Chapter 10

Dynamic Friction

Having defined a model of friction, we now consider the complexity of computing valid contact forces in the presence of friction. Most of the complexity results in this thesis apply to configurations with dynamic friction only. Complexity results for static friction are presented in Chapter 14. Until then, only configurations of bodies with dynamic friction are considered.

Given that no static friction forces act at any contact points, the contact force direction at each contact point is given by Equation (9–6), since μ and the tangential velocity are known quantities. Thus, as in Chapter 8, we are interested in computing only the vector $\boldsymbol{f}_{\mathbf{N}}$ of constraint force magnitudes. Solutions for $\boldsymbol{f}_{\mathbf{N}}$ for configurations with just dynamic friction can still be expressed as the LCP of Equation (8–7). However, as a result of the contact force having a tangential component, the matrix $\mathbf{A}$ in Equation (8–7) is no longer necessarily symmetric PSD and Equation (8–7) is no longer a convex LCP.

As a result, several nice properties are lost. Although solutions to convex LCP's can be found in polynomial time, non-convex LCP's do not share this property; finding solutions for non-convex LCP's is *NP*-hard. What then, is the complexity

of solving the non-convex LCP's corresponding to configurations of bodies with dynamic friction? Also, the uniqueness and existence properties associated with the convex LCP's of Chapter 8 no longer always hold. With dynamic friction, it is possible that there may be no solution $\boldsymbol{f_N}$, and thus no valid contact forces. We call a configuration exhibiting this behavior *inconsistent*.[1] It is also possible that the acceleration produced by solutions of the LCP is not unique; such a configuration is called *indeterminate*.

Configurations with one contact point that exhibit inconsistency and indeterminacy have been discussed by Erdmann[12], Lötstedt[28], and Mason and Wang[35]. Lötstedt[30], realizing that indeterminacy and inconsistency major present difficulties for a simulation process, proposes a modification of the Coulomb friction law that eliminates both inconsistency and indeterminacy, and further causes the LCP to always be convex. We will discuss Lötstedt's modification in Chapter 14, when we consider static friction. In this chapter, we discuss two well known one contact point configurations that exhibit inconsistency and indeterminacy. Chapter 11 uses these configurations as building blocks in a proof about the computational complexity of solving dynamic friction problems.

10.1 Indeterminacy

In Figure 10.1, body A is a thin rod of length two with a symmetric mass distribution that contacts body B at a single contact point. Body B, the "base", is fixed. The

[1] The use of the word "inconsistent" throughout this thesis requires some qualification. The fact that configurations exist which have no valid contact forces is counter-intuitive to most people at first. These configurations are "inconsistent" with the initial expectation that a valid set of (finite) contact forces must exist for any configuration that can be physically realized. However, we do not use "inconsistent" to mean "unnatural", nor do we mean to imply that inconsistent configurations are in any way extraordinary physical occurrences. Readers who are familiar with "self-locking" behavior of assemblies due to friction will find "inconsistent" configurations to be nothing special at all.

example is two-dimensional, so the tangent direction is written simply as $\hat{t}$. The particular values of I, θ and μ given in Figure 10.1 are chosen to simplify later computations. A gravity force of $-mg\hat{n}$ acts on A. Let the angular velocity ω and gravity constant g be such that A is rotating clockwise and

$$|\omega|^2 \sin\theta - g = 1. \tag{10–1}$$

Given these values, let the linear velocity of A be chosen so that the point p_a of the rod is sliding tangent to B as shown. Since A and B are neither colliding nor separating at p_a, a contact force is presumed to act. Since the tangential velocity is non-zero, a dynamic friction force acts on A in the $\hat{t}$ direction. From Chapter 9, the net contact force acting on A is

$$f_N\hat{n} + \mu f_N\hat{t} = f_N(\hat{n} + \mu\hat{t}). \tag{10–2}$$

Because B is fixed and A is thin, the normal acceleration a_N is simply

$$a_N = \hat{n} \cdot \ddot{p}_a. \tag{10–3}$$

To compute a_N, let a and α denote the linear and angular acceleration of A, ω, the angular velocity of A, and r, the displacement of p_a from the center of mass of A. Although this configuration is two-dimensional, we will regard a, p_a, $\ddot{p}_a$, $\hat{n}$ and r as 3-space vectors lying in the xy plane, and ω and α as 3-space vectors parallel to the z axis. From Figure 10.1, $r = (-\cos\theta, -\sin\theta, 0)$. The acceleration $\ddot{p}_a$ may be expressed as the sum of three terms: the linear acceleration a, the tangential acceleration $\alpha \times r$, and the centripetal acceleration $\omega \times (\omega \times r)$. The linear acceleration, a, is

$$\begin{aligned} a &= \frac{f_N\hat{n} + \mu f_N\hat{t} + (-mg)\hat{n}}{m} \\ &= \frac{f_N\hat{n} + \mu f_N\hat{t}}{m} - g\hat{n}. \end{aligned} \tag{10–4}$$

Variables

p_a	contact point	$\dot{p}_a$	contact point velocity
$\hat{n}$	unit surface normal	$\hat{t}$	unit surface tangent
m	A's mass	I	A's moment of inertia
ω	A's angular velocity	μ	coefficient of friction
g	gravitational acceleration		

Relations

$$I = \frac{m}{16} \qquad \theta = 72^\circ \qquad 16(\cos^2\theta - \mu\cos\theta\sin\theta) = -2 \qquad (\mu \approx 3/4)$$

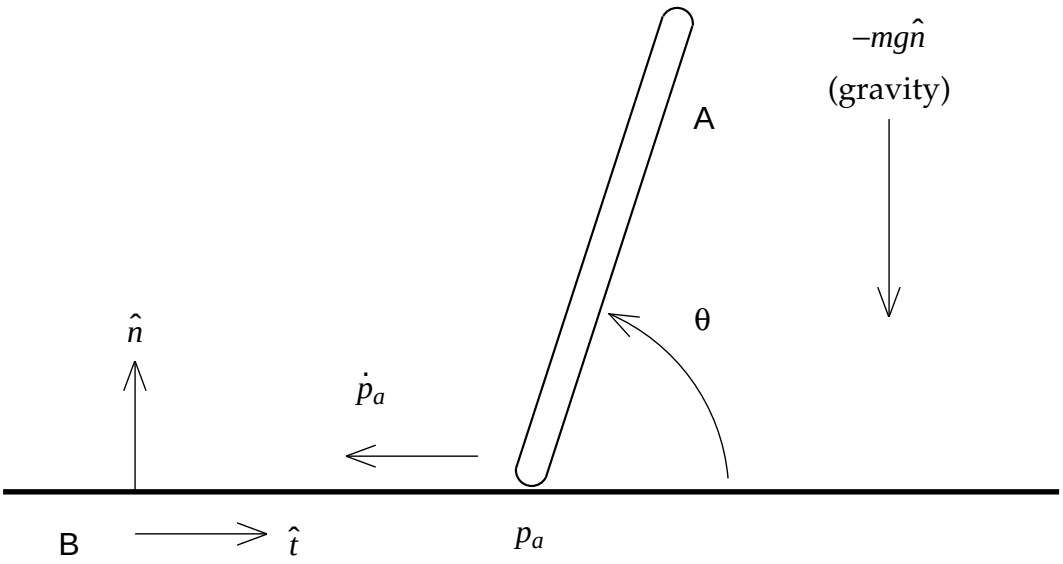

Figure 10.1: A one contact point configuration between a thin rod A and a fixed base B.

The torque on A about its center of mass is

$$r \times (f_N \hat{n} + \mu f_N \hat{t})$$

which yields an angular acceleration of

$$\alpha = \frac{r \times (f_N \hat{n} + \mu f_N \hat{t})}{I}. \tag{10–5}$$

Then

$$\begin{aligned} \ddot{p}_a &= a + \alpha \times r + \omega \times (\omega \times r) \qquad (10–6) \\ &= \frac{f_N \hat{n} + \mu f_N \hat{t}}{m} - g\hat{n} + \frac{r \times (f_N \hat{n} + \mu f_N \hat{t})}{I} \times r + \omega \times (\omega \times r). \end{aligned}$$

Taking the dot product of Equation (10–4) with $\hat{n}$,

$$\hat{n} \cdot a = \frac{f_N \hat{n} \cdot \hat{n} + \mu f_N \hat{n} \cdot \hat{t}}{m} - g\hat{n} \cdot \hat{n} = \frac{f_N}{m} - g. \tag{10–7}$$

Taking the dot product of the tangential acceleration $\alpha \times r$ with $\hat{n}$ and using the geometry of Figure 10.1,

$$\hat{n} \cdot \omega \times (\omega \times r) = |\omega|^2 \sin\theta \tag{10–8}$$

and

$$\begin{aligned} \hat{n} \cdot (\alpha \times r) &= \hat{n} \cdot \left(\frac{r \times (f_N \hat{n} + \mu f_N \hat{t})}{I} \times r \right) \\ &= \frac{f_N(\cos^2\theta - \mu \cos\theta \sin\theta)}{I}. \qquad (10–9) \end{aligned}$$

Then, from the relations in Figure 10.1,

$$\begin{aligned} \hat{n} \cdot \ddot{p}_a &= \frac{f_N}{m} - g + \frac{f_N(\cos^2\theta - \mu \cos\theta \sin\theta)}{I} + |\omega|^2 \sin\theta \\ &= \frac{f_N}{m}(1 + 16(\cos^2\theta - \mu \cos\theta \sin\theta)) + |\omega|^2 \sin\theta - g \\ &= \frac{f_N}{m}(1 - 2) + |\omega|^2 \sin\theta - g \qquad (10–10) \\ &= -\frac{f_N}{m} + (|\omega|^2 \sin\theta - g). \end{aligned}$$

Finally, from Equation (10–1),

$$\begin{aligned} \hat{n} \cdot \ddot{p}_a &= -\frac{f_N}{m} + (|\omega|^2 \sin\theta - g) \\ &= -\frac{f_N}{m} + 1. \end{aligned} \tag{10–11}$$

From Equation (8–7), the constraints on f_N are the LCP

$$f_N \geq 0, \quad -\frac{f_N}{m} + 1 \geq 0 \qquad \text{and} \qquad f_N(-\frac{f_N}{m} + 1) = 0. \tag{10–12}$$

This LCP yields two solutions for f_N; one solution is $f_N = 0$, and the other is $f_N = m$. For the $f_N = 0$ solution, $\hat{n} \cdot \ddot{p}_a = 1$. In this solution, the centripetal acceleration of $\ddot{p}_a$ is stronger than the force of gravity pulling A down; thus, A merely continues its rotation and the point p_a moves off of B (Figure 10.2a).

In the second solution, $f_N = m$ and $\hat{n} \cdot \ddot{p}_a = 0$. A normal force of $m\hat{n}$ and a friction force of $\mu m\hat{t}$ act on A. The torque generated by friction balances the centripetal acceleration of p_a; as a result, A and B do not break contact (Figure 10.2b). Note that the *only* valid values of f_N for this configuration are $f_N = 0$ or $f_N = m$. Since the solutions produce different accelerations for A, the configuration is indeterminate.

The fact that there is more than one possible behavior for A should not be of great concern. We have already seen that the rigid body assumption allows for non-unique distributions of contact forces, as in the case of a chair resting on four legs. When the rigid body assumption is relaxed to allow deformation, the indeterminacy in the force distribution is resolved. In the case of the sliding rod, sufficient knowledge about the flexibility and compressibility of the rod would serve to determine the actual behavior of the rod. Since we have chosen to treat bodies as perfectly rigid, we are in essence stating that this knowledge is not available to us.

