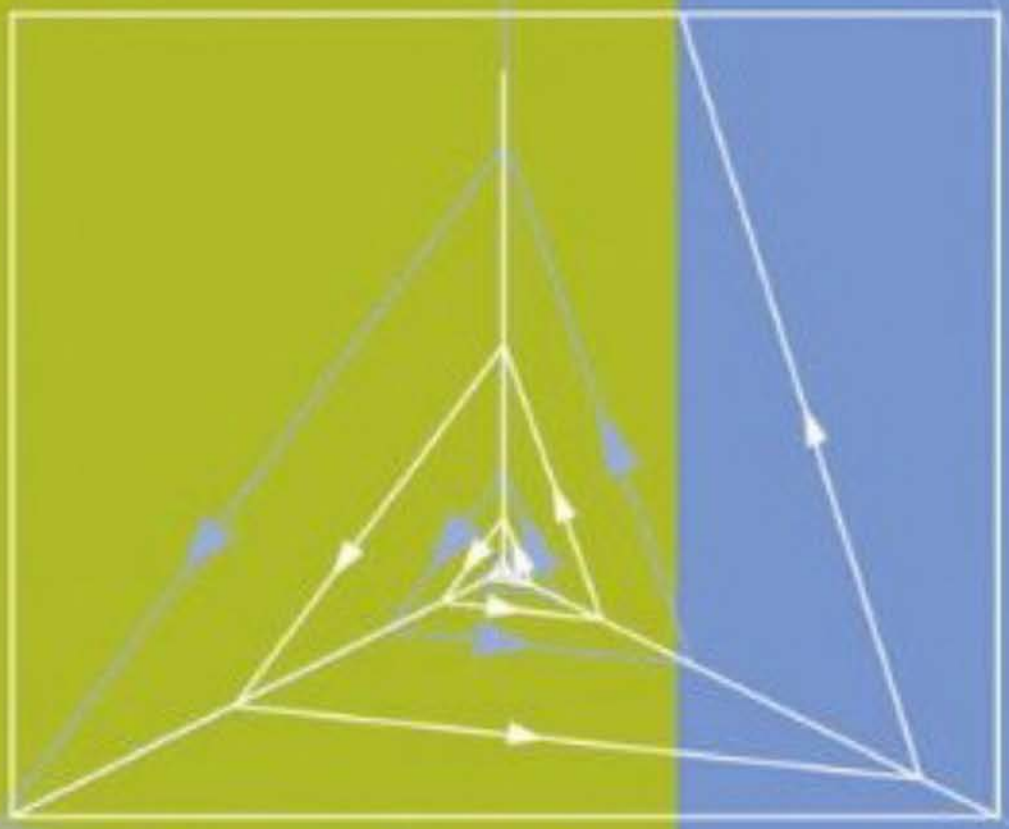
Dynamics with Inequalities

Impacts and Hard Constraints



David E. Stewart

Dynamics with Inequalities

Dynamics with Inequalities
Impacts and Hard Constraints

David E. Stewart

University of Iowa
Iowa City, Iowa

siam®

Society for Industrial and Applied Mathematics
Philadelphia

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data
Stewart, David, 1961-
Dynamics with inequalities : impacts and hard constraints / David E. Stewart.
p. cm. -- (Applied mathematics)
Includes bibliographical references and index.
ISBN 978-1-611970-70-8 (pbk. : alk. paper) 1. Dynamics. I. Title.
QA851.S84 2011
512.9'7--dc22

2011007832

Contents

Preface

Hard constraints are represented by inequalities. Hard constraints occur frequently in practical systems and their models; these hard constraints are often used in optimization since optima are frequently found at these hard limits. In optimal control theory, this can be seen in the prevalence of "bang-bang" solutions—the "bang" represents a control at a hard limit.

In spite of this, hard limits are eschewed in most dynamical models. There are a number of reasons for this. One is the lack of a suitable or "nice" theory for such systems. Another is that numerical methods do not handle this situation well. A third is that it is often not clear what should happen in a differential equation when a hard limit is reached.

Most commonly, when researchers in the sciences, engineering, or economics come across a dynamic system with a hard constraint, the usual instinct is to smooth out the hard constraint, often by adding a "penalty" term for either violating, or approaching, the hard constraint. There are several reasons why this is not necessarily wise:

- This complicates an otherwise simple model, and furthermore, the strength of the penalty is a new parameter that should somehow be calibrated to fit the situation. And if the calibration indicates that the value should be zero (or infinity) for all practical purposes, then we are back to a system with a hard constraint.

- Numerically solving a penalized differential equation with a small penalty term (so as to well approximate a hard constraint) results in a stiff differential equation, which must be solved either with extremely small step sizes or using implicit methods that require the solution of nonlinear equations that are nearly equivalent to the hard constraint.

This book aims to fill this gap: hard limits are natural models for many dynamic phenomena, and there are ways of creating "differential equations with hard constraints" that are natural and provide accurate models of many physical, biological, and economic systems. The models that are described here have roots in optimization theory, and so we will see Lagrange multipliers and complementarity principles not only as methods to enable us to minimize functions subject to constraints but also as ways of formulating dynamic models to systems with hard constraints.

A central idea here is the idea of *index*. This represents the number of differentiations between a hard constraint and the state variables of the dynamic system. The higher the index, the more difficult it is to solve. This index is closely related to the index used in the area of *differential algebraic equations* (DAEs), which can be seen as differential equations with *equality constraints*. This will provide an organizing principle for much of this book.

Connections to related dynamical systems with hard constraints will be mentioned, such as linear complementarity systems (LCSs), projected dynamical systems (PDSs), differential inclusions, and more general concepts such as hybrid systems and variable structure systems. In contrast to hybrid and variable structure systems, the systems described here are more limited, but they do not suffer from the same theoretical limbo where solutions might or might not exist or be meaningful depending on the precise structure of the system.

Differential variational inequalities (DVIs) form a focal point in this book. These are not the most general class of dynamics that can represent hard constraints or impacts. Differential inclusions (also discussed) can be more general. But with such great generality it becomes difficult to turn the abstract formulation into a practical or computationally useful form. DVIs, on the other hand, provide a general means both of modeling and of carrying out computations. The connections with more abstract theories, such as differential inclusions, are fleshed out so that the reader can compare the properties and strengths of the two means of modeling such systems.

I have tried to make this book mathematically self-contained, so that it is accessible to engineers, economists, and others with a strong mathematical background. The material contained in this book, however, does involve considerable technical development, especially for problems involving partial differential equations, as these involve spaces of functions with infinite dimensions. Results are given which include infinite-dimensional cases wherever possible. If the reader does not have a background in functional analysis, regarding statements such as "$\ldots X$ is a Banach space $\ldots$," "$\ldots X$ is a Hilbert space $\ldots$," "$\ldots X$ has the Radon–Nikodym property $\ldots$," the reader should think of $\mathbb{R}^n$, the space of n-dimensional (column) vectors. The dual space X' would then be the space of n-dimensional row vectors. The duality pairing between a vector $x \in X$ and $y^T \in X'$ is $\langle y^T, x \rangle = y^T x$, the usual inner product for n-dimensional vectors. The mysterious function J_X (which takes vectors in a Hilbert space X to the dual space X') can be thought of as the transpose operation for $\mathbb{R}^n$. Weak and strong convergence are identical for finite-dimensional spaces, but they can be different in important ways in infinite-dimensional spaces. The relevant concepts and theorems are detailed in the appendices, along with relevant material on convex analysis and differential equations.

The results of mine here are often improvements of my published results. While I have aimed for generality, I have often not given the most general possible formulation. As this is a book, the target is readability and ease of understanding as well as completeness of the technical results.

I would also like to take this occasion to thank the many people who have contributed to this work or commented on it, including (in alphabetical order) Jeongho Ahn, Mihai Anitescu, Kendall Atkinson, Bernard Brogliato, Kanat Çamlıbel, Alan Champneys, Marius Cocou, Muddappa S. Gowda, Lanshan Han, Weimin Han, João A. C. Martins (who has sadly passed away), Manuel Monteiro-Marques, Boris Mordukhovich, Jong-Shi Pang, Laetitia Paoli, Adrien Petrov, Michelle Schatzman, J. M. (Hans) Schumacher, Roger Temam, and Theodore Wendt. Suely Oliveira also contributed ideas which have strengthened the presentation of this work. I would also like to thank the Instituto de Ciências Matemáticas e de Computação at the Universidade de São Paulo, São Carlos, for their hospitality while I stayed there, especially José Cuminato, Alexandre N. Cavalho, and Antonio Castelo. The University of Iowa Mathematics Department has been my mathematical home for many years, and I would like to thank many people there including Ken Atkinson,

Weimin Han (who also works on contact mechanics and has helped improve the book), Yi Li, and Lihe Wang, for supporting me in my work over many years. The University of Iowa also has allowed me to take a leave which I have used to finish this book.

David Stewart *Iowa City and São Carlos, 2010.*

Chapter 1

Some Examples

> ... there is no philosophy which is not founded upon knowledge of the phenomena, but to get any profit from this knowledge it is absolutely necessary to be a mathematician.
>
> *Daniel Bernoulli*

In this chapter we see how a number of differential equations with inequalities arise. The inequalities that arise are naturally in the form of complementarity problems or in variational inequalities. These may arise from variational principles or from simple principles of the form "if this is positive, then that must be zero." There are more general formulations of hybrid systems which consist of a system of differential equations together with rules of the kind "when we are in this (discrete) state and we reach this set, then we change our state to another and continue with the corresponding differential equation." Such models may appear attractive to engineers, because this corresponds closely to how a system might be programmed. But there are a number of problems with this approach.

Often hybrid systems like this *chatter*; that is, they cross and recross the sets that define the transitions of the (discrete) state variable(s) in arbitrarily short times. For example, a bouncing ball might be modeled like this with a transition whenever the ball touches the ground. We would probably use Newton's rule, where the upward velocity just after contact is a constant multiple (designated e) of the downward velocity just before contact. If $0 < e < 1$, then there will be infinitely many bounces in finite time. Discrete-state systems cannot handle this. Alternatively, a popular approach to controller design has been to use *sliding mode controllers*. These controllers deliberately build discontinuities into the controller in order to force the system into a particular subset of state space. However, if there is any delay in the controller, then the system will jump from one side of the discontinuity to the other and back again with a frequency inversely proportional to the response time of the controller. Trying to model these systems from the point of view of general hybrid systems is at best misleading.

Rather, the purpose of this chapter is to show how the apparently special structure of rules based on complementarity problems or variational inequalities are in fact very common and natural. Complementarity problems typically have the following form: Given

a function $f : \mathbb{R}^n \to \mathbb{R}^n$, find a vector $z \in \mathbb{R}^n$, where

$$0 \le z \perp f(z) \ge 0;$$

note that "$a \ge b$" for vectors a and b means that "$a_i \ge b_i$ for all i" and that "$a \perp b$" means that $a^T b = 0$. Variational inequalities have the following form: Given $f : \mathbb{R}^n \to \mathbb{R}^n$ and a closed convex set K, find a vector z such that

$$z \in K \quad \& \quad (\widetilde{z} - z)^T f(z) \ge 0 \qquad \text{for all } \widetilde{z} \in K.$$

Both complementarity problems and variational inequalities are closely related to optimization and variational principles. More will be said about this in Chapter 2.

1.1 Mechanical impact

1.1.1 Ball and table

Consider a rigid ball falling onto a rigid table, as illustrated in Figure 1.1.

The equations of motion of the ball, assuming no air resistance, are given by Newton's second law of motion:

$$m\frac{d^2y}{dt^2} = -mg + N(t),$$

where m is the mass of the ball, and g (downward) is the gravitational acceleration. But sooner or later we must face the hard constraint that $y(t) - r \ge 0$ for all times t.

What happens at the time τ of impact when $y(\tau) = r$? Then we must have a reaction force $N(t)$ to prevent penetration. In fact, since we expect that $dy/dt(\tau^-) < 0$ at that time,

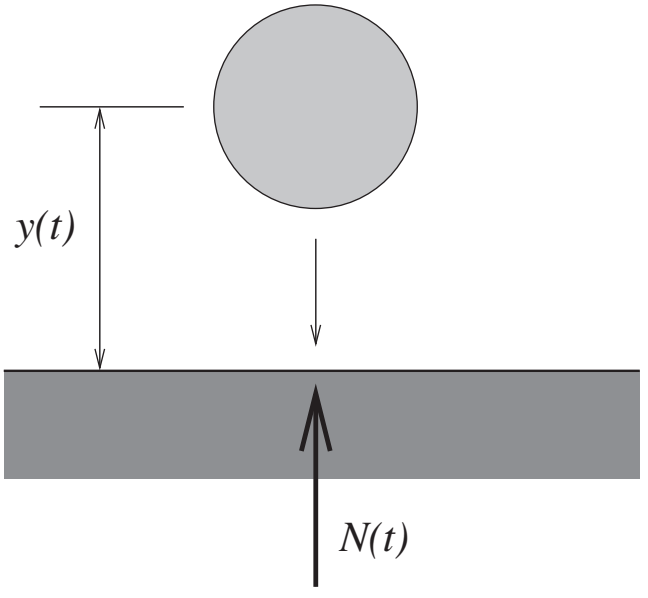


Figure 1.1: Ball and table example.

we must have a jump discontinuity in the velocity so that $dy/dt(\tau^+) \geq 0$. This means that $N(t)$ must contain a Dirac-δ function, or impulse, at time τ: $N(t) = N^*\delta(t-\tau) + g(t)$, where $g(t)$ is "nice," or at least "nicer" than a δ-function near $t = \tau$.

When there is contact, we expect the normal contact force to be upwards. Downward normal contact forces $N(t) < 0$ would indicate some kind of "glue" stopping the two bodies from separating. So we assume that $N(t) \geq 0$ for all t, and in particular, our impulse should also be upward: $N^* \geq 0$.

The relationship between the position $y(t)$ and the normal contact force $N(t)$ is given by the following relations:

$$\begin{aligned} 0 &\leq y(t) - r \qquad \text{for all } t, \\ 0 &\leq N(t) \qquad \text{for all } t, \\ 0 &= (y(t) - r)N(t) \quad \text{for all } t. \end{aligned}$$

This is a *complementarity condition*. It can be written more succinctly in the form

$$0 \leq y(t) - r \quad \perp \quad N(t) \geq 0 \qquad \text{for all } t. \tag{1.1}$$

The "$a \perp b$" sign means that a and b are orthogonal as vectors; that is, the dot product of a and b is zero. For scalar a and b, this amounts to saying "$ab = 0$."

Unfortunately, these conditions together with initial conditions are not sufficient to determine the trajectory uniquely. Even in this simple model we need an extra condition, usually given in terms of a coefficient of restitution $0 \leq e \leq 1$. The value of e determines how much "bounce" we expect to see in a collision. The standard (Newtonian) version is like this: if $y(\tau) - r = 0$, then

$$\frac{dy}{dt}(\tau^+) = -e\frac{dy}{dt}(\tau^-).$$

That is, the postimpact normal velocity should be $-e$ times the preimpact normal velocity. On a superficial level this is necessary to distinguish between the behavior of (for example) Play-Doh, which exhibits nearly completely inelastic behavior ($e \approx 0$), and solid rubber "superballs," which are nearly perfectly elastic ($e \approx 1$). Deeper investigation of these questions has pointed out a number of difficulties with these models (such as impacts resulting in increased energy) and a proliferation of alternative models (such as Poisson models and energy-based models of restitution). Ultimately, these issues should be resolved by a better understanding of elastic bodies in impact (which shall be introduced in Section 1.1.3).

With a *coefficient of restitution* e given we can determine the impulse strength N^*: $dy/dt(\tau^+) = dy/dt(\tau^-) + N^*/m = -e\,dy/dt(\tau^-)$, and so $N^*/m = -(1+e)dy/dt(\tau^-)$. Even with a given coefficient of restitution, solutions are not necessarily unique, although finding such examples is not easy [24].

1.1.2 More complex rigid-body systems with impact

Dealing with more complex situations we might have many bodies which can potentially make contact with many others at any point on the boundary. The equations of motion are also more complex, as the orientation of the body can change, and we need to take account of quantities such as angular momentum as well as ordinary momentum. There are

different ways of representing the orientation of a body (some use Euler angles, some unit quaternions, some orthogonal matrices, and some more obscure systems such as Rodriguez coordinates). The configuration of a hinged body, such as our arms and legs, or of robots, can often be expressed most efficiently in terms of the angles of the hinges. To avoid making the treatment dependent on a particular choice, we will take the configuration of the body to be represented by a vector $q \in \mathbb{R}^n$ which contains all the necessary information to describe the configuration of the mechanical system.

If we use Lagrangian mechanics to describe a mechanical system, then we would write the Lagrangian

$$L(q,v) = T(q,v) - V(q),$$

where $v = dq/dt$ is the (generalized) velocity, $T(q,v)$ is the kinetic energy function, and $V(q)$ is the potential energy function. Typically $T(q,v) = \frac{1}{2}v^T M(q)v$, where $M(q)$ is a symmetric positive definite matrix referred to as the *mass matrix*. The mass matrix will often contain quantities such as moments of inertia as well as actual masses, though. The constraints on the configuration of the system can be represented by means of a system of inequalities $\varphi_i(q(t)) \geq 0$, $i = 1, 2, \ldots, m$, for all t. Adding this to the Lagrangian by means of Lagrange multipliers $\lambda(t)$ gives

$$\begin{aligned} L(q,v,\lambda) &= \frac{1}{2}v^T M(q)v - V(q) - \sum_{i=1}^{m} \lambda_i \varphi_i(q) \\ &= \frac{1}{2}v^T M(q)v - V(q) - \lambda^T \varphi(q). \end{aligned}$$

However, since we have inequality constraints, we should not use the Lagrange multiplier principle itself, but we should use the corresponding principle for inequality-constrained optimization: the (Karush–)Kuhn–Tucker condition. Doing this formally with the principle of least action gives the following conditions:

$$M(q)\frac{dv}{dt} = -\nabla V(q) + k(q,v) - \sum_{i=1}^{m} \lambda_i \nabla \varphi_i(q), \tag{1.2}$$

$$\frac{dq}{dt} = v, \tag{1.3}$$

$$0 \leq \lambda_i \perp \varphi_i(q) \geq 0 \qquad \text{for all } i \text{ and } t. \tag{1.4}$$

The function $k(q,v)$ gives the pseudoforces (for example, Coriolis forces) that arise in the system, and this has the form

$$k_i(q,v) = -\frac{1}{2}\sum_{r,s}\left[\frac{\partial m_{ir}}{\partial q_s} + \frac{\partial m_{is}}{\partial q_r} - \frac{\partial m_{rs}}{\partial q_i}\right] v_r v_s.$$

If this model is analyzed, we again see that we must have impulses in the solution in general. Also, we need to have a coefficient of restitution for each pair of bodies that impact each other.

The theoretical questions that arise with these models having impulsive solutions which satisfy complementarity conditions are formidable, and they will be discussed later in Chapter 6.

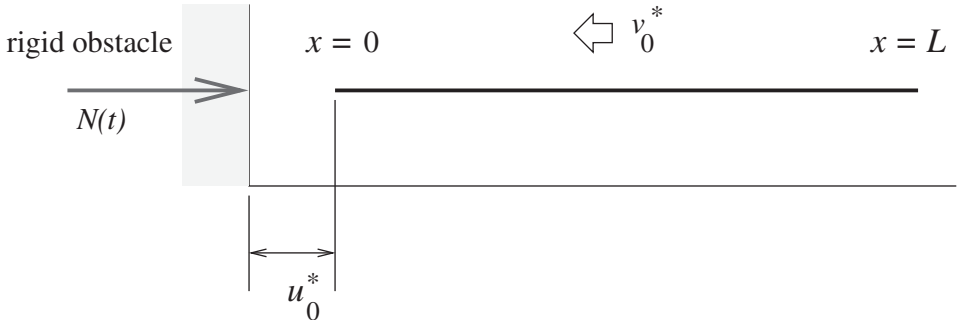


Figure 1.2: Routh's rod problem.

1.1.3 Elastic bodies in impact

The problems of elastic bodies in impact without friction are partial differential equations with complementarity conditions. These can either be "boundary thin" obstacle problems where the complementarity conditions apply to the boundary conditions or "thick" problems where the complementarity conditions apply over part of the domain of the solution.

For fully three-dimensional elastic bodies, contact occurs over the boundary, so this problem is a "boundary thin" problem. The simplest example of an elastic impact problem was first investigated over 150 years ago by Routh [216, pp. 442–444]. He considered the problem of a rod impacting a rigid obstacle (see Figure 1.2) and two elastic rods colliding. The equations of motion consist of the wave equation in the domain, complementarity conditions at the contacting boundary, and zero force on the free end. The equations of motion inside each rod are given by the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \qquad \text{for } 0 < x < L, \quad t > 0. \tag{1.5}$$

The boundary conditions are

$$\begin{aligned} -\frac{\partial u}{\partial x}(0,t) &= N(t), \\ \frac{\partial u}{\partial x}(L,t) &= 0, \\ 0 \le N(t) &\perp u(0,t) \ge 0. \end{aligned}$$

Note that our contact conditions are again complementarity conditions. These contact conditions are examples of Signorini conditions which were first proposed by Signorini [225] to describe contact between a static elastic body and a rigid obstacle.

In addition there are the initial conditions $u(x,0) = u_0(x)$ and $\partial u/\partial t(x,0) = v_0(x)$ with u_0 and v_0 given functions. In his treatise, Routh gave the exact solution for this problem with $u_0(x) \equiv u_0^*$ and $v_0(x) \equiv -v_0^*$: once contact is made, an elastic compression wave travels from $x = 0$ to $x = L$; on reflection at $x = L$ the wave becomes an expansion wave which then travels back to $x = 0$. When the expansion wave reaches $x = 0$ the entire rod has velocity $+v_0^*$, and the rod then separates from the obstacle. During the period in which the rod is in contact with the obstacle, the normal contact force $N(t)$ is constant.

Note that even though this is a second order problem in time, the contact force does not have any impulses; instead the normal contact force has a jump discontinuity at the time of impact and a jump discontinuity at the time of separation.

The rod problem of Routh is just the beginning of problems involving impact for elastic and viscoelastic bodies. Usually these involve three-dimensional bodies with contact possible on part or all of the boundary, with elastic or viscoelastic behavior of the material of the body. Furthermore, Coulomb friction may occur at the contact region. All of these variations make for a range of interesting and difficult problems.

1.2 Coulomb friction

A second example of how inequalities arise is in dry or Coulomb friction. The basic principles of friction, that

- the force of friction is directly proportional to the applied load, and
- the force of friction is independent of the apparent area of contact,

were first noted by da Vinci [70] in the late 1400s, but da Vinci never published these results (see [82, p. 99]). It was not until Amontons [9] in 1699 that these principles were published. Coulomb's experimental observations extended Amontons' and da Vinci's laws to dynamic friction [71], although with a different and smaller coefficient of friction. Incidentally, the two-coefficient model of friction (one for static friction where there is no slip, and another for dynamic friction where there is slip) was considered by Euler [92, 93]. The currently used "Coulomb friction laws" are actually a simplification of Coulomb's experimentally derived results which allowed for a small but nonzero adhesive term in the friction force, and Coulomb also noted that the coefficient of friction tended to increase as two surfaces stayed in contact without slip.

The engineering tribology literature has numerous modifications and additions to Coulomb's basic laws, but however they are formulated, the basic Coulomb laws of friction are taken to be the starting point for any theory of dry or unlubricated friction.

Coulomb's basic laws can be summarized as follows:

- the magnitude of the friction force is proportional to the load (the normal contact force) when there is slip;
- the friction force is bounded by a quantity proportional to the load when there is no slip; and
- the direction of the friction force during slip is opposite to the direction of slip.

The constant of proportionality of the friction force to the normal contact force is usually denoted by μ, a convention that dates back to Kotel'nikov, a student of Euler in Saint Petersburg [82, p. 213]. If we wish to distinguish between dynamic and static friction coefficients, we write μ_s for the static coefficient and μ_d for the dynamic coefficient. More modern theories assume $\mu = \mu(\|\mathbf{v}\|)$, where $\mathbf{v}$ is the slip velocity with a graph similar to that shown in Figure 1.3.

For simplicity, we will assume that $\mu(\|\mathbf{v}\|)$ is constant. With this simplification, consider a simple system of a brick on a ramp, as shown in Figure 1.4.

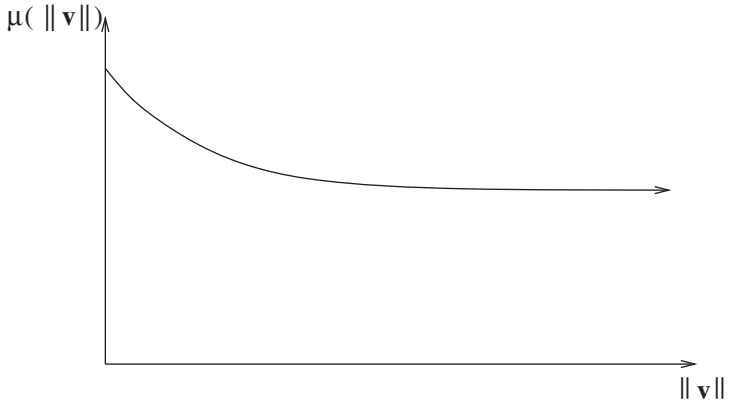


Figure 1.3: Possible dependence of the friction coefficient on the slip velocity.

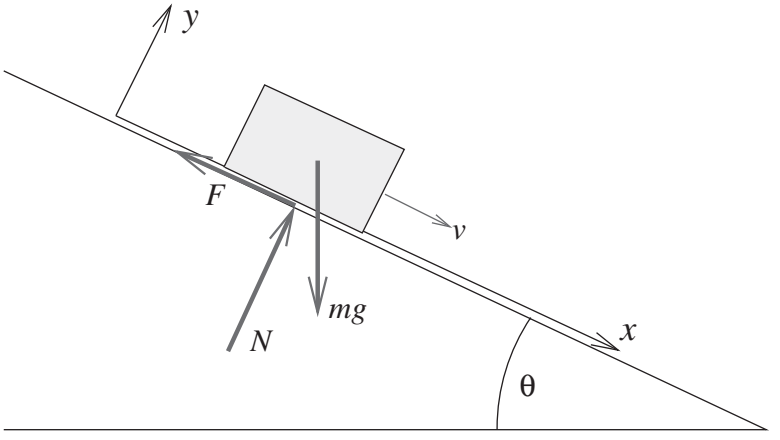


Figure 1.4: Brick on a ramp with dry friction. This figure first appeared in [239].

The system of equations to solve is

$$m\frac{d^2x}{dt^2} = -F + mg\sin\theta,$$
$$|F| \le \mu N = \mu mg\cos\theta$$

with F having opposite sign to the velocity $v = dx/dt$. A simple way of writing this uses the sgn function:

$$m\frac{dv}{dt} = -\mu mg\cos\theta\,\mathrm{sgn}(v) + mg\sin\theta,$$

where $\mathrm{sgn}(v) = +1$ if $v > 0$, $\mathrm{sgn}(0) = 0$, and $\mathrm{sgn}(v) = -1$ if $v < 0$. The trouble with this differential equation is that it has a discontinuity at $dx/dt = 0$; furthermore, if $|\tan\theta| < \mu$, the discontinuity is reached in finite time. If we take $\mathrm{sign}(0) = 0$, and $\sin\theta \neq 0$, there is no

solution to this differential equation [235, 236]. However, there is a solution in the sense of differential inclusions [19, 102, 104] if we replace $\operatorname{sgn}(v)$ with

$$\operatorname{Sgn}(v) = \begin{cases} \{+1\}, & v > 0, \\ [-1, +1], & v = 0, \\ \{-1\}, & v < 0. \end{cases}$$

The problem with taking $\operatorname{sign}(0) = 0$ is that we should be allowing F to be *any* value in the range $[-\mu N, +\mu N]$ if $v = 0$, and not just $F = 0$.

