.* T. W. and T. Ruddimans, Edinburgh.

Sean McNamara and W. R. Young. 1994. Inelastic collapse in two dimensions. *Phys. Rev. E* 50, 1 (Jul 1994), R28–R31. https://doi.org/10.1103/PhysRevE.50.R28

J. J. Moreau. 1985. Standard Inelastic Shocks and the Dynamics of Unilateral Constraints. In *Unilateral Problems in Structural Analysis: Proceedings of the Second Meeting on Unilateral Problems in Structural Analysis, Ravello, September 22–24, 1983*, Gianpietro Del Piero and Franco Maceri (Eds.). Springer Vienna, Vienna, 173–221. https://doi.org/10.1007/978-3-7091-2632-5_9

Pieter J. Mosterman. 2001. On the Normal Component of Centralized Frictionless Collision Sequences. *J. Appl. Mech.* 74, 5 (2001), 908–915.

N. Nguyen and B. Brogliato. 2014. *Multiple Impacts in Dissipative Granular Chains.* Springer Heidelberg.

X. Provot. 1997. Collision and Self-collision Handling in Cloth Model Dedicated to Design. In *Computer Animation and Simulation '97.* 177–190.

Breannan Smith, Danny M. Kaufman, Etienne Vouga, Rasmus Tamstorf, and Eitan Grinspun. 2012. Reflections on Simultaneous Impact. *ACM Trans. Graph.* 31, 4, Article 106 (July 2012), 12 pages. https://doi.org/10.1145/2185520.2185602

David E. Stewart. 2000. Rigid-Body Dynamics with Friction and Impact. *SIAM Rev.* 42, 1 (2000), 3–39. https://doi.org/10.1137/S0036144599360110

David E Stewart. 2011. *Dynamics with Inequalities: Impacts and Hard Constraints.* Society for Industrial and Applied Mathematics.

Dan Stoianovici and Yildirim Hurmuzlu. 1996. A critical study of the applicability of rigid-body collision theory. *Journal of Applied Mechanics* 63, 2 (1996), 307–316.

Jan Telgen. 1983. Identifying Redundant Constraints and Implicit Equalities in Systems of Linear Constraints. *Management Science* 29, 10 (1983), 1209–1222. https://doi.org/10.1287/mnsc.29.10.1209

Albert Anton ten Dam. 1997. *Unilaterally constrained dynamical systems.* Ph.D. Dissertation. Rijksuniversiteit Groningen. http://hdl.handle.net/11370/0c2036b1-cab4-49e1-a602-13e417923985

T. Uchida, M. Sherman, and S. Delp. 2015. Making a meaningful impact: modelling simultaneous frictional collisions in spatial multibody systems. *Proc. Math. Phys. Eng. Sci.* 471, 2177 (2015), 20140859.

Roger J.-B. Wets and Christoph Witzgall. 1967. Algorithms for Frames and Lineality Spaces of Cones. *Journal of Research of the National Bureau of Standards, B Mathematics and Mathematical Physics* 71B, 1 (January-March 1967), 1–7. https://doi.org/10.6028/jres.071B.001

A. Wouterse, S. Luding, and A. P. Philipse. 2009. On contact numbers in random rod packings. *Granular Matter* 11 (2009), 169–177.

Tianxiang Zhang, Sheng Li, Guoping Wang, Dinesh Manocha, and Hanqiu Sun. 2015. Quadratic Contact Energy Model for Multi-impact Simulation. *Computer Graphics Forum* 34 (2015), 133–144. Issue 7.