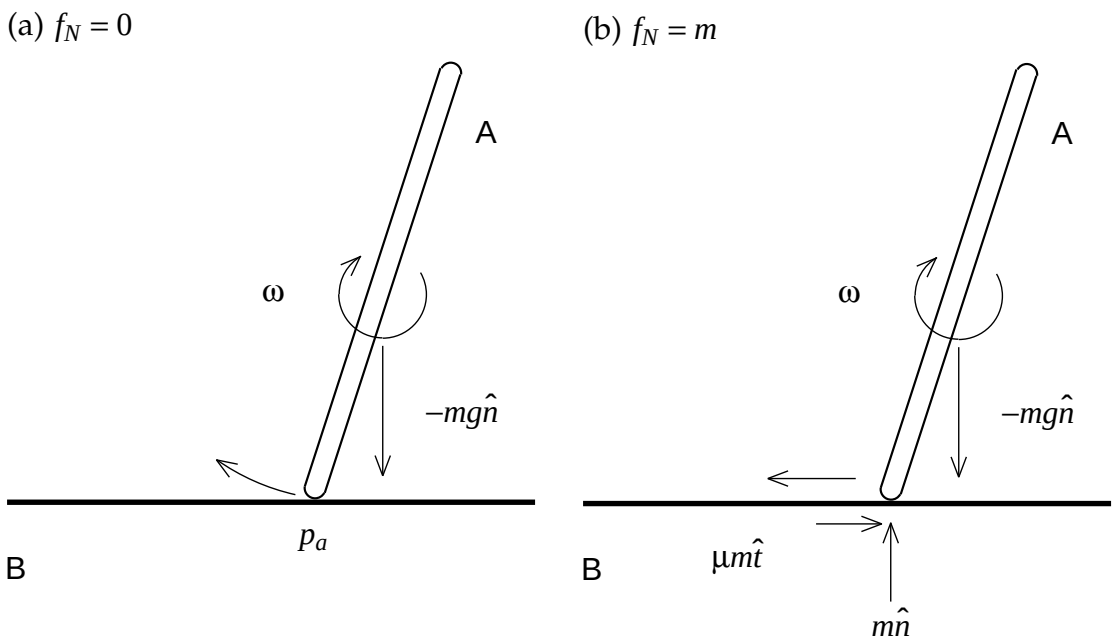

Figure 10.2: An indeterminate configuration. (a) The contact force between A and B is zero. The point p_a rotates to the left and up, breaking contact with B. (b) The normal and friction forces balance gravity and centripetal acceleration; p_a moves horizontally and maintains contact with B.

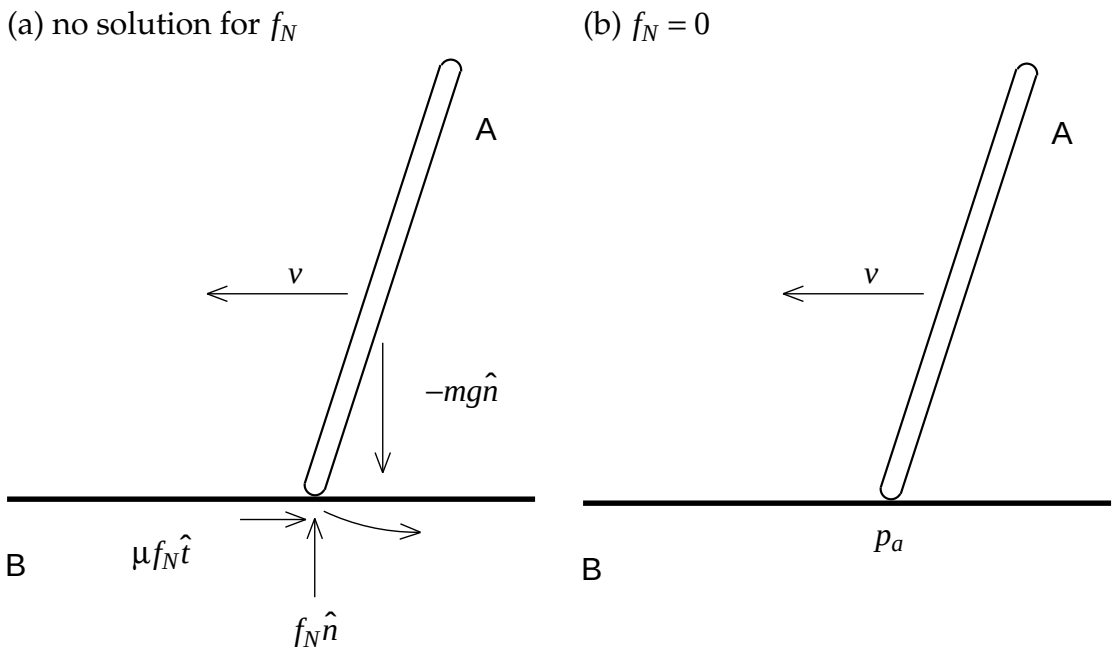

Figure 10.3: A potentially inconsistent configuration. (a) For any $f_N \geq 0$, the point p_a is accelerated downwards into B. This configuration is inconsistent. (b) The configuration has a unique solution of $f_N = 0$ when gravity is removed; A skims along the surface of B. This configuration is not inconsistent.

10.2 Inconsistency

To demonstrate inconsistency, consider the same sliding rod configuration, but with a different angular velocity. Suppose that A has an angular velocity of zero ($|\omega| = 0$), and a linear velocity v opposite $\hat{t}$, as shown in Figure 10.3. Then the condition $a_N = \hat{n} \cdot \ddot{p}_a \geq 0$ is

$$\hat{n} \cdot \ddot{p}_a = -\frac{f_N}{m} - g \geq 0. \tag{10–13}$$

However, if $g > 0$ (Figure 10.3a), then $f_N \geq 0$ implies that $\hat{n} \cdot \ddot{p}_a < 0$. Thus, no (positive) value of f_N is sufficient to prevent p_a from accelerating downwards and into B. This configuration is inconsistent.

Note that the value of g here is crucial. If $g = 0$, so that no external force acts

on A, then $f_N = 0$ becomes the (unique) valid contact force (Figure 10.3b). Any positive value of f_N for this configuration results in inter-penetration. Figure 10.3b corresponds to to p_a "skimming" horizontally over B, with neither a normal force nor a friction force exerted on A. If g becomes even slightly positive however, the configuration is inconsistent.

How can we interpret this phenomenon? Although our mathematical model admits no solution for this configuration, we know that there must be some physical answer. Still, how can we get around the fact that no matter what value we assume for f_N, the point p_a accelerates downwards? The resolution of the issue lies in how we view the subject of inconsistency. When two rigid bodies collide, no force, no matter how large, is sufficient to prevent some degree of inter-penetration. For rigid body mechanics, the need to prevent inter-penetration requires the extension of the rigid body model to include the concept of impulsive forces which can discontinuously change velocity to prevent inter-penetration. Impulsive forces are added to the rigid body model to avoid the inconsistency that non-impulsive forces are not always sufficient to prevent inter-penetration.

For the sliding inconsistent rod of Figure 10.3a, non-impulsive forces cannot prevent inter-penetration because the constraints of dynamic friction result in a fixed direction for the contact force, no matter how strong. In contrast, a contact impulse applied to A and B can prevent inter-penetration because the direction of the contact impulse is not fixed. After the impulse is applied, the inconsistency is removed and a non-impulsive contact force can be found. Mason and Wang[35] discuss the computation of this frictional impulse and show that it is consistent with the way in which frictional impulses are determined for collisions between rigid bodies.

Interestingly, what appears to disturb people most about the configuration of Figure 10.3a is that an impulse is necessary even though the relative normal velocity

between A and B is zero. Because of this, the configuration is seen as pathological, and the application of an impulse is seen as an action of last resort. In the next chapter, we will consider the computational complexity of finding non-impulsive solutions, so as to avoid applying impulsive solutions as much as possible.

Chapter 11

An *NP*-complete Class of Configurations

We define the *frictional consistency* problem as the problem of deciding if a given configuration with dynamic friction is consistent. In this section, we prove that the frictional consistency problem is *NP*-complete. We first show that the frictional consistency problem lies in *NP* and then establish the frictional consistency problem as *NP*-hard. Although the configurations constructed in this section seem contrived (and arguably are), the *NP*-hardness result has grave implications even when inconsistency is not a concern during simulation.

DEFINITION. *An instance of the frictional consistency problem is a configuration* C *of bodies that contact at* n *distinct contact points. The physical properties of each body (mass, moment of inertia, linear and angular velocity, position and orientation, and external forces) are described by rational numbers. The specifics of a contact point (position, coefficient of friction, surface normal) are also described by rational numbers. The relative motion between bodies at contact points with friction is non-zero in the direction tangent to the contact surface and zero in the direction normal to*

the contact surface. The notation $|\mathrm{C}| = \mathrm{k}$ *means that configuration* C *is describable in* k *bits. Clearly* $\mathrm{n} < \mathrm{k}$.

THEOREM 1. *The frictional consistency problem lies in NP.*

PROOF. Given an instance of C, an LCP of size n with the following two properties exists.

(1) If C is consistent, then an n-vector $\boldsymbol{f}_{\mathbf{N}}$ that is a solution to the LCP exists. The set of contact forces such that the magnitude of the normal force at the ith contact point is f_{N_i} is a valid set of contact forces for the configuration C.

(2) If C is inconsistent, the LCP has no solution.

The LCP is the LCP of Equation (8–7). The numerical quantities in the LCP are computed from the rational entries of C in a total of $O(n^3)$ arithmetical operations. The LCP can therefore be constructed in time polynomial to k. The problem of deciding if an LCP has a solution lies in *NP*[40]. It follows from this that deciding frictional consistency is also in *NP*. □

In order to show that deciding frictional consistency is *NP*-hard, we will reduce the *NP*-complete problem "subset sum" to the frictional consistency problem.

DEFINITION. *An instance of the subset sum problem is a pair* (A, S) *where* $\mathrm{A} = \{\mathrm{a}_1, \ldots, \mathrm{a}_\mathrm{n}\}$ *is a set of positive integers and* S *is a single positive integer. A subset sum instance* (A, S) *is* satisfiable *if there exists a subset* $\mathrm{A}' \subset \mathrm{A}$ *such that*

$$\sum_{\mathrm{a} \in \mathrm{A}'} \mathrm{a} = \mathrm{S}. \tag{11–1}$$

Deciding if an instance of the subset sum problem is satisfiable is an NP-complete problem[16].

To show that deciding frictional consistency is *NP*-hard we will take an arbitrary instance (A, S) of the subset sum problem and construct (in polynomial time) a configuration of bodies C. The configuration C will have the property that C is consistent if and only if (A, S) is satisfiable.

THEOREM 2. *Deciding frictional consistency is NP-hard.*

PROOF. We begin by presenting a slightly modified version of the configuration of Figure 10.3. In Figure 11.1, body B is free to move vertically, either up or down, but it is still not allowed to rotate or move horizontally. Let the mass M, of body B, be $M = 100m$ so that B is still massive compared to A. Assume that there is no external gravity force acting on A, but that B is subject to a vertical external force $f_b\hat{n}$, though its center of mass (Figure 11.1). A positive value for f_b indicates the force acts upwards, pushing B towards A.

Since B may now move vertically, Equation (10–13) is modified. Since there is no external force on A, $g = 0$. In place of $-f_N/m - g \geq 0$,

$$-\frac{f_N}{m} + \frac{f_N}{M} - \frac{f_b}{M} \geq 0. \tag{11–2}$$

The term f_N/M is the linear acceleration of B away from A due to the contact force. The term f_b/M is the upwards linear acceleration of B due to the external force. Since $M = 100m$,

$$-\frac{f_N}{m} + \frac{f_N}{M} - \frac{f_b}{M} = -\frac{99}{100}\frac{f_N}{m} - \frac{f_b}{M}. \tag{11–3}$$

Equation (10–13) is then rewritten as

$$-\frac{99}{100}\frac{f_N}{m} - \frac{f_b}{M} \geq 0. \tag{11–4}$$

As before, if $f_b > 0$, an inconsistency occurs. If $f_b = 0$, then $f_N = 0$ is the unique valid solution and the configuration is consistent.

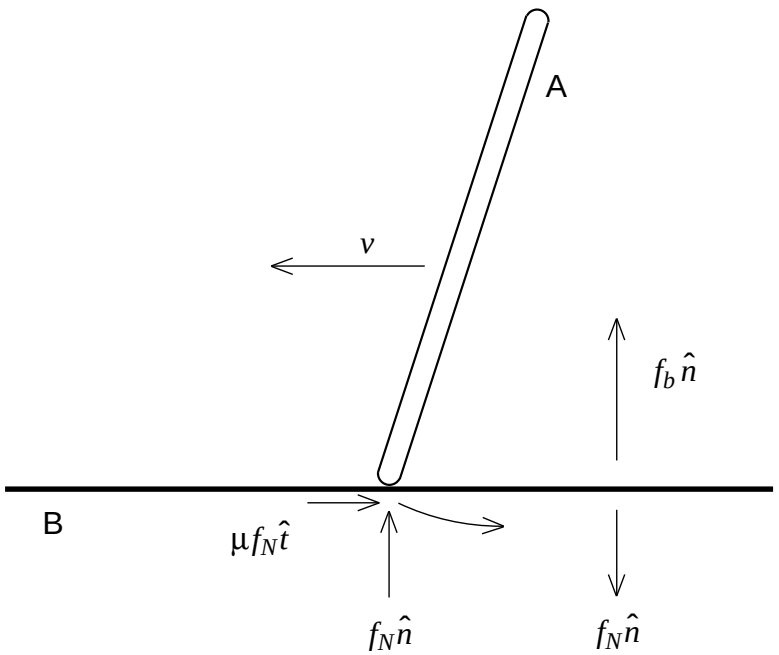

Figure 11.1: A possibly inconsistent configuration. Body B may move vertically, but is not allowed to move horizontally or rotate. If the external vertical force magnitude f_b is positive, the configuration is inconsistent. The contact force $f_N\hat{n}$ on A acts equally and opposite on B.