An alternative is based on optimization principles, specifically the principle of maximum dissipation [91, 248]: the friction force F should be the force that maximizes the rate of energy dissipation:

$$\begin{aligned} &\max_F -v\,F \\ &\text{subject to} \quad |F| \le \mu N. \end{aligned}$$

Since the problem is a linear program, the friction force is the solution to a variational inequality

$$\begin{aligned} v \cdot \left(\widetilde{F} - F\right) \ge 0 \qquad &\text{for all } \widetilde{F} \in [-\mu N, +\mu N], \\ F \in [-\mu N, &+\mu N]. \end{aligned}$$

Alternatively, using the Kuhn–Tucker conditions (see Appendix B), F is the solution of a mixed complementarity problem

$$\begin{aligned} v &= \lambda_+ - \lambda_-, \\ 0 \le \lambda_+ &\perp \mu N + F \ge 0, \\ 0 \le \lambda_- &\perp \mu N - F \ge 0. \end{aligned}$$

A difference between the Coulomb friction problem (with N given) and the frictionless impact problem is that there are no impulsive forces unless $N(t)$ has impulses. Also, the friction force depends on $v = dx/dt$, so that there is only one differentiation needed to go from the velocity to the acceleration and the friction forces. This means that these problems have index one, while impact problems have index two.

One of the most important, and often surprising, aspects of friction is the possibility of frictional instabilities. This can arise with velocity-dependent friction coefficients (as illustrated in Figure 1.3). The reduction of the friction coefficient with increasing slip velocity means that once slip starts, it tends to accelerate. This phenomenon means that when car brakes begin to skid, it can be important to release the brake and then reapply it. But if you have an ABS system in your car, you should not release and reapply the brake, as the ABS system already does it for you.

If variations in the normal contact force have to be taken into account, there can be feedback effects between the friction and normal contact forces. While the normal and frictional contact forces can each be considered to be "monotone" in that the forces behave monotonically in the normal displacement and the slip separately, this no longer holds when they are linked. This effect can be so strong as to require the notion of solution to change to allow impulsive forces even without collisions. This is an example of the Painlevé paradox (see Section 6.1.6). If you push a stick of chalk the "wrong way" on a blackboard and see the chalk jittering along, leaving a trail of dots, you have witnessed the paradox in action.

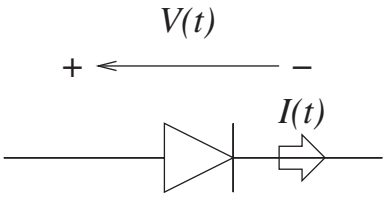


Figure 1.5: Diode configuration.

1.3 Diodes and transistors

1.3.1 Diode circuits

Diodes are passive electrical devices that allow current to flow in one direction but not the other. They are very useful for rectifying alternating current: changing electrical current from alternating its direction to always going in the same direction. Modern diodes are usually made of semiconductors, in which case they are created by joining together an n-type and a p-type semiconductor [251]. The junction allows current to pass easily only in one direction, as illustrated in Figure 1.5.

Transistors are active devices that are used to amplify signals and for other reasons. There are several different types of transistors, but we will focus on bipolar junction transistors. These are formed by layering n-type and a p-type semiconductors to form either an npn or a pnp sandwich. Each junction is a diode-type junction, but a small current flowing to the middle of the sandwich can control much larger currents flowing between the ends of the sandwich.

For a diode the standard Shockley model [251] gives a relationship between voltage and current of

$$I = I_0 \left(e^{qV/kT} - 1 \right),$$

where q is the magnitude of the charge on an electron, k is Boltzmann's constant, and T is the absolute temperature. The parameter I_0 is dependent on the specific details of the diode but is usually very small. For ordinary temperatures $kT/q \approx 26\,\mathrm{mV}$, which is quite small. A simple model that approximates this well for voltages where $V/26\,\mathrm{mV}$ is large is the complementarity formulation:

$$0 \geq V \quad \perp \quad I \geq 0, \tag{1.6}$$

which can be considered an ideal diode characteristic: if the diode is reverse biased (the voltage is trying to push current the "wrong" direction through the diode), then no current flows; if the diode is forward biased, then the effective resistance of the diode is zero. This model can be improved by allowing a "threshold voltage" $V_T \approx 0.7$ volts, where there is no current until the diode is forward biased by at least V_T:

$$0 \geq V - V_T \quad \perp \quad I \geq 0. \tag{1.7}$$

Related to diodes and transistors are thyristors, which are used in large power systems such as for inverters (which turn direct current into alternating current). In such systems the

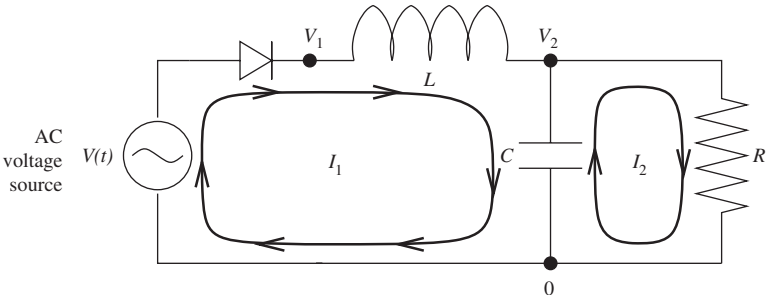


Figure 1.6: Half-wave rectifier.

thyristors can behave either as diodes or as short circuits; the precise model is hysteretic, as once it enters a "short circuit" it can resume its state as a diode only if the current passing through it goes below a threshold value.

The simplest interesting class of circuits that can be analyzed using complementarity approaches are circuits made from passive components (linear resistors, capacitors, and inductors, and perhaps transformers) and diodes. Complementarity formulations of such circuits have been analyzed [47, 122].

For example, Figure 1.6 shows a half-wave rectifier for converting alternating current into direct current with smoothing by a capacitor and an inductor.

The full system of equations and complementarity conditions for this half-wave rectifier is

$$\begin{aligned} C\frac{dV_2}{dt}(t) &= I_1(t) - I_2(t), \\ L\frac{dI_1}{dt}(t) &= V_1(t) - V_2(t), \\ V_2(t) &= R\,I_2(t), \\ 0 \le I_1(t) &\perp V_1(t) - V(t) \ge 0. \end{aligned}$$

Furthermore, we can incorporate threshold effects into this model by replacing "$V_1(t) - V(t)$" in the last line with "$V_1(t) - V(t) + V_T$."

The equations for the half-wave rectifier are an example of a *linear complementarity system* (LCS) [48, 124, 223]. These have the form (allowing for external inputs $f(t)$)

$$\frac{dx}{dt}(t) = A\,x(t) + B\,u(t) + E\,f(t), \qquad x(t_0) = x_0, \tag{1.8}$$

$$y(t) = C\,x(t) + D\,u(t) + F\,f(t), \tag{1.9}$$

$$0 \le y(t) \perp u(t) \ge 0 \qquad \text{for all } t. \tag{1.10}$$

Circuits with linear passive components and diodes have index one or index zero regardless of the number of circuit elements or how they are connected. This makes their analysis much easier.

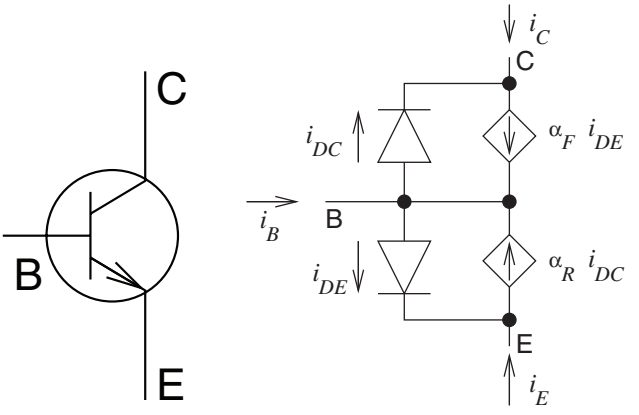


Figure 1.7: Ebers–Moll transistor model (*npn* transistor).

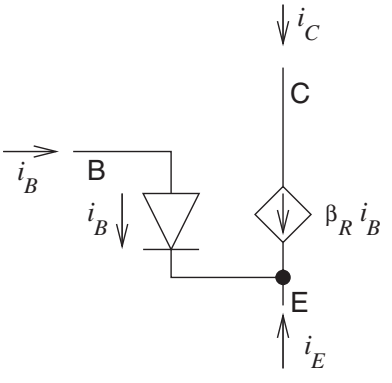


Figure 1.8: Simplified transistor model (*npn* transistor).

1.3.2 Bipolar junction transistors

As yet, complementarity models of bipolar junction transistors have not yet been properly analyzed, but these have a form similar to the diode models. For example, an *npn* bipolar junction transistor can be modeled using a diode model, as shown in Figure 1.7.

Typically $\alpha_F, \alpha_R \approx 1$, so that provided the transistor is forward biased ($V_C \geq V_B \geq V_E$) we assume $i_{DE} = 0$ and so $i_B + i_C = -i_E = \alpha_R i_{DC} = \alpha_R i_C$. Thus $i_C = i_B/(1-\alpha_R)$ and $-i_E = \alpha_R i_B/(1-\alpha_R)$. The coefficient $\alpha_R/(1-\alpha_R)$ is denoted β_R and is the nominal current gain of the transistor. The value of β_R is the main parameter of a simplified transistor model shown in Figure 1.8. This simplified model ignores the problems of saturation that can occur when the current source ($\beta_R i_B$) results in the voltage of the collector exceeding the supplied voltage at the collector (C). This can be especially important when a transistor is used as a switch, as is the case in transistor power or digital circuits.

Models of bipolar junction transistors clearly can be based on diode models, and so we can use complementarity-based models for transistors. However, this introduces a new element: current sources. This means that we have some new feedback in transistor

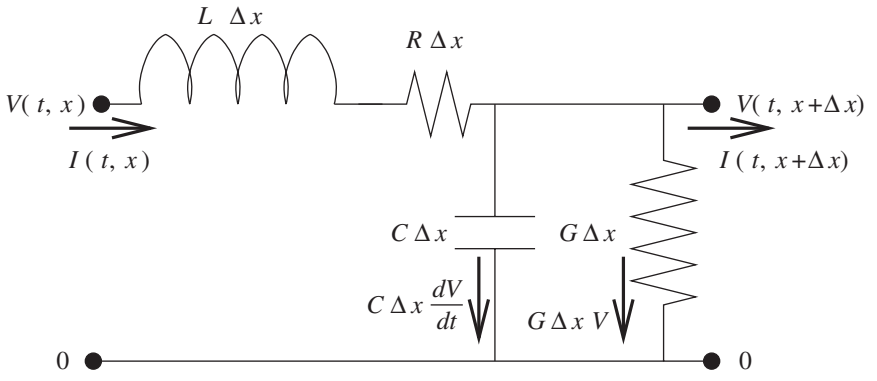


Figure 1.9: Segment of transmission line (Heaviside's model).

circuits that cannot occur in diode circuits, and some additional analysis is needed to make sure that this does not cause problems in existence or uniqueness theory, or in numerical methods used to solve the problem. In particular, there are static circuits (that is, without either capacitors or inductors) that do not have unique solutions, such as "flip-flops"—bistable circuits made up of two transistors which are normally in a state with one transistor conducting and the other not. Complementarity problems have already been found to be very useful in enumerating the possible steady states of circuits with transistors [265]. But this lack of uniqueness of solutions for the static problem has important consequences for both the theory and computation of solutions to problems with active elements.

1.3.3 Transmission lines with diodes

Partial differential equations can arise in diode models if they are connected to transmission lines, for example. The equations describing the behavior of the transmission line can be taken from Heaviside's model:

$$\frac{\partial V}{\partial x} + L\frac{\partial I}{\partial t} = -R\,I, \tag{1.11}$$

$$\frac{\partial I}{\partial x} + C\frac{\partial V}{\partial t} = -G\,V, \tag{1.12}$$

where R, C, L, and G are the resistance, capacitance, inductance, and leakage conductance[1] per unit length of the transmission line (see Figure 1.9).

Consider the circuit shown in Figure 1.10. Taking Laplace transforms in time of (1.11)–(1.12) gives an ordinary differential equation in x to solve for the Laplace transforms of the voltage and current, $\mathcal{L}V(s,0)$ and $\mathcal{L}I(s,0)$,

$$\mathcal{L}V(s,0) = \frac{Cs+G}{Ls+R}\tanh\left(\ell\sqrt{(Cs+G)(Ls+R)}\right)\mathcal{L}I(s,0) + \widetilde{q}(s), \tag{1.13}$$

[1] Conductance is the inverse of resistance.

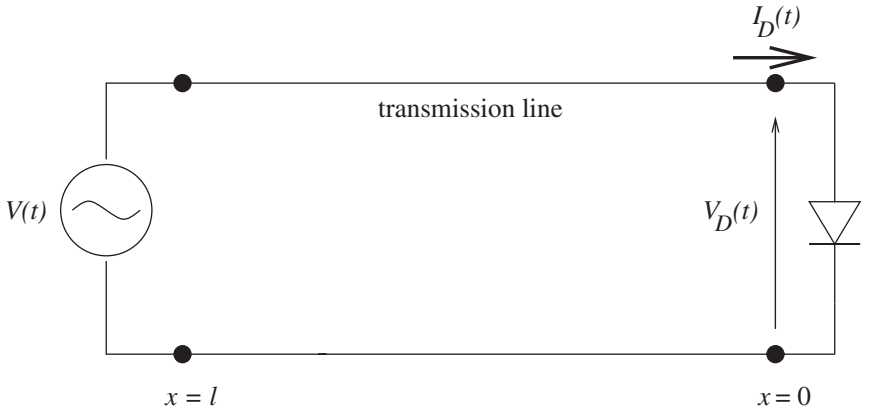


Figure 1.10: Circuit with diode and transmission line.

where $\widetilde{q}(s)$ is obtained from the initial conditions and the voltage source $V(t)$. The solution of the differential equations can then be expressed in terms of a convolution giving $V(t,0)$ in terms of $I(\tau,0)$ for $0 < \tau < t$:

$$V(t,0) = \int_0^t k(t-\tau)\, I(\tau,0)\, d\tau + q(t) = (k * I(\cdot,0)) + q(t), \tag{1.14}$$

$$k(t) = \frac{L}{C}\delta(t) + k_1(t), \qquad k_1 \text{ bounded.} \tag{1.15}$$

The model for the circuit in Figure 1.10 is then

$$\begin{aligned} V_D(t) &= V(t,0) + R_D\, I(t,0) \\ &= R_D\, I(t,0) + (k * I(\cdot,0)) + q(t), \\ I_D(t) &= I(t,0), \\ 0 \geq I_D(t) &\perp V_D(t) \leq 0. \end{aligned}$$

This is an index-zero convolution complementarity problem (CCP), which has solutions as shown in Section 4.6 on CCPs. These solutions can be found by numerical methods directly based on the CCP or by using an implicit time-stepping method.

1.4 Queues and resource limits

1.4.1 Queues

A simple continuous deterministic model of a queue is to have a queue length $\ell(t)$, with service rate $s(t) \geq 0$ (the rate at which customers are served and then leave the queue) and arrival rate $a(t) \geq 0$ (the rate at which new customers arrive at the queue). If $\ell(t) > 0$, then we can model the dynamics of the queue by

$$\frac{d\ell}{dt}(t) = a(t) - s(t).$$

However, if $a(t)$ drops to zero and $s(t)$ remains high, then eventually we will reach $\ell(t) = 0$. The length of the queue clearly cannot go negative, so it stops at $\ell(t) = 0$. Once this happens, if the (maximum) service rate exceeds the arrival rate, no queue can develop: each customer is served as soon as they arrive. To model this, we set $r(t)$ to be the actual rate at which customers are served at time t. Clearly the actual rate of service cannot exceed the maximum rate of service, so $r(t) \le s(t)$. If $\ell(t) > 0$, then we have $r(t) = s(t)$, as the servers are (or should be!) trying to help the waiting customers as fast as possible. This leads to the following complementarity formulation:

$$\begin{aligned} \frac{d\ell}{dt}(t) &= a(t) - r(t), \\ 0 \le \ell(t) \perp s(t) - r(t) &\ge 0. \end{aligned}$$

Naturally, we expect that if $\ell(t) = 0$ and $a(t) \le s(t)$, then $r(t) = a(t)$; that is, the actual service rate matches the arrival rate, but this is not immediately apparent from this model. In fact, it is true for this model, as if $r(t) < a(t)$ for any length of time, then we will have $\ell(\tau) > 0$ for any $\tau > t$ shortly afterward. This in turn means that $r(\tau) = s(\tau)$ for $\tau > t$ shortly afterward, which would drive $\ell(\tau)$ to zero in an arbitrarily short time. Thus $r(t) = a(t)$ if $\ell(t) = 0$ and $a(t) \le s(t)$.

Models of this kind can also model the flow of inventory through a factory: the queues are items waiting to be processed by a particular machine. Factories viewed in this way can be seen as a network of queues: when items are processed and leave one queue, they join another for further processing. After final processing, they leave the factory, to be used or sit on a shelf (as part of another queue).

Stochastic models of this kind can also be developed. In fact, these models are known as *Skorokhod problems*. While the theory of stochastic differential equations is beyond the scope of this monograph, they can be applied to this problem. We can, at least formally, write these problems in complementarity form: for a single queue with random arrive rate given by a random process dA_t, we have the stochastic differential complementarity problem

$$\begin{aligned} dL_t &= dA_t - R(t)\,dt, \\ 0 \le L_t \perp S(t) - R(t) &\ge 0. \end{aligned}$$

Although stochastic problems are beyond the scope of this book, there are plenty of interesting issues regarding the deterministic versions, as can be seen in Chapter 5.

1.4.2 Traffic flow

Traffic flow is an example of an economic network [83, 187]. It is an *economic* network because each person in the network is trying to optimize something, typically the time to get to work or the time to get home. The combined action of each individual in the network drives the behavior of the network as a whole, and it can lead to such things as congestion and traffic jams, where individually optimal choices can lead to behavior that is far from optimal for the system as a whole.

Traffic modeling involves a network or roads or transportation links connected to a *graph* or *network*, which consists of nodes together with edges connecting these nodes. Each edge is a road or transportation link, while each node is an intersection of roads, or

a point where different transportation links meet. Graphs or networks are a basis for many models of economic behavior [255, 186].

Most analyses of traffic flow are essentially static. The most common models are based on the notion of a *Wardrop equilibrium* [267]: each person in the network takes the quickest path home given the prevailing conditions, while the time needed to pass through an edge is a function of the number of cars attempting to do so. This has the features expected of an equilibrium formulation: it is static, where participants have complete knowledge of the properties of the network. It is reasonable to assume drivers have complete knowledge after having an infinite amount of time to explore alternatives or hearing of other routes. Such systems can be made dynamic to reflect the fact that traffic conditions do change with time (yes, traffic jams do eventually end, at least in most cities). The knowledge that a person has about the state of the network can (and does) change with time. Various assumptions can be made about the knowledge that an individual has about the state of the network: Are they all listening in to the traffic reports on the radio? Or is it only when they get to an intersection that they can see how congested the exiting roads are? Knowledge of the traffic flow is nonanticipative: individuals cannot infallibly predict future network conditions.

Whatever models are used for determining network flows, there are a number of important issues that must be addressed by any dynamic traffic model:

- flow quantities must be preserved at each node: nodes (such as intersections) cannot "store" a significant amount of traffic;
- flow quantities must be conserved along each edge: although edges can "store" a significant amount of traffic, the total number of vehicles in an edge must balance the number of vehicles entering and leaving the edge;
- traffic flows in an edge cannot become negative, nor can the number of individuals on an edge;
- each individual in the network has the opportunity to make decisions at nodes and will presumably use some optimality principle to decide what edge to leave on.

The optimality conditions for each individual in the network can be expressed in complementarity or variational inequality form. The resulting dynamic system is a differential equation with complementarity or variational inequality conditions imposed and is therefore a differential complementarity problem or differential variational inequality. For more on this topic, see Section 5.5.

1.4.3 Biological resource limits

There are a number of ways in which hard resource constraints can lead to differential equations with inequalities. Consider, for example, a tissue of cells in, say, a human body as illustrated in Figure 1.11. The cells in the tissue will take up oxygen for the cell's metabolic processes and release waste products such as CO_2 to be taken away by nearby blood vessels or capillaries.

One model is that the rate of oxygen uptake is essentially constant until the concentration of oxygen in the tissue reaches very low levels. In such a case, the cells could be

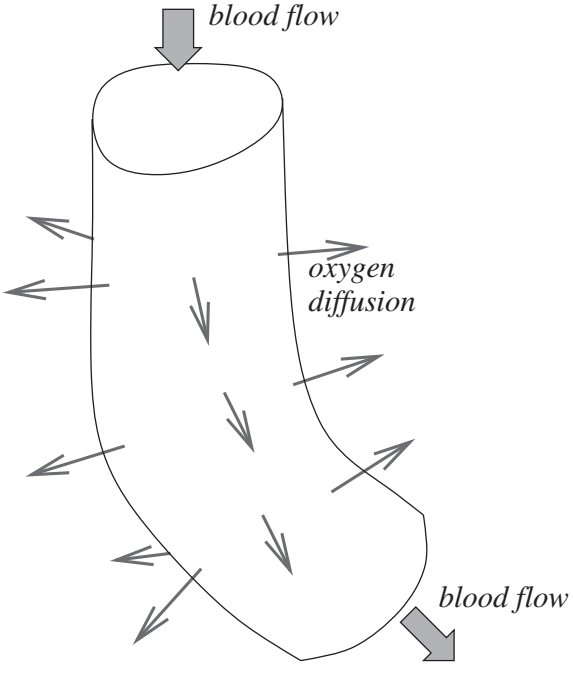


Figure 1.11: Oxygen diffusion in biological tissues.

assumed to "turn off" most of their metabolic processes. This can be modeled as a diffusion process with absorption rate which becomes zero when the concentration becomes zero. These are typically partial differential equations of a diffusion kind to reflect the dependence of the oxygen concentration on position as well as time. Let $u(t,\mathbf{x})$ be the oxygen concentration at time t and position $\mathbf{x}$. Then the evolution of the oxygen concentration can be modeled as

$$\frac{\partial u}{\partial t} = \nabla \cdot (D\,\nabla u) - r(t,\mathbf{x}),$$
$$r(t,\mathbf{x}) = \arg \max_{r\in[0,r_{max}]} r\,u(t,\mathbf{x}).$$

Note that D is the coefficient of diffusion of oxygen in the tissue, and r_{max} is the maximum rate of uptake of oxygen by the tissue. The optimality principle used here, that cells take up oxygen as rapidly as possible up to the limit r_{max}, can be turned into a variational inequality:

$$r(t,\mathbf{x}) \in [0,r_{max}] \quad \& \quad (r(t,\mathbf{x}) - \widetilde{r})\,u(t,\mathbf{x}) \geq 0 \qquad \text{for all } \widetilde{r} \in [0,r_{max}].$$

Such problems can also be treated as *parabolic variational inequalities* (see Section 4.7).

Chapter 2

Static Problems

> ... a utopia is always static; it is always descriptive and has no, or almost no, plot dynamics.
>
> *H.G. Wells*

In this chapter we set out the main ideas and tools for static complementarity and variational inequality problems, as well as for set-valued functions. The first section is on basic tools and deals with aspects of convex analysis and set-valued functions. Readers may wish to review the material in the appendices for background information on abstract spaces, and convex sets and functions.

2.1 Basic tools

The spaces in which we look for solutions are Banach or Hilbert spaces, with the particular example of finite-dimensional spaces $\mathbb{R}^n$ of n-dimensional vectors. All Banach and Hilbert spaces considered here are defined over the real numbers $\mathbb{R}$. Properties of these spaces can be found in Appendix A. The n-dimensional Euclidean space $\mathbb{R}^n$ of n-dimensional vectors has the usual inner product $(x, y)_{\mathbb{R}^n} = x^T y$ and norm $\|x\|_{\mathbb{R}^n} = \sqrt{x^T x}$. We use the notation $B_X = \{ x \in X \mid \|x\|_X < 1 \}$ for the open unit ball in a Banach space X.

Vectors in $\mathbb{R}^n$ are usually considered to be column vectors, but the dual space $(\mathbb{R}^n)'$ can be considered as the space of corresponding row vectors and identified with $\mathbb{R}^n$ if desired (so that $F\colon X \to X'$ includes the case $F\colon \mathbb{R}^n \to \mathbb{R}^n$).

The notation $(u, v)_H$ is used for inner products if u and v are elements of the same Hilbert space H. We drop the subscript and write (u, v) if H is understood. For example, if $u, v \in \mathbb{R}^n$, $(u, v) = u^T v = \sum_{i=1}^n u_i v_i$.

The dual space X' of a Banach space X is the vector space of continuous linear functions $X \to \mathbb{R}$, which are called functionals. Similar notation is used for duality pairing between a Banach space X and its dual space X': $\langle u, v \rangle$ is the result of applying the functional $u \in X'$ to $v \in X$: $\langle u, v \rangle = u(v)$ explicitly uses the fact that u is a function $X \to \mathbb{R}$. If the space X is ambiguous, or the interpretation is unclear, we denote the duality pairing by $\langle u, v \rangle_{X' \times X}$ to indicate that $u \in X'$ and $v \in X$.

Many readers may be used to identifying X and its dual space X' if X is a Hilbert space. They are, after all, isometrically isomorphic under the map $J_X : X \to X'$ given by $J_X(u) = (u, \cdot)_X$. If we consider $\mathbb{R}^n$ to be the space of n-dimensional *column* vectors and $(\mathbb{R}^n)'$ to be the space of n-dimensional *row* vectors, then $J_X(u) = u^T$. However, in many applications this is not appropriate to identify the two spaces. Consider, for example, the real *weighted* Hilbert space

$$\ell^2_{\mathbf{w}} = \left\{ \mathbf{x} = (x_1, x_2, x_3, \ldots) \mid \sum_{i=1}^{\infty} w_i \, x_i^2 < +\infty \right\}$$

for a positive weight vector $\mathbf{w} = (w_1, w_2, w_3, \ldots)$; this has the inner product

$$(\mathbf{x}, \mathbf{y})_{\mathbf{w}} = \sum_{i=1}^{\infty} w_i \, x_i \, y_i .$$

The dual space to $\ell^2_{\mathbf{w}}$ is most easily identified not with $\ell^2_{\mathbf{w}}$, but with $\ell^2_{\mathbf{v}}$, where $v_i = 1/w_i$. If $w_i \to +\infty$ as $i \to \infty$, then $\ell^2_{\mathbf{w}}$ is a much smaller space than $\ell^2_{\mathbf{v}}$. The map J_X for $X = \ell^2_{\mathbf{w}}$ is far from the identity map:

$$J_X(\mathbf{x}) = \mathbf{u}, \qquad \text{where } u_i = w_i \, x_i .$$

This choice of representation $\left(\ell^2_{\mathbf{w}}\right)' = \ell^2_{\mathbf{v}}$ is chosen to make the duality pairing

$$(\mathbf{x}, \mathbf{u}) = \sum_{i=1}^{\infty} x_i \, u_i$$

natural and independent of the weight vector, the relationship between these being that $\langle \mathbf{x}, J_X(\mathbf{y}) \rangle = (\mathbf{x}, \mathbf{y})_X$ for all $\mathbf{x}, \mathbf{y} \in X$.

Most often we will deal with Hilbert spaces, but there are occasions to work with Banach spaces that are not Hilbert spaces. Most of these are *reflexive* spaces where we can naturally identify a Banach space X with its second dual X''. This identification uses the *natural map* $\natural : X \to X''$ given by

$$\langle \natural(x), \xi \rangle_{X'' \times X'} = \langle \xi, x \rangle_{X' \times X} . \tag{2.1}$$

Note that $\natural : X \to X''$ is one-to-one and continuous for any Banach space X; it is surjective if and only if X is a reflexive space. Reflexive spaces include all Hilbert spaces, $L^p(\Omega)$ for $1 < p < \infty$, and $\mathbb{R}^n$, but not $L^1(\Omega)$, $L^\infty(\Omega)$, or the space of measures of bounded variation over a set Ω: $\mathcal{M}(\Omega)$.

2.1.1 Convex analysis

Convex sets are sets C, where for any $x, y \in C$ and $0 \le \theta \le 1$, we have $\theta x + (1-\theta) y \in C$; *convex functions* are functions $f : X \to \mathbb{R} \cup \{\infty\}$, where

$$f(\theta x + (1-\theta) y) \le \theta \, f(x) + (1-\theta) \, f(y) \qquad \text{for all } x, y \text{ and } 0 \le \theta \le 1,$$

but we assume that f is *proper*; that is, $f(x) < \infty$ for some x. Usually we assume that f is also *lower semicontinuous,* so that if $x_k \to x$, we have $\liminf_{k\to\infty} f(x_k) \geq f(x)$. Every convex function has an associated *epigraph*:

$$\operatorname{epi} f = \{(x,s) \mid s \geq f(x)\}. \tag{2.2}$$

If f is convex, then $\operatorname{epi} f$ is a convex set; if f is lower semicontinuous, then $\operatorname{epi} f$ is a closed set. The main results of convex analysis that we need are given in Appendix B. We summarize some of the results here. In a Hilbert space X containing a closed convex set K, we have *projection* operators $\Pi_K : X \to X$, where $\Pi_K(x)$ is the point in K closest to x. If K is a subspace of X, then $\Pi_K(x)$ is the orthogonal projection of x onto K. In Hilbert spaces, Π_K is characterized by

$$(x - \Pi_K(x), z - \Pi_K(x))_X \leq 0 \qquad \text{for all } z \in K.$$

The projection operator is Lipschitz with Lipschitz constant one: $\|\Pi_K(x) - \Pi_K(y)\| \leq \|x - y\|$.

A *cone* C is a set such that whenever $x \in C$ and $\alpha \geq 0$, we have $\alpha x \in C$. Convex cones are important for understanding both local and global structure of convex sets and even functions. Given a closed convex cone K, the *dual cone* K^* and the *polar cone* K° are defined by

$$K^* = \left\{ y \in X' \mid \langle y, x\rangle \geq 0 \text{ for all } x \in K \right\} \subseteq X', \tag{2.3}$$

$$K^\circ = -K^*. \tag{2.4}$$

If X is a reflexive space, then $K = K^{**} = K^{\circ\circ}$, identifying X and X'' via the natural map. If X is a Hilbert space and we identify X with its dual space X', we say a cone K is *self-dual* if $K = K^*$. For example, $K = \mathbb{R}^n_+$ (the nonnegative orthant of n-dimensional vectors with nonnegative entries) is a self-dual cone.

A convex cone K is *pointed* if $K \cap (-K) = \{0\}$. In finite dimensions, K is pointed if and only if K^* (or K°) contains an open set. In infinite dimensions this may not hold; instead we say K is *strongly pointed* if K^* (or K°) contains an open set.

Another cone associated with a closed convex K set is the *recession* or *asymptotic cone* K_∞ given by

$$K_\infty = \left\{ \lim_{k\to\infty} t_k x_k \mid t_k \downarrow 0,\ x_k \in K \ \text{ for all } k \right\}. \tag{2.5}$$

Recession cones give information about the K "at infinity."

The *tangent cone* of a convex set K at a point $x \in K$ is

$$T_K(x) = \left\{ \lim_{k\to\infty} (x_k - x)/t_k \mid t_k \downarrow 0,\ x_k \in K \text{ for all } k \right\}, \tag{2.6}$$

which is also a closed convex cone as illustrated in Figure 2.1. Closely associated with this cone is the *normal cone* for a convex set at a point $x \in K$:

$$N_K(x) = T_K(x)^\circ. \tag{2.7}$$

Polyhedral sets are closed convex sets that are the intersection of a finite number of half-spaces $\bigcap_{k=1}^M \{x \mid \langle d_k, x\rangle \geq \alpha_k\} \subseteq \mathbb{R}^n$. *Polyhedral cones* are polyhedral sets that are also cones. Polyhedral sets have the important property that every real linear function over such a set either is unbounded below or has a minimum. Related to this, the sum of two polyhedral sets $P + Q = \{x + y \mid x \in P,\ y \in Q\}$ is also a (closed) polyhedral set.

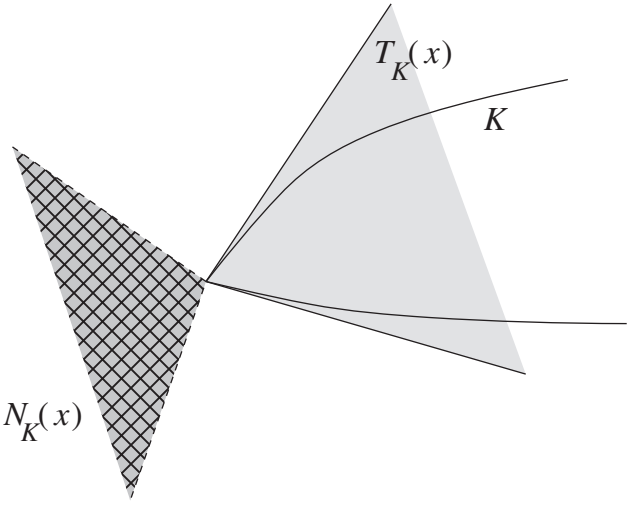


Figure 2.1: Tangent and normal cones.

2.1.2 Set-valued functions

Set-valued functions are functions $\Phi\colon X \to \mathcal{P}(Y)$, where $\mathcal{P}(A)$ is the collection of subsets of A. These generalize ordinary functions since any ordinary function $\phi\colon X \to Y$ can be represented by $\Phi(x) = \{\phi(x)\}$ for all X. The *domain* of a set-valued function is

$$\operatorname{dom}\Phi = \{x \mid \Phi(x) \neq \emptyset\}. \tag{2.8}$$

Often the values $\Phi(x)$ have some special characteristics (such as being closed and convex), and often the graph of Φ, given by

$$\operatorname{graph}\Phi = \{(x, y) \mid y \in \Phi(x)\}, \tag{2.9}$$

should be a closed subset of $X \times Y$. A closely related concept to Φ having a closed graph is that of *upper semicontinuity* (see Section 2.1.3). Questions of integration of set-valued functions involve issues of measurability, which are discussed below. Here we will concentrate on some of the issues for set-valued functions $\Phi\colon \Omega \to \mathcal{P}(X)$, where X is a Banach space and Ω a metric space. An essential reference for set-valued analysis is Aubin and Frankowska [21].

Later we will consider maximal monotone operators which are a special class of set-valued functions that generalize monotonicity for ordinary functions:

$$0 \leq \langle \phi(z_2) - \phi(z_1), z_2 - z_1 \rangle \qquad \text{for all } z_1, z_2 \in X.$$

Maximal monotone operators are closely connected to subdifferentials of convex functions: for a lower semicontinuous proper convex function $\phi\colon X \to \mathbb{R} \cup \{\infty\}$, the subdifferential $\partial\phi$ generalizes the gradient function, and it is a maximal monotone operator. See Section 2.3.6 for more details.

Unlike single-valued functions $\varphi\colon \Omega \to X$, there are several notions of continuity for set-valued functions $\Phi\colon \Omega \to \mathcal{P}(X)$. The simplest is to use a metric, and the most

commonly used metric for sets is the *Hausdorff metric*: suppose A and B are bounded sets; then we define

$$d_H(A, B) = \max(\sup_{a\in A} d(a, B), \sup_{b\in B} d(b, A)), \tag{2.10}$$

$$d(x, C) = \inf_{c\in C} d(x,c) = \inf_{c\in C} \|x - c\|. \tag{2.11}$$

Note that this can really be a metric only for *closed* sets A and B, as the distance between an open interval and its closure is $d_H((a,b),[a,b]) = 0$. This definition ensures that d_H satisfies the requirements of a metric for closed and bounded sets:

1. $d_H(A, A) = 0$ and $d_H(A, B) = 0$ implies $A = B$;
2. $d_H(A, B) = d_H(B, A)$;
3. $d_H(A, C) \le d_H(A, B) + d_H(B, C)$.

We say that $\Phi\colon \Omega \to \mathcal{P}(X)$ with nonempty closed bounded values is continuous with respect to the Hausdorff metric if for any $x \in \Omega$ and $\epsilon > 0$ there is a $\delta > 0$ such that

$$d(x, y) < \delta \quad \text{implies} \quad d_H(\Phi(x), \Phi(y)) < \epsilon.$$

In addition, there is upper and lower semicontinuity.

Upper semicontinuity of $\Phi\colon \Omega \to \mathcal{P}(X)$ means that if $\mathcal{V}$ is an open set containing $\Phi(x)$, then there is a neighborhood $\mathcal{U}$ of x where $\Phi(y) \subset \mathcal{V}$ for all $y \in \mathcal{U}$. This can be defined in terms of the "one-sided metric"

$$\delta_H(A, B) = \sup_{a\in A} d(a, B). \tag{2.12}$$

If $\Phi(x)$ is compact, upper semicontinuity at x then means that for any $\epsilon > 0$ there is a $\delta > 0$ such that for all y,

$$d(x, y) < \delta \quad \text{implies} \quad \delta_H(\Phi(y), \Phi(x)) < \epsilon.$$

Lower semicontinuity of $\Phi\colon \Omega \to \mathcal{P}(X)$ means that for any $z \in \Phi(x)$, if $\mathcal{V}$ is an open set containing z, then there is a neighborhood $\mathcal{U}$ of x where $\Phi(y) \cap \mathcal{V} \neq \emptyset$ for all $y \in \mathcal{U}$. In other words, if $x_k \to x$ in Ω and $z \in \Phi(x)$, then there is a sequence $z_k \in \Phi(x_k)$ such that $z_k \to z$.

Similarly, lower semicontinuity then means that for any $x \in \Omega$ and $\epsilon > 0$ we have $\delta > 0$ such that

$$d(x, y) < \delta \quad \text{implies} \quad \delta_H(\Phi(x), \Phi(y)) < \epsilon.$$

A related concept to upper semicontinuity is that of having a closed graph. A set-valued function $\Phi\colon \Omega \to \mathcal{P}(X)$ has a *closed graph* if the graph of Φ

$$\operatorname{graph}\Phi = \{(x, z) \in \Omega \times X \mid z \in \Phi(x)\} \tag{2.13}$$

is closed in $\Omega \times X$.

A set-valued function $\Phi\colon \Omega \to \mathcal{P}(X)$, where X is a reflexive Banach space, is *hemicontinuous* if $\operatorname{graph}\Phi$ is closed in the strong $\times$ weak topology of $\Omega \times X$. That is, if $y_k \in \Phi(x_k)$, $x_k \to x$ strongly, and $y_k \rightharpoonup y$ weakly, then $y \in \Phi(x)$. Hemicontinuity is equivalent to having a closed graph if X is finite dimensional, as the weak topology of $\mathbb{R}^n$ is equivalent to its strong topology.

2.1.3 Upper semicontinuity and closed graphs

An alternative condition that a set-valued function $\Phi\colon X \to \mathcal{P}(Y)$ is upper semicontinuous at $x \in X$ is that for any $\epsilon > 0$ there is a $\delta > 0$ where

$$\|x - z\| < \delta \quad \text{implies} \quad \Phi(z) \subseteq \Phi(x) + \epsilon\, B_Y, \tag{2.14}$$

B_Y being the open unit ball $\{\, y \in Y \mid \|y\| < 1 \,\}$. This condition can be expressed more succinctly as

$$\Phi(x + \delta\, B_X) \subseteq \Phi(x) + \epsilon\, B_Y,$$

where for a set A,

$$\Phi(A) = \bigcup_{a \in A} \Phi(a).$$

Upper semicontinuity can be represented in terms of a type of inverse to a set-valued function. For set-valued functions there are two kinds of inverse functions:

$$\Phi^-(U) = \{\, x \mid \Phi(x) \cap U \neq \emptyset \,\}, \tag{2.15}$$

$$\Phi^+(U) = \{\, x \mid \Phi(x) \subseteq U \,\}. \tag{2.16}$$

Note that $\Phi^-(U)$ is called the *weak inverse image* of U while $\Phi^+(U)$ is called the *strong inverse image* of U. Then Φ is upper semicontinuous if and only if $\Phi^+(U)$ is open for all open sets $U \subseteq Y$. Since $\Phi^-(Y \backslash U) = X \backslash \Phi^+(U)$, by taking complements of this definition of upper semicontinuity, we see that Φ is upper semicontinuous if and only if $\Phi^-(C)$ is closed for all closed sets $C \subseteq Y$. These inverse functions have some, but not all, the properties of inverses of the standard (set-valued) inverse to single-valued functions:

$$f^{-1}(U) = \{\, x \mid f(x) \in U \,\}. \tag{2.17}$$

If Φ is the set-valued version of a single-valued function f (where $\Phi(x) = \{f(x)\}$ for all x), then $f^{-1}(U) = \Phi^-(U) = \Phi^+(U)$ for all $U \subseteq Y$. The following properties hold for the weak inverse images:

$$\Phi^-(U \cap V) = \Phi^-(U) \cap \Phi^-(V),$$
$$\Phi^-(U \cup V) \subseteq \Phi^-(U) \cup \Phi^-(V).$$

Corresponding properties hold for the strong inverse image Φ^+ by taking complements.

If $f\colon X \to Y$ is a single-valued function, then $\Phi(x) := \{f(x)\}$ is upper semicontinuous if and only if f is continuous. However, other set-valued functions are upper semicontinuous, such as

$$\operatorname{Sgn}(x) = \begin{cases} \{+1\}, & x > 0, \\ [-1, +1], & x = 0, \\ \{-1\}, & x < 0. \end{cases} \tag{2.18}$$

This is the usual set-valued version of the sgn function $\operatorname{sgn}(x) = +1$ for $x > 0$, -1 for $x < 0$, and zero for $x = 0$. To ensure that Sgn is upper semicontinuous it would be sufficient to have the value $\{-1, +1\}$ at $x = 0$. However, it is desirable to have set-valued functions which have closed *convex* values, as will be seen in Section 4.1.

The property of $\Phi\colon X \to Y$ being upper semicontinuous is closely related to the property that the graph of Φ is closed.

Lemma 2.1. *If a set-valued function $\Phi\colon X \to Y$ is upper semicontinuous with closed values, then it has a closed graph. If Φ is upper semicontinuous with closed convex values, then it is hemicontinuous provided X is a reflexive Banach space. Conversely, if $\Phi\colon X \to Y$ has a closed graph with $\Phi(X)$ compact, then Φ is upper semicontinuous.*

Proof. Suppose $\Phi\colon X \to Y$ is upper semicontinuous. Suppose we have a sequence $(x_m, y_m) \in \operatorname{graph}\Phi$ where $(x_m, y_m) \to (x, y)$. By upper semicontinuity, for any $\epsilon > 0$ there is a $\delta > 0$ such that $\Phi(x + \delta B_X) \subseteq \Phi(x) + \epsilon B_Y$. Thus for m sufficiently large, $y_m \in \Phi(x_m) \subseteq \Phi(x + \delta B_X) \subseteq \Phi(x) + \epsilon B_Y$. Thus the limit $y \in \overline{\Phi(x) + \epsilon B_Y} \subseteq \Phi(x) + \epsilon \overline{B_Y}$. Since this is true for all $\epsilon > 0$ and $\Phi(x)$ is closed, $y \in \Phi(x)$ and so $(x, y) \in \operatorname{graph}\Phi$. That is, Φ has a closed graph.

Now suppose that Φ is upper semicontinuous with closed convex values, and that $z_k \in \Phi(x_k)$, and $x_k \to x$ in Ω and $z_k \rightharpoonup z$ in X. For every $\epsilon > 0$ there is a $\delta < 0$ such that $d(x_k, x) < \delta$ implies that $\Phi(x_k) \subseteq \Phi(x) + \epsilon \overline{B_X}$. Since $\overline{B_X}$ is weakly compact in a reflexive Banach space and $\Phi(x)$ is weakly closed (as it is a strongly closed convex set), $\Phi(x) + \epsilon \overline{B_X}$ is weakly closed. Thus $z \in \Phi(x) + \epsilon \overline{B_X}$, or equivalently, $d(z, \Phi(x)) \leq \epsilon$. Since this is true for all $\epsilon > 0$, and $\Phi(x)$ a closed set, $z \in \Phi(x)$. So if Φ is upper semicontinuous with closed convex values, $\operatorname{graph}\Phi$ is closed in the strong $\times$ weak topology; that is, Φ is hemicontinuous.

Now suppose that $\Phi\colon X \to Y$ has a closed graph and $\Phi(X)$ is compact. Let $x \in X$. First note that $\Phi(x)$ is closed since $\{x\} \times \Phi(x) = (\operatorname{graph}\Phi) \cap (\{x\} \times Y)$, the intersection of closed sets, is closed; furthermore, $\Phi(x)$ is compact since it is a closed subset of the compact set $\Phi(X)$. Let $\epsilon > 0$ be given. If there is no $\delta > 0$ such that $\Phi(x + \delta B_X) \subseteq \Phi(x) + \epsilon B_Y$, there must be a sequence $(x_k, y_k) \in \operatorname{graph}\Phi$ such that $x_k \to x$ but $y_k \notin \Phi(x) + \epsilon B_Y$ for all k. Since $\Phi(X)$ is a compact set, there is a convergent subsequence $y_k \to y$ as $k \to \infty$ in the subsequence. Since Φ has a closed graph, $(x, y) \in \operatorname{graph}\Phi$, and so $y \in \Phi(x)$ and $y_k \in \Phi(x) + \epsilon B_Y$ for sufficiently large k, contradicting our assumption. Thus Φ is upper semicontinuous. □

Without having compactness, the converse, that closed graph functions are upper semicontinuous, is, however, false. A simple example with closed graph but not upper semicontinuous is the "lighthouse function" as illustrated by Figure 2.2:

$$\Phi(\theta) = \mathbb{R}_+(\cos\theta, \sin\theta). \tag{2.19}$$

This is a ray going from the origin at angle θ to the horizontal. The graph can be seen to be closed. The graph is $\{(\theta, x, y) \mid x = \alpha\cos\theta,\ y = \alpha\sin\theta,\ \alpha \geq 0\}$. Suppose that $\operatorname{graph}\Phi \ni (\theta_k, x_k, y_k) \to (\widehat{\theta}, \widehat{x}, \widehat{y})$ as $k \to \infty$. Since $\alpha = \sqrt{x^2 + y^2}$ on the graph, then we get $\alpha_k = \sqrt{x_k^2 + y_k^2} \to \widehat{\alpha} := \sqrt{\widehat{x}^2 + \widehat{y}^2}$. So

$$\begin{aligned}
\widehat{x} &= \lim_{k\to\infty} x_k = \lim_{k\to\infty} \alpha_k \cos\theta_k = \widehat{\alpha}\cos\widehat{\theta},\\
\widehat{y} &= \lim_{k\to\infty} y_k = \lim_{k\to\infty} \alpha_k \sin\theta_k = \widehat{\alpha}\sin\widehat{\theta}.
\end{aligned}$$

Thus $(\widehat{\theta}, \widehat{x}, \widehat{y}) \in \operatorname{graph}\Phi$, and so $\operatorname{graph}\Phi$ is closed.

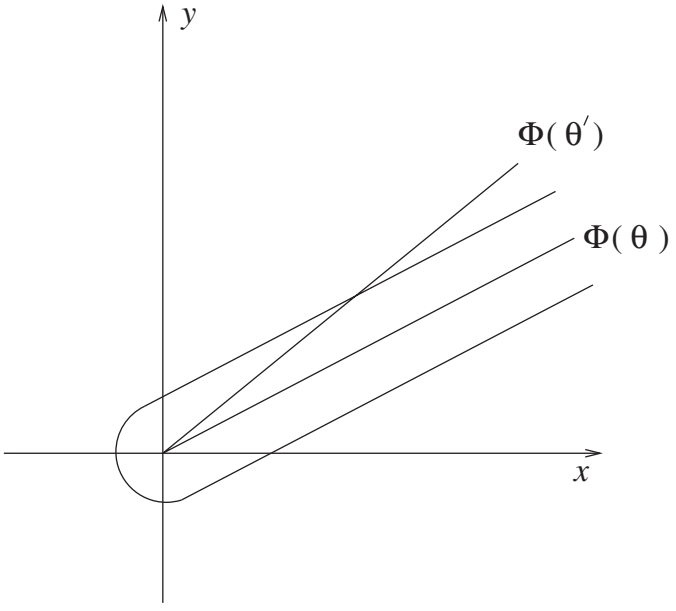


Figure 2.2: Lighthouse function: a set-valued function that has a closed graph but is not upper semicontinuous.

On the other hand, $\Phi(\theta) + B_{\mathbb{R}^2}$ is an open set containing $\Phi(\theta)$; but for any $\theta' \not\equiv \theta \pmod{2\pi}$, $\Phi(\theta') \not\subseteq \Phi(\theta) + B_{\mathbb{R}^2}$. Thus Φ is not upper semicontinuous.

Even though having a closed graph is not enough to imply that a set-valued function is upper semicontinuous, there are other approximation properties that can be shown, at least under some mild assumptions. Our particular concern is with set-valued functions $\Phi\colon \Omega \to \mathcal{P}(X)$ which have closed, convex values.

Pointed cones and strongly pointed cones are important for approximating cones. For the case of closed convex *cone*-valued functions $\Phi\colon \Omega \to \mathcal{P}(X)$ with a closed graph, if $K_0 = \Phi(x_0)$ is a *strongly pointed* cone, then there is a family of cones K_η for $\eta > 0$ (also strongly pointed) with $K_0 = \bigcap_{\eta>0} K_\eta$. Furthermore, for every $\eta > 0$ there is a $\delta > 0$ where $d(x, x_0) < \delta$ implies $\Phi(x) \subseteq K_\eta$. These cones K_η can be constructed as follows. Let S_X be the unit sphere in X:

$$S_X = \{ x \in X \mid \|x\| = 1 \}. \tag{2.20}$$

Lemma 2.2. *If K_0 is a strongly pointed cone in a reflexive Banach space X, then we have that $d\,(0, \overline{\text{co}}(K_0 \cap S_X)) > 0$, and for any $0 < \eta < d\,(0, \overline{\text{co}}(K_0 \cap S_X))$,*

$$K_\eta := \text{cone}\left(\left[\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}\right] \cap S_X\right)$$

is a nested family of closed pointed cones, and

$$K_0 = \bigcap_{\eta>0} K_\eta.$$

Proof. If K_0 is strongly pointed, then $\overline{\text{co}}(K_0 \cap S_X)$ does not contain zero. To see this, suppose we have a sequence $x_\ell \in \text{co}(K_0 \cap S_X)$ where $x_\ell \to 0$ as $\ell \to \infty$. Now $x_\ell = \sum_{i=1}^{n_\ell} \theta_{\ell,i} z_{\ell,i}$, where $\theta_{\ell,i} \geq 0$, $\sum_{i=1}^{n_\ell} \theta_{\ell,i} = 1$, and $z_{\ell,i} \in K_0 \cap S_X$ for all ℓ and i. Now K_0 is a strongly pointed cone, so there is a $\nu \in K^*$ where $\langle \nu, x \rangle \geq \|x\|_X$ for all $x \in K$ by Lemma B.3. Now $\|x_\ell\| \leq 1$ for all ℓ, so by Alaoglu's theorem there is a weak* convergent subsequence, which is weakly convergent if X is reflexive. Let $\widehat{x}$ be the weak limit of this subsequence. Then

$$\begin{aligned}
\langle \nu, \widehat{x} \rangle &= \lim_{\ell \to \infty} \langle \nu, x_\ell \rangle \quad \text{(in the subsequence)} \\
&= \lim_{\ell \to \infty} \sum_{i=1}^{n_\ell} \theta_{\ell,i} \langle \nu, z_{\ell,i} \rangle \\
&\geq \lim_{\ell \to \infty} \sum_{i=1}^{n_\ell} \theta_{\ell,i} \|z_{\ell,i}\| = 1.
\end{aligned}$$

Thus $\widehat{x} \neq 0$. Thus $x_\ell \rightharpoonup \widehat{x} \neq 0$ in the subsequence, which contradicts the strong convergence of $x_\ell \to 0$ as $\ell \to \infty$.

This establishes the fact that K_0 strongly pointed implies $d(0, \overline{\text{co}}(K_0 \cap S_X)) > 0$. For any $0 < \eta < \eta_{max} := d(0, \overline{\text{co}}(K_0 \cap S_X))$, we set

$$K_\eta = \text{cone}\left(\left[\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}\right]\right).$$

Clearly K_η is a convex cone, as it is a cone generated by a convex set. To show that K_η is closed, suppose that $x_\ell \in K_\eta$ and $x_\ell \to x$ in X. If $x = 0$, then clearly $x \in K_\eta$. Otherwise, for sufficiently large ℓ, $x_\ell \neq 0$ and so $x_\ell / \|x_\ell\| \in \overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}$. The set $\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}$ is a weakly closed set, as it is a sum of two weakly compact sets (being bounded closed convex sets in a reflexive Banach space). Since $x_\ell \to x \neq 0$, $x_\ell / \|x_\ell\| \to x / \|x\|$ strongly, we have $x / \|x\| \in \left[\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}\right]$. Thus $x \in K_\eta$, as desired.

We now show that $K_0 = \bigcap_{\eta > 0} K_\eta$. Now $K_0 \cap S_X \subset \left[\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}\right]$ and K_0 is a cone, so $K_0 = \text{cone}(K_0 \cap S_X) \subset K_\eta$ for all $\eta > 0$. We then have to show the reverse inclusion: $\bigcap_{\eta > 0} K_\eta$ is a nested intersection of closed convex sets, and so it is a closed convex set. It is also a cone, as it is an intersection of cones. Suppose $x^* \in \bigcap_{\eta > 0} K_\eta$. Since $\bigcap_{\eta > 0} K_\eta$ and K_0 are cones, we can assume without loss of generality that $\|x^*\| = 1$ by scaling. Then

$$\begin{aligned}
x^* &\in \bigcap_{\eta > 0} \left(K_\eta \cap S_X\right) \\
&= \bigcap_{\eta > 0} \left[\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}\right] \cap S_X \\
&\subset \bigcap_{\eta > 0} \left[\overline{\text{co}}(K_0 \cap S_X) + \eta \overline{B_X}\right].
\end{aligned}$$

Thus $d(x^*, \overline{\text{co}}(K_0 \cap S_X)) \leq \eta$ for all $\eta > 0$. Now K_0 is closed and convex, so $\overline{\text{co}}(K_0 \cap S_X) \subset K_0$, and thus $d(x^*, K_0) \leq \eta$ for all $\eta > 0$. This implies that $d(x^*, K_0) = 0$, and $x^* \in K_0$. That is, $K_0 = \bigcap_{\eta > 0} K_\eta$, as desired. ∎

Without pointedness, this result fails in finite dimensions. Consider, for example, the half-space $K_0 := \{\mathbf{x} = [x, y]^T \in \mathbb{R}^2 \mid x \geq 0\}$. The only convex cone that strictly contains a half-space is the whole space, so $K_\eta = \mathbb{R}^2$ for any $\eta > 0$.

Without *strong* pointedness, this result fails in infinite dimensions. For example, we could take $K_0 := \{\mathbf{x} \in \ell^2 \mid x_1 \geq x_j/j \text{ for } j = 2, 3, \ldots\}$. Now $0 \in \overline{\text{co}}\left(K_0 \cap S_{\ell^2}\right)$ since $\mathbf{x}_j = \left(\mathbf{e}_1/j + \mathbf{e}_j\right)/\sqrt{1 + 1/j^2} \in K_0 \cap S_{\ell^2}$, which converges weakly to zero in ℓ^2. By Mazur's lemma (Lemma A.3), there is a strongly convergent subsequence in $\text{co}\left(K_0 \cap S_{\ell^2}\right)$ which converges to zero, and so $0 \in \overline{\text{co}}\left(K_0 \cap S_{\ell^2}\right)$. The cone generated by any open set containing $\overline{\text{co}}\left(K_0 \cap S_{\ell^2}\right)$ would then contain the entire space.