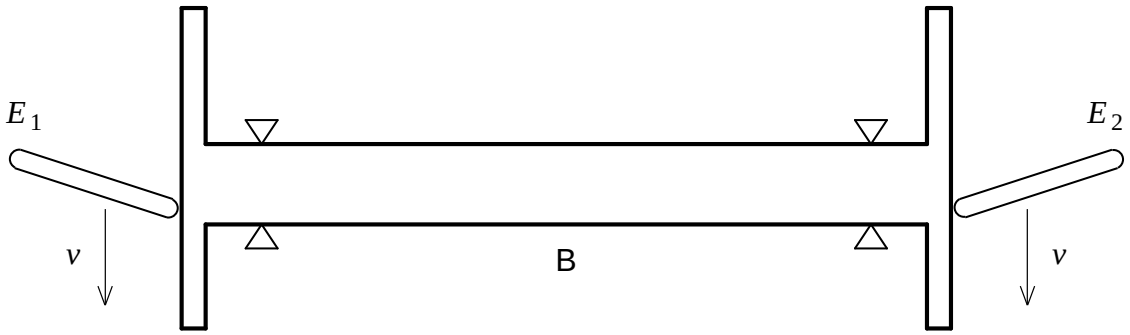

Figure 11.2: A configuration whose consistency depends on an external force. Body B is constrained by the fixed wedges and can only move horizontally. The configuration is consistent only if B is not subject to a net horizontal force.

Now, consider the configuration of Figure 11.2. Body B of Figure 11.2 is initially at rest and is positioned by four fixed triangular wedges that contact B without friction. Body B is therefore free to move horizontally, but can neither rotate nor move vertically. On either side of body B are thin rods E_1 and E_2 that have no angular velocity, and a linear velocity as indicated. The rods E_1 and E_2 contact B in the same manner as the configuration of Figure 11.1 except that the frames of reference for E_1 and E_2 are rotated by 90° with respect to Figure 11.1. In the configuration of Figure 11.1, inconsistency occurred if f_b was positive, pushing B towards A. The same holds true for Figure 11.2. If B has an acceleration leftwards (towards E_1), then inconsistency occurs. Likewise, if B has an acceleration rightwards (towards E_2), then inconsistency also occurs. Thus, the configuration of Figure 11.2 is consistent only if the net horizontal acceleration of B is zero. In this case, the rods E_1 and E_2 skim along the surface of B as in Figure 10.3b.

Now consider Figure 11.3, where a collection of thin rods $R_1, \ldots, R_n$ have been added. In addition, an external horizontal force with magnitude μS acts on B,

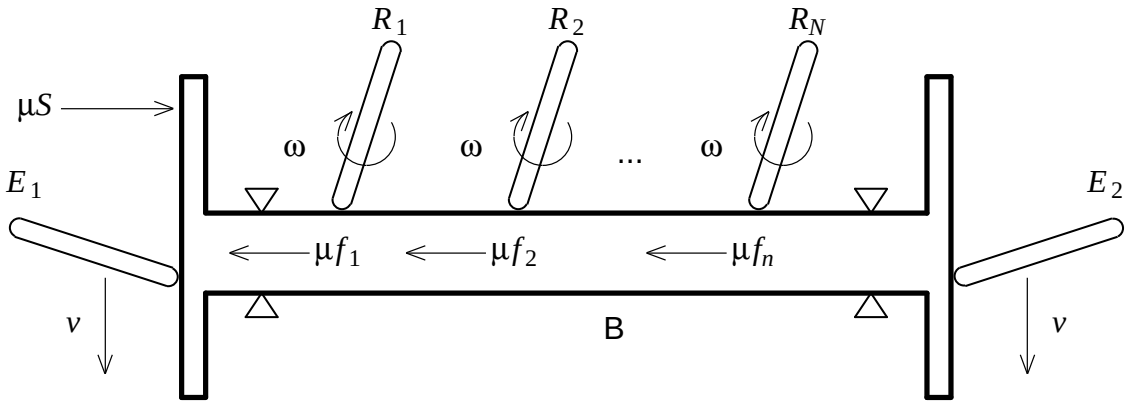

Figure 11.3: A configuration that is consistent if and only if the friction forces on B sum to μS.

trying to accelerate B to the right. Each rod R_i has mass m_i. The configuration between each rod R_i and B is the same as the configuration of Figure 10.2; thus each rod R_i has angular velocity ω and is subject to an external gravity force. Let $f_{N\,i}$ be the magnitude of the normal force between R_i and B. As in Figure 10.2, the only valid solutions for $f_{N\,i}$ are $f_{N\,i} = 0$ and $f_{N\,i} = m_i$. If $f_{N\,i} = 0$, then no friction force acts between R_i and B. Otherwise, $f_{N\,i} = m_i$ and a friction force of magnitude μm_i acts between R_i and B. The friction force pushes R_i to the right and B to the left, with magnitude μm_i. The friction force on B therefore acts to oppose the external force of magnitude μS.

In order for the configuration of Figure 11.3 to be consistent, B must have no net horizontal acceleration. This means that the friction forces exerted on B from the n rods must sum to μS, balancing the external force applied to B. Thus, the configuration is consistent if and only if

$$\sum_{i=1}^{n} \mu f_{N\,i} = \mu S. \tag{11–5}$$

Since each $f_{N\,i}$ is either 0 or m_i, the configuration is consistent if and only if some subset of $\{m_1, \ldots, m_n\}$ sums to S.

We can now perform the reduction from subset sum to show *NP*-hardness. Given any set $A = \{a_1, \ldots, a_n\}$ and any target sum S, construct the configuration of Figure 11.3. The values chosen in the previous section for μ, θ and ω may introduce irrational quantities into the descriptions of the configurations. Since the frictional consistency problem must be described in terms of rational quantities, we may assume that μ, θ and ω are suitably perturbed so that all quantities are rational, without otherwise disturbing the solution properties of the configurations. This can be done since the set of rational numbers is dense with respect to the set of real numbers. In the constructed configuration, assign $m_i = a_i$ for $1 \leq i \leq n$, and let an external horizontal force of μS act on B as shown in Figure 11.3. By the above discussion, the configuration is consistent if and only if there exists a subset of $\{m_1, \ldots, m_n\}$ that sums to S. But since $A = \{m_1, \ldots, m_n\}$, the configuration is consistent if and only if (A, S) is satisfiable. We conclude that the problem of deciding frictional consistency is *NP*-hard. □

THEOREM 3. *Deciding frictional consistency is NP-complete.*

PROOF. The result follows immediately from Theorem 1 and Theorem 2. □

COROLLARY 1. *Computing contact forces (if they exist) for a configuration is NP-hard.*

PROOF. Since deciding if a set of contact forces exists is an *NP*-complete problem, computing the contact forces (if they exist) is an *NP*-hard problem. □

A reasonable response to this corollary is to question its practical applicability to the problem of computing valid contact forces. Suppose the possibility of an inconsistent configuration occurring in a simulation is so unlikely that it can be discounted completely. Can we construct an efficient algorithm that computes contact

forces only for the set of consistent configurations? The following corollary asserts that a polynomial time algorithm for computing contact forces, even restricted to consistent configurations, does not exist unless $P = NP$.

COROLLARY 2. *A polynomial time algorithm for computing valid contact forces for consistent configurations exists if and only if* $P = NP$.

PROOF. Suppose that $P = NP$. Since deciding if an LCP has a solution lies in NP, $P = NP$ implies a polynomial time algorithm for finding the solution to an LCP. Since valid contact forces for a consistent configuration of bodies can be found by solving an associated LCP, valid contact forces are computable in polynomial time if $P = NP$.

Conversely, suppose that contact forces for consistent configurations can be computed in polynomial time. Then there exists a machine M and a polynomial p with the following behavior. Whenever M is given a consistent configuration C as input, M outputs a valid set of contact forces within time $p(|C|)$. M's behavior when C is inconsistent is undefined. Given *any* configuration C, not necessarily consistent, M can be used to decide consistency in polynomial time as follows.

Let C be input to M and run for $p(|C|)$ time. If M fails to output within this time, then C is inconsistent. Otherwise, M has produced some output. Since deciding frictional consistency is in NP, the validity of M's output can be decided in an additional amount of time that is also a polynomial function of $|C|$. If M's output is a valid set of contact forces, then clearly C is consistent. If M's output is invalid, then C must be inconsistent (else M would have output a valid answer). In any event, the consistency of C has been decided in polynomial time.

Since deciding consistency is NP-complete, we conclude that the existence of a polynomial time algorithm for computing contact forces on consistent configurations

would imply that $P = NP$. $\Box$

From a simulation viewpoint, this seems to be a depressing result. However, there has been an implicit, unchallenged assumption in the above discussion. The assumption is that we must determine whether or not a configuration is consistent, so that we will know whether impulsive or non-impulsive forces should be applied. This involves a further assumption: it is invalid to apply impulsive forces to configurations that are consistent (and thus have a non-impulsive force solution). In the next chapter, the argument will be made that this second assumption is false. Based on this argument, we will reformulate the problem of computing contact forces so that it is neither necessary nor even desirable to determine whether or not a configuration is consistent. Following that, it will be shown how avoiding the need to determine consistency leads to a practical polynomial time algorithm for computing contact forces for dynamic friction.

Chapter 12

Physical Models

In this section, a physical model for both inconsistency and indeterminacy is presented. We will use this model to develop a computationally practical method of dealing with dynamic friction in Chapter 13. The model in this section was developed in order to understand the behavior of inconsistent and indeterminate configurations. The insights gained from considering this model also suggest a general principle regarding indeterminacy during simulation.

The motivation of a physical model stems from the need to answer the following basic questions. First, what should be the result of a simulation when inconsistency is encountered? As we have seen, the only resolution of inconsistent configurations, such as Figure 10.3a, is the application of an impulsive contact force. What is the physical process which produces such a force? Second, is there a simple physical explanation for the type of indeterminacy that the configuration of Figure 10.2 exhibits?

In the physical world, there is of course no such thing as a perfectly rigid body. For near rigid bodies, contact forces arise as a result of small elastic deformations in the neighborhood of the contact area. The penalty method, as described in Chap-

ter 7, is based on a mathematical method from numerical optimization. However, the penalty method can equally well be viewed (and is, by many) as an attempt to model elastic deformations in a very simple form. In the penalty method, small amounts of inter-penetration model the amount of deformation between bodies, and the penalty force accompanying the inter-penetration models the elastic restoring force. Typically, the normal force is modeled as a linear spring force $-Kd$, where K is the spring constant and d is the amount of inter-penetration. As Chapter 7 shows, the penalty method converges to a correct result as K is increased. It therefore makes sense to consider applying the penalty model, conceptually, to situations involving friction, and studying the limiting behavior. This thesis does not propose the penalty model as embodying a specific physical process. Rather, the penalty model is used as an intuition to help visualize why impulses are necessary in some situations, and to help guide our development of a definition of exactly when and how impulses can be applied to a system without colliding contact.