It would be tempting to believe that if $\Phi\colon \Omega \to \mathcal{P}(X)$ has closed convex values and has a closed graph, then the recession cone function $x \mapsto \Phi(x)_\infty$ also has a closed graph. However, this is not the case, even in finite dimensions with $\Phi(x)_\infty$ (strongly) pointed. Consider, for example, the set-valued function $\Phi\colon \mathbb{R} \to \mathcal{P}(\mathbb{R})$ given by $\Phi(x) = [1/x, \infty)$ for $x > 0$ and $\Phi(x) = \{0\}$ for $x \leq 0$. It can be easily checked that the graph of Φ is closed, with closed convex values. However, $\Phi(x)_\infty = [0, \infty) = \mathbb{R}_+$ for $x > 0$ and $\Phi(x)_\infty = \{0\}$ for $x \leq 0$, so $x \mapsto \Phi(x)_\infty$ does not have a closed graph.

Part of the problem with this example is that the minimum norm point of $\Phi(x)$, $\Pi_{\Phi(x)}(0)$, is unbounded as $x \downarrow 0$. If $\min\{\|y\| \mid y \in \Phi(x)\} \leq R$ for some real R, then we at least have a closed graph for the recession cone if Φ has a closed graph and convex values.

Lemma 2.3. *If $\Phi\colon \Omega \to \mathcal{P}(X)$, X a reflexive Banach space, and Φ is hemicontinuous with closed convex values and $\min_{y \in \Phi(x)} \|y\| \leq R$ for all $x \in \Omega$, then the map $x \mapsto \Phi(x)_\infty$ is also hemicontinuous.*

Proof. Suppose that $x_k \to x$ in Ω and $w_k \in \Phi(x_k)_\infty$, where $w_k \rightharpoonup w$. Suppose also that $\overline{y}_k \in \Phi(x_k)$ with $\|\overline{y}_k\| \leq R$. Since $\Phi(x_k) + \Phi(x_k)_\infty \subseteq \Phi(x_k)$, for any $\tau \geq 0$ we have $\overline{y}_k + \tau\, w_k \in \Phi(x_k)$ for all k. Since the y_k are bounded and X is reflexive, by Alaoglu's theorem, there is a weakly convergent subsequence (which we also denote by $\overline{y}_k$) such that $\overline{y}_k \rightharpoonup \overline{y}$. Thus $\overline{y}_k + \tau\, w_k \rightharpoonup \overline{y} + \tau\, w$. By hemicontinuity, $\overline{y} + \tau\, w \in \Phi(x)$. Since this is true for all $\tau \geq 0$, it follows that $w \in \Phi(x)_\infty$. Hence $x \mapsto \Phi(x)_\infty$ is hemicontinuous. ∎

Hemicontinuity by itself is not a strong condition. For example, consider the convex cone-valued map $\Phi\colon \mathbb{R} \to \mathcal{P}(\ell^2)$ given by

$$\Phi(t) = \begin{cases} \mathbb{R}_+\mathbf{e}_j, & t_{j+1} < t < t_j, \\ \mathbb{R}_+\mathbf{e}_j + \mathbb{R}_+\mathbf{e}_{j+1}, & t = t_{j+1}, \\ \{0\}, & t \leq 0, \end{cases}$$

where $t_j \downarrow 0$ as $j \to \infty$ and $t_1 = +\infty$. This is a hemicontinuous set-valued map at zero since for any sequence $y_k \in \Phi(s_k)$ with $s_k \to 0$ and $y_k \rightharpoonup y$ we must have $y = 0$. And yet, $\Phi(0) = \{0\}$ gives essentially no information about $\Phi(t)$ for small nonzero t. In particular, outer approximations to $\Phi(t)$ for small t must use more information than can be found in $\Phi(0)$. Strong pointedness of $\Phi(0)$ in particular is insufficient to construct suitable outer approximations for $\Phi(t)$. However, if we know that $\Phi(t)$ has a suitable outer approximation (no matter how "big"), we can use this to give arbitrarily close approximations (in a suitable sense).

Define $H_{\xi,\alpha}$ for $\xi \in X'$ and $\alpha \in \mathbb{R}$ to be the half-space

$$H_{\xi,\alpha} = \{ y \in X \mid \langle \xi, y \rangle + \alpha \geq 0 \}.$$

Our basic result for outer approximations of set-valued functions is as follows, using the support function $\sigma_K(\eta) = \sup_{x \in K} \langle \eta, x \rangle$.

Theorem 2.4. *Suppose that $\Phi\colon \Omega \to \mathcal{P}(X)$ is hemicontinuous with closed convex values in X, a reflexive Banach space. Suppose also that for any $x_0 \in \Omega$ and a neighborhood $\mathcal{U}$ of x_0,*

$$\Phi(x) \subseteq L + R\overline{B_X}$$

with L a strongly pointed closed convex cone and $R \geq 0$, and that $\min_{y \in \Phi(x)} \|y\| \leq R$ for all $x \in \mathcal{U}$. Let $K = \Phi(x_0)$. Then for any $-\xi \in \operatorname{int}\operatorname{dom}\sigma_K$ and $\alpha > \sigma_K(-\xi)$, there is a $\delta > 0$ such that

$$d(x, x_0) < \delta \Rightarrow \Phi(x) \subseteq H_{\xi,\alpha}.$$

Proof. Suppose otherwise. Then there is a sequence $x_k \to x_0$ as $k \to \infty$ in Ω and $y_k \in \Phi(x_k)$ with $y_k \notin H_{\xi,\alpha}$ for some $-\xi \in \operatorname{int}\operatorname{dom}\sigma_K$ and $\alpha > \sigma_K(-\xi)$. Since $-\xi \in \operatorname{int}\operatorname{dom}\sigma_K$ and σ_k is a convex lower semicontinuous function, there is a closed neighborhood of $-\xi + \theta\overline{B_X}$ on which σ_K is continuous. By choosing $\theta > 0$ sufficiently small, we can ensure that for $-\xi' \in -\xi + \theta\overline{B_X}$ we have

$$\left|\sigma_K(-\xi') - \sigma_K(-\xi)\right| \leq \frac{1}{2}(\alpha - \sigma_K(-\xi)),$$

and so $\sigma_K(-\xi') \leq \frac{1}{2}(\alpha + \sigma_K(-\xi)) < \alpha$ for all such ξ'. Let $\alpha' := \frac{1}{2}(\alpha + \sigma_K(-\xi)) < \alpha$.

Since $y_k \notin H_{\xi,\alpha}$we have $\langle \xi, y_k \rangle + \alpha < 0$ for all k.

Suppose first that y_k is a bounded sequence. Then by Alaoglu's theorem and reflexivity of X, there is a weakly convergent subsequence (also denoted y_k) such that $y_k \rightharpoonup y$. As Φ is hemicontinuous, $y \in \Phi(x_0) = K$.

From weak convergence,

$$\langle \xi, y_k \rangle + \alpha \to \langle \xi, y \rangle + \alpha \leq 0 \qquad \text{for } k \to \infty.$$

So

$$\begin{aligned} \langle -\xi, y \rangle &\geq \alpha > \sigma_K(-\xi) \\ &= \sup_{w \in K} \langle -\xi, w \rangle \geq \langle -\xi, y \rangle, \end{aligned}$$

which is a contradiction.

Now suppose that y_k is an unbounded sequence; by choosing a suitable subsequence we can ensure that $\|y_k\| \uparrow \infty$ as $k \to \infty$. By Lemma B.3, pick $\zeta \in \operatorname{int} L^*$ such that $\langle \zeta, w \rangle \geq \|w\|$ for all $w \in L$. We will need this later.

Suppose that $\eta \in X'$ with $\|\eta\|_{X'} \leq \theta$, so that $\sigma_K(-\xi + \eta) \leq \alpha'$. Thus $\langle \xi - \eta, y_k \rangle + \alpha' \geq 0$. Now $y_k / \|y_k\|$ is a bounded sequence in a reflexive Banach space X, and so it has a

weakly convergent subsequence $y_k/\|y_k\| \rightharpoonup \widehat{y}$. So taking limits, in the subsequence, of

$$\left\langle \xi - \eta, \frac{y_k}{\|y_k\|} \right\rangle \geq \frac{\alpha'}{\|y_k\|} \qquad \text{gives}$$
$$\langle \xi - \eta, \widehat{y} \rangle \geq 0.$$

Since this is true for all $\eta \in X'$ with $\|\eta\|_{X'} \leq \theta$, we have $\langle \xi, \widehat{y} \rangle \geq \theta \, \|\widehat{y}\|$ for some $\theta > 0$.

If $\widehat{y} = 0$, we would not be able to obtain a contradiction. We need to use $\Phi(x_k) \subseteq L + R\,\overline{B_X}$ and strong pointedness of L to show that $\widehat{y} \neq 0$. For each k write $y_k = u_k + v_k$, $u_k \in L$, and $\|v_k\| \leq R$. Now

$$\begin{aligned} \langle \zeta, y_k \rangle &= \langle \zeta, u_k \rangle + \langle \zeta, v_k \rangle \\ &\geq \|u_k\| - R\,\|\zeta\|_{X'} \\ &\geq \|y_k\| - 2R\,\|\zeta\|_{X'}. \end{aligned}$$

Thus

$$\left\langle \zeta, \frac{y_k}{\|y_k\|} \right\rangle \geq 1 - \frac{\|\zeta\|_{X'}}{\|y_k\|} 2R \to 1 \qquad \text{as } k \to \infty$$

in the subsequence. Taking weak limits, $\langle \zeta, y_k/\|y_k\| \rangle \to \langle \zeta, \widehat{y} \rangle \geq 1$, so $\|\widehat{y}\| \geq 1/\|\zeta\|_{X'} > 0$.

For each k we can choose $\overline{y}_k \in \Phi(x_k)$ with $\|\overline{y}_k\| \leq R$. By Alaoglu's theorem and reflexivity of X, there is a weakly convergent subsequence to which we restrict our attention so that $\overline{y}_k \rightharpoonup \overline{y}$ in the subsequence. By convexity of $\Phi(x_k)$ for all k, for any $0 \leq \beta_k \leq 1$ we have $\overline{y}_k + \beta_k \left(y_k - \overline{y}_k\right) \in \Phi(x_k)$. In particular, for a given $\tau \geq 0$ we can set $\beta_k = \min(1, \tau/\|y_k\|)$. Then as $\|y_k\| \to \infty$, for sufficiently large k,

$$\overline{y}_k + \frac{\tau}{\|y_k\|} \left(y_k - \overline{y}_k\right) \in \Phi(x_k).$$

Taking weak limits on the left and using hemicontinuity of Φ, we see that

$$\overline{y} + \tau\, \widehat{y} \in \Phi(x_0).$$

As this is true for all $\tau \geq 0$, $\widehat{y} \in \Phi(x_0)_\infty$. Since $\Phi(x_0) = K \subseteq H_{\xi,\alpha}$, $\widehat{y} \in \left(H_{\xi,\alpha}\right)_\infty = H_{\xi,0}$; that is, $\langle \xi, \widehat{y} \rangle \leq 0$. However, we have already seen that $\langle \xi, \widehat{y} \rangle \geq \theta \, \|\widehat{y}\| > 0$. This is a contradiction.

Thus there must be a $\delta > 0$ such that

$$d(x, x_0) < \delta \Rightarrow \Phi(x) \subseteq H_{\xi,\alpha}. \quad \square$$

In finite dimensions, hemicontinuity is equivalent to having a closed graph. The condition "$\Phi(x) \subseteq L + R\,\overline{B_X}$, L strongly pointed, for all x in a neighborhood of x_0" looks like a difficult condition to check, but in finite dimensions this can be reduced to simply requiring that "$\Phi(x_0)_\infty$ is a pointed cone."

Lemma 2.5. *Suppose that $\Phi\colon \Omega \to \mathcal{P}(\mathbb{R}^n)$ has a closed graph with closed convex values and $\min_{y \in \Phi(x)} \|y\| \leq R$ for all $x \in \mathcal{U}$, $\mathcal{U}$ a neighborhood of x_0. If, in addition, $\Phi(x_0)_\infty$ is a pointed cone, then there is a (strongly) pointed cone L, $R > 0$, and $\delta > 0$ such that*

$$d(x, x_0) < \delta \Rightarrow \Phi(x) \subseteq L + R\,\overline{B_{\mathbb{R}^n}}.$$

Proof. Take $K = \Phi(x_0)_\infty$ and $L = K_\eta$ for some $\eta > 0$ as given in Lemma 2.2. We prove the result by contradiction. Suppose that $x_k \to x_0$ in Ω and there exist $y_k \in \Phi(x_k)$ such that $\|y_k - \Pi_L(y_k)\| \to \infty$ as $k \to \infty$. Then $\|y_k\| \to \infty$. Now $y_k / \|y_k\|$ are in a bounded closed set, and so there is a convergent subsequence. Restrict attention to this subsequence, and let $y_k / \|y_k\| \to \widehat{y}$ in the subsequence. By the same arguments as in Theorem 2.4, $\widehat{y} \in K$. Now $L = K_\eta$ contains a neighborhood of $K \cap S_{\mathbb{R}^n} \ni \widehat{y}$. Thus $y_k / \|y_k\| \in L$ for sufficiently large k, and thus $y_k \in L$ for sufficiently large k. Hence $\|y_k - \Pi_L(y_k)\| = 0$ for sufficiently large k, which contradicts $\|y_k - \Pi_L(y_k)\| \to \infty$. □

These approximations will be particularly useful in dealing with measure differential inclusions (MDIs). See Section 4.4.

2.1.4 Measurability considerations

For matters of integration and differential equations, it is necessary to deal with matters of measurability of set-valued functions. Measurability is such a common property that these considerations tend to be rather technical with little practical impact. Nevertheless, for the *existence* of solutions, it can be important that the desired functions are shown to be measurable so that their integrals are meaningful. Section A.4 contains basic material on measurability for single-valued functions and on σ-algebras.

Let X be a measure space (with its own σ-algebra of measurable sets $\mathcal{A}$) and Y be a topological space with a σ-algebra of measurable sets $\mathcal{B}$. Recall that $f\colon X \to Y$ is measurable if $f^{-1}(E)$ is measurable in X for every measurable set E in Y; that is, for all $E \in \mathcal{B}$, $f^{-1}(E) \in \mathcal{A}$. Often we consider Y merely a topological space, in which case we take $\mathcal{B}$ to be the σ-algebra of Borel sets in Y. In that case, $f\colon X \to Y$ is measurable if $f^{-1}(U)$ is measurable in X for every open set U in Y.

We say a set-valued function $\Phi\colon X \to \mathcal{P}(Y)$ is *strongly measurable* if $\Phi^-(C)$ is measurable in X for every closed set C in Y; we say Φ is *weakly measurable* if $\Phi^-(U)$ is measurable in X for every open set U in Y. We define $\mathcal{A} \otimes \mathcal{B}$ to be the σ-algebra generated by the Cartesian products $E \times F$ with $E \in \mathcal{A}$ and $F \in \mathcal{B}$. Then we have the characterization theorem of measurability (see, for example, [21, Thm. 8.1.4] or [129, Thm. 2.4]).

Theorem 2.6. *For a set-valued function $\Phi\colon X \to \mathcal{P}(Y)$, where X is a measure space and Y a complete separable metric space with $\mathcal{A}$ the σ-algebra of measurable sets of X and $\mathcal{B}$ the σ-algebra of Borel sets of Y, the following are equivalent:*

1. *Φ is strongly measurable.*
2. *Φ is weakly measurable.*
3. *$\operatorname{Graph}\Phi$ is measurable in $X \times Y$; that is, $\operatorname{graph}\Phi \in \mathcal{A} \otimes \mathcal{B}$.*
4. *$\Phi^-(E)$ is measurable for all Borel $E \subseteq Y$.*
5. *For all $y \in Y$, the function $x \mapsto d_Y(y, \Phi(x))$ is a measurable function $X \to \mathbb{R}$.*

Theorem 2.6 can be used to create an "arithmetic" of measurable functions: unions, intersections, Cartesian products, and compositions of measurable set-valued functions to

separable spaces that are also measurable. Also, upper semicontinuous and lower semicontinuous functions are measurable. If the conditions of Theorem 2.6 hold, we drop the qualifiers "weak" and "strong" and simply say that Φ is measurable. An important consequence of measurability of a set-valued function is the existence of a single-valued *selection* of Φ.

Lemma 2.7. *If $\Phi\colon X \to \mathcal{P}(Y)\setminus\{\emptyset\}$ is measurable and Y is a separable metric space, then there is a measurable selection $f\colon X \to Y$ such that $f(x) \in \Phi(x)$ for all $x \in X$.*

A proof can be found in, for example, [21, Thm. 8.1.3] or in [4, Cor. 18.14]. Another important consequence of measurability of set-valued functions is the Filippov lemma below. If A is a measurable space and X and Y are topological spaces, then a function $f\colon A \times X \to Y$ is a *Carathéodory function* if for each $a \in A$, $x \mapsto f(a,x)$ is continuous and for each $x \in X$, $a \mapsto f(a,x)$ is measurable. Carathéodory functions are measurable functions on $A \times X$ with the σ-algebra of measurable sets $\mathcal{A}\otimes\mathcal{B}$, where $\mathcal{A}$ is the collection of measurable sets of A and $\mathcal{B}$ is the collection of Borel sets of X (see, for example, [4, Lem. 4.51]). From this it is easy to show the following lemma.

Lemma 2.8 (Filippov implicit function lemma). *Suppose that A is a measurable space, X and Y are separable metric spaces, $\Phi\colon A \to \mathcal{P}(X)$ and $g\colon A \to Y$ are measurable, and $f\colon A \times X \to Y$ is a Carathéodory function. Then the set-valued function $\Psi\colon A \to \mathcal{P}(X)$ given by*

$$\Psi(a) = \{x \in \Phi(x) \mid f(a,x) = g(a)\}$$

is a measurable function, and so it has a measurable selection $h\colon A \to X$, $h(a) \in \Psi(a)$ for all $a \in A$.

Proofs can be found in, for example, [4, 21, 101, 167]. This lemma is important to avoid problems of nonmeasurability when there is nonuniqueness in a representation, such as in differential inclusions (see Section 4.1).

2.2 Complementarity problems

Complementarity problems (CPs) have the following form: Given $F\colon \mathbb{R}^n \to \mathbb{R}^n$, find $z \in \mathbb{R}^n$ such that

$$0 \le z \quad \perp \quad F(z) \ge 0. \tag{2.21}$$

Note that "$a \ge 0$" for a vector a means that the components $a_i \ge 0$ for all i, and "$a \perp b$" means that $a^T b = 0$, or that the inner or dot product of a and b is zero. For all our CPs, we will assume that F is a continuous function. We denote the problem (2.21) by CP(F).

If F is an affine function $F(z) = Mz + q$, then we call (2.21) a *linear complementarity problem* (LCP) [67]: Given $M \in \mathbb{R}^{n\times n}$ and $q \in \mathbb{R}^n$, find $z \in \mathbb{R}^n$ such that

$$0 \le z \quad \perp \quad Mz + q \ge 0. \tag{2.22}$$

This is denoted LCP(q, M).

Generalized complementarity problems (GCPs) replace the componentwise inequality "$a \geq 0$" with a more general condition "$a \in K$," where K is a closed convex cone. The GCP is the problem of finding z satisfying

$$K \ni z \quad \perp \quad F(z) \in K^*. \tag{2.23}$$

Here K^* is the *dual cone* to K (see (B.8)).

We denote the problem (2.23) by CP(F, K). Note that CP(F) = CP($F, \mathbb{R}^n_+$) for $F\colon \mathbb{R}^n \to \mathbb{R}^n$. Again, if F is affine ($F(z) = Mz + q$), then we have a *generalized linear complementarity problem* (GLCP), which is denoted by LCP(q, M, K).

CPs date back to the early 1960s with the work of Lemke and Howson [157] and Cottle and Dantzig [66], who worked essentially with LCPs. The connections with quadratic programming with inequality constraints were soon identified [271].

CPs can be obtained from constrained optimization problems via the Kuhn–Tucker conditions. For example, consider the problem of minimizing $f(x)$ subject to $c_i(x) \geq 0$ for $i = 1, 2, \ldots, m$. Then if we write $L(x, \lambda) = f(x) - \sum_{i=1}^m \lambda_i c_i(x)$, the Kuhn–Tucker conditions become

$$0 = \nabla_x L(x, \lambda), \tag{2.24}$$

$$0 \leq \lambda_i \perp c_i(x) \geq 0 \qquad \text{for all } i, \tag{2.25}$$

provided a suitable constraint qualification is satisfied. One possible constraint qualification is that $\{\nabla c_i(x) \mid i = 1, 2, \ldots, m, \text{ and } c_i(x) = 0\}$ is a linearly independent set for any x. This is known as the linear independence constraint qualification (LICQ). A refined version is the Mangasarian–Fromowitz constraint qualification (MFCQ), which for the case of only inequality constraints requires that for any x there be a vector d where $\nabla c_i(x) d < 0$ for all i, where $c_i(x) = 0$. In the case where $-c_i$ is convex for all i, there is the Slater constraint qualification which simply requires the existence of a point x_0, where $c_i(x_0) > 0$ for all i.

2.2.1 Lemke's algorithm

For an LCP with $K = \mathbb{R}^n_+$, Lemke's method is the most common method of computing a solution. It is also an excellent technique for proving the existence of solutions. Let us consider LCP(q, M): Given $M \in \mathbb{R}^{n \times n}$ and $q \in \mathbb{R}^n$, find $z \in \mathbb{R}^n$ such that

$$0 \leq z \quad \perp \quad Mz + q \geq 0.$$

This can be solved using methods similar to the simplex method for linear programming.

A quick outline of the simplex method

Linear programming is the problem of minimizing a linear function subject to linear inequality constraints. This can be put into the standard form:

$$\min_x c^T x + d \qquad \text{subject to} \tag{2.26}$$

$$Ax = b, \qquad x \geq 0. \tag{2.27}$$

As usual "$a \geq b$" for vectors a and b means that "$a_i \geq b_i$ for all i." We will assume that $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ so that A is an $m \times n$ matrix ($n \geq m$). The simplex method uses a

simplex tableau, which is a representation of the linear program:

$$\begin{array}{c|c} b & A \\ \hline & c^T \end{array}$$

This tableau has a *basis* B which is a subset of m indexes of the variables x_i: $B \subseteq \{1, 2, \ldots, n\}$, where for each $i \in B$, the ith column of A is a column of the identity matrix, and the matrix $\left[a_{ij} \mid i \in B, j = 1, 2, \ldots, m\right]$ is a *permutation matrix* (that is, a matrix formed by permuting either rows or columns of the identity matrix). Writing $B = \{i_1, i_2, \ldots, i_m\}$ so that the i_jth column of A is the jth column of the identity matrix, we can read off the values of x_i associated with the simplex tableau: $x_{i_j} = b_j$ for $j = 1, 2, \ldots, m$, and $x_i = 0$ if $i \notin B$. For this to be a feasible point ($x_i \geq 0$ for all i) we need $b_j \geq 0$ for all j.

To deal with the cost vector c associated with the linear program, we suppose that $c_i = 0$ for all $i \in B$. If $c_i < 0$ for some $i \notin B$, then we have an opportunity to reduce the cost associated with the simplex tableau by means of an operation called *pivoting*. Let us suppose that $b_j > 0$ for all j; the other case will be considered later.

If $c_i < 0$, then the point x associated with the current simplex tableau has $x_i = 0$, and $i \notin B$. If we increase x_i from the value zero, then we will have to change the values of the x_k variables for $k \in B$, but we will leave the values of x_k for $k \notin B$ and $k \neq i$ unchanged. For $k \in B$ we will have to change x_k from $x_k = b_k$ to

$$x_k = b_k - a_{ki} x_i.$$

The value of the objective function decreases by $c_i x_i$. If $a_{ki} \leq 0$ for all k, then there is no limit to how much we can increase x_i while staying feasible, and so there is no limit to how much we can decrease the objective function. Such a linear program has no solution: it has just an infimum of $-\infty$.

If some $a_{ki} > 0$, then we cannot increase x_i without limit. Instead, we have to ensure that $b_k - a_{ki} x_i \geq 0$. That is, we cannot make x_i larger than b_k / a_{ki} for $a_{ki} > 0$. Since this must hold for *all* $k = 1, 2, \ldots, m$, then the most we can make x_i is $\min_{k:a_{ki}>0} b_k / a_{ki}$. Pick ℓ as the minimizing value of k, and let j be the index, where $a_{\ell j} = 1$; that is, x_j is the ℓth basis variable. We will assume that this is unique for now. If we increase x_i to this upper limit, then we have $x_\ell = 0$ after the increase, and we should take ℓ out of the basis B. Bringing i into the basis and taking j out of the basis can be carried out by means of adding multiples of row ℓ to other rows, which bring to zero all rows of column i except row ℓ, and we scale row ℓ to make $a_{i\ell} = 1$. At the same time, in order to keep the same feasible set, whatever row operations we do to the A matrix should also be done to the vector b. Finally, for the cost vector c to reflect these changes, we should subtract a multiple of row ℓ from c to set $c_j = 0$. Finally the objective value at the new point must be updated: $d \leftarrow d + c_i x_i$.

If we always have $b_j > 0$ for all j, then at each stage of the simplex method, either we strictly reduce the objective function value $c_j \geq 0$ for all j and we are at a minimum, or we discover that there is no minimum. Since there is only a finite number of possible basis sets B, and the basis set determines the simplex tableau, then the simplex method cannot cycle, and so it must terminate. If we get $b_j = 0$ for some j during the simplex method, then there is the possibility that there is no reduction in the objective function value. This is known as *degeneracy*. In this case the method can cycle. Even though degeneracy is destroyed by small, random perturbations to the data, this is an important practical issue, and cycling can occur in practical problems using practical implementations of the simplex method unless steps are taken to prevent this. The most commonly presented method for

handling degeneracy is lexicographical degeneracy resolution. Although this is not the best performing method computationally, it at least resolves the theoretical questions of existence of minima. Details of how lexicographical degeneracy resolution works can be found in, for example, [106]. More practical methods can be found in, for example, [192].

The basic idea of lexicographical degeneracy resolution is that for determining the variable to remove from the basis, instead of choosing $k = \ell$ to be the minimizer of b_k/a_{ki}, where i is the index of the variable entering the basis, we use the lexicographical ordering of the vectors $[b_k, a_{k1}, a_{k2}, \ldots, a_{kn}]/a_{ki}$ for $a_{ki} > 0$: in the lexicographic ordering $u <_L v$ for vectors $u, v \in \mathbb{R}^n$ if $u \neq v$ and for $p = \min\{ j \mid u_j \neq v_j \}$, $u_p < v_p$. We say $u \leq_L v$ if $u <_L v$ or $u = v$. Note that "$<_L$" is a complete ordering of $\mathbb{R}^n$; that is, for any two vectors $u, v \in \mathbb{R}^n$ either $u <_L v$, or $u = v$, or $v <_L u$. If we arrange for the initial tableau to have $b \geq 0$ and the initial variables in the basis to be $x_1, x_2, \ldots, x_m$, then the tableau has the form

$$\left[b \mid I, A' \right]$$

and the rows of the initial tableau are lexicographically positive. That is,

$$0 <_L [b_k, a_{k1}, a_{k2}, \ldots, a_{kn}] \qquad \text{for all } k.$$

By choosing the lexicographical minimizer of $[b_k, a_{k1}, a_{k2}, \ldots, a_{kn}]/a_{ki}$ over k with $a_{ki} > 0$, we ensure that the subsequent tableau has lexicographically positive rows. Furthermore, for there to be a tie (two rows giving the same lexicographical minimum), two of the rows of the tableau have to be linearly dependent, which is impossible since at each stage A contains an $m \times m$ permutation matrix associated with the columns in the basis B.

The lexicographical degeneracy resolution method enables us to prove the following *reversibility lemma* for simplex tableau pivoting.

Lemma 2.9. *Suppose $[b \mid A]$ is a simplex tableau with lexicographically positive rows and basis B if we perform a simplex pivot to bring a variable x_p ($p \notin B$) into the basis, removing x_q according to the lexicographical rule and producing tableau $\left[b' \mid A' \right]$ with basis $B' = (B \setminus \{q\}) \cup \{p\}$. Then bringing variable x_q into the basis for tableau $\left[b' \mid A' \right]$ produces tableau $[b \mid A]$ with basis B.*

This result turns out to be essential for understanding the Lemke method described in the next section.

Proof. For $1 \leq i \leq m$, let $\pi(i)$ be the index of the basic variable $x_{\pi(i)}$ associated with row i in tableau $[b \mid A]$. Let k be the row associated with variable x_q which is removed from the basis B in tableau $[b \mid A]$; $\pi(k) = q$. Thus $a_{kp} > 0$ and

$$[b_k, a_{k1}, \ldots, a_{kn}]/a_{kp} <_L [b_i, a_{i1}, a_{i2}, \ldots, a_{in}]/a_{ip} \qquad \text{for all } i \neq k.$$

After the simplex pivot step, A' has entries $a'_{kp} = 1$, $a'_{ip} = 0$ for $i \neq k$, and

$$\begin{aligned} \left[b'_k, a'_{k1}, \ldots, a'_{kn} \right] &= [b_k, a_{k1}, \ldots, a_{kn}]/a_{kp}, \\ \left[b'_i, a'_{i1}, \ldots, a'_{in} \right] &= [b_i, a_{i1}, \ldots, a_{in}] - \frac{a_{ip}}{a_{kp}} [b_k, a_{k1}, \ldots, a_{kn}]. \end{aligned}$$

If $\pi'(i)$ is the index of the basic variable associated with row i in tableau $\left[b' \mid A' \right]$, then $\pi'(k) = p$ and $\pi'(i) = \pi(i)$ for $i \neq k$.

Now we want to show that if we bring variable x_q into the basis B' of tableau $\left[b' \mid A'\right]$, we must remove x_q from the basis; that is, we want to show that row k gives the lexicographical minimum of $\left[b'_i, a'_{i1}, \ldots, a'_{in}\right]/a'_{iq}$ over i, where $a'_{iq} > 0$. To do this, note that $a'_{kq} = a_{kq}/a_{kp} = 1/a_{kp} > 0$, and for $i \neq k$, $a'_{iq} = a_{iq} - a_{ip}a_{kq}/a_{kp} = -a_{ip}/a_{kp}$ since $a_{kq} = 1$ and $a_{iq} = 0$ if $i \neq k$. Then

$$\left[b'_k, a'_{k1}, \ldots, a'_{kn}\right]/a'_{kq} = [b_k, a_{k1}, \ldots, a_{kn}],$$
$$\left[b'_i, a'_{i1}, \ldots, a'_{in}\right]/a'_{iq} = [b_k, a_{k1}, \ldots, a_{kn}] - \frac{a_{kp}}{a_{ip}}[b_i, a_{i1}, \ldots, a_{in}]$$

for $i \neq k$. But we consider such rows for the lexicographical minimum only if $a'_{iq} = -a_{ip}/a_{kp} > 0$. Since $[b_i, a_{i1}, \ldots, a_{in}]$ is lexicographically positive, it follows that

$$\left[b'_k, a'_{k1}, \ldots, a'_{kn}\right]/a'_{kq} <_L \left[b'_i, a'_{i1}, \ldots, a'_{in}\right]/a'_{iq}$$

whenever $a'_{iq} > 0$. Thus if we bring x_q into the basis in tableau $\left[b' \mid A'\right]$, we must remove x_p. Elementary calculations show that the resulting tableau is $[b \mid A]$, as desired. □

Lemke's method via simplex tableaus

Lemke's method is based on the simplex method, but without the cost vector c. Instead we rewrite the LCP as

$$\begin{aligned} Iw - Mz &= q, \\ z, w &\geq 0, \\ z^T w &= 0. \end{aligned}$$

So we start with an initial simplex tableau with the variables $x_i = w_i$ for $i = 1, 2, \ldots, n$, and $x_{n+i} = z_i$ for $i = 1, 2, \ldots, n$. If the vector $q \geq 0$, then the point associated with the tableau $w = q$, $z = 0$ is feasible and we have a solution of the LCP. Unfortunately, this is rarely the case: usually some $q_i < 0$.

To handle this we add an extra variable $s \geq 0$ and a vector d with $d_i > 0$ for all i called the *covering vector*, and s is called a *slack variable*. The system of the tableau then becomes

$$\begin{aligned} Iw - Mz - sd &= q, \\ s, z, w &\geq 0, \\ z^T w &= 0. \end{aligned}$$

To start Lemke's algorithm, we do the operations to bring s into the basis and make the tableau feasible so that the vector on the right $(q + sd)$ is nonnegative. That is, we increase s until $q + sd \geq 0$. In fact, we increase s until we reach the smallest value where this is true: $s = \min_{k:q_k<0} -q_k/d_k \geq 0$. The value $\ell = k$ which gives the minimum indicates that the variable w_ℓ must be removed from the basis.

Complementarity is then used to decide which variable must next be brought into the basis. After bringing s into the basis and removing w_ℓ, the only variable that can be

brought into the basis without violating complementarity ($z^T w = 0$) is z_ℓ. In general, if z_j is removed from the basis in one simplex step, then we must (try to) bring in w_j in the next step; conversely if w_j is removed at the end of one simplex step, then we must (try to) bring in z_j in the next step. There are two ways in which this process can stop. One is if s is removed from the basis in a simplex step. In the resulting simplex tableau, the associated point has $w - Mz = q$, $z, w \geq 0$ and $z^T w = 0$. In other words, we have found a solution to the LCP. The other is if we find that we have an unbounded ray of feasible points: $(s, z, w) = (s_0, z_0, w_0) + \alpha(s_\infty, z_\infty, w_\infty)$, $(s_\infty, z_\infty, w_\infty) \neq 0$, and

$$\begin{aligned} I\,w - M\,z - s\,d &= q, \\ s, z, w &\geq 0, \\ z^T w &= 0 \qquad \text{for all } \alpha \geq 0. \end{aligned}$$

Since we have not brought s out of the basis, we must have $s_0 > 0$ in this case. Clearly (taking $\alpha = 0$), $s_0, z_0, w_0 \geq 0$, and also (taking $\alpha \to \infty$) $s_\infty, z_\infty, w_\infty \geq 0$. Also, $w_0 = Mz_0 + s_0 d + q$, and $w_\infty = Mz_\infty + s_\infty d$. From the complementarity conditions $z^T w = 0$, we see that $(z_0 + \alpha z_\infty)^T (w_0 + \alpha w_\infty) = 0$ for $\alpha > 0$. Since all vectors in this inner product are nonnegative, this implies that $z_0^T w_0 = z_\infty^T w_0 = z_0^T w_\infty = z_\infty^T w_\infty = 0$.

If we focus on what happens as $\alpha \to \infty$, we remove q from consideration and focus only on the matrix. In linear complementarity theory, there are a wide range of matrix classes that are important. We will have a look at these in the next section.

Matrix classes and Lemke's algorithm

Some LCPs do not have solutions, and those for which we can guarantee existence usually have some kind of "positivity" property. The first is the property of being *copositive*: A matrix $M \in \mathbb{R}^{n \times n}$ is copositive if

$$z \geq 0 \Longrightarrow z^T M z \geq 0. \tag{2.28}$$

The matrix M being copositive is not sufficient to ensure the existence of solutions of LCP(q, M) for all q. Two conditions are known to be sufficient for this in addition to copositivity. A matrix $M \in \mathbb{R}^{n \times n}$ is *strictly copositive* if

$$\left[\, z \geq 0 \;\&\; z \neq 0 \,\right] \Longrightarrow z^T M z > 0. \tag{2.29}$$

A matrix $M \in \mathbb{R}^{n \times n}$ is *copositive plus* if it is copositive and

$$\left[z \geq 0 \;\&\; z^T M z = 0 \right] \Longrightarrow \left(M + M^T \right) z = 0 \qquad \text{for all } z. \tag{2.30}$$

To compare with other well-known matrix classes, note that any positive semidefinite matrix is copositive and any positive definite matrix is strictly copositive. However, matrices with nonnegative entries are copositive; if, in addition, every row or every column of M has a strictly positive entry, then M is strictly copositive. Symmetric positive semidefinite matrices are copositive plus.

The set of copositive $n \times n$ matrices is a closed convex cone in the space of $n \times n$ matrices.

To see how these matrix classes relate to Lemke's algorithm and the solution of LCPs, note that if Lemke's algorithm terminates at an unbounded ray $(s, z, w) = (s_0, z_0, w_0) + \alpha(s_\infty, z_\infty, w_\infty) \geq 0$ with $\alpha \geq 0$, then we have the following properties:

$$\begin{aligned} w_0 &= Mz_0 + s_0 d + q, \\ w_\infty &= Mz_\infty + s_\infty d, \\ 0 &= z_0^T w_0 = z_\infty^T w_0 = z_0^T w_\infty = z_\infty^T w_\infty. \end{aligned}$$

Recall that $0 \leq (s_\infty, z_\infty, w_\infty) \neq 0$ and $s_0 > 0$. So, for copositive M,

$$\begin{aligned} 0 = z_\infty^T w_\infty &= z_\infty^T (Mz_\infty + s_\infty d) \\ &\geq s_\infty z_\infty^T d \geq 0. \end{aligned}$$

Since d is a vector of strictly *positive* entries, this means that either $s_\infty = 0$ or $z_\infty = 0$. If $s_\infty > 0$, then $z_\infty = 0$, and so $w_\infty = s_\infty d$, and our unbounded ray corresponds to the first feasible basis of the tableau. By the reversibility lemma (Lemma 2.9), this is impossible. Thus we must conclude that $s_\infty = 0$.

If M is strictly copositive, then we have $z_\infty^T M z_\infty = 0$, which implies that $z_\infty = 0$. With both s_∞ and z_∞ zero, we see that $w_\infty = 0$, and so we do not really have a ray at all: $(s_\infty, z_\infty, w_\infty) = 0$. So if M is strictly copositive, then Lemke's method cannot terminate at an unbounded ray. The only possibility left is that Lemke's method finds a solution of the LCP.

If M is copositive plus, then the arguments are a little more complicated, and we prove a weaker result: *Lemke's algorithm finds a solution if and only if a solution exists*. So we start out by assuming that Lemke's algorithm fails. Again, we have $s_\infty = 0$ and $z_\infty^T M z_\infty = 0$ for terminating at a ray. Thus $w_\infty = Mz_\infty \geq 0$. Again, note that we cannot have $z_\infty = 0$, for then $(s_\infty, z_\infty, w_\infty) = 0$ and there is no unbounded ray. For M copositive plus, $z_\infty^T M z_\infty = 0$ implies $\left(M + M^T\right) z_\infty = 0$. Thus $M^T z_\infty = -Mz_\infty = -w_\infty \leq 0$. Then

$$\begin{aligned} 0 = z_0^T w_\infty &= z_0^T \left(-M^T z_\infty\right) = -z_\infty^T M z_0, \\ 0 = z_\infty^T w_0 &= z_\infty^T (Mz_0 + s_0 d + q) = s_0 z_\infty^T d + z_\infty^T q. \end{aligned}$$

Since $s_0 > 0$ and $z_\infty^T d > 0$, then $z_\infty^T q < 0$. It then turns out that the feasible set $\{(z, w) \mid w = Mz + q,\ z, w \geq 0\}$ must be empty. Suppose that there is a feasible z and w. Then $0 \leq z_\infty^T w = z_\infty^T (Mz + q) = (-Mz_\infty)^T z + z_\infty^T q = -w_\infty^T z + z_\infty^T q < 0$, which is impossible.

Another important class of matrices related to Lemke's method is the class of *P-matrices*. A P-matrix $M \in \mathbb{R}^{n \times n}$ is a square matrix where every principal submatrix has positive determinant [67, Def. 3.3.1]. That is, for every subset $I \subseteq \{1, 2, \ldots, n\}$ the submatrix $M_{II} := \left[m_{ij} \mid i, j \in I\right]$ has positive determinant. An equivalent (and for our purposes more useful) characterization of P-matrices [67, Thm. 3.3.4] is that

$$z_i (Mz)_i \leq 0 \qquad \text{for all } i \quad \text{implies} \quad z = 0. \tag{2.31}$$

For every P-matrix we can easily show that solutions must be unique: If $z^{(1)}$ and $z^{(2)}$ are two solutions to $\mathrm{CP}(M(\cdot) + q, \mathbb{R}^n_+)$, then

$$\begin{aligned} 0 \leq z_i^{(1)} \perp \left(Mz^{(1)} + q\right)_i \geq 0 \qquad &\text{for all } i, \\ 0 \leq z_i^{(2)} \perp \left(Mz^{(2)} + q\right)_i \geq 0 \qquad &\text{for all } i. \end{aligned}$$

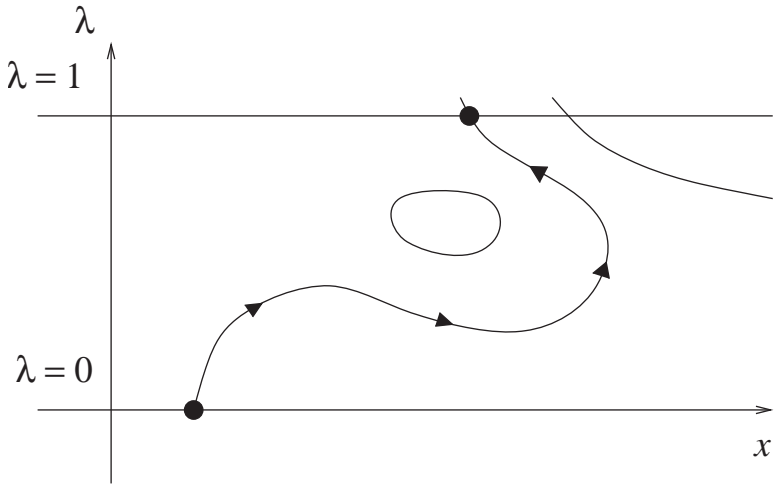


Figure 2.3: Illustration of homotopy methods for solving nonlinear equations.

Subtracting and cross-multiplying give

$$\begin{aligned}
&\left(z_i^{(1)} - z_i^{(2)}\right)\left(M\left(z^{(1)} - z^{(2)}\right)\right)_i \\
&\quad = z_i^{(1)}\left(Mz^{(1)} + q\right)_i - z_i^{(2)}\left(Mz^{(1)} + q\right)_i \\
&\qquad - z_i^{(1)}\left(Mz^{(2)} + q\right)_i + z_i^{(2)}\left(Mz^{(2)} + q\right)_i \\
&\quad = -z_i^{(2)}\left(Mz^{(1)} + q\right)_i - z_i^{(1)}\left(Mz^{(2)} + q\right)_i \;\leq\; 0 \qquad \text{for all } i.
\end{aligned}$$

Then it is clear that if M is a P-matrix, then $z^{(1)} - z^{(2)} = 0$ and solutions are unique. Further, the solution can be computed via Lemke's algorithm.

2.2.2 Lemke's method and homotopy methods

Despite its appearance, Lemke's method has behind it an important topological idea, which relates it closely to *homotopy* or *continuation methods* for solving nonlinear systems of equations [5, 6, 108, 109]. The basic idea of homotopy methods is that to solve a difficult system of equations $f(x) = 0$ with $f\colon \mathbb{R}^n \to \mathbb{R}^n$, we find an easy-to-solve system $g(x)$ with $g\colon \mathbb{R}^n \to \mathbb{R}^n$ and a sufficiently "nice" homotopy $h\colon [0,1] \times \mathbb{R}^n \to \mathbb{R}^n$, where

$$\begin{aligned}
h(0, x) &= g(x), \\
h(1, x) &= f(x).
\end{aligned}$$

We then follow the solution of $h(\lambda, x) = 0$ from $\lambda = 0$ to $\lambda = 1$, as illustrated in Figure 2.3. When we reach $\lambda = 1$, we have a solution of the problem.

To make things more concrete, we can assume that f is smooth and take, for example, $g(x) = x - a$ with a piecewise linear homotopy

$$h(\lambda, x) = \lambda\, f(x) + (1 - \lambda)(x - a).$$

Assume that the $(n+1)\times n$ Jacobian matrix $\nabla_{(\lambda,x)}h(\lambda,x)$ has full rank whenever $h(\lambda,x)=0$. Then the set $\{(\lambda,x)\mid h(\lambda,x)=0\}$ is a union of smooth curves in $\mathbb{R}^{n+1}$. The idea then is to follow one of these smooth curves from $\lambda=0$ until $\lambda=1$. The older continuation algorithms had the simpler strategy of increasing λ by a small amount $\Delta\lambda$ and then solving $h(\lambda+\Delta\lambda,x)=0$ for x, using Newton's method, for example. If Newton's method failed, then $\Delta\lambda$ is reduced, and the process is repeated; otherwise, update $\lambda\leftarrow\lambda+\Delta\lambda$ and continue following the curve.

The trouble with continuation methods is that sometimes the curve "doubles back" and to follow the curve, λ must be reduced rather than increased. So the more modern homotopy algorithms were developed [231, 268] which treated the curves in terms of arc-length continuation: $(\lambda(s),x(s))$ and $\|(d\lambda/ds(s),dx/ds(s))\|=1$. Care must be taken to prevent the algorithm from reversing direction along the curve that it is tracking and to avoid jumping from one curve to another. In general, curves given by equations $h(x,\lambda)=0$ are not necessarily smooth and can have bifurcations. The way to avoid this is to note that if we have an extra parameter such as a in $h(x,\lambda;a)=\lambda f(x)+(1-\lambda)(x-a)$, provided the Jacobian matrix $\nabla_a h(x,\lambda;a)$ is nonsingular, then for *almost all* a, the curves $h(x,\lambda;a)=0$ are smooth and $\nabla_{(x,\lambda)}h(x,\lambda;a)$ has full rank on these curves. This can be proved from a generalization of the *Morse–Sard theorem* [7, 272].

These methods can be very effective for highly nonlinear systems of equations and even for LCPs [269]. However, Lemke's method does not involve smooth functions.

Lemke's method is a piecewise affine version of homotopy path following. In Lemke's method, s takes the role of λ, although s does not go from zero to one. Instead, s goes from a large value (such that $sd+q\geq 0$) down to zero to obtain a solution to LCP(q,M). Indeed, s is not guaranteed to be reduced at each step of Lemke's method, but it may increase at times before eventually being brought to zero when Lemke's algorithm succeeds. The homotopy can be considered as changing LCP(q,M) to LCP$(sd+q,M)$, which is easy for large $s>0$ because $sd+q\geq 0$ implies that $z=0$ is a solution. The reduction of Lemke's method to a homotopy method for a piecewise linear functions can be carried out by using an equivalent nonlinear system of equations, such as

$$z \text{ solves LCP}(q,M) \Longleftrightarrow \min(z, Mz+q)=0,$$

since for any vectors $a,b\in\mathbb{R}^n$,

$$0\leq a\perp b\geq 0 \Longleftrightarrow \min(a,b)=0.$$

The minimum is understood to be a componentwise minimum:

$$\min(a,b)_i=\min(a_i,b_i).$$

The connection with homotopy methods has also been used in other contexts, such as to prove the existence of solutions to linear and nonlinear complementarity problems and GCPs [137, 138, 140, 197].

2.2.3 Polyhedral cones

Polyhedral cones are cones generated by a finite number of vectors. These are important for many applications, and the standard cone $\mathbb{R}^n_+$ is an example of a polyhedral cone.

The general form for a polyhedral cone is

$$\begin{aligned} K &= \operatorname{cone}\{v_1, v_2, \ldots, v_m\} \\ &= \left\{ \sum_{i=1}^{m} \alpha_i v_i \mid \alpha_i \geq 0 \text{ for all } i \right\}. \end{aligned} \tag{2.32}$$

For $K \subseteq \mathbb{R}^n$, if we let V be the $n \times m$ matrix $[v_1, v_2, \ldots, v_m]$, then $K = V(\mathbb{R}^n_+)$. We would like to use Lemma B.8(2) to find the dual cone, but V need not be invertible, especially if $m > n$. A general formula for the dual cone can be found using the ideas of Lemma B.8(2).

Lemma 2.10. *If L is a closed convex cone in X and $V : X \to Y$ is a linear operator, then the dual cone $V(L)^* = \{x \mid V^* x \in L^*\}$.*

Proof. This is a straightforward calculation:

$$\begin{aligned} V(L)^* &= \left\{ z \in Y' \mid \langle z, Vw \rangle \geq 0 \text{ for all } w \in L \right\} \\ &= \left\{ z \in Y' \mid \left\langle V^* z, w \right\rangle \geq 0 \text{ for all } w \in L \right\} \\ &= \left\{ z \in Y' \mid V^* z \in L^* \right\}, \end{aligned}$$

as desired. □

In the particular case where $L = \mathbb{R}^n_+$, we can use the Moore–Penrose pseudoinverse V^+ of V to get

$$\left(V\,\mathbb{R}^n_+\right)^* = \left(V^T\right)^+ \mathbb{R}^n_+ + \operatorname{null}\left(V^T\right).$$

If $m \leq n$ and V has full rank (that is, $\operatorname{rank}(V) = \min(m,n)$), then $\operatorname{null}\left(V^T\right) = \{0\}$, so

$$\left(V\,\mathbb{R}^n_+\right)^* = \left(V^T\right)^+ \mathbb{R}^n_+.$$

LCP$(q, M, V\,\mathbb{R}^n_+)$ becomes the following: Given q, M, and V, find z such that

$$V\,\mathbb{R}^n_+ \ni z \perp Mz + q \in \left(V^T\right)^+ \mathbb{R}^n_+.$$

Writing $z = Vx$, $x \in \mathbb{R}^n_+$, we have $Mz + q = MVx + q = \left(V^T\right)^+ w$, where $w \geq 0$. If V is square and nonsingular, then $w = V^T(MVx + q)$ with $x, w \in \mathbb{R}^n_+$. We can then represent the CP over $V\,\mathbb{R}^n_+$ as a standard CP over $\mathbb{R}^n_+$.

2.2.4 Special structure

Copositivity can be generalized to general closed convex cones:

$$M \text{ is } K\text{-copositive if } \langle z, Mz \rangle \geq 0 \text{ for all } z \in K. \tag{2.33}$$

Existence for GCPs CP$(M(\cdot)+q, K)$ can be shown if M is K-copositive and $\left(M + M^T\right) z = 0$ implies $\langle z, q \rangle > 0$.

A *strongly K-copositive* matrix M is one where there is an $\eta > 0$ such that for all $z \in K$ we have $\langle z, Mz \rangle \geq \eta \|z\|^2$. This echoes the definition of strongly monotone, but it is restricted to the cone used for complementarity. Strong copositivity can be used to obtain bounds on the solution of an LCP $K \ni z \perp Mz + q \in K^*$: $0 = \langle z, Mz + q \rangle \geq \eta \|z\|^2 - \|q\| \|z\|$ so $\|z\| \leq \|q\| / \eta$. However, strong copositivity does not guarantee uniqueness.

Uniqueness for $K = \mathbb{R}^n_+$ occurs for all $q \in \mathbb{R}^n$ if and only if M is a P-matrix (see, for example, [67, Thm. 3.3.7]). Related properties can be developed for other structures, such as where $K = K_1 \times K_2$ is a Cartesian product. Suppose that if we break up M and q in a consistent way, they have the form

$$M = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix}, \qquad q = \begin{bmatrix} q_1 \\ q_2 \end{bmatrix}.$$

Provided LCP(q_2, M_{22}, K_2) has a unique solution z_2, then we obtain the subproblem LCP($q_1 - M_{12} z_2$, M_{11}, K_1). If this in turn also has a unique solution, then we have found the unique solution of LCP(q, M, $K_1 \times K_2$). Conversely, if LCP(q, M, $K_1 \times K_2$) has a unique solution, we see that these subproblems must also have unique solutions.

Thus, e.g., if M_{11} and M_{22} are positive definite (so that $\langle z_1, M_{11} z_1 \rangle$, $\langle z_2, M_{22} z_2 \rangle > 0$ for all nonzero z_1, z_2), then solutions of LCP(q, M, K) exist and are unique, even though M itself might not be positive definite.

A generalization of the P-matrix property can be applied to a general Cartesian product of cones $K = K_1 \times K_2 \times \cdots \times K_m = \prod_{i=1}^m K_i$. If we partition M into blocks M_{ij} consistent with this Cartesian product, we say that M is a P(K)-matrix if

$$0 \geq \langle z_i, (Mz)_i \rangle = \sum_{j=1}^m \langle z_i, M_{ij} z_j \rangle \quad \text{implies}$$
$$z = 0.$$

Other examples of special cones that have received particular attention include the *Lorentz cone* (also called the *ice cream cone*) in $\mathbb{R}^n$ with $n \geq 2$:

$$L_n := \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \mid x \in \mathbb{R}, y \in \mathbb{R}^{n-1}, x \geq \|y\|_2 \right\}. \tag{2.34}$$

This is a self-dual cone. To see this, suppose $\left[u, v^T\right]^T \in L_n^*$. Then

$$\left[u, v^T\right] \left[x, y^T\right]^T = ux + v^T y \geq 0$$

for all $\left[x, y^T\right]^T \in L_n$. Taking the minimum of $ux + v^T y$ over all y satisfying $\|y\|_2 \leq x$, we get $ux - \|v\|_2 x = x(u - \|v\|_2) \geq 0$. Since $x \geq 0$ we have $u \geq \|v\|_2$, and so $\left[u, v^T\right]^T \in L_n$. Thus $L_n^* \subseteq L_n$. Conversely, it is easy to show that if $\left[u, v^T\right]^T \in L_n$, then $\left[u, v^T\right]^T \in L_n^*$: for $\left[x, y^T\right]^T \in L_n$,

$$\begin{aligned} \left[u, v^T\right] \left[x, y^T\right]^T &\geq ux - \|v\| \|y\| \\ &\geq ux - ux \geq 0. \end{aligned}$$

Another cone that has been considered is the cone of semidefinite symmetric $n \times n$ matrices:

$$S_n := \left\{ A \in \mathbb{R}^{n \times n} \mid A^T = A, A \text{ positive semidefinite} \right\}.$$

This is also a self-dual cone under the inner product $\langle A, B \rangle = \text{trace}(A^T B) = \sum_{i,j=1}^{n} A_{ij} B_{ij}$. This inner product is called the *Frobenius inner product* and is often denoted by $A \bullet B$ in the optimization literature, or $A : B$ in the continuum mechanics literature. The standard nonnegative cone $\mathbb{R}^n_+$, the Lorenz cone L_n, and the cone of semidefinite matrices S_n are all examples of symmetric cones. *Symmetric cones* are self-dual cones K that are homogeneous; that is, for every pair x, y in the interior of K there is a matrix A such that $AK = K$ and $Ax = y$. All such symmetric cones are generated by *Euclidean Jordan algebras* [97, 114]. Euclidean Jordan algebras are finite-dimensional vector spaces V with a bilinear product $\circ\colon V \times V \to V$ and an inner product $(\cdot, \cdot)_V$ on V with the following properties:

$$x \circ y = y \circ x, \tag{2.35}$$

$$x \circ \left(x^2 \circ y \right) = x^2 \circ (x \circ y), \qquad \text{where } x^2 = x \circ x, \tag{2.36}$$

$$(x \circ y, z)_V = (y, x \circ z)_V \tag{2.37}$$

for all x, y, and $z \in V$. The cone generated by the algebra is the cone of squares: $K = \left\{ x^2 \mid x \in V \right\}$. For all such cones K generated by a Euclidean Jordan algebra, not only is K convex, but K is self-dual (that is, $K = K^*$) in the inner product $(\cdot, \cdot)_V$ for V [97].

For example, if $K = \mathbb{R}^n_+$, we can take $x \circ y$ to be the componentwise or Hadamard product $(x \circ y)_i = x_i y_i$. For the Lorentz cone L_n we use

$$\begin{bmatrix} x_0 \\ x \end{bmatrix} \circ \begin{bmatrix} y_0 \\ y \end{bmatrix} = \begin{bmatrix} x_0 y_0 + x^T y \\ x_0 y + y_0 x \end{bmatrix}. \tag{2.38}$$

For the cone of symmetric semidefinite $n \times n$ matrices we use $A \circ B = \frac{1}{2}(AB + BA)$. In all finite-dimensional Euclidean Jordan algebras there is an element e where $e \circ x = x$ for all x in the algebra. For the Hadamard product e is the vector of ones of the appropriate dimension. For the Lorentz product (2.38), e is $[1, 0]^T$. For $A \circ B = \frac{1}{2}(AB + BA)$, e is just the identity matrix I.