12.1 A model of inconsistency

Figure 12.1 shows the behavior of the inconsistent configuration of Figure 10.3a when the penalty method is applied. At time t_0, consider the tip of the rod, p_a, to be resting exactly on B, with zero inter-penetration. Since there is no inter-penetration, the normal force is zero. Even though p_a is sliding along B, the friction force is zero since the normal force is zero. Since the only force acting on A is the external gravity force $-mg\hat{n}$, p_a accelerates downwards. At time $t_0 + \Delta t$, p_a has inter-penetrated B by an amount d_1, so a normal force $Kd_1\hat{n}$ acts on A. Since p_a is still sliding, a friction force of $\mu Kd_1\hat{t}$ also acts on A. The net result, from Equation (10–11), is that this causes p_a to accelerate downwards (even faster than before). As the penalty force continues to increase, it causes more inter-penetration

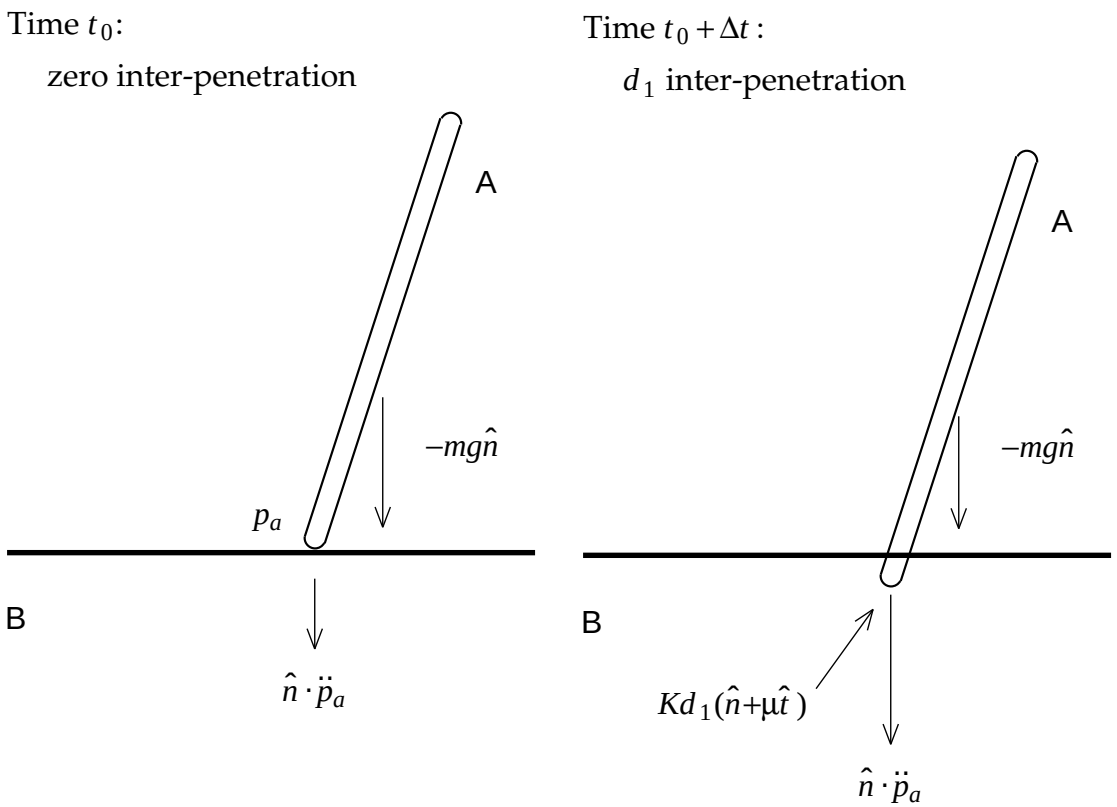

Figure 12.1: A model of indeterminacy. At time t_0, only gravity acts on A. At time $t_0 + \Delta t$, the inter-penetration distance is d_1 and both a penalty and a gravity force act on A, causing p_a's downwards acceleration to increase.

between A and B; a form of positive feedback. Accordingly, both the friction and the normal force increase, and the cycle continues. Since we are trying to model A and B as rigid bodies, the penalty constant K must be allowed to be arbitrarily large. The larger K is, the faster inter-penetration increases and the faster the normal and friction forces build.

Recall that the friction force opposes the sliding motion of A across B. By making K arbitrarily large, the friction force brings p_a to rest (horizontally) in an arbitrarily short time. For example, suppose K is adjusted so that p_a comes to rest within time Δt. Then the amount of inter-penetration is $O(\Delta t^2)$, since the distance traveled by p_a depends quadratically on the time for which it travels. In the limit as K goes to infinity, the contact force on A acts as an impulse and instantaneously brings p_a to rest horizontally, without inter-penetration occurring.

The impulse that brings p_a to rest horizontally produces a configuration in which A and B are colliding (in the normal direction) at p_a. This requires the application of a second impulse, to counter the downwards velocity acquired by p_a. Once the impulses have been applied and p_a is no longer sliding across B, dynamic friction is replaced with static friction. As a result, the net contact force direction is no longer fixed, and the inconsistency no longer exists.

12.2 A model of indeterminacy

Consider the indeterminate configuration of Figure 10.2, which has solutions $f_N = 0$ and $f_N = m$. Using the penalty method, the indeterminacy can be removed by assuming some amount of initial inter-penetration between A and B. If the initial inter-penetration between A and B is zero (Figure 12.2) then no normal force exists at time t_0 and the net acceleration of p_a is upwards. Contact is immediately broken; this behavior is shown in Figure 10.2a. However, if the initial inter-penetration

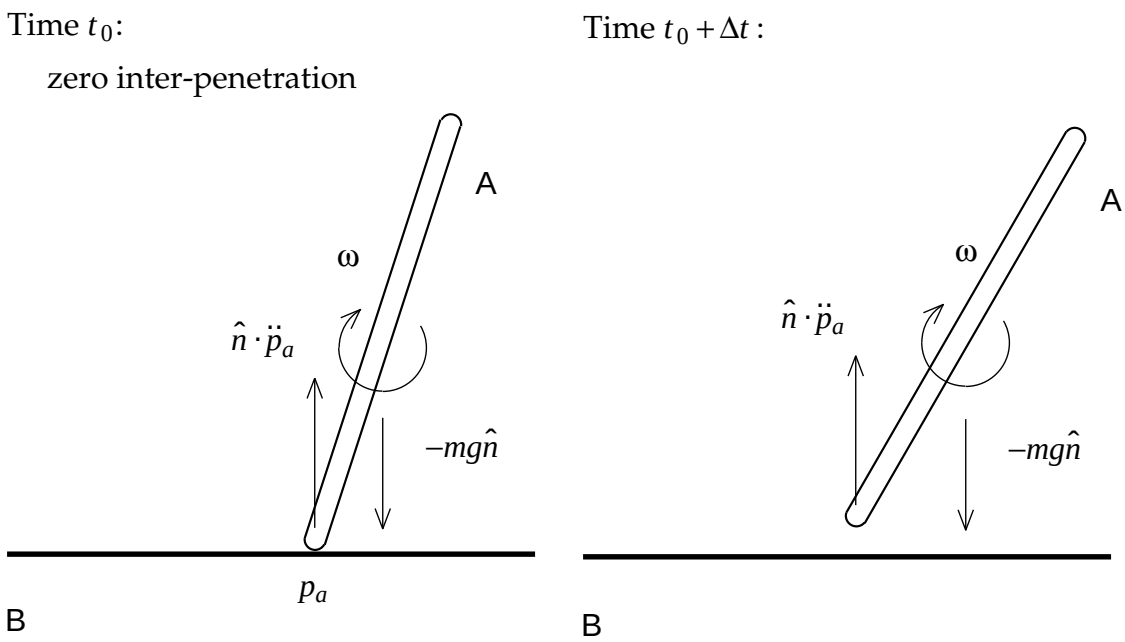

Figure 12.2: A configuration with no initial inter-penetration. Only gravity acts on A initially. The centripetal acceleration of A pulls p_a away from B and contact is broken.

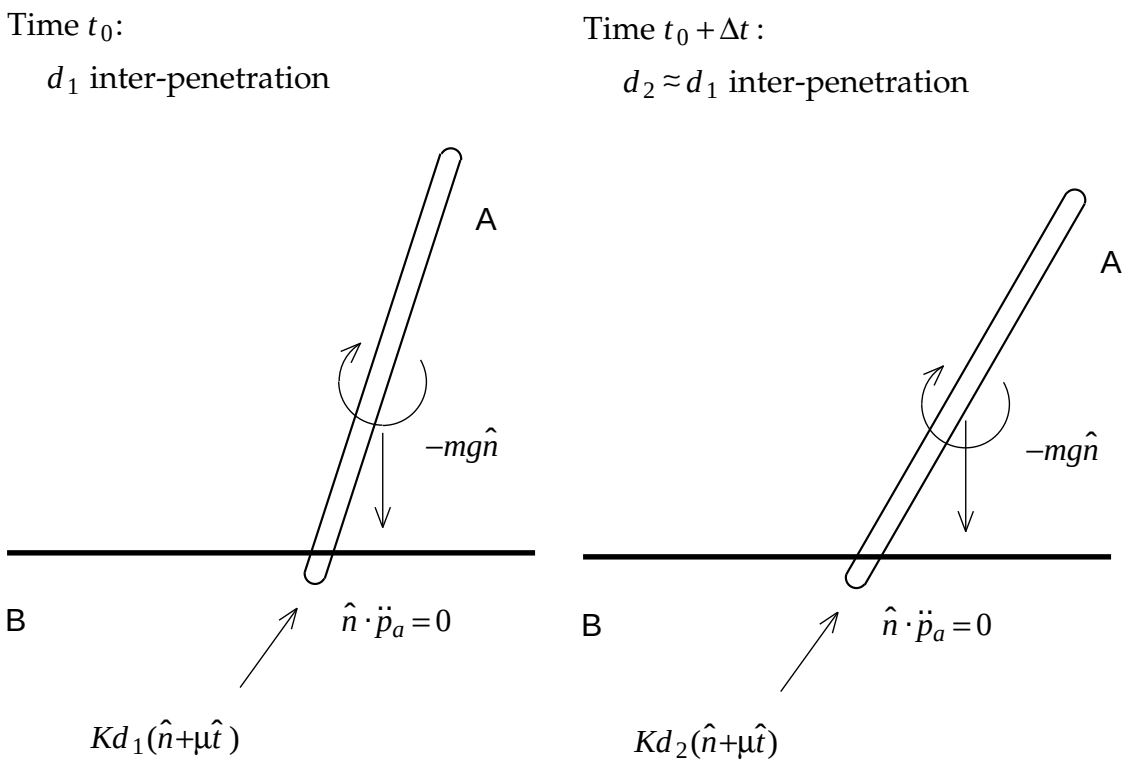

Figure 12.3: A configuration with an initial inter-penetration depth of d_1. Both gravity and a penalty force act on A. Body A slides and falls without breaking contact with B.

produces a normal force magnitude of m, then the normal and friction forces prevent A from breaking contact with B. In Figure 12.3, let the initial inter-penetration d_1 be m/K. Then the normal force magnitude at time t_0 is m. Since p_a is sliding on B, a friction force acts on A as shown. As A falls, maintaining contact with B, the inter-penetration varies smoothly, produce a varying normal force. At time $t_0 + \Delta t$, body A still inter-penetrates B by an amount $d_1 \approx d_2$, and the behavior of the configuration is that of Figure 10.2b. Thus, the initial amount of inter-penetration determines which behavior occurs.

The numerical methods we will use to compute forces and impulses do not

model inter-penetration. Instead of determining behavior by initial choice of inter-penetration, we can consider the normal force that existed at a contact point at a time t immediately before t_0 as the initial conditions at time t_0. These initial conditions are used to determine the behavior of bodies. This brings up the obvious question of what initial conditions to use for the first time step of a simulation, or after a discontinuous velocity change of bodies caused by a collision. The choice of initial conditions depends upon the application; if there is no basis for preferring one set of initial normal forces over another, the solution found by the numerical routines solving the contact force equations arbitrarily determines the behavior simulated. We will consider, in a moment, the validity of arbitrarily choosing a path for the simulation to follow. A much deeper question that we will not attempt to address is whether or not following only this one path to the end of the simulation reveals useful information.

12.3 Indeterminacy and inconsistency combined

Let us now consider a configuration that combines indeterminacy and inconsistency so that it is indeterminate whether or not an impulsive solution is required. Consider Figure 12.4. Once again, the combination of gravity and the angular velocity of R_1 is the same as in Figure 11.2 and Figure 11.3. Similarly, a horizontal acceleration of B results in inconsistency. If E_1 is ignored for the moment, then both $f_N = 0$ and $f_N = m_1$ are valid solutions for f_N. The only valid solution for the configuration as a whole though, is $f_N = 0$; a magnitude $f_N = m_1$ pushes B to the left, causing inconsistency.

If we subscribe to the physical model presented in this chapter, then it is clear that whether or not inconsistency occurs depends solely on the initial inter-penetration between R_1 and B. The implication of our physical model then

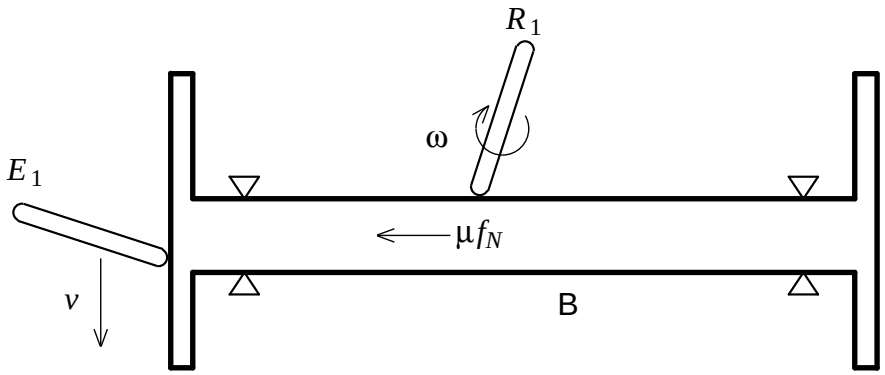

Figure 12.4: A configuration with a single impulse-free solution. The magnitude f_N must be either 0 or m_1 to be valid. However, $f_N = m_1$ causes inconsistency. The only valid solution is $f_N = 0$.

is the following: even though the configuration is consistent, that is, it has an impulse-free solution, there is no *a priori* reason to prefer this impulse-free behavior to the non-impulse-free behavior for this configuration. Stated differently, if we have no bias on the initial conditions of the configuration (that is, the initial inter-penetration distances), then the inconsistency resulting from $f_N = m_1$, and subsequent application of impulsive contact forces must be considered as acceptable a behavior as the application of non-impulsive contact forces resulting from $f_N = 0$. Even though the configuration in Figure 12.4 has only one valid solution of contact forces, ($f_N = 0$), it has two possible behaviors and is thus indeterminate.