The effect of this structure on the solution of CPs is discussed in [114, 115], for example.

There are two properties that are important for future developments. The first is that a matrix $M \in \mathbb{R}^{n \times n}$ is a GUS(K) matrix, where K is a closed convex cone, if for all $q \in \mathbb{R}^n$ the CP

$$K \ni z \perp Mz + q \in K^*$$

has a unique solution. A matrix M is an LS(K) matrix if the solution map $q \mapsto z$ for LCP(q, M, K),

$$K \ni z \perp Mz + q \in K^*,$$

is well defined and single valued for all $q \in K$, and is a Lipschitz map. If $M \in \text{LS}(K)$, then clearly $M \in \text{GUS}(K)$, since the solution operator is already single valued and everywhere defined. For polyhedral cones, by a result of Gowda [113], if $M \in \text{GUS}(K)$, then the solution operator is Lipschitz as well, so $M \in \text{LS}(K)$; thus $\text{GUS}(K) = \text{LS}(K)$ for polyhedral cones, but this is not necessarily true for general K.

2.2.5 Complementarity in infinite dimensions

CPs in infinite dimensions arise in many situations, such as in connection with partial differential equations. The framework for CPs starts with a Banach space X and its dual space X'. Then for K a closed convex cone we have the dual cone given in terms of the duality pairing between X and X'. We need the function $F\colon X \to X'$ (rather than $F\colon X \to X$) to be continuous. Actually requiring $F\colon X \to X'$ is an advantage, since for second order elliptic partial differential equations we can take $X = H^1(\Omega)$ for Ω to be a bounded open set in $\mathbb{R}^d$, and $X' = H^{-1}(\Omega)$. Then we can take $F(u) = -\nabla^2 u + a\,u$, where ∇^2 is the usual Laplacian operator ($\partial^2/\partial x^2 + \partial^2/\partial y^2$ on $\mathbb{R}^2$).

If we consider the problem of preventing penetration into an obstacle given by $u(\mathbf{x}) \geq \varphi(\mathbf{x})$ for all $\mathbf{x} \in \Omega$, then we have the following *obstacle problem*: find $N(\mathbf{x})$ and $u(\mathbf{x})$ such that

$$\begin{aligned} 0 \leq N(\mathbf{x}) \perp u(\mathbf{x}) - \varphi(\mathbf{x}) \geq 0 \qquad & \text{with} \\ -\nabla^2 u + a\,u = N(\mathbf{x}) + f(\mathbf{x}) \qquad & \text{in } \Omega \text{ and} \\ u(\mathbf{x}) = 0 \qquad & \text{on } \partial\Omega. \end{aligned}$$

Here we take $X = H_0^1(\Omega)$, which incorporates the boundary conditions $u(\mathbf{x}) = 0$ for $\mathbf{x} \in \partial\Omega$, and so $X' = H_0^1(\Omega)'$ is the dual space. Since the operator $-\nabla^2$ is an elliptic operator $H_0^1(\Omega) \to H_0^1(\Omega)'$, we can show that there exists a unique solution to this CP. The techniques to prove this are outlined in the next section. A more detailed example will be given in Section 2.6.

2.3 Variational inequalities

Around the same time as CPs were being created and analyzed in finite-dimensional situations, variational inequalities (VIs) were being applied to infinite-dimensional situations. The first application was to the frictionless contact of an elastic body with a rigid obstacle. This problem was first posed by Signorini [225] in 1933 and first resolved in a theoretical sense by Fichera [98] in 1963. The general idea and applications of VIs was developed further by Lions and Stampacchia [160]. More information about VIs from the point of view of partial differential equations can be found in [23, 84]. VIs can also be used for finite-dimensional problems; see [95, 96].

The precise formulation of VIs requires a closed convex set K (but not necessarily a cone) in a Banach space X (which can be $\mathbb{R}^n$ or a suitable Hilbert space) and a continuous function $F\colon K \to X'$. The problem is then to find a z such that

$$z \in K \quad \text{and} \qquad 0 \leq \langle \widetilde{z} - z, F(z) \rangle \quad \text{for all } \widetilde{z} \in K. \tag{2.39}$$

We denote this problem by VI(F, K). If K is a cone as well, then the VI is, in fact, a CP.

Lemma 2.11. *If K is a closed convex cone, then*

$$z \in K \quad \textit{and} \qquad 0 \leq \langle \widetilde{z} - z, F(z) \rangle \quad \textit{for all } \widetilde{z} \in K$$

if and only if

$$K \ni z \quad \perp \quad F(z) \in K^*.$$

That is, if K is a closed convex cone, then VI(F, K) is equivalent to CP(F, K).

Proof. Suppose z solves VI(F, K). We show first that $F(z) \in K^*$. Let $w \in K$. We want to show that $\langle w, F(z)\rangle \geq 0$. This is obviously true if $w = 0$, so suppose that $w \neq 0$. Then, for $\alpha > 0$, $\alpha w \in K$ since K is a cone, and thus setting $\widetilde{z} = \alpha w$, $0 \leq \langle \widetilde{z} - z, F(z)\rangle = \alpha \langle w, F(z)\rangle - \langle z, F(z)\rangle$. Dividing by α and taking $\alpha \to \infty$ give $\langle w, F(z)\rangle \geq 0$. Thus $F(z) \in K^*$.

Now $z \perp F(z)$: we can take $\widetilde{z} = 0 \in K$ to get $\langle -z, F(z)\rangle \geq 0$; on the other hand, since $z \in K$ and $F(z) \in K^*$, $\langle z, F(z)\rangle \geq 0$. The only way both these inequalities can be true is if $\langle z, F(z)\rangle = 0$, as desired. Thus z solves CP(F, K).

Now suppose that z satisfies CP(F, K). Then $z \in K$, and if $\widetilde{z} \in K$ as well,

$$\langle \widetilde{z} - z, F(z)\rangle = \langle \widetilde{z}, F(z)\rangle - \langle z, F(z)\rangle \geq 0$$

since $\langle z, F(z)\rangle = 0$ and $F(z) \in K^*$. Thus z solves VI(F, K). □

VIs are most useful in dealing with certain optimization problems over closed convex sets. Consider the problem of minimizing $f(x)$ over $x \in K$. If f is continuously (or Frèchet) differentiable, then

$$f(y) = f(x) + \langle y - x, \nabla f(x)\rangle + o(\|y - x\|) \qquad \text{as } y \to x.$$

If x^* minimizes f over K, then for any $y \in K$, $\theta y + (1-\theta)x = x + \theta(y - x) \in K$, so

$$f(x) \leq f(x + \theta(y - x)) = f(x) + \theta\langle y - x, \nabla f(x)\rangle + o(\theta\|y - x\|).$$

Subtracting $f(x)$, dividing by $\theta > 0$, and taking $\theta \downarrow 0$, we get

$$x \in K \quad \text{and} \qquad 0 \leq \langle y - x, \nabla f(x)\rangle \quad \text{for all } y \in K, \tag{2.40}$$

which is a VI. Conversely, if x satisfies (2.40) and f is convex, then as $f(y) \geq f(x) + \langle y - x, \nabla f(x)\rangle$, for $y \in K$ we have $f(y) \geq f(x) + \langle y - x, \nabla f(x)\rangle \geq f(x)$. Thus x is the global minimizer of f over K.

There is a more geometric viewpoint to describe the solution to a VI. Consider again (2.39):

$$z \in K \quad \text{and} \qquad 0 \leq \langle \widetilde{z} - z, F(z)\rangle \quad \text{for all } \widetilde{z} \in K.$$

Since $T_K(z) = \overline{\text{cone}(K - z)}$ for $z \in K$, $\langle w, F(z)\rangle \geq 0$ for all $w \in T_K(z)$. Thus

$$F(z) \in -N_K(z). \tag{2.41}$$

That is, $-F(z)$ points in the direction of the normal cone at z. Conversely, if $F(z) \in -N_K(z)$ for closed convex K, then $z \in K$ and $\langle \widetilde{z} - z, F(z)\rangle \geq 0$ for all $\widetilde{z} \in K$, so that z solves the VI.

2.3.1 VIs of the second kind

VIs of the second kind have the following form: Given $F\colon X \to X'$, K a closed convex subset of X, and a convex and lower semicontinuous[2] function $j\colon X \to \mathbb{R}$, find z satisfying

$$z \in K \qquad \text{and} \qquad \langle \widetilde{z} - z, F(z)\rangle \geq j(z) - j(\widetilde{z}) \quad \text{for all } \widetilde{z} \in K. \tag{2.42}$$

[2]That is, if $x_n \to x$, then $\liminf_{n\to\infty} f(x_n) \geq f(x)$. This is a weaker condition than requiring f to be continuous. See also Appendix A.

This problem is denoted by $\mathrm{VI}_2(F, K, j)$. This formulation can be represented as a standard VI: Let

$$L = \left\{ \begin{bmatrix} z \\ s \end{bmatrix} \mid z \in K,\ s \geq j(z) \right\}, \tag{2.43}$$

$$G(z,s) = \begin{bmatrix} F(z) \\ 1 \end{bmatrix}. \tag{2.44}$$

Lemma 2.12. *$\mathrm{VI}_2(F, K, j)$ is equivalent to $\mathrm{VI}(G, L)$, provided j is convex and lower semicontinuous and L and G are given by* (2.43)–(2.44)*.*

Proof. Suppose that z is a solution of $\mathrm{VI}_2(F, K, j)$ and that $j\colon X \to \mathbb{R} \cup \{\infty\}$ is a convex lower semicontinuous function, K is closed and convex in X, and $F\colon X \to X'$. Since j is convex, L is a convex set; since j is lower semicontinuous and K is closed, L is also closed. Note that $j(z) < \infty$. We show that $x := \left[\begin{smallmatrix} z \\ j(z) \end{smallmatrix}\right]$ solves $\mathrm{VI}(G, L)$. Since z solves $\mathrm{VI}_2(F, K, j)$, $z \in K$, then $\left[\begin{smallmatrix} z \\ j(z) \end{smallmatrix}\right] \in L$. Now suppose that $\widetilde{x} := \left[\begin{smallmatrix} \widetilde{z} \\ \widetilde{s} \end{smallmatrix}\right] \in L$. Then $\widetilde{z} \in K$ and $\widetilde{s} \geq j(\widetilde{z})$. So

$$\begin{aligned} \left\langle \begin{bmatrix} \widetilde{z} \\ \widetilde{s} \end{bmatrix} - \begin{bmatrix} z \\ j(z) \end{bmatrix}, \begin{bmatrix} F(z) \\ 1 \end{bmatrix} \right\rangle &= \langle \widetilde{z} - z, F(z) \rangle + (\widetilde{s} - j(z)) \\ &\geq \langle \widetilde{z} - z, F(z) \rangle + j(\widetilde{z}) - j(z) \geq 0. \end{aligned}$$

Thus x solves $\mathrm{VI}(G, L)$.

Conversely, suppose $x = \left[\begin{smallmatrix} z \\ s \end{smallmatrix}\right]$ solved $\mathrm{VI}(G, L)$. Then, for any $\widetilde{z} \in K$ and $\widetilde{s} \geq j(\widetilde{z})$,

$$\begin{aligned} 0 &\leq \left\langle \begin{bmatrix} \widetilde{z} \\ \widetilde{s} \end{bmatrix} - \begin{bmatrix} z \\ s \end{bmatrix}, \begin{bmatrix} F(z) \\ 1 \end{bmatrix} \right\rangle \\ &= \langle \widetilde{z} - z, F(z) \rangle + \widetilde{s} - s. \end{aligned}$$

Note that if we choose $\widetilde{z} = z$ and $\widetilde{s} = j(\widetilde{z})$, then we get

$$0 \leq j(z) - s \qquad \text{and} \qquad s \geq j(z),$$

so $s = j(z)$. Now, for any $\widetilde{z} \in K$, choosing $\widetilde{s} = j(\widetilde{z})$ we have

$$0 \leq \langle \widetilde{z} - z, F(z) \rangle + j(\widetilde{z}) - j(z),$$

and z solves $\mathrm{VI}_2(F, K, j)$, as desired. □

Thus the class of second kind VIs is actually no larger than the class of standard VIs.

2.3.2 Equivalent formulations

Existence of solutions can be shown by means of equivalent nonlinear (in fact, nonsmooth) systems of equations. There are two equivalent equations $\Phi(x) = 0$ of particular importance: the *normal map*

$$\Phi_{\mathrm{nor}}(x; K, F) = F(\Pi_K(x)) + J_X(x - \Pi_K(x))$$

and the *natural map*

$$\Phi_{\text{nat}}(x; K, F) = x - \Pi_K\left(x - J_X^{-1}(F(x))\right).$$

If K and F are both understood from context, we drop these parameters from Φ_{nor} and Φ_{nat}. For both the normal and natural maps, $\Pi_K(x)$ is the nearest point in K to x. The function Π_K is well defined and continuous, in fact Lipschitz continuous, where X is a Hilbert space. After we have established the basic properties of Π_K we will show that solving VI(F, K) is equivalent to finding zeros of the normal and natural maps. First we set out connections with more geometric properties. The proof uses the characterization of $\Pi_K(x)$ in (B.6).

Lemma 2.13. *If $K \subseteq X$, with X a Hilbert space, is closed and convex, then*

$$J_X(x - \Pi_K(x)) \in N_K(\Pi_K(x))$$

holds for all $x \in X$. Furthermore, for $z \in K$, then $J_X(y) \in N_K(z)$ if and only if $\Pi_K(z + y) = z$.

Proof. Let $z = \Pi_K(x)$. Then $z \in K$ and from (B.6), $(x - z, z - w)_X \leq 0$ for all $w \in K$. Thus $J_X(x - z) \in T_K(z)^\circ = N_K(z)$. That is, $J_X(x - \Pi_K(x)) \in N_K(\Pi_K(x))$, as desired.

For the second statement, suppose first that $z \in K$ and that $J_X(y) \in N_K(z)$. Then for any $w \in K$, $((z + y) - z, z - w)_X = (y, z - w)_X \leq 0$. Thus by (B.6), $z = \Pi_K(z + y)$. Conversely, if $z = \Pi_K(z + y)$, then $(y, z - w)_X = ((z + y) - z, z - w)_X \leq 0$ for all $w \in K$. By (B.14), this implies that $J_X(y) \in N_K(z)$. □

Now $z \in K$ is a solution of VI(F, K) if for all $w \in K$, $\langle w - z, F(z)\rangle \geq 0$, or equivalently that $\langle w - z, -F(z)\rangle \leq 0$ for all $w \in K$. Then by (B.14), this is equivalent to $-F(z) \in N_K(z)$. This gives a third equivalent condition for z solving VI(F, K):

$$0 \ni F(z) + N_K(z). \tag{2.45}$$

We can now return to the normal and natural maps.

Lemma 2.14. *The* VI (2.39) *for z holds if and only if $\Phi_{\text{nor}}(x) = 0$, where $z = \Pi_K(x)$, and if and only if $\Phi_{\text{nat}}(x) = 0$, where $z = x$.*

Proof. Suppose that z solves VI(F, K):

$$z \in K \quad \text{and} \quad \langle \widetilde{z} - z, F(z)\rangle \geq 0 \quad \text{for all } \widetilde{z} \in K.$$

We show that this is equivalent to $\Phi_{\text{nat}}(z) := z - \Pi_K(z - J_X^{-1}(F(z))) = 0$: Note that (2.39) is equivalent to $z \in K$ and $0 \geq \langle \widetilde{z} - z, -F(z)\rangle = (\widetilde{z} - z, (z - J_X^{-1}(F(z))) - z)_X$ for all $\widetilde{z} \in K$. By (B.6), this is equivalent to $z = \Pi_K(z - J_X^{-1}(F(z)))$ and thus $\Phi_{\text{nat}}(z) = 0$. Now suppose that $\Phi_{\text{nat}}(z) = 0$ so that $z - \Pi_K(z - J_X^{-1}(F(z))) = 0$; that is, $z = \Pi_K(z - J_X^{-1}(F(z)))$. Then $-F(z) \in N_K(z)$, and $0 \in F(z) + N_K(z)$, and so z solves VI(F, K).

For the normal map, we show that z being a solution of VI(F, K) (2.39) implies $\Phi_{\text{nor}}(x) = 0$, where $x = z - J_X^{-1}(F(z))$ or $J_X(x - z) = -F(z)$. From the previous

paragraph we see that (2.39) is equivalent to $z = \Pi_K(z - J_X^{-1}(F(z))) = \Pi_K(x)$. Thus (2.39) implies that $\Phi_{\text{nor}}(x) = F(\Pi_K(x)) + J_X(x - \Pi_K(x)) = F(z) + J_X(x - z) = F(z) - F(z) = 0$. Conversely, suppose that $\Phi_{\text{nor}}(x) = 0$. Then set $z = \Pi_K(x) \in K$. From (B.6), $(x - z, y - z)_X \leq 0$ for all $y \in K$. Also, substituting for $\Pi_K(x)$ in $\Phi_{\text{nor}}(x) = F(\Pi_K(x)) + J_X(x - \Pi_K(x)) = F(z) + J_X(x - z) = 0$ we see that $J_X(x - z) = -F(z)$, and so $\langle -F(z), y - z\rangle \leq 0$ for all $y \in K$. That is, $z \in K$ and $\langle y - z, F(z)\rangle \geq 0$ for all $y \in K$, and so z satisfies (2.39). □

The ability to formulate the VI as a nonlinear equation means that we can apply, for example, techniques from topology to prove existence of solutions.

2.3.3 Complementarity bounds

The simplest bounds we can obtain are for strongly monotone VIs, as we have seen. However, at least in finite dimensions we can obtain some bounds for VIs with linear functions in terms of related CPs using strong copositivity over the recession cone.

Lemma 2.15. *Let $K \subseteq \mathbb{R}^n$ be a closed, convex set with recession cone K_∞. If $M \in \mathbb{R}^{n\times n}$ is K_∞-strongly copositive, then there are a constant C and a neighborhood $\mathcal{V}$ of M such that for any solution of a linear VI with $\widetilde{M} \in \mathcal{V}$,*

$$z \in K \quad \& \quad 0 \leq \left\langle \widetilde{z} - z, \widetilde{M}z + q \right\rangle \qquad \text{for all } \widetilde{z} \in K,$$

we have the bound

$$\|z\| \leq C\,(1 + \|q\|)\,. \tag{2.46}$$

Note that C depends on K and $\mathcal{V}$ but not on q.

Proof. Pick a fixed $z^* \in K$. Let $\eta > 0$ be the constant for strong K_∞-copositivity of M, so that $\langle w, Mw\rangle \geq \eta \|w\|^2$ for all $w \in K_\infty$. Suppose that there are no such C and $\mathcal{V}$. Then there must be a sequence of $q^\ell \in \mathbb{R}^n$, $M^\ell \in \mathbb{R}^{n\times n}$, and $z^\ell \in K$ such that $\left\langle z^* - z^\ell, M^\ell z^\ell + q^\ell \right\rangle \geq 0$ and $M^\ell \to M$, $\left\|z^\ell\right\| / \left(1 + \left\|q^\ell\right\|\right) \to \infty$ as $\ell \to \infty$. This implies that $\left\|z^\ell\right\| \to \infty$ and $\left\|q^\ell\right\| / \left\|z^\ell\right\| \to 0$ as $\ell \to \infty$. Now $\widehat{z}^\ell := z^\ell / \left\|z^\ell\right\|$ is bounded in $\mathbb{R}^n$, and so it has a convergent subsequence; denote such a subsequence by $\widehat{z}^\ell$ and let $\widehat{z}$ be its limit. Taking limits in the subsequence of

$$0 \leq \left\langle \frac{z^*}{\left\|z^\ell\right\|} - \frac{z^\ell}{\left\|z^\ell\right\|}, M^\ell \frac{z^\ell}{\left\|z^\ell\right\|} + \frac{q^\ell}{\left\|z^\ell\right\|} \right\rangle$$

gives the inequality

$$0 \leq -\langle \widehat{z}, M\widehat{z}\rangle\,.$$

But $z^\ell \in K$ for all ℓ, so $\widehat{z} = \lim_{\ell\to\infty} z^\ell / \left\|z^\ell\right\| \in K_\infty$ (taking limits in the subsequence); also $\|\widehat{z}\| = 1$. Therefore, $0 \leq -\eta \|\widehat{z}\|^2$, implying that $\|\widehat{z}\| = 0$, which is impossible.

Thus there is such a C independent of q for which the above bound holds. □

The following bound will be useful later on for proving existence of solutions for index-one differential VIs.

Lemma 2.16. *If $K = C + L$, where C and K are closed and convex, with C bounded and L a cone, and M strongly L-copositive with constant $\eta > 0$, then there is a constant γ (depending only on C, L, and M) where for any solution z of the VI*

$$z \in K \quad \& \quad 0 \le \langle \widetilde{z} - z, Mz + q \rangle \qquad \text{for all } \widetilde{z} \in K$$

we have

$$\|z\| \le \gamma \left(1 + \|q - \Pi_{L^*}(q)\|\right). \tag{2.47}$$

Furthermore, γ depends continuously on $\eta > 0$, $\|M\|$, and $\max_{u \in C} \|u\|$.

Proof. For any solution $z \in K = C + L$ we have $z = u + v$ with $u \in C$ and $v \in L$. Set $\widetilde{z} = u \in C \subseteq C + L = K$. Then

$$\begin{aligned} 0 \le \langle \widetilde{z} - z, Mz + q \rangle &= -\langle v, Mu + Mv + q \rangle \\ &= -\langle v, Mv \rangle - \langle v, Mu \rangle - \langle v, q - \Pi_{L^*}(q) + \Pi_{L^*}(q) \rangle \\ &\le -\eta \|v\|^2 - \langle v, Mu \rangle - \langle v, \Pi_{L^*}(q) \rangle - \langle v, q - \Pi_{L^*}(q) \rangle . \end{aligned}$$

But $\langle v, \Pi_{L^*}(q) \rangle \ge 0$ as $v \in L$. Hence

$$0 \le -\eta \|v\|^2 - \|v\| \, \|M\| \, \|u\| - \|v\| \, \|q - \Pi_{L^*}(q)\| .$$

Rearranging and dividing by $\|v\|$ give

$$\begin{aligned} \|v\| &\le \left(\|M\| \max_{u \in C} \|u\| + \|q - \Pi_{L^*}(q)\| \right) / \eta \\ &\le \beta \left(1 + \|q - \Pi_{L^*}(q)\|\right) \end{aligned}$$

for $\beta = \max\left(\|M\| \max_{u \in C} \|u\|, 1\right)/\eta$. Then

$$\begin{aligned} \|z\| \le \|u\| + \|v\| &\le \max_{u \in C} \|u\| + \beta \left(1 + \|q - \Pi_{L^*}(q)\|\right) \\ &\le \gamma \left(1 + \|q - \Pi_{L^*}(q)\|\right) \end{aligned}$$

for $\gamma = \beta + \max_{u \in C} \|u\|$, as desired. This formula is clearly continuous in M and η; in turn, $\eta := \inf_{v \in L : \|v\| = 1} \langle v, Mv \rangle$ depends continuously on M, as we are taking the infimum of a continuous function over a compact set. □

Remark 2.17. It might be tempting to think that if $K \subset C + L$, where C and L are closed and convex with C bounded and L a cone, then we can still obtain the same bound (2.47). However, this is not so. An example follows. See also Figure 2.4.

We start with a function $\psi : [0, \infty) \to [0, \infty)$ that is smooth, except at $t = 0$, where $\psi'(0^+) = +\infty$, and that $\psi''(x) \le 0$ for all $x > 0$. Let

$$K = \{ (x, y) \mid |y| \le \psi(x) \} .$$

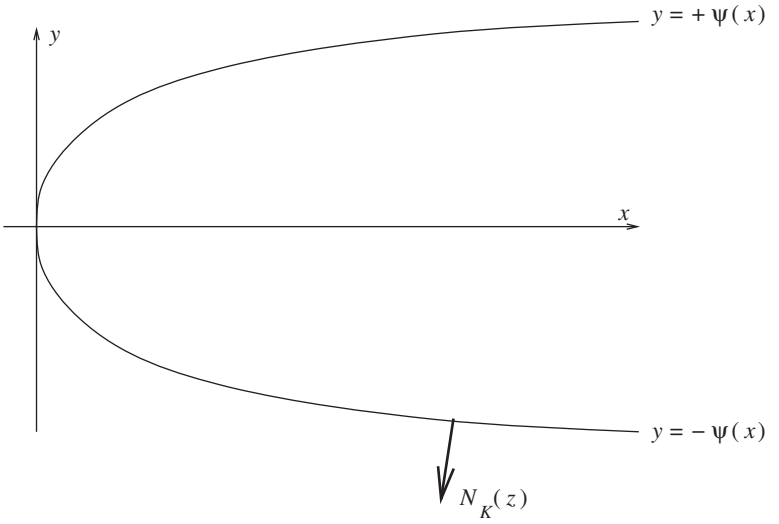


Figure 2.4: The set $K \subset C + L$ for a counterexample to (2.47).

Since $\psi''(x) \leq 0$ for all x, K is a convex set. Furthermore, $\psi'(x) \geq 0$ for all $x \geq 0$. If this were not so, suppose that $\psi'(x^*) < 0$. Then, as ψ' is a decreasing function, for $x \geq x^*$ we have $\psi(x) \leq \psi(x^*) + \psi'(x^*)(x - x^*)$, which must eventually become negative, contradicting $\psi(x) \geq 0$ for $x \geq 0$.

We will also assume that $\psi'(x) > 0$ for all x and that $\lim_{x\to\infty} \psi'(x) = 0$. The first of these conditions ensures that $\psi'(x)$ is not eventually zero; in this case, Lemma 2.16 would apply. The second condition ensures that the recession cone $L = K_\infty = \mathbb{R}_+ \times \{0\}$, the positive x-axis. Choosing ψ so that $\psi(x) \leq R$ for all $x > 0$ implies that $K \subset K_\infty + R\overline{B}$, where B is the unit ball in $\mathbb{R}^2$.

For our VI, we set

$$M = \begin{bmatrix} +1 & 0 \\ 0 & -1 \end{bmatrix},$$

which is strongly copositive on $L = \mathbb{R}_+ \times \{0\}$. We want to find a family of vectors $q \in L^*$ such that the solutions of the VIs

$$z \in K \quad \& \quad 0 \leq (\widetilde{z} - z)^T \left[Mz + q\right] \qquad \text{for all } \widetilde{z} \in K$$

are unbounded. This would contradict a bound of the form (2.47). To carry out this task, we use the formulation of the VI

$$-\left[Mz + q\right] \in N_K(z),$$

which can be interpreted in a more geometric way. Let $z = \left[x, y\right]^T$. If $|y| < \psi(x)$, then $N_K(z) = \{0\}$; if $y = \pm\psi(x)$ and $x > 0$, then $N_K(z) = \mathbb{R}_+ \left[-\psi'(x), \pm 1\right]^T$; if $x = y = 0$, then $N_K(z) = [-1, 0]^T$. There are three possibilities for a solution of this VI:

- $x = y = 0$:

$$\begin{bmatrix} -q_1 \\ -q_2 \end{bmatrix} = \alpha \begin{bmatrix} -1 \\ 0 \end{bmatrix} \qquad \text{for } \alpha \geq 0,$$

- $|y| < \psi(x)$:

$$\begin{bmatrix} -x - q_1 \\ +y - q_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \text{or}$$

- $y = \pm\psi(x)$ and $x > 0$:

$$\begin{bmatrix} -x - q_1 \\ +y - q_2 \end{bmatrix} = \alpha \begin{bmatrix} -\psi'(x) \\ \pm 1 \end{bmatrix} \qquad \text{for } \alpha \geq 0.$$

Now $q \in L^* = \mathbb{R}_+ \times \mathbb{R}$ means that $q_1 \geq 0$. We will focus on the cases $q_1 = 0$ and $q_2 \to +\infty$. The latter case for solutions is the one that we will consider. After all, we do not have to show that *all* solutions violate (2.47), only that there is *one* that does. Even more specifically, we will assume that $y = -\psi(x)$, so that $y - q_2 = -\alpha$; that is, $\alpha = q_2 - y = q_2 + \psi(x)$. Then

$$\begin{aligned} -x &= -\alpha\, \psi'(x) \\ &= -(q_2 + \psi(x))\, \psi'(x). \end{aligned}$$

That is, $x = (q_2 + \psi(x))\, \psi'(x)$. Since $\psi'(x) > 0$ for all x, if we take $q_2 \to +\infty$, $x \to +\infty$ as well. In particular, for any value of x sufficiently large we can take $q_2 = x/\psi'(x) - \psi(x) > 0$. Thus we have a family of solutions to the VI that is unbounded for $q \in L^*$, even though $K \subset C + L$ with C closed, convex, and bounded and L a closed convex cone with M strongly L-copositive. Thus (2.47) does not hold in this case.