12.4 Preferred solutions

By accepting the physical model based on the penalty method, we accept the fact that the two behaviors of Figure 12.4 must be considered equally valid. Let us contrast this model with a principle proposed by Kilmister and Reeve[25] called the *principle of constraints*. The principle states that when computing forces for

a configuration of bodies, impulsive forces should be used only if non-impulsive forces do not exist for the configuration. In essence, the principle of constraints is an assertion that non-impulsive solutions are preferable to solutions involving impulses. This principle seems (at first glance) sound; however, let us consider arguments against the principle of constraints.

From the discussion of the previous section, we can view the *NP*-hardness result of computing contact forces in a different light. By reasoning that solutions involving impulsive contact forces are acceptable even when a non-impulsive solution exists, we have extended the possible behaviors of configurations of bodies. (A formal definition of validity with regard to contact impulses has not been given yet, and will be presented in the next chapter.) If we subscribed to the principle of constraints however, we would first attempt to solve the LCP of Equation (8–7), in order to find a solution of non-impulsive contact forces. This is a search problem. Since some configurations (such as the configuration of Figure 11.3) may possess exponentially many impulsive solutions and only a single non-impulsive solution, it is not surprising that an attempt to search for this single solution leads to *NP*-hardness. Thus, the first argument we might advance against the principle of constraints is that by accepting a bias against solutions with impulsive forces, we are required to perform possibly an exponential amount of work. In contrast, disregarding this bias and accepting both impulsive and non-impulsive solutions turns out to require far less work (Chapter 13).

This is not really an argument, but more of a statement of the consequences of preferring non-impulsive solutions to impulsive solutions. A better argument is that the physical model proposed in this section, which is logically consistent with frictionless behavior, and explains both indeterminacy and inconsistency, has no such preference. In comparison, the principle of constraints is nothing more than a statement of the preference of non-impulsive behavior, without any justifying

explanation. Still, one might reject the physical model presented in this section, and argue from a rational mechanics points of view. That is, impulses have been added to our model of rigid body mechanics to deal with situations in which non-impulsive forces cannot prevent inter-penetration. The principle of constraints is therefore a good principle of constraints is preferable because it always seeks the "simplest" solution; that is, it only extends the axioms of rigid body mechanics to admit impulses when necessary.

One can refute this rational mechanics viewpoint very strongly with the following argument. Suppose that at time $t = 0$, we are faced with an indeterminate configuration, and suppose that there are two behaviors of the system past time zero. Let us call these behaviors of the systems *trajectories* A and B (Figure 12.5). The principle of constraints asserts that we must always avoid impulsive behavior, whenever possible. As we have seen, it is *NP*-hard to determine which trajectory (if either) of A or B avoids impulsive behavior at time zero. But suppose that neither trajectory A and B involves impulsive behavior at time zero, and that we arbitrarily choose A. If at time $t > 0$ along trajectory A we encounter inconsistency, what should we do? According to the principle of constraints, it would logical to back up our simulation to time $t = 0$ and consider trajectory B, for trajectory B might not require impulsive solutions for all $t > 0$. If so, then we see that in fact we really should have chosen trajectory B to start with, for by doing do, we avoid impulses forever, past time zero.

Clearly, this is enormously expensive, computationally, because it forces us to explore potential trajectories, which might themselves branch, to find the trajectory that is most free of impulses. The possible expense is for all practical purposes unbounded, since we can make the choice of the correct trajectory depend upon the outcome of some physical process which is arbitrarily hard to compute. Again, it can be argued that computational expense should not be a factor in rejecting

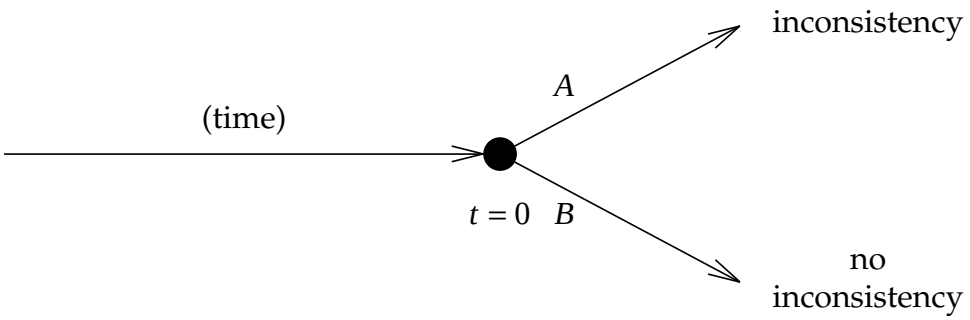

Figure 12.5: A configuration with two different non-impulsive solutions at time zero. If trajectory A is chosen, inconsistency is encountered for some later time $t > 0$, but if B is chosen, no inconsistency is encountered for $t > 0$.

the principle of constraints. However, suppose that an external time-varying force $F(t)$ acts during the course of the simulation. The decision as to whether trajectory A or B best avoids impulses past time zero would then require knowledge of the behavior of $F(t)$ for time $t > 0$. If we chose A or B based on $F(t)$ for $t > 0$, then we violate causality, since events at times $t > 0$ determine our course of action at time zero. In fact, we can create a paradox, by letting the force $F(t)$ be determined according to which trajectory is chosen, which is itself determined by $F(t)$. Thus, the principle of constraints in seeking to always use the smallest axiom set for rigid body mechanics, is an "unnatural" principle: it requires a physical system to be able to look forward in time to determine its action at the present instant.

The above argument clearly generalizes to any physical system with indeterminacy. Any principle that asserts that one sort of behavior is more preferable than another over the course of time leads to a violation of causality, and is thus an unnatural principle. That is not to say however that we cannot prefer one behavior for a particular *instant*. For instance, when we encounter an indeterminacy in a simulation, we might be told that one trajectory is more probable than another. If we regard indeterminacy as simply inadequate knowledge of a system's state to the

point that we cannot predict its actions, we might in fact be given a "tip" as to the correct behavior of the system by an oracular observer who has more knowledge of the system's state than we do. Although we should regard all trajectories as equally possible, they need not all be equally likely.

Given that we are willing to accept both impulsive and non-impulsive solutions as valid solutions for a configuration with dynamic friction, the following question arises. We know that searching for a particular type of solution, namely solutions without impulsive forces, is *NP*-hard. How difficult is it to find either a solution with or without impulsive forces? That is, how much time is required of an algorithm that computes either an impulsive or a non-impulsive solution for a configuration, without preference? Such an algorithm could not be used to decide if a configuration was consistent, since the algorithm could return an impulsive solution whether or not the configuration was consistent. The use of such an algorithm to compute contact forces (and thus choose an arbitrary trajectory for the simulation) would avoid the *NP*-complete problem of deciding consistency. In the next chapter, we will consider an algorithm with this property.

Chapter 13

Computing Valid Contact Forces and Impulses

Before computational methods can be considered, a definition of validity for contact impulses is needed. In order for a definition of validity to be useful, all inconsistent configurations should have a set of contact impulses satisfying the definition. An analysis of the computational method presented in Section 13.2 will show that the definition of validity presented satisfies this condition.

13.1 Defining valid contact impulses

In the penalty method interpretation of Figure 12.1 an impulse occurred because no matter how strong the normal force became, it was insufficient to prevent interpenetration. As a result, after the contact impulse was applied, the relative velocity of the bodies at the contact point was directed inwards. Since contact impulses may need to be applied to configurations involving more than one contact point, validity must be defined for a set of contact impulses. For example, in Figure 12.4, if the $f_N = m_1$ behavior is chosen, a contact impulse should occur between E_1 and B.

However, there should be no contact impulse between R_1 and B.

We will call a set of contact impulses valid under the following two conditions. The first condition is that the contact impulses must convert at least one of the contact points with dynamic friction to static friction. The reasoning here is that the impulse has occurred because the initial conditions of normal forces have caused inconsistency, as in Figure 12.4, and the inconsistency is removed by converting a contact point from dynamic to static friction.

The second condition is that every contact point at which a contact impulse occurs must end up with a non-positive relative normal velocity; that is, after the contact impulses are applied, bodies should *not* be separating wherever contact impulses occurred. The justification for this is that the contact impulses occur only when the normal force grows without bound to oppose inter-penetration. Intuitively, valid contact impulses are the limiting result of increasing normal forces without bound under the penalty method. If bodies are separating at a contact point after contact impulses are applied, then the normal force at the contact point should not have grown without bound into a contact impulse.[1]

[1] For planar configurations, this definition is valid, because a contact with dynamic friction cannot change its direction of sliding without having zero sliding velocity. For three-dimensional configurations, the sliding velocity direction can change without a contact changing from dynamic to static friction. Thus, for three-dimensional configurations, we would want to find contact impulses that satisfy the second condition. We would then need to differentially apply the contact impulses, and account for the change of sliding velocity direction; essentially the same problem as dealing with frictional impulses for three-dimensional collisions. Thus, the results of this chapter are valid for planar configurations, and are a starting point for three-dimensional systems. If one adopts the *ad hoc* assumption that the sliding velocity direction does not change during a collision for three-dimensional systems, then the results of this chapter can be applied without complication to three-dimensional systems.

13.2 Lemke's algorithm

Having defined validity for contact impulses, we can consider computational methods for computing contact forces and impulses. Given that computing contact forces alone is hard, how can either contact forces or impulses be computed efficiently? We begin by considering the LCP of Equation (8–7). One of the first algorithms for solving linear complementarity problems was introduced by Lemke[27] and is known as Lemke's algorithm. Lemke's algorithm is a pivoting method, similar to the simplex method of linear programming. The algorithm is exponential in the worst case, but has an expected running time polynomial in n. Lemke's algorithm progresses, like the simplex method, by trying various descent directions. If an LCP is PSD and has no solution then Lemke's algorithm will at some point encounter an *unbounded ray*; that is, a descent direction along which one can travel infinitely far without making any progress. Otherwise, if a PSD LCP has a solution, then no unbounded ray exists for that LCP, and Lemke's algorithm terminates by finding a solution to the LCP. The algorithm is viewed as a practical solution method to the problem of solving PSD LCP's.

However, for non-PSD LCP's, Lemke's algorithm is not guaranteed to terminate correctly (although it still takes only expected polynomial time to do so). For a non-PSD LCP, if there is no solution, Lemke's algorithm terminates by encountering an unbounded ray. Unfortunately, if there is a solution, the algorithm is not guaranteed to find it. For non-PSD LCP's with solutions, Lemke's algorithm terminates either by finding a solution or by encountering an unbounded ray.[2] As a result, Lemke's algorithm is not suitable for solving non-PSD LCP's.

However, whenever Lemke's algorithm terminates by encountering an unbounded

[2] Encountering an unbounded ray when there is a solution is roughly analogous to getting stuck at a non-global minimum in a non-convex minimization problem.

ray, it has found an n-vector $\mathbf{z}$ with the property

$$\mathbf{z} \geq \mathbf{0} \qquad \text{and} \qquad z_i > 0 \;\text{ implies }\; (\mathbf{Az})_i \leq 0 \qquad (1 \leq i \leq n) \tag{13–1}$$

where $(\mathbf{Az})_i$ is the ith component of the vector $\mathbf{Az}$. Why is this property of interest?

Suppose we let $\mathbf{z}$ represent a set of contact impulses in the same way that $\boldsymbol{f_N}$ represents a set of contact forces, and apply those contact impulses to the configuration. By the definition of impulse, the change in relative normal velocity at each contact point is $(A\mathbf{z})_i$. Since the relative normal velocity at each contact point is assumed to be zero before application of the impulse, the vector of relative normal velocities after applying the impulse is simply $\mathbf{Az}$.

Thus, the contact impulses indicated by $\mathbf{z}$ satisfy the second condition of validity: every contact point subject to a non-zero contact impulse $z_i > 0$ ends up with a non-positive relative normal velocity $(\mathbf{Az})_i \leq 0$. In order to completely satisfy the definition of validity, $\mathbf{z}$ must be scaled upwards from zero until it causes a contact point with dynamic friction to be converted to static friction. A subtle point to note is that if $\mathbf{z}$ specified contact impulses only at the *frictionless* contact points, then $\mathbf{z}$ could not satisfy the first condition condition for valid contact impulses (although it might satisfy the second). In order for $\mathbf{z}$ to represent valid contact impulses, $\mathbf{z}$ must have the property that for at least one $1 \leq i \leq n$, the ith contact point has dynamic friction and $z_i > 0$. A proof that Lemke's algorithm either solves the LCP or finds a vector $\mathbf{z}$ such that $\alpha\mathbf{z}$ represents valid contact impulses for some $\alpha > 0$ is given in Appendix B. After application of the contact impulse represented by $\alpha\mathbf{z}$, some real collisions (that is, collisions with positive normal approach velocity) occur and must be dealt with.

The behavior of Lemke's algorithm exactly matches our new view of the problem of computing contact forces. If the configuration has no valid contact impulse solutions, Lemke's algorithm cannot terminate with a ray $\mathbf{z}$ and must therefore

find a valid contact force solution. For inconsistent configurations, no valid contact force solution exists, so Lemke's algorithm must terminate with a ray $\mathbf{z}$, providing a contact impulse solution. For configurations with both a valid force and impulse solution, Lemke's algorithm will terminate by computing one or the other.

Note that the definition of valid contact impulses does not add any new solutions to frictionless systems. To see this, consider an arbitrary frictionless system, and suppose that a vector $\mathbf{z}$ satisfies the condition of Equation (13–1). Then

$$\mathbf{z}^T\mathbf{A}\mathbf{z} = \sum_{i=1}^{n} z_i(\mathbf{A}\mathbf{z})_i \leq 0. \tag{13–2}$$

But since the system is frictionless, $\mathbf{A}$ is PSD. Since a PSD matrix satisfies $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for any vector $\mathbf{x}$, Equation (13–2) implies that

$$\mathbf{z}^T\mathbf{A}\mathbf{z} = 0. \tag{13–3}$$

Since $\mathbf{A}$ is PSD, this in turn implies that $\mathbf{A}\mathbf{z} = \mathbf{0}$. Thus, any vector $\mathbf{z}$ that fits the first condition of valid contact impulses is a vector in the null-space of A. Since frictionless systems are determinate, given two vectors $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{y}$, application of the impulses (or forces) indicated by $\mathbf{x}$ has the same effect as the application of the impulses (or forces) indicated by $\mathbf{y}$. Since $\mathbf{A}\mathbf{z} = \mathbf{0}$, this implies that the application of the impulses indicated by $\mathbf{z}$ has absolutely no effect at all on the system. Thus, the definition of valid contact impulses is such that it adds no new solutions to frictionless systems.

Although Lemke's algorithm runs, practically speaking, in polynomial time, this is not a proof that finding either valid contact forces or impulses is a polynomial time problem. From a practical standpoint, though, Lemke's algorithm provides an efficient algorithm for computing valid contact forces or impulses. The computational complexity of either solving an LCP or finding a vector $\mathbf{z}$ satisfying Equation (13–1) is unknown; but it would be surprising if this could not be done in polynomial time.

13.3 Continuity of solutions

One implementation issue we need to be concerned with is the continuity of the solutions chosen for $\boldsymbol{f_N}$. Consider a configuration with two possible solutions $\boldsymbol{f_N}$ that produce different accelerations of the bodies. Suppose that we (perversely) alternate between the two solutions at each time step, causing the accelerations of the bodies to be discontinuous. This will cause the ordinary differential equation solver to break down, since it expects to be integrating a continuous function.

In Section 8.1 where configurations were frictionless, this problem did not occur because all solutions of Equation (8–7) produced the same acceleration. When friction is introduced this is no longer true, and care must be taken to make $\boldsymbol{f_N}$ (at least) piecewise continuous over time. Recall that Section 8.2 introduced the concept of computing $\boldsymbol{f_N}$ by solving a non-singular linear system of equations as indicated by an index set L. This computation method for $\boldsymbol{f_N}$ may be regarded as a time saving luxury for frictionless configurations, but it is essential in simulating configurations with friction. Since $\mathbf{A}$ and $\mathbf{b}$ depend on spatial and velocity variables, they are continuous functions of time (in the absence of impulses). As a result, the system of equations produced from L is continuous as long as L remains constant. Because of this, the $\boldsymbol{f_N}$ vectors computed with this method are piecewise continuous, with discontinuities occurring whenever L changes. The integrator is of course informed whenever L changes.

Chapter 14

Static Friction

The results in this thesis concerning the complexity of static friction are much less comprehensive than the results for dynamic friction. We will indulge in some speculation concerning complexity results for static friction, and then discuss some methods for approximating static friction forces for simulation purposes. The approximation methods are suitable for some applications (such as computer animation), but may be unsuitable with regards to other applications.

14.1 Complexity of static friction

Consider once more the ith contact point of a configuration. The normal force and tangential friction force magnitudes at this point are f_{Ni}, f_{xi} and f_{y_i}. The corresponding acceleration magnitudes are a_{Ni}, a_{xi} and a_{y_i}. Since we already know that the presence of dynamic friction leads to *NP*-hardness, let us study configurations with only static friction and no dynamic friction.

In Chapter 9, the conditions for static friction were presented. These conditions

can be written as

$$\left\{\begin{array}{c} f_{N\,i} \geq 0 \text{ and} \\ a_{N\,i} \geq 0 \text{ and} \\ f_{N\,i} a_{N\,i} = 0 \text{ and} \\ f_{x\,i}^2 + f_{y_i}^2 \leq (\mu f_{N\,i})^2 \end{array}\right. \quad \text{and either} \quad \left\{\begin{array}{c} a_{x\,i} = a_{y_i} = 0 \text{ or} \\ f_{x\,i}^2 + f_{y_i}^2 = (\mu f_{N\,i})^2 \text{ and} \\ f_{x\,i} a_{x\,i} + f_{y_i} a_{y_i} \leq 0. \end{array}\right. \tag{14–1}$$

The first questions we might ask are: does Equation (14–1) always have a solution? Does Equation (14–1) always have a unique solution? Lötstedt[28] gives an example of a one contact point configuration with static friction that does not have a unique solution. The example given is the sliding rod of Figure 10.2, but with the linear velocity adjusted so that the tip of the rod is motionless with respect to the ground. Once again, the rod is free to either rotate off the ground or remain in contact. Thus, Equation (14–1) does not necessarily have a unique solution. Whether Equation (14–1) always has even one solution is unknown. So far, no examples of configurations with static friction resulting in inconsistency are known (excepting the trivial case when the configuration has unsatisfiable kinematic constraints). This is particularly striking in light of the fact that inconsistency for dynamic friction can occur for one contact point configurations. For static friction, we can however prove that every one contact point configuration has a solution. The proof technique makes use of geometrical constructs based on work by Wang and Mason[34].

THEOREM 4. *All two-dimensional one contact point configurations with static friction have valid solutions.*

PROOF. For two dimensional configurations, we can describe the frictional force and the tangential accelerations in terms of just $f_{x\,i}$ and $a_{x\,i}$. The analogue of

Equation (14–1) in two dimensions is therefore

$$\left\{\begin{array}{c} f_{Ni} \geq 0 \text{ and} \\ a_{Ni} \geq 0 \text{ and} \\ f_{Ni} a_{Ni} = 0 \text{ and} \\ -\mu f_{Ni} \leq f_{xi} \leq \mu f_{Ni} \end{array}\right. \quad \text{and either} \quad \left\{\begin{array}{c} a_{xi} = 0 \text{ or} \\ |f_{xi}| = \mu f_{Ni} \text{ and} \\ f_{xi} a_{xi} \leq 0. \end{array}\right. \tag{14–2}$$

For a one contact configuration, we can express the normal acceleration a_N and the frictional acceleration a_x in terms of the normal force f_N and the frictional force f_x as

$$\begin{aligned} a_N &= B f_N + C f_x + F \\ a_x &= C f_N + E f_x + G \end{aligned} \tag{14–3}$$

where B, C, D, and E satisfy[28]

$$B > 0, \quad E > 0 \qquad \text{and} \qquad BE > C^2. \tag{14–4}$$

Additionally, we can assume $C < 0$ by reversing the sense of the tangent direction along which f_x and a_x are measured, if necessary. (If $C = 0$ then Equation (14–3) decouples and it is trivial to show Equation (14–2) always has a solution.) From this, it is easily shown that

$$\frac{-E}{C} > \frac{-C}{B}. \tag{14–5}$$

Using these facts, we show that Equation (14–2) always has a solution.

If $F \geq 0$ then Equation (14–2) has the trivial solution $f_N = f_x = 0$. Otherwise, $F < 0$. In this case, we claim a solution with $f_N > 0$, $a_N = 0$ and one of the following three cases always exists:

(i) $a_x = 0$ and $-\mu f_N \leq f_x \leq \mu f_N$

(ii) $a_x > 0$ and $f_x = -\mu f_N$

(iii) $a_x < 0$ and $f_x = \mu f_N$.

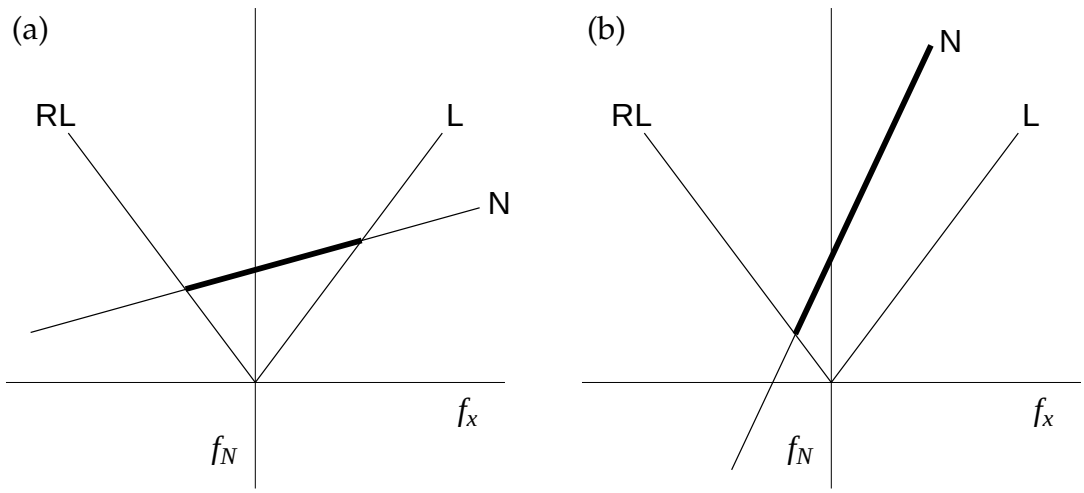

Figure 14.1: Force space. (a) Possible solutions for P lie on N between L and RL. (b) Possible solutions for P lie on N to the right of N's intersection with RL.

To verify this claim, we use a geometrical method based on the impulse space described by Wang and Mason[34].

Consider Figure 14.1 where f_N is plotted vertically and f_x is plotted horizontally. We are searching for a point $P = (f_N, f_x)$ in this space that satisfies both Equation (14–2), and $f_N > 0$ and $a_N = 0$. Several lines are identified in this space. First, the line $Bf_N + Cf_x + F = a_N = 0$ is labeled N. Next, the lines $f_x = \mu f_N$ and $f_x = -\mu f_N$ are labeled L and RL respectively. Since $F < 0$, $B > 0$ and $C \leq 0$, N is either horizontal or has positive slope and crosses the f_N axis above the fx axis. Because of this, N always intersects RL and may (Figure 14.1a) or may not (Figure 14.1b) intercept L. In order for a point P to satisfy $f_N > 0$, $a_N = 0$ and $-\mu f_N \leq f_x \leq f_N$, P must lie on N above the horizontal axis, and on or between L and RL. Points satisfying these conditions are indicated by the bold portion of line N in Figure 14.1.