2.3.4 Existence and uniqueness in finite dimensions

Existence results for solutions to VIs can be obtained by a number of means, although the most common is to use *coercivity* or *semicoercivity* of F. We say that $F\colon X \to X'$ is semicoercive on K if there are a $z_0 \in K$ and $R > 0$ such that $\langle x - z_0, F(x)\rangle > 0$ for all $x \in K$, where $\|x\| \geq R$.

To establish the existence of solutions for semicoercive F (Lemma 2.18) we use some topological arguments. In particular we use the tools of *degree theory* [106, 162]. The basic properties that we use are as follows: pick a number $R > 0$. We consider continuous functions $f\colon \mathbb{R}^n \to \mathbb{R}^n$ so that whenever $\|x\| \geq R$, then $f(x) \neq 0$. Every such function has a well-defined degree, which is an integer. If the degree of f is not zero, then $f(x^*) = 0$ for some x^* with $\|x^*\| < R$. If f is smooth and has finitely many zeros with the Jacobian matrix $\nabla f(x)$ nonsingular at each zero x, then the degree of f is

$$\deg f = \sum_{x:\, f(x)=0} \operatorname{sgn} \det(\nabla f(x));$$

$\operatorname{sgn}(s) = +1$ if $s > 0$, -1 if $s < 0$, and zero if $s = 0$. In particular, if f is the identity function, $\deg f = +1$.

The property of degrees that makes degree theory particularly useful is the *homotopy* property: if we have a continuous function $h\colon [0,1]\times\mathbb{R}^n \to \mathbb{R}^n$ such that $h(s,x)\neq 0$ whenever $\|x\|\geq R$, and $f(x)=h(0,x)$, $g(x)=h(1,x)$, then $\deg f = \deg g$. The function h is a *homotopy between f and g*; we say that f and g are *homotopic*. More information on using homotopies to solve systems of equations can be found in Section 2.2.2.

A strategy for showing that solutions exist to a problem is to reduce the problem to solving a certain system of equations $f(x)=0$. Then show that the function f is homotopic to a simpler one g where we can prove that $\deg g\neq 0$. This is a strategy that can be applied to prove the existence of solutions to VIs. An alternative approach is to use the fixed point theorem of Brouwer.

Lemma 2.18. *Assume $K\subseteq X=\mathbb{R}^n$ and K is closed and convex. Suppose that F is continuous and is semicoercive in the sense that there are a $z_0\in K$ and $R>0$ such that $\langle x-z_0, F(x)\rangle > 0$ for $\|x\|\geq R$ and $x\in K$. Then solutions exist for the VI*

$$z\in K \quad \textit{and} \qquad \langle \widetilde{z}-z, F(z)+q\rangle \geq 0 \quad \textit{for all } \widetilde{z}\in K.$$

Proof. Suppose G is semicoercive: $\langle x-z_0, G(x)\rangle > 0$ for $\|x\|\geq R$ and $x\in K$. Then it is impossible for VI(G,K) to have a solution z with $\|z\|\geq R$: $\langle z_0-z, G(z)\rangle < 0$ for $z\in K$ and $\|z\|\geq R$. We will assume that $R>\|z_0\|$.

Let $G(x):=J_X(x-z_0)$. Now we can construct a homotopy between the function $x\mapsto \Phi_{\text{nor}}(x;G,K)$ and $x\mapsto\Phi_{\text{nor}}(x;F,K)$ where $\langle x-z_0,F(x)\rangle>0$ for $\|x\|\geq R$ and $x\in K$. Let $H(s,x)=s\,F(x)+(1-s)\,J_X(x-z_0)$. Then, for $x\in K$, $\|x\|\geq R$,

$$\langle x-z_0, H(s,x)\rangle = s\,\langle x-z_0, F(x)\rangle + (1-s)\,(x-z_0, x-z_0)_X > 0$$

since both $\langle x-z_0,F(x)\rangle>0$ and $(x-z_0,x-z_0)_X=\|x-z_0\|_X^2>0$. Thus

$$x\mapsto \Phi_{\text{nor}}(x;H(s,\cdot),K) = H(s,\Pi_K(x)) + J_X(x-\Pi_K(x))$$

is a suitable homotopy between $\Phi_{\text{nor}}(\cdot;F,K)$ and $\Phi_{\text{nor}}(\cdot;G,K)$.

What remains now is to find the degree of $\Phi_{\text{nor}}(\cdot;G,K)$. Now

$$\begin{aligned}\Phi_{\text{nor}}(x;G,K) &= G(\Pi_K(x)) + J_X(x-\Pi_K(x))\\ &= J_X(\Pi_K(x)-z_0+x-\Pi_K(x)) = J_X(x-z_0).\end{aligned}$$

There is only one solution to $\Phi_{\text{nor}}(x)=0$: $x=z_0$. The Jacobian matrix $\nabla\Phi_{\text{nor}}(x;G,K)=I$ for all x, so $\deg\Phi_{\text{nor}}(\cdot;G,K)=\operatorname{sgn}\det\nabla\Phi_{\text{nor}}(z_0;G,K)=+1$. Thus $\deg\Phi_{\text{nor}}(\cdot;F,K)=+1$, and so there is at least one zero $\Phi_{\text{nor}}(z)=0$, and so VI(F,K) has a solution $z\in K$, $\|z\|\leq R$. □

For infinite-dimensional problems we also need some compactness property, weak continuity, or similar property of F in order to establish the existence of solutions. See, for example, Section 2.5 on pseudomonotonicity.

Uniqueness holds for solutions of VIs if F is *strictly monotone*: for any $z_1, z_2\in K$ with $z_1\neq z_2$,

$$\langle z_1-z_2, F(z_1)-F(z_2)\rangle > 0. \tag{2.48}$$

Lemma 2.19. *If $F\colon X\to X'$ is strictly monotone, then there is at most one solution of* VI(F,K) *for any closed, convex $K\subset X$.*

Proof. Suppose $z_1, z_2 \in K$ are two solutions of VI(F, K). Then, since $z_1, z_2 \in K$,

$$0 \le \langle z_2 - z_1, F(z_1)\rangle,$$
$$0 \le \langle z_1 - z_2, F(z_2)\rangle.$$

Adding these inequalities gives

$$\langle z_2 - z_1, F(z_1) - F(z_2)\rangle \ge 0,$$

which contradicts strict monotonicity of F unless $z_1 = z_2$. Thus solutions are unique. □

Existence and uniqueness follow from a single condition, at least for finite-dimensional problems: $F\colon X \to X'$ is *strongly monotone* if there is a constant $\eta > 0$ such that

$$\langle z_1 - z_2, F(z_1) - F(z_2)\rangle \ge \eta \|z_1 - z_2\|^2 \qquad \text{for all } z_1, z_2 \in X. \tag{2.49}$$

Strong monotonicity implies semicoercivity as described above.

Lemma 2.20. *If $F\colon X \to X'$ is strongly monotone with constant $\eta > 0$ and $K \subseteq X = \mathbb{R}^n$ is closed and convex, then solutions for* VI(F, K) *exist and are unique. Furthermore, if z_1 solves* VI($F + q_1$, K) *and z_2 solves* VI($F + q_2$, K), *then $\|z_1 - z_2\| \le \|q_1 - q_2\|/\eta$.*

Proof. Pick $z_0 \in K$. Then

$$\langle z - z_0, F(z) - F(z_0)\rangle \ge \eta \|z - z_0\|^2,$$

so $\langle z - z_0, F(z)\rangle \ge \eta \|z - z_0\|^2 + \langle z - z_0, F(z_0)\rangle \ge \|z - z_0\|(\eta\|z - z_0\| - \|F(z_0)\|)$, which is positive for $\|z\| > \|z_0\| + \|F(z_0)\|/\eta$. Then by Lemma 2.18 solutions of (2.39) exist. Solutions are unique since strong monotonicity implies strict monotonicity.

To show $F + q$ monotone, note that $(F+q)(z_1) - (F+q)(z_2) = F(z_1) - F(z_2)$ so that $F + q$ also satisfies (2.49) with the same $\eta > 0$. Now, for any $\widetilde{z} \in K$,

$$\langle \widetilde{z} - z_1, F(z_1) + q_1\rangle \ge 0,$$
$$\langle \widetilde{z} - z_2, F(z_2) + q_2\rangle \ge 0.$$

Putting $\widetilde{z} = z_2$ in the first inequality and $\widetilde{z} = z_1$ in the second and adding the inequalities give

$$\langle z_2 - z_1, F(z_1) - F(z_2)\rangle + \langle z_2 - z_1, q_1 - q_2\rangle \ge 0.$$

Thus

$$\|z_2 - z_1\|\,\|q_1 - q_2\| \ge \langle z_2 - z_1, F(z_2) - F(z_1)\rangle \ge \eta \|z_2 - z_1\|^2.$$

Dividing by $\|z_2 - z_1\|$ gives the desired bound on $\|z_2 - z_1\|$. □

The final conclusion of this lemma shows that if F is strongly monotone, then the solution operator is not only well defined but is also Lipschitz with Lipschitz constant $\le 1/\eta$. This carries over to the infinite-dimensional case, which we now consider.

2.3.5 Existence of solutions for infinite-dimensional problems

Proofs of existence of solutions for infinite-dimensional VIs are usually based on either monotone operator arguments or on compactness arguments.

We start with a lemma, which shows how we can use strong monotonicity in the context of infinite-dimensional problems to show how to solve "nearby" VIs.

Lemma 2.21. *Suppose $F\colon X \to X'$ is strongly monotone in the sense that*

$$\langle z_1 - z_2,\, F(z_1) - F(z_2)\rangle \geq \eta \,\|z_1 - z_2\|_X^2, \qquad \eta > 0, \tag{2.50}$$

and that solutions exist for VI$(F+q, K)$ *for all $q \in X'$. Suppose also that $G\colon X \to X'$ is Lipschitz with constant $L < \eta$. Then solutions exist for* VI$(F+G+q, K)$ *for all $q \in X'$.*

Proof. Let $\mathrm{sol}_{F,K}\colon X' \to X$ be the map such that $\mathrm{sol}_{F,K}(q) = z$, where z solves VI$(F+q, K)$. Since F is strongly monotone, $\mathrm{sol}_{F,K}$ is well defined and Lipschitz with constant $1/\eta$, as can be seen from the last conclusion of Lemma 2.20.

Now for $q \in X'$ consider the iteration where $z^{(k+1)}$ is the solution of VI$(F+G(z^{(k)})+q, K)$. That is, $z^{(k+1)} = \mathrm{sol}_{F,K}\left(G(z^{(k)})+q\right)$. The mapping $T\colon X \to X$ given by $T(z) = \mathrm{sol}_{F,K}\left(G(x)+q\right)$ is a contraction mapping since it is Lipschitz with constant $L/\eta < 1$. Thus there is a fixed point; call it z^*. Then z^* is the (unique) solution of VI$(F+G(z^*)+q, K)$; that is,

$$z^* \in K \quad \text{and} \qquad \left\langle \widetilde{z} - z^*,\, F(z^*)+G(z^*)+q\right\rangle \geq 0 \quad \text{for all } \widetilde{z} \in K.$$

That is, z^* solves VI$(F+G+q, K)$. □

This can be used to show that solutions exist for VI(F, K), where F is strongly monotone and Lipschitz, and $K \subseteq X$, with X a Hilbert space. Note that we do not identify X with X', even though X is a Hilbert space.

Theorem 2.22. *Suppose $F\colon X \to X'$ is strongly monotone in the sense of* (2.50) *and Lipschitz with X a Hilbert space. Then, for any closed convex set $K \subseteq X$,* VI$(F+q, K)$ *has a unique solution for any $q \in X'$.*

Proof. Let $J_X\colon X \to X'$ be the duality map $J(x) = (x, \cdot)_X$. Since it is a continuous linear map, it is also Lipschitz. Also, J_X is strongly monotone in the sense of (2.50) since for all $x \in X$

$$\langle J_X(x), x\rangle_{X'\times X} = (x, x)_X = \|x\|_X^2 \qquad (\eta = 1).$$

First we show that VI(J_X+q, K) has a solution for all $q \in X'$. That is, we want to find $z \in K$ such that $\langle \widetilde{z} - z,\, J_X(z)+q\rangle_{X\times X'} \geq 0$ for all $\widetilde{z} \in K$. From the definition of J_X, this amounts to requiring that $(\widetilde{z} - z,\, z + J_X^{-1}(q))_X \geq 0$ for all $\widetilde{z} \in K$. From (B.6), $z = \Pi_K(-J_X^{-1}(q))$, so that VI(J_X+q, K) has a solution for all $q \in X'$. Let L_J be the Lipschitz constant of J_X, and L_F be the Lipschitz constant of F.

We now create a homotopy between J_X and F:

$$H(s,x) = (1-s)\, J_X(x) + s\, F(x).$$

For all $0 \le s \le 1$, $H(s,\cdot)$ is Lipschitz continuous with the Lipschitz constant bounded by $\max(L_J, L_F)$. The Lipschitz constant of $H(s,\cdot) - H(s',\cdot)$ is bounded by $|s - s'| \max(L_J, L_F)$. Now $H(s,\cdot)$ is strongly monotone in the sense of (2.50) for all $0 \le s \le 1$ since

$$\begin{aligned}
&\langle H(s,x_1) - H(s,x_2), x_1 - x_2\rangle \\
&\quad = (1-s)\,\langle J_X(x_1) - J_X(x_2), x_1 - x_2\rangle + s\,\langle F(x_1) - F(x_2), x_1 - x_2\rangle \\
&\quad \ge (1-s)\,\|x_1 - x_2\|_X^2 + s\,\eta\,\|x_1 - x_2\|_X^2 \\
&\quad \ge \min(1,\eta)\,\|x_1 - x_2\|_X^2 .
\end{aligned}$$

We now show that $\mathrm{VI}(H(s,\cdot)+q, K)$ has a solution for all $q \in X'$ and all $0 \le s \le 1$. Let $\delta = \min(1,\eta)/(2(L_J + L_F))$; we choose this value so that if $0 \le s, s' \le 1$, and $|s - s'| \le \delta$, then the Lipschitz constant of $H(s,\cdot) - H(s',\cdot)$ is less than or equal to $\min(1,\eta)/2 < \min(1,\eta)$, and so by Lemma 2.20, if $\mathrm{VI}(H(s,\cdot)+q, K)$ has a solution for all $q \in X'$, then $\mathrm{VI}(H(s',\cdot)+q', K)$ has a solution for all $q' \in X'$. Now $\mathrm{VI}(H(0,\cdot)+q, K) = \mathrm{VI}(J_X + q, K)$ has a solution for all $q \in X'$ by the preceding paragraph. We can show by induction that $\mathrm{VI}(H(k\delta,\cdot)+q, K)$ has a solution for $k = 0, 1, 2, \ldots, \lfloor 1/\delta \rfloor$ and any $q \in X'$. For any $0 \le s \le 1$ there is a $k = 0, 1, 2, \ldots, \lfloor 1/\delta \rfloor$ such that $|s - k\delta| \le \delta/2$, so that $\mathrm{VI}(H(s,\cdot), K)$ has a solution, as desired.

In particular, we can conclude that for $s = 1$, $\mathrm{VI}(H(1,\cdot)+q, K) = \mathrm{VI}(F+q, K)$ has a solution for any $q \in X'$. That the solution is unique follows as $F + q$ is strongly monotone. □

Theorem 2.22 applies to strongly monotone functions $F\colon X \to X'$, and yet our finite-dimensional results in Lemma 2.18 require only coercivity. Here is one way in which we can extend Theorem 2.22 to cover situations of this kind.

Theorem 2.23. *Suppose that $F\colon X \to X'$ is strongly monotone in the sense of* (2.50)*, $G\colon X \to X'$ is a compact operator, and $\langle z, F(z) + G(z)\rangle > 0$ for all $z \in K$, where $\|z\| \ge R$. Then* $\mathrm{VI}(F+G, K)$ *has a solution.*

Proof. Let $\mathrm{sol}_{F,K}\colon X' \to X$ be the map $\mathrm{sol}_{F,K}(q) = z$, where z solves $\mathrm{VI}(F+q, K)$. Since F is strongly monotone with monotonicity constant $\eta > 0$, $\mathrm{sol}_{F,K}$ is Lipschitz with Lipschitz constant $1/\eta$. To solve $\mathrm{VI}(F+G, K)$, we need to find z^* such that z^* solves $\mathrm{VI}(F + G(z^*), K)$; in other words, we want to find z^* such that $z^* = \mathrm{sol}_{F,K}(G(z^*))$. From the inequality $\langle z, F(z) + G(z)\rangle > 0$ for all $z \in K$, where $\|z\| \ge R$, there can be no solution z with $\|z\| > R$. Thus we can restrict our attention to the ball $B = \{z \in X \mid \|z\|_X \le R\}$. Now $z \mapsto \mathrm{sol}_{F,K}(G(z))$ is a map $B \to B$.

Since G is a compact operator, it maps bounded sets to precompact sets, and $\mathrm{sol}_{F,K}$ maps precompact sets to precompact sets by continuity of $\mathrm{sol}_{F,K}$. Thus any z^* must lie in the closure $\overline{\mathrm{sol}_{F,K}(G(B))}$, which is a compact set. Furthermore, its closed convex hull $\overline{\mathrm{co}}\,\overline{\mathrm{sol}_{F,K}(G(B))}$ is also compact (by Mazur's lemma), and we can consider the restriction of $z \mapsto \mathrm{sol}_{F,K}(G(z))$ to this compact, convex set. We can then apply the Leray–Schauder fixed point theorem (Proposition A.13) to show that there must exist a $z^* \in \overline{\mathrm{co}}\,\overline{\mathrm{sol}_{F,K}(G(B))}$ such that $z^* = \mathrm{sol}_{F,K}(G(z^*))$, and therefore that z^* solves $\mathrm{VI}(F+G, K)$, as desired. □

For more approaches to VIs in infinite-dimensional spaces, see the section on pseudomonotone operators (Section 2.5).

2.3.6 Convex functions and subdifferentials

Maximal monotone operators can be obtained from convex functions: a function $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is *convex* if for all $x, y \in X$ and $0 \le \theta \le 1$, we have $\phi(\theta x + (1-\theta)y) \le \theta\,\phi(x) + (1-\theta)\,\phi(y)$. Note that we allow ∞ as a value: $0 \times \infty = 0$; $\alpha + \infty = \infty$ and $\alpha < \infty$ for any real α; if $\alpha > 0$, then $\alpha \times \infty = \infty$. However, we do not simultaneously allow $-\infty$ as a value since $\infty - \infty$ is undefined. The function ϕ is a *proper convex function* if it is convex and $\phi(x) < \infty$ for some $x \in X$. The domain of ϕ is

$$\operatorname{dom}\phi = \{\, x \in X \mid \phi(x) < \infty \,\}.$$

Usually we also require that ϕ be lower semicontinuous. Associated with a convex function ϕ is the *epigraph* of ϕ, which is a convex set:

$$\operatorname{epi}\phi = \left\{ \begin{bmatrix} x \\ s \end{bmatrix} \mid s \ge \phi(x) \right\} \subset X \times \mathbb{R}.$$

If ϕ is lower semicontinuous, then $\operatorname{epi}\phi$ is a closed set. Epigraphs are important, as they allow us to apply results for convex sets (such as the separating hyperplane theorem) to prove things about convex functions.

Associated with a proper lower semicontinuous convex function $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is its Fenchel dual $\phi^*\colon X' \to \mathbb{R} \cup \{\infty\}$:

$$\phi^*(\xi) = \sup_{x \in X} \langle \xi, x \rangle - \phi(x),$$

which is also proper, convex, and lower semicontinuous. Properties of Fenchel duals are given in Theorem B.15.

The *subdifferential* of a convex function ϕ is a set-valued function $\partial\phi\colon X \to \mathcal{P}(X')$ given by

$$\partial\phi(x) = \left\{\, w \in X' \mid \phi(y) \ge \phi(x) + \langle w, y - x \rangle \text{ for all } y \in X \,\right\}. \tag{2.51}$$

Subdifferentials generalize the notion of gradient or derivative for smooth convex functions in that $\partial\phi(x) = \{\nabla\phi(x)\}$ if ϕ is differentiable at x. An overview of properties of subdifferentials is given in Appendix B. The most important of these are given in Theorem B.14. The subdifferential already has a number of the important properties of maximal monotone operators: $\partial\phi(x)$ is always closed and convex (Theorem B.14(2)) and $\partial\phi$ has a closed graph (Theorem B.14(3)).

2.4 Maximal monotone operators

The theory of maximal monotone operators was initially developed by Minty [172, 173] as part of developing methods for solving electrical network problems with nonlinear resistors [171]. This theory was taken up and extended by Brézis [41] from static problems to dynamic problems.

2.4.1 Main properties

A *monotone operator* on a Hilbert space X is a set-valued function $\Phi\colon X \to \mathcal{P}(X')$ such that

$$\begin{aligned} &\langle y_1 - y_2, x_1 - x_2\rangle_{X'\times X} \geq 0 \\ &\qquad \text{for all } y_1 \in \Phi(x_1),\ y_2 \in \Phi(x_2). \end{aligned} \tag{2.52}$$

A set-valued function $\Phi\colon X \to \mathcal{P}(X')$ is a *maximal monotone operator* if Φ is a monotone operator, and the only monotone operator $\widetilde{\Phi}$, where $\Phi(x) \subseteq \widetilde{\Phi}(x)$ for all x, is $\widetilde{\Phi} = \Phi$. If X is a Hilbert space, we often identify X with X' and consider set-valued maps $\Phi\colon X \to \mathcal{P}(X)$ as (maximal) monotone operators. We also refer to single-valued functions $\phi\colon X \to X'$ as being (maximal) monotone operators by identifying ϕ with the set-valued map $\Phi(x) = \{\phi(x)\}$ for all x.

The function $s \mapsto \mathrm{Sgn}(s)$ given by

$$\mathrm{Sgn}(s) = \begin{cases} \{+1\}, & s > 0, \\ [-1,+1], & s = 0, \\ \{-1\}, & s < 0, \end{cases}$$

is a maximal monotone operator $\mathbb{R} \to \mathcal{P}(\mathbb{R})$. That it is monotone is easy to establish; that it is maximal is also fairly easy. Suppose $\widetilde{\Phi}\colon \mathbb{R} \to \mathcal{P}(\mathbb{R})$ is a monotone extension of Sgn: $\mathrm{Sgn}(s) \subseteq \widetilde{\Phi}(s)$ for all s. If there is a $y \in \widetilde{\Phi}(s)\backslash\mathrm{Sgn}(s)$, for some s, we consider the three different cases: $s > 0$, $s = 0$, and $s < 0$. First, for $s > 0$, $y \neq +1$. If $y > +1$, then choose $0 < s < s'$ so that $+1 \in \mathrm{Sgn}(s') \subseteq \widetilde{\Phi}(s')$. Then $(y - (+1))\big(s - s'\big) < 0$, contradicting monotonicity. If $y < +1$, then choose $0 < s' < s$ so that again $+1 \in \mathrm{Sgn}(s') \subseteq \widetilde{\Phi}(s')$. Then $(y - (+1))\big(s - s'\big) < 0$, again contradicting monotonicity. Thus we cannot have any extension $y \in \widetilde{\Phi}(s)\backslash\mathrm{Sgn}(s)$ for $s > 0$. Similar arguments show that we cannot have $y \in \widetilde{\Phi}(s)\backslash\mathrm{Sgn}(s)$ for $s < 0$. Now we consider $s = 0$. If $y \in \widetilde{\Phi}(0)\backslash\mathrm{Sgn}(0)$, then either $y > +1$ or $y < -1$. In the first case, choose $s' > 0$ and $y' = +1 \in \mathrm{Sgn}(s') \subseteq \widetilde{\Phi}(s')$. Then $(y - (+1))\big(0 - s'\big) < 0$, contradicting monotonicity. On the other hand, if $y < -1$, choose $s' < 0$ and $y' = -1 \in \mathrm{Sgn}(s') \subseteq \widetilde{\Phi}(s')$. Then $(y - (-1))\big(0 - s'\big) < 0$, again contradicting monotonicity. Thus $\widetilde{\Phi} = \mathrm{Sgn}$, and so Sgn is maximal monotone.

It is possible for a maximal monotone function to have the empty set as a value. Consider, for example, $\Phi\colon \mathbb{R} \to \mathcal{P}(\mathbb{R})$ given by

$$\Phi(x) = \begin{cases} \{0\} & \text{if } x > 0, \\ -\mathbb{R}_+ & \text{if } x = 0, \\ \emptyset & \text{if } x < 0. \end{cases}$$

The graph of this function is shown in Figure 2.5. This is maximal monotone, as can easily be checked: $x + \Phi(x) = y$ has one and only one solution in x: $x = \max(y, 0)$.

Every monotone set-valued function has a maximal monotone extension, thanks to Zorn's lemma.

Maximal monotone operators have some important properties for differential inclusions.

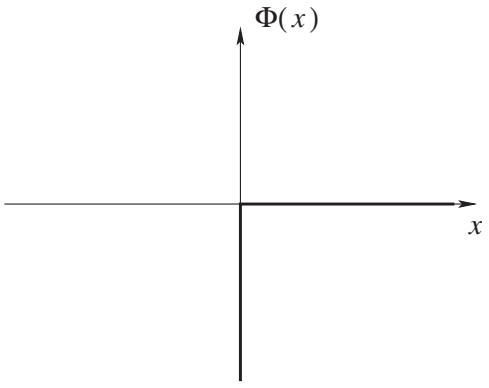


Figure 2.5: Example of maximal monotone function with a domain that is not the whole space, $\mathbb{R}$.

Lemma 2.24. *If $\Phi\colon X \to \mathcal{P}(X)$ is a maximal monotone operator, then*

1. *Φ has a closed graph, even in the strong $\times$ weak topology;*
2. *$\Phi(x)$ is a closed, convex set for all $x \in X$;*
3. *$\Phi(x) \neq \emptyset$ for some $x \in X$ (that is, $\operatorname{dom}\Phi \neq \emptyset$).*

Proof.

1. Consider the set-valued map $\widetilde{\Phi}$ where $\operatorname{graph}\widetilde{\Phi} = \overline{\operatorname{graph}\Phi}$, which is an extension of Φ. But $\widetilde{\Phi}$ is also monotone. To see this, suppose $y_1 \in \widetilde{\Phi}(x_1)$, $y_2 \in \widetilde{\Phi}(x_2)$. Then there are sequences $(x_1^{(k)}, y_1^{(k)}) \to (x_1, y_1)$ and $(x_2^{(k)}, y_2^{(k)}) \to (x_2, y_2)$ in the strong $\times$ weak topology as $k \to \infty$ where $(x_1^{(k)}, y_1^{(k)}), (x_2^{(k)}, y_2^{(k)}) \in \operatorname{graph}\Phi$. Since $\langle y_1^{(k)} - y_2^{(k)}, x_1^{(k)} - x_2^{(k)}\rangle \geq 0$ as Φ is monotone, taking limits as $k \to \infty$ we see that $\langle y_1 - y_2, x_1 - x_2\rangle \geq 0$. Thus $\widetilde{\Phi}$ is a monotone extension of Φ. Since Φ is maximal monotone, $\widetilde{\Phi} = \Phi$, and so $\operatorname{graph}\Phi = \overline{\operatorname{graph}\Phi}$ is closed.