Suppose now that N crosses both L and RL, and consider Figure 14.2. The line

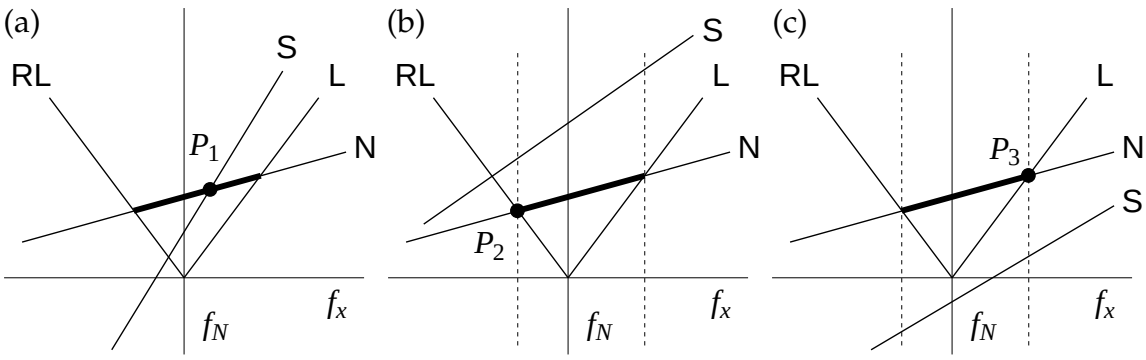

Figure 14.2: Line N intersects L and RL. (a) S intersects N and point P_1 satisfies case (i). (b) S lies above N between L and RL. Points lying below S have $a_x > 0$; point P_2 satisfies case (ii). (c) S lies below N between L and RL. Points lying above S have $a_x < 0$; point P_3 satisfies case (iii).

$Cf_N + Ef_x + G = a_x = 0$ is labeled S. Any point P on S results in zero tangential acceleration. From the definition of S and since $C < 0$, points lying *above* S satisfy $a_x < 0$ while points lying *below* S satisfy $a_x > 0$. Suppose S intersects N between L and RL (Figure 14.2a). Then the intersection point P_1 falls under case (i) and is a solution. If however S does not intersect N between L and RL, then either S lies above N as in Figure 14.2b, or below N as in Figure 14.2c. For Figure 14.2b, the point P_2 is a solution under case (ii) since it satisfies $f_x < 0$ and $a_x > 0$. In Figure 14.2c, P_3 satisfies $f_x > 0$ and $a_x < 0$ and is a solution under case (iii).

The only situation left to consider is if N intersects only RL and not L (Figure 14.3). Suppose that S lies above N's intersection with RL (Figure 14.3a). Since points below S satisfy $a_x > 0$, the point P_4 is a solution under case (ii). Otherwise, S lies below N's intersection with RL (Figure 14.3b). Since S has slope $-E/C$ and N has slope $-C/B$, from Equation (14–5), S has a steeper slope. Therefore, S must intersect N somewhere between L and RL at P_5; this solution

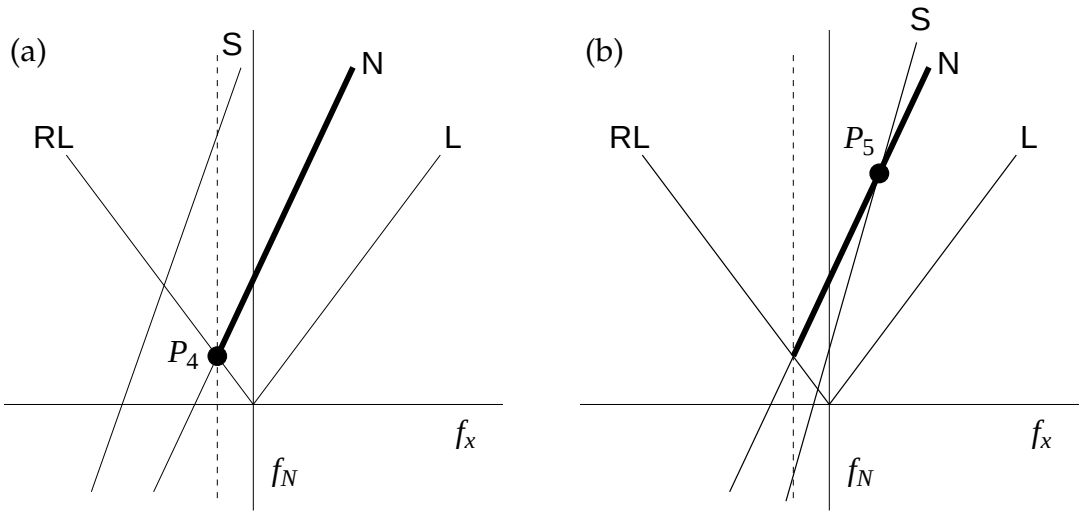

Figure 14.3: Line N intersects RL, but not L. (a) S lies above the intersection of N and RL; point P_4 satisfies case (ii). (b) S lies below the intersection of N and RL. Since S's slope is greater than N's, S must intersect N somewhere between L and RL. The intersection point, P_5 satisfies case (i).

falls under case (i).

Having verified our claim that that for $F < 0$ one of cases (i), (ii) or (iii) always holds, we conclude that all two-dimensional one contact point configurations with static friction have valid solutions. □

This result is easily extendible to three dimensions. The same geometric intuition is used, except that N and S are planes and the lines L and RL are replaced with the cone $f_x^2 + f_y^2 = (\mu f_N)^2$.

We can speculate that Equation (14–1) does in fact always have a solution. Having shown that all one contact point configurations with static friction have a solution, we can contemplate an inductive proof that all n contact point configurations with static friction have a solution. Theorem 4 would be the base case, and the inductive step might be as follows. Assume all n contact point configurations with friction have solutions, and consider an $n + 1$ contact point configuration.

Let one contact point be arbitrarily called the "new" contact point. We will insert an external force at this new contact point that acts equally and opposite on the contacting bodies at this point. Henceforth we ignore this new contact point, and consider the configuration to be simply an n contact point configuration. For any external force we insert at the new contact point, there is a valid solution of contact forces at the other n contact points, since all n contact point configurations have solutions.

The idea then is that we slowly increase the strength of the external force acting at the new contact point. As we do this, the contact forces at the other n contact points continually readjust to form a valid solution. Eventually, we should reach a point where the external force at the new contact point is just sufficient to prevent inter-penetration without being strong enough to cause separation at the new contact point. At this point, if we treat the external force as a contact force at the new contact point, we will have a valid solution for the $n+1$ contact point configuration. This will complete the inductive step. However, the above is only speculation; there are still some unresolved difficulties associated with this proof.

What can we actually say about the complexity of computing contact forces? First, note that the problem of computing contact forces for two-dimensional configurations with friction lies in *NP*. Just as any solution to the LCP of Equation (8–7) could be written in terms of the solution of a linear system, the same is true for a solution to Equation (14–2). As a result, any solution of Equation (14–2) must be rational and of polynomial length[52]. Thus, finding a solution to Equation (14–2) lies in *NP* since a solution to the equation can be verified in polynomial time.

Unfortunately, for three-dimensional configurations, the Coulomb friction constraint $f_{x\,i}^2 + f_{y_i}^2 \leq (\mu f_{N\,i})^2$ brings up the possibility that a solution might involve irrational quantities. We therefore cannot say that finding a solution to Equation (14–1) lies in *NP*. Practically speaking however, we would not expect

finding a solution to Equation (14–1) to be any harder than finding a solution to Equation (14–2). Laying aside the fact that solving Equation (14–1) might not lie in *NP*, can we at least express solutions to Equation (14–1) in terms of a convex minimization problem, as was done for frictionless configurations? Convex minimization problems, even when not lying in *NP*, usually admit practical solutions by some sort of descent procedure. Even if we could not exactly classify the difficulty of solving Equation (14–1), phrasing it as a convex minimization problem would go a long way towards describing the inherent complexity. The answer to this question is, unfortunately, "no". The configuration described by Lötstedt[28] that does not possess a unique solution has in fact exactly two distinct solutions. Since the solution set of a convex minimization problem is always a single connected component, the solution set of Equation (14–1) cannot be posed in terms of a convex minimization problem.

Should we believe then that computing contact forces for configurations with static friction is *NP*-hard? We cannot say, but we can again speculate. If we were able to find a configuration that had no solution, then we could probably form a reduction similar to the reduction of Chapter 11 to show that computing contact forces is *NP*-hard. Suppose however that our speculation that configurations with static friction always have a solution is correct. As previously mentioned, computing contact forces for two-dimensional configurations lies in *NP*. If all configurations with static friction have solutions, the decision problem of whether or not a two-dimensional configuration with static friction has a solution is trivially in P (polynomial time), since the answer is always "yes". Would this mean that contact forces could also be computed in polynomial time?

So far, an example of an *NP* decision problem lying in P that requires more than polynomial time to actually find a witness is unknown. Indeed the existence of such a problem would constitute a proof that $P \neq NP$, the reasoning being that

if $P = NP$, all NP witness-finding problems can be solved in polynomial time. If in fact all configurations with static friction had solutions, showing that contact forces could not be computed in polynomial time would be equivalent to showing that $P \neq NP$. Thus, if all configurations had solutions, it would probably be more fruitful to attempt to construct a polynomial time algorithm for computing contact forces than to attempt to disprove the existence of such an algorithm.

14.2 Approximation methods

We will conclude our discussion of static friction by considering three different methods for approximating static friction forces. The first method is due to Lötstedt[30] and appears to be the only published approximation method for contact forces with static friction that does not involve the penalty method. The second two methods are new. For all three methods, we express tangential and normal accelerations as linear functions of the contact forces by writing

$$\mathbf{a} = \mathbf{A}\boldsymbol{f} - \mathbf{b} \tag{14–6}$$

where

$$\mathbf{a} = (a_{N1}, a_{x1}, a_{y_1}, \ldots, a_{Nn}, a_{xn}, a_{y_n})^T \tag{14–7}$$

and

$$\boldsymbol{f} = (f_{N1}, f_{x1}, f_{y_1}, \ldots, f_{Nn}, f_{xn}, f_{y_n})^T. \tag{14–8}$$

The matrix $\mathbf{A}$ is a $3n \times 3n$ matrix and $\mathbf{b}$ is a $3n$-vector.

14.2.1 Lötstedt's approximation

Lötstedt's approximation is a very simple modification of the Coloumb law of friction, which very cleverly achieves two difficult goals. First, it allows dynamic

friction problems to be solved by convex QP's or LCP's. Second, it converts the static friction conditions into a form that is also solvable by quadratic programming; moreover, the quadratic program produced is convex. This is a very elegant transformation!

For dynamic friction, Lötstedt sets the dynamic friction force to have a magnitude of μ times the normal force magnitude *of the previous time step*. This makes the dynamic friction force a known external force at the current time step. Since only unknown normal forces and known external forces act on the bodies now, the configuration has a convex LCP. Thus, Lötstedt's modification does away with both indeterminacy and inconsistency, and produces an LCP that can be efficiently solved (since the LCP is convex).

For static friction, Lötstedt sets the upper bound on the magnitude of the static friction force to μ times the normal force magnitude of the previous time step. This allows static friction and normal forces to be solved by quadratic programming in the following manner. For a given time step, let $\tilde{f}_{N\,i}$ denote the normal force at the ith contact point from the *previous* time step. Lötstedt's modification produces a system of equations of the form

$$\left\{\begin{array}{c} f_{N\,i} \geq 0 \text{ and} \\ a_{N\,i} \geq 0 \text{ and} \\ f_{N\,i} a_{N\,i} = 0 \text{ and} \\ -\mu \tilde{f}_{N\,i} \leq f_{x\,i} \leq \mu \tilde{f}_{N\,i} \text{ and} \\ -\mu \tilde{f}_{N\,i} \leq f_{y\,i} \leq \mu \tilde{f}_{N\,i} \text{ and} \\ a_{x\,i}(\mu \tilde{f}_{N\,i} - |f_{x\,i}|) = 0 \quad \text{and} \quad f_{x\,i} a_{x\,i} \leq 0 \text{ and} \\ a_{y_i}(\mu \tilde{f}_{N\,i} - |f_{y_i}|) = 0 \quad \text{and} \quad f_{y_i} a_{y_i} \leq 0. \end{array}\right. \tag{14–9}$$

The approximation is appealing because it still enforces a very important property of static friction: the friction force achieves its upper bound and is opposite the tangential acceleration whenever it is unable to prevent sliding. Still, the above

conditions still appear fairly complex, and certainly not the conditions of a quadratic program.

It turns out however that the QP

$$\underset{\boldsymbol{f}}{\text{minimize}} \ \tfrac{1}{2}\boldsymbol{f}^T\mathbf{A}\boldsymbol{f} - \mathbf{b}^T\boldsymbol{f} \quad \text{subject to} \quad \left\{ \begin{array}{c} f_{N\,i} \geq 0 \\ -\mu\tilde{f}_{N\,i} \leq f_{x\,i} \leq \tilde{f}_{N\,i} \\ -\mu\tilde{f}_{N\,i} \leq f_{y_i} \leq \tilde{f}_{N\,i} \end{array} \right\} \qquad (14\text{–}10)$$

has KKT conditions that match the conditions of Equation (14–9), so a solution of Equation (14–10) satisfies the conditions of Equation (14–9). Thus, the fairly complex conditions of Equation (14–9) have the highly desirable property of being solvable by quadratic programming. (Note that Equation (14–9) does not appear to be solvable as an LCP though.) Unfortunately though, when $\mathbf{A}$ is not symmetric, Equation (14–10) has an entirely different set of KKT conditions that do not match Equation (14–9). For configurations with dynamic friction, even if $\mathbf{A}$ is still PSD (as often happens when the frictional coefficient μ is small) it is rarely (if ever) symmetric. This means that we cannot easily extend this approximation to handle the unmodified dynamic friction law, because dynamic friction causes $\mathbf{A}$ to become asymmetric.

Additionally, there is the question of the decoupling of the $\hat{t}_x$ and $\hat{t}_y$ axes, and the replacement of $f_{N\,i}$ with $\tilde{f}_{N\,i}$. The effects of decoupling of the axes can possibly be minimized by a wise choice of vectors $\hat{t}_x$ and $\hat{t}_y$ for each contact point (so that one of $f_{x\,i}$ or f_{y_i} is near zero, for each contact point), but the replacement of $f_{N\,i}$ with $\tilde{f}_{N\,i}$ is more serious. It is unclear how one produces a suitable $\tilde{f}_{N\,i}$ for the initial time step (or after a discontinuity in the configuration's state). One possibility is to pick arbitrary initial values for the $\tilde{f}_{N\,i}$ (such as zero), solve the system, and update $\tilde{f}_{N\,i}$ with the newly computed $f_{N\,i}$, and iterate. However, this is method will not necessarily converge; for the one-point configuration of Figure 10.1, the magnitude $f_{N\,i}$ will in fact diverge to infinity. Likewise, consider Figure 10.1 with a reversed

tangential velocity, so that the friction force is in effect accelerating the contact point p_a "upwards". Suppose $\tilde{f}_{Ni}$ is set to zero, and f_{Ni} is solved for; on the next iteration, we set $\tilde{f}_{Ni} = f_{Ni}$. Depending on the value of g and the value of θ, this new value of $\tilde{f}_{Ni}$ could be sufficiently large that it causes contact between the rod and the ground to break. If this happens, the new value of f_{Ni} will be zero, since contact is breaking, and the following iteration will start with $\tilde{f}_{Ni}$ set to zero as well. The iterative process then continues, with f_{Ni} and $\tilde{f}_{Ni}$ oscillating between two distinct values.

It is difficult to quantitatively describe the effects of the approximation. In situations where neither indeterminacy nor inconsistency occurs, perhaps the approximation has little effect on the outcome of a simulation. As a trivial example, if we set $\mu = 0$, then the approximation has no effect at all. We can imagine that as we slowly increase μ, the effects of the approximation might not be apparent until μ takes on a suitably large value. But for situations with inconsistency, Lötstedt's approximation produces a non-impulsive solution where such a solution did not exist previously. In this situation, the approximation must clearly have a large effect on the simulation.

14.2.2 Dynamic friction approximation

The second approximation method we consider has the following virtues. It is extremely simple to implement, very robust, and handles dynamic friction. The approximation is similar in concept to the penalty method of computing normal forces based on inter-penetration depth. It avoids some of the difficulties that make the penalty method unsuitable as a simulation method, but it does share the major flaw of all penalty methods: it is an *ad hoc* description of a physical process.

The second method is as follows. In order to determine whether static friction or

dynamic friction should occur at a contact point, the simulation program must have some threshold value ϵ. If the tangential velocity $(v_{xi}^2 + v_{y_i}^2)^{1/2} \geq \epsilon$, then dynamic friction occurs. Otherwise static friction occurs. Since the dynamic friction force is calculated by

$$f_{xi} = \mu f_{N\,i} \frac{-v_{xi}}{\sqrt{v_{xi}^2 + v_{y_i}^2}} \qquad \text{and} \qquad f_{y_i} = \mu f_{N\,i} \frac{-v_{y_i}}{\sqrt{v_{xi}^2 + v_{y_i}^2}} \tag{14–11}$$

we approximate static friction as

$$f_{xi} = \frac{\sqrt{(v_{xi}^2 + v_{y_i}^2)}}{\epsilon} \mu f_{N\,i} \frac{-v_{xi}}{\sqrt{v_{xi}^2 + v_{y_i}^2}} = \mu f_{N\,i} \frac{-v_{xi}}{\epsilon} \tag{14–12}$$

and similarly for f_{y_i}. Thus, we approximate static friction as a dynamic friction force that varies in magnitude from zero to an upper limit of $\mu f_{N\,i}$ as the tangential speed varies from 0 to ϵ. The method of Section 13.2 can be used to compute the contact forces, since this is now entirely a dynamic friction problem.

In this approximation however, static friction occurs only when the tangential velocity is non-zero. Bodies must acquire some small amount of "crawl" in order to maintain a static friction force. This is similar to the penalty method, where bodies must acquire some degree of inter-penetration for a sufficient normal force to exist. However, in the penalty method, it is necessary to increase the spring constant K (Chapter 12) without bound as the mass of bodies increases. Our approximation method does not suffer from this problem. If ϵ is made small enough, the "crawling" behavior of bodies is not visible, no matter what masses or forces or exist. If ϵ is made excessively small, the differential equations of motion may become stiff; otherwise, the method has a reasonable performance. The graphical simulations produced using this approximation are fairly realistic.

14.2.3 Quadratic programming approximation

The last method we present attempts to model static friction by quadratic programming. The basic idea is that by guessing the signs of the variables f_{xi} and f_{y_i} in a solution, we can solve for $\boldsymbol{f}$ by solving a QP. Given a guess of the values

$$\text{sgn}(f_{xi}) \qquad \text{and} \qquad \text{sgn}(f_{y_i}) \tag{14–13}$$

where $\text{sgn}(x) = 1$ if $x \geq 0$ and -1 otherwise, we will try to find a solution $\boldsymbol{f}$ that agrees with these signs. Thus, we treat $\text{sgn}(f_{xi})$ and $\text{sgn}(f_{y_i})$ as known and fixed quantities for now. We will describe later how we produce a guess of these values, and then discuss some shortcomings of the method.

Because we cannot represent the constraint $f_{xi}^2 + f_{y_i}^2 \leq (\mu f_{Ni})^2$ in a QP, we decouple the $\hat{t}_x$ and $\hat{t}_y$ axes as Lötstedt does, replacing $f_{xi}^2 + f_{y_i}^2 \leq (\mu f_{Ni})^2$ with

$$-\mu f_{Ni} \leq f_{xi} \leq \mu f_{Ni} \qquad \text{and} \qquad -\mu f_{Ni} \leq f_{y_i} \leq \mu f_{Ni}. \tag{14–14}$$

Note however that for two-dimensional configurations, no decoupling of axes is needed and this method does not making *any* approximations of the laws of static friction.

We will also replace the condition $a_{xi} f_{xi} + a_{y_i} f_{y_i} \leq 0$ with the conditions

$$\begin{aligned} f_{xi}\,\text{sgn}(f_{xi}) \geq 0 \qquad &\text{and} \qquad a_{xi}\,\text{sgn}(f_{xi}) \leq 0 \\ f_{y_i}\,\text{sgn}(f_{y_i}) \geq 0 \qquad &\text{and} \qquad a_{y_i}\,\text{sgn}(f_{y_i}) \leq 0. \end{aligned} \tag{14–15}$$

This condition ensures that $a_{xi} f_{xi} + a_{y_i} f_{y_i} \leq 0$ (and is actually somewhat stronger). Note that the (seeming) tautologies $f_{xi}\,\text{sgn}(f_{xi}) \geq 0$ and $f_{y_i}\,\text{sgn}(f_{y_i}) \geq 0$ are actually constraints that f_{xi} and f_{y_i} have the same signs as $\text{sgn}(f_{xi})$ and $\text{sgn}(f_{y_i})$ (which we are regarding as known quantities). Next, we add the standard constraint that

$$f_{Ni} \geq 0 \qquad \text{and} \qquad a_{Ni} \geq 0. \tag{14–16}$$

Since we are treating $\mathrm{sgn}(f_{x_i})$ and $\mathrm{sgn}(f_{y_i})$ as known quantities, and $a_{N\,i}$, $a_{x\,i}$ and a_{y_i} are linear functions of the contact forces, all of the above conditions can be expressed as linear constraints involving $f_{x\,i}$, f_{y_i} and $f_{N\,i}$.

The condition that the static friction force attains its upper bound when slipping begins is written as

$$\begin{aligned}(\mu f_{N\,i} - f_{x\,i}\,\mathrm{sgn}(f_{x\,i}))(-a_{x\,i}\,\mathrm{sgn}(f_{x\,i})) = 0 \\ (\mu f_{N\,i} - f_{y_i}\,\mathrm{sgn}(f_{y_i}))(-a_{y_i}\,\mathrm{sgn}(f_{y_i})) = 0.\end{aligned} \tag{14–17}$$

Finally, we add the standard constraint on the normal forces that

$$(f_{N\,i})(a_{N\,i}) = 0. \tag{14–18}$$

Since each of $f_{N\,i}$, $a_{N\,i}$, $\mu f_{N\,i} - f_{x\,i}\,\mathrm{sgn}(f_{x\,i})$, $\mu f_{N\,i} - f_{y_i}\,\mathrm{sgn}(f_{y_i})$, $-a_{x\,i}\,\mathrm{sgn}(f_{x\,i})$, and $-a_{y_i}\,\mathrm{sgn}(f_{y_i})$ are constrained to be positive, we can express the fact that Equations (14–17) and (14–18) hold for all i as

$$\sum_{i=1}^{n} \left(\begin{array}{c} f_{N\,i} a_{N\,i} + (\mu f_{N\,i} - f_{x\,i}\,\mathrm{sgn}(f_{x\,i}))(-a_{x\,i}\,\mathrm{sgn}(f_{x\,i})) \\ +(\mu f_{N\,i} - f_{y_i}\,\mathrm{sgn}(f_{y_i}))(-a_{x\,i}\,\mathrm{sgn}(f_{y_i})) \end{array} \right) = 0. \tag{14–19}$$

Since this is a quadratic function of $f_{x\,i}$, f_{y_i} and $f_{N\,i}$, we can find a solution to Equation (14–19) satisfying the linear constraints of Equations (14–14), (14–15) and (14–16) by solving a QP.

We still must describe how to guess the signs of $f_{x\,i}$ and f_{y_i}. Iterative methods for quadratic programming exist that are very similar to the Gauss-Seidel or Jacobi iterative methods used to solve linear systems[40]. These iterative routines are much simpler than routines that solve QP's by direct methods; in particular, it is extremely simple to modify these iterative routines to find a solution to Equation (14–1). The reason that we do not use these modified iterative routines is that they have poor convergence properties. However, we can run one of these modified iterative routines for some period of time and examine the (hopefully) partially converged variables

f_{xi} and f_{y_i} to guess their signs. (Appendix C discusses how to formulate an iterative routine to solve Equation (14–1).) Using the signs of the variables, we then form and solve a quadratic program as described above. Coherence can be exploited (and must, to provide continuity over time steps) by using the same concept of basic solutions presented in sections 8.2 and 13.3.

This method can break down at any number of places. First, we still do not know that a solution for static friction always exists. Even if there is a solution, it is possible that the iterative algorithm may not converge to a solution, so our guess of the signs of f_{xi} and f_{y_i} may not form a QP that leads to a solution. Furthermore, the QP may be non-convex, and therefore impractical to solve (even if only static friction occurs). Last, if we attempt to incorporate both dynamic and static friction, we still need to be concerned about the *NP*-hardness due to just the dynamic friction. Unfortunately, the form of the linear constraints of the QP do not allow us to solve it as an LCP using Lemke's method. No quadratic programming algorithm is known that has the property of Lemke's algorithm: returning a valid non-impulsive solution or indicating a valid impulsive solution. It is possible that whenever the iterative algorithm fails to converge, it does so in such a way that an analysis of the divergence indicates a valid set of contact impulses; it is not clear how to perform such an analysis.

However, this last method yields a very acceptable visual result, when it works. For large numbers of contact points ($n \approx 40$), the last method sometimes breaks down; either the iteration does not converge, or it converges to an incorrect solution. For configurations with approximately 40 contact points or less, the method is reasonably successful. Using this method, simulations with $n \approx 40$, and involving thousands of time steps have been run successfully. Commented out for TechReport version. Images from several such simulations are shown in Appendix D.