2. Suppose that $\Phi(x^*)$ is not convex. Then there must be a pair $\widetilde{y}_1, \widetilde{y}_2 \in \Phi(x^*)$ and $0 < \theta < 1$ such that $y^* = \theta\widetilde{y}_1 + (1-\theta)\widetilde{y}_2 \notin \Phi(x^*)$. Let $\widetilde{\Phi}(x) = \Phi(x)$ for all $x \neq x^*$ and $\widetilde{\Phi}(x^*) = \Phi(x) \cup \{y^*\}$. We show that $\widetilde{\Phi}$ is monotone: The monotonicity condition $\langle y_1 - y_2, x_1 - x_2\rangle \geq 0$ for $y_1 \in \widetilde{\Phi}(x_1)$, $y_2 \in \widetilde{\Phi}(x_2)$ can fail only if one of x_1 and x_2 is x^* and one of y_1 and y_2 is y^*. Without loss of generality, suppose that $x_2 = x^*$ and $y_2 = y^*$. Then the monotonicity condition becomes

$$\begin{aligned} 0 &\leq \langle y_1 - y^*, x_1 - x^*\rangle \\ &= \theta\langle y_1 - \widetilde{y}_1, x_1 - x^*\rangle + (1-\theta)\langle y_1 - \widetilde{y}_2, x_1 - x^*\rangle, \end{aligned}$$

   which is true since $\widetilde{y}_1, \widetilde{y}_2 \in \Phi(x^*)$ and Φ is monotone.

   Since Φ is maximal monotone, $\widetilde{\Phi} = \Phi$ and so $y^* \in \Phi(x^*)$, contradicting our assumption. Thus $\Phi(x)$ must be convex, no matter what $x \in X$ is chosen.

3. If $\Phi(x) = \emptyset$ for all x, then we can set $\widetilde{\Phi}(x) = \{0\}$ for all x, which is a strict monotone extension, so Φ cannot be maximal monotone. □

Showing that a set-valued map $\Phi\colon X \to \mathcal{P}(X')$ is monotone is usually straightforward. Showing the "maximal" part is harder. The following theorem due to Minty [173] and Browder gives a characterization of monotone maps that are also maximal monotone. The method of proof is based on Borwein [34] and Borwein and Zhu [36], rather than the approach of Brézis [41]. The approach here is based on convex functions, subdifferentials, and Fenchel duality, which are described in Appendix B, which the reader may wish to review.

Theorem 2.25 (Minty–Browder). *If X is a Hilbert space and $\Phi\colon X \to \mathcal{P}(X')$ is monotone, then Φ is maximal monotone if and only if for each $y \in X'$ there is a unique solution x to the inclusion $y \in J_X(x) + \Phi(x)$, where $J_X\colon X \to X'$ is the standard duality map for X.*

This result in the case when X is a reflexive space and J_X is the associated duality map given by $J_X = \partial(\frac{1}{2}\|\cdot\|_X^2)$ is known as *Rockafellar's theorem*.

First, if Φ is monotone and $J_X + \Phi$ is surjective, then Φ is maximal monotone. To see this, suppose that $\widetilde{\Phi}$ is an extension to Φ but $J_X + \Phi$ is surjective. Suppose that $\xi \in \widetilde{\Phi}(x)\backslash\Phi(x)$ and let $\eta = J_X(x) + \xi$. Since $J_X + \Phi$ is surjective, there is a $z \in X$ where $\eta = J_X(z) + \zeta \in J_X(z) + \Phi(z)$. Now, if $\widetilde{\Phi}$ were also monotone, then

$$\begin{aligned} 0 = \langle \eta - \eta, x - z\rangle &= \langle J_X(x - z), x - z\rangle + \langle \xi - \zeta, x - z\rangle \\ &\geq \|x - z\|_X^2 . \end{aligned}$$

This implies $z = x$; but then $\zeta = \xi$, which contradicts $\xi \in \widetilde{\Phi}(x)\backslash\Phi(x)$. So $\widetilde{\Phi}$ cannot be monotone, and so Φ is maximal monotone.

The converse is much harder. The proof used here is based on the *Fitzpatrick function* $F_\Phi\colon X \times X' \to \mathbb{R} \cup \{\infty\}$ (see [36, 105]) for a maximal monotone operator Φ:

$$\begin{aligned} F_\Phi(x, \xi) &= \sup_{y,\eta:\eta\in\Phi(y)} \big[\langle \eta, x\rangle + \langle \xi, y\rangle - \langle \eta, y\rangle\big] \\ &= \langle \xi, x\rangle - \inf_{y,\eta:\eta\in\Phi(y)} \langle \eta - x, y - x\rangle . \end{aligned} \tag{2.53}$$

From (2.53) it is clear that F_Φ is a convex lower semicontinuous function. Since Φ is monotone, $\langle \eta - \xi, y - x\rangle \geq 0$ whenever $\eta \in \Phi(y)$ and $\xi \in \Phi(x)$. Thus

$$F_\Phi(x, \xi) \geq \langle \xi, x\rangle . \tag{2.54}$$

Equality holds in (2.54) if and only if $\langle \eta - \xi, y - x\rangle \geq 0$ for every pair (y, η) with $\eta \in \Phi(y)$. Note that we have equality if $y = x$ or $\xi = \eta$. By maximality of Φ, this is equivalent to $\xi \in \Phi(x)$. Thus equality holds in (2.54) if and only if $\xi \in \Phi(x)$.

We also need the following "sandwich lemma" of Borwein [34].

Lemma 2.26. *Suppose $f, -g\colon X \to \mathbb{R} \cup \{\infty\}$ are proper convex lower semicontinuous functions, where $f(x) \geq g(x)$ for all $x \in X$. Suppose also that*

$$0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom} g) .$$

Then there is a $\psi \in X'$ such that

$$f(x) - g(y) \geq \langle \psi, x - y \rangle .$$

Proof. Let $h(u) = \inf_{x \in X} f(x) - g(x-u) = (f \boxdot (-g(-\cdot)))(u)$, where $\psi \boxdot \chi$ is the inf-convolution of convex functions ψ and χ. Thus h is proper and convex. Note that $h(0) \geq 0$. If $0 \in \operatorname{int}(\operatorname{dom} f - \operatorname{dom} g)$, then $\operatorname{dom} h$ contains an open ball around zero (see Lemma B.19); thus h is Lipschitz continuous near zero and $\partial h(0) \neq \emptyset$. Choose $\psi \in \partial h(0)$, so $h(u) \geq h(0) + \langle \psi, u \rangle$. Unwrapping the definition of h we get $f(x) - g(x-u) \geq h(0) + \langle \psi, u \rangle \geq \langle \psi, u \rangle$. Setting $u = x - y$ gives the desired result. ∎

Rather than simply prove surjectivity of $J_X + \Phi$ for maximal monotone Φ, we show surjectivity of $J_X + \partial f + \Phi$ for f, a proper convex lower semicontinuous function. This extension is not much harder than showing surjectivity of $J_X + \Phi$, and it can be used to show maximal monotonicity of $\Phi + \Psi$ for maximal monotone Φ and Ψ under some mild but important conditions. Recall from Section B.2 that a proper lower semicontinuous convex function $\varphi \colon X \to \mathbb{R} \cup \{\infty\}$ has a dual function $\varphi^* \colon X' \to \mathbb{R} \cup \{\infty\}$ where $\varphi(x) + \varphi^*(\xi) \geq \langle \xi, x \rangle$ with equality if and only if $\xi \in \partial\varphi(x)$, or equivalently, $x \in \partial\varphi^*(\xi)$.

Theorem 2.27. *If $\Phi \colon X \to \mathcal{P}(X')$ is maximal monotone, $f \colon X \to \mathbb{R} \cup \{\infty\}$ is a proper convex lower semicontinuous function, and X is a Hilbert space, then $J_X + \partial f + \Phi$ is surjective, provided that*

$$0 \in \operatorname{int}\big[\operatorname{dom}\Phi - \operatorname{dom}\partial f\big].$$

Proof. We show that zero is in the image of $J_X + \partial f + \Phi$; for any $\eta \in X'$, we simply repeat the arguments with Φ replaced by $\Phi - \eta$ to show that η is in the image of $J_X + \partial f + \Phi$. Note that since f is a proper convex lower semicontinuous function, by the separating hyperplane theorem applied to epi f and the point $(x, f(x) - 1)$ with $x \in \operatorname{dom} f$, there are a $\psi \in X'$ and $\beta \in \mathbb{R}$ where $f(x) \geq \langle \psi, x \rangle + \beta$.

Let $f_J(x) = f(x) + \frac{1}{2}\|x\|_X^2$ for all x. Now f_J is also a proper convex lower semicontinuous function. Let $G(x,\xi) = -f_J(x) - f_J^*(-\xi)$, which is a concave function $G \colon X \times X' \to \mathbb{R} \cup \{-\infty\}$. By Theorem B.15, $-G(x,\xi) \geq \langle -\xi, x \rangle = -\langle \xi, x \rangle$ for all x and ξ, with equality if and only if $-\xi \in \partial f_J(x) = \partial f(x) + J_X(x)$. Note that $\operatorname{dom} f_J^* = X'$ since for any $\xi \in X'$,

$$\begin{aligned} f_J^*(\xi) &= \sup_x \langle \xi, x \rangle - f_J(x) \\ &\leq \sup_x \|\xi\|_{X'} \|x\|_X + |\beta| + \|\psi\|_{X'} \|x\|_X - \frac{1}{2}\|x\|_X^2 \\ &\leq |\beta| + \frac{1}{2}\left(\|\xi\|_{X'} + \|\psi\|_{X'}\right)^2 < +\infty. \end{aligned}$$

Thus $F_\Phi(x,\xi) \geq \langle \xi, x \rangle \geq G(x,\xi)$ for all x and ξ. Now $\operatorname{dom} G \supseteq \operatorname{dom} f \times X'$. On the other hand, $\operatorname{dom} F_\Phi \supseteq \operatorname{graph}\Phi$ since $F_\Phi(x,\xi) = \langle \xi, x \rangle < +\infty$ whenever $\xi \in \Phi(x)$. So

$$\operatorname{dom} F_\Phi - \operatorname{dom} G \supseteq (\operatorname{dom}\Phi - \operatorname{dom} f) \times X',$$

which contains zero in its interior by assumption. Then, by Lemma 2.26, there is $(\zeta, z) \in$

$X' \times X = \big(X \times X'\big)'$ (since X is reflexive), where

$$\begin{aligned} 0 &\le F_\Phi(x,\xi) - G(y,\eta) - \langle (\zeta, z), (x,\xi) - (y,\eta) \rangle \\ &= F_\Phi(x,\xi) - G(y,\eta) - \langle \zeta, x - y \rangle - \langle z, \xi - \eta \rangle \end{aligned} \tag{2.55}$$

for all $x, y \in X$ and $\xi, \eta \in X'$.

Our task now is to show that $\zeta \in \Phi(z)$ and $-\zeta \in \partial f_J(z)$, as then $0 = \zeta - \zeta \in J_X(z) + \partial f(z) + \Phi(z)$. To do this we show that we have equalities $F_\Phi(z,\zeta) = \langle \zeta, z \rangle = G(z,\zeta)$.

Suppose $\xi \in \Phi(x)$. Then $F_\Phi(x,\xi) = \langle \xi, x \rangle$, so (2.55) becomes $\langle \xi, x \rangle - G(y,\eta) \ge \langle \zeta, x - y \rangle + \langle z, \xi - \eta \rangle$. Since $\operatorname{dom} f_J^* = X'$, there is a $v \in X$ where $v \in \partial f_J^*(-\zeta)$. Then $G(v,\zeta) = -f_J(v) - f_J^*(-\zeta) = -\langle v, -\zeta \rangle$ by Theorem B.15. Substituting $y = v$ and $\eta = \zeta$ into (2.55) give

$$\langle \xi, x \rangle - \langle \zeta, v \rangle \ge \langle \zeta, x - v \rangle + \langle z, \xi - \zeta \rangle .$$

Rearranging this gives $\langle \xi - \zeta, x - z \rangle \ge 0$. Since this is true for all $(x,\xi) \in \operatorname{graph} \Phi$, it follows that $\zeta \in \Phi(z)$.

On the other hand, now substitute $x = z$ and $\xi = \zeta$ into (2.55):

$$\langle \zeta, z \rangle - G(y,\eta) \ge \langle \zeta, z - y \rangle + \langle z, \zeta - \eta \rangle .$$

Subtracting $2\langle \zeta, z \rangle$ gives $\langle -\zeta, z \rangle - G(y,\eta) \ge \langle -\zeta, y \rangle + \langle -\eta, z \rangle$. Adding $G(y,\eta)$ to both sides and substituting the definition of G gives

$$\langle -\zeta, z \rangle \ge \langle -\zeta, y \rangle - f_J(y) + \langle -\eta, z \rangle - f_J^*(-\eta).$$

Taking the supremum over both y and η gives

$$\langle -\zeta, z \rangle \ge f_J^*(-\zeta) + f_J^{**}(z) \;=\; f_J^*(-\zeta) + f_J(z).$$

Since we have the reverse inequality by Theorem B.15, we have $f_J(z) + f_J^*(-\zeta) = \langle -\zeta, z \rangle$ and $-\zeta \in \partial f_J(z) = J_X(z) + \partial f(z)$. Thus $0 \in J_X(z) + \partial f(z) + \Phi(z)$, as desired. □

The Minty–Browder theorem (Theorem 2.25) follows from Theorem 2.27 by taking $f(x) = 0$ for all $x \in X$, as then $\operatorname{dom} f = X$ and thus $\operatorname{dom} \Phi - \operatorname{dom} \partial f = X$, which contains zero in its interior.

The Minty–Browder theorem is also very important, as it immediately leads to the *resolvent operator* $R_\lambda \colon X' \to X$ given by

$$R_\lambda = (J_X + \lambda \Phi)^{-1} \qquad \text{for } \lambda > 0. \tag{2.56}$$

For maximal monotone Φ, R_λ is a well-defined, single-valued, and Lipschitz function $X' \to X$ with Lipschitz constant one. Identifying X and X' so that $J_X = I$ (the identity operator), the resolvent can be used to construct Lipschitz approximations Φ_λ, called *Yosida approximations*, to Φ:

$$\Phi_\lambda = \frac{I - R_\lambda}{\lambda} \qquad \text{for } \lambda > 0. \tag{2.57}$$

If we cannot identify X and X', then

$$\Phi_\lambda = J_X \frac{J_X^{-1} - R_\lambda}{\lambda} J_X = \frac{J_X - J_X R_\lambda J_X}{\lambda} \qquad \text{for } \lambda > 0. \tag{2.58}$$

These resolvents and approximations can be used to prove a large number of properties of maximal monotone operators and are crucial for much of the application to differential equations. The relationship between the resolvents and the Yosida approximations involves the minimum-norm points of $\Phi(x)$:

$$\begin{aligned}\Phi^0(x) = y, \qquad & \text{where } \|y\| = \min\{\|w\| \mid w \in \Phi(x)\} \\ &= \Pi_{\Phi(x)}(0)\end{aligned}$$

using the projection operator $\Pi_K(x)$ of x onto a closed convex K.

Lemma 2.28. *The resolvent R_λ is a monotone Lipschitz function $X' \to X$ with Lipschitz constant one; Φ_λ is a monotone Lipschitz function $X \to X'$ with Lipschitz constant $1/\lambda$. The second Yosida approximation $\left(\Phi_\mu\right)_\lambda = \Phi_{\mu+\lambda}$. Also, whenever $x \in \operatorname{dom}\Phi$, $\lim_{\lambda\downarrow 0}\Phi_\lambda(x) = \Phi^0(x)$ and $\|\Phi_\lambda(x)\| \uparrow \left\|\Phi^0(x)\right\|$ as $\lambda \downarrow 0$.*

Proof. To show that R_λ is monotone and Lipschitz for $\lambda > 0$, suppose $R_\lambda y_1 = x_1$ and $R_\lambda y_2 = x_2$. Equivalently, $y_1 \in J_X(x_1) + \lambda\,\Phi(x_1)$ and $y_2 \in J_X(x_2) + \lambda\,\Phi(x_2)$. Since $\lambda\,\Phi$ is monotone,

$$\begin{aligned}0 &\le \langle (y_1 - J_X(x_1)) - (y_2 - J_X(x_2)), x_1 - x_2\rangle \\ &= \langle y_1 - y_2, x_1 - x_2\rangle - \|x_1 - x_2\|_X^2,\end{aligned}$$

so

$$\|x_1 - x_2\|_X^2 \le \langle y_1 - y_2, x_1 - x_2\rangle \le \|y_1 - y_2\|_{X'}\,\|x_1 - x_2\|_X.$$

It is clear that $0 \le \langle y_1 - y_2, x_1 - x_2\rangle$, so R_λ is monotone. Dividing by $\|x_1 - x_2\|_X$ gives $\|x_1 - x_2\|_X \le \|y_1 - y_2\|_{X'}$, so R_λ is Lipschitz with constant one.

To show that Φ_λ is monotone and Lipschitz, suppose $\Phi_\lambda x_1 = y_1$ and $\Phi_\lambda x_2 = y_2$. This is equivalent to $y_i = J_X(J_X^{-1} - R_\lambda)J_X x_i/\lambda$. Now $R_\lambda J_X x_i = w_i$ means that $J_X x_i \in J_X w_i + \lambda\,\Phi(w_i)$, so $J_X w_i \in J_X x_i - \lambda\,\Phi(w_i)$. Then $y_i = (J_X x_i - J_X w_i)/\lambda \in \Phi(w_i)$. Thus

$$\begin{aligned}\langle y_1 - y_2, x_1 - x_2\rangle &= \langle y_1 - y_2, w_1 - w_2\rangle + \langle y_1 - y_2, (x_1 - x_2) - (w_1 - w_2)\rangle \\ &\ge \langle y_1 - y_2, (x_1 - w_1) - (x_2 - w_2)\rangle \\ &= \lambda\left\langle y_1 - y_2, J_X^{-1}y_1 - J_X^{-1}y_2\right\rangle = \lambda\,\|y_1 - y_2\|_{X'}^2.\end{aligned}$$

This shows that Φ_λ is both monotone and (after division by $\|y_1 - y_2\|_{X'}$) Lipschitz with constant $1/\lambda$.

To show that $\left(\Phi_\mu\right)_\lambda = \Phi_{\lambda+\mu}$, we note that $\left(\Phi_\mu\right)_\lambda = \widetilde{\Phi}_\lambda$, where

$$\begin{aligned}&\widetilde{\Phi}_\lambda = \left(J_X - J_X\widetilde{R}_\lambda J_X\right)/\lambda, \widetilde{R}_\lambda = \left(J_X + \lambda\Phi_\mu\right)^{-1}, \\ &\Phi_\mu = \left(J_X - J_X R_\mu J_X\right)/\mu, \\ &R_\mu = (J_X + \mu\Phi)^{-1}.\end{aligned}$$

We can unwrap the equation $y = \widetilde{\Phi}_\lambda x$ to show that it is equivalent to $y = \Phi_{\lambda+\mu}x$. Note that in the following, J_X is linear, but the other operators usually are not. First $y = \widetilde{\Phi}_\lambda x = \left(J_X x - J_X\widetilde{R}_\lambda(J_X x)\right)/\lambda$. Let $u = \widetilde{R}_\lambda(J_X x)$, so $y = (J_X x - J_X u)/\lambda$. From the definition of

$\widetilde{R}_\lambda$, $J_X u + \lambda\Phi_\mu(u) = J_X x$, so $J_X x - J_X u = \lambda\Phi_\mu(u)$. (Recall that Φ_μ is a single-valued function.) But $\Phi_\mu(u) = \big(J_X u - J_X R_\mu(J_X u)\big)/\mu$. Let $v = R_\mu(J_X u)$, so $J_X x - J_X u = \lambda\Phi_\mu(u) = (\lambda/\mu)(J_X u - J_X v)$. From the definition of R_μ, $J_X v + \mu\Phi(v) \ni J_X u$, or equivalently $J_X u - J_X v \in \mu\Phi(v)$.

We proceed by eliminating u in terms of x and v: since J_X is a linear isomorphism, $x - u = (\lambda/\mu)(u - v)$. Then we can write $u = (\lambda v + \mu x)/(\lambda + \mu)$. Substituting this into $J_X u - J_X v \in \mu\Phi(v)$ gives $\mu(J_X x - J_X v)/(\lambda + \mu) \in \mu\Phi(v)$; that is, $J_X x \in J_X v + (\lambda + \mu)\Phi(v)$, which means $v = R_{\lambda+\mu}(J_X x)$. Finally,

$$\begin{aligned} y &= (J_X x - J_X u)/\lambda = (J_X x - J_X v)/(\lambda + \mu) \\ &= \big(J_X x - J_X R_{\lambda+\mu}(J_X x)\big)/(\lambda + \mu) = \Phi_{\lambda+\mu}(x), \end{aligned}$$

as desired.

For the results concerning $\Phi^0(x)$, let $y_0 = \Phi^0(x) \in \Phi(x)$. Noting that $\Phi_\lambda(x) \in \Phi(R_\lambda(J_X x))$, by monotonicity of Φ,

$$\begin{aligned} 0 &\le \langle y_0 - \Phi_\lambda(x), x - R_\lambda(J_X x)\rangle \\ &= \lambda\Big\langle y_0 - \Phi_\lambda(x), J_X^{-1}\Phi_\lambda(x)\Big\rangle \\ &= \lambda\left(\Big\langle y_0, J_X^{-1}\Phi_\lambda(x)\Big\rangle - \|\Phi_\lambda(x)\|_{X'}^2\right). \end{aligned}$$

Thus

$$\|y_0\|_{X'}\,\|\Phi_\lambda(x)\|_{X'} \ge \Big\langle y_0, J_X^{-1}\Phi_\lambda(x)\Big\rangle \ge \|\Phi_\lambda(x)\|_{X'}^2;$$

dividing by $\|\Phi_\lambda(x)\|_{X'}$ gives $\|\Phi_\lambda(x)\|_{X'} \le \|y_0\|_{X'} = \big\|\Phi^0(x)\big\|_{X'}$. Since $\Phi_{\lambda+\mu} = (\Phi_\lambda)_\mu$, $\big\|\Phi_{\lambda+\mu}(x)\big\|_{X'} \le \big\|\Phi_\lambda^0(x)\big\|_{X'} = \|\Phi_\lambda(x)\|_{X'}$, as Φ_λ is single valued. Thus $\|\Phi_\lambda(x)\|$ increases as λ decreases.

Similarly, $\big\|\Phi_{\lambda+\mu}(x)\big\|_{X'}^2 \le \langle\Phi_\lambda(x), J_X^{-1}\Phi_{\lambda+\mu}(x)\rangle$. So

$$\begin{aligned} \big\|\Phi_{\lambda+\mu}(x) - \Phi_\lambda(x)\big\|_{X'}^2 &= \big\|\Phi_{\lambda+\mu}(x)\big\|_{X'}^2 + \|\Phi_\lambda(x)\|_{X'}^2 - 2\Big\langle\Phi_\lambda(x), J_X^{-1}\Phi_{\lambda+\mu}(x)\Big\rangle \\ &\le \|\Phi_\lambda(x)\|_{X'}^2 - \big\|\Phi_{\lambda+\mu}(x)\big\|_{X'}^2. \end{aligned}$$

Since $\|\Phi_\lambda(x)\|_{X'}$ is bounded as $\lambda \downarrow 0$ (by $\big\|\Phi_\lambda^0(x)\big\|_{X'}$) and increasing, it has a limit. This means that $\Phi_\lambda(x)$ is a Cauchy sequence as $\lambda \downarrow 0$, and thus it has a limit which we call y. Now $\lambda\Phi_\lambda(x) = J_X x - J_X R_\lambda(J_X x) \to 0$ as $\lambda \downarrow 0$, so $R_\lambda(J_X x) \to x$ as $\lambda \downarrow 0$. Now $\Phi_\lambda(x) \in \Phi(R_\lambda(J_X x))$, so the limit $y \in \Phi(x)$ as Φ has a closed graph. But from the arguments regarding $\|\Phi_\lambda(x)\|_{X'}$, $\|y\|_{X'} \le \big\|\Phi^0(x)\big\|_{X'}$. But the only way that this can be true is if $y = \Phi^0(x)$. That is, $\Phi_\lambda(x) \to \Phi^0(x)$ as $\lambda \downarrow 0$, as desired. □

While these results may seem somewhat technical, they will be very useful in understanding solutions of differential equations with maximal monotone operators.

2.4.2 More examples of maximal monotone operators

Maximal monotone functions $\Phi\colon \mathbb{R} \to \mathbb{R}$ are relatively easy to describe. They are essentially single-valued monotone functions on an interval (a, b) with the jump discontinuities

"filled in." Except for the case $\Phi(x) = \mathbb{R}$ for $x = x^*$ and $\Phi(x) = \emptyset$ otherwise, for each maximal monotone $\Phi\colon \mathbb{R} \to \mathbb{R}$ there is an interval (a, b) (a or b possibly $\pm\infty$) and a monotone function $\phi\colon (a,b) \to \mathbb{R}$ such that if x is a point of continuity of ϕ, then $\Phi(x) = \{\phi(x)\}$, and if x is a point of discontinuity of ϕ, then $\Phi(x) = \big[\phi(x^-), \phi(x^+)\big]$, where $\phi(x^-) = \lim_{z\uparrow x} \phi(z)$ and $\phi(x^+) = \lim_{z\downarrow x} \phi(z)$. If $\phi(b^-)$ is finite, then $\Phi(b) = [\phi(b^-), \infty)$; otherwise $\Phi(b) = \emptyset$. A similar rule applies at a: if $\phi(a^+)$ is finite, then $\Phi(a) = (-\infty, \phi(a^+)]$; otherwise $\Phi(a) = \emptyset$. If $x \notin [a, b]$, then $\Phi(x) = \emptyset$.

The Sgn function can be obtained from the ordinary sgn function in this way, and the function of Figure 2.5 can be obtained from the zero function on the interval $(0, +\infty)$.

As will be noted in the next section, subdifferentials of proper lower semicontinuous convex functions also provide a class of maximal monotone operators. For a closed convex set K the normal cone operator N_K is a maximal monotone operator that can be represented as a subdifferential.

Lemma 2.29. *For any nonempty closed convex set $K \subseteq X$, where X is a Hilbert space, the normal cone operator N_K is a maximal monotone operator.*

The proof of this will wait until the representation of N_K as a subdifferential is given in the next section. However, the normal cone operator is very useful, as it often gives a way to impose hard constraints.

Other examples of maximal monotone operators include, for example, elliptic partial differential operators such as $-\nabla^2\colon H^1(\Omega) \to H^{-1}(\Omega)$. Often we should be careful about the choice of space X on which the operator acts so that we have an operator $X \to X'$. Since $-\nabla^2$ is a linear operator, all that is required to show that it is maximal monotone is that it is defined on all of $H^1(\Omega)$ and that

$$\left\langle u, -\nabla^2 u \right\rangle = \int_\Omega u\left(-\nabla^2 u\right) dx \geq 0 \qquad \text{for all } u \in H^1(\Omega). \tag{2.59}$$

To show this, we just need to note that since $-\nabla^2$ is bounded $H^1(\Omega) \to H^{-1}(\Omega)$, it is enough to show that (2.59) holds for a dense subset of $H^1(\Omega)$, such as for smooth functions u. In fact, we can use smooth functions u with $\partial u/\partial n = 0$ on $\partial\Omega$ since making this requires just a small change to u (and an $\mathcal{O}(1)$ change to ∇u) in a small neighborhood of the boundary $\partial\Omega$. Then we can use the classical divergence theorem:

$$\begin{aligned}
\int_\Omega u\left(-\nabla^2 u\right) dx &= \int_\Omega \left[\nabla\cdot(-u\nabla u) + \nabla u\cdot\nabla u\right] dx \\
&= \int_{\partial\Omega} -u\,\frac{\partial u}{\partial n}\, dS + \int_\Omega \nabla u\cdot\nabla u\, dx \\
&= \int_\Omega \nabla u\cdot\nabla u\, dx \;\geq\; 0.
\end{aligned}$$

If ϕ is a proper lower semicontinuous convex function, then $\partial\phi$ is a maximal monotone operator $X \to \mathcal{P}(X')$. Monotonicity is easy to check: if $g_1 \in \partial\phi(x_1)$ and $g_2 \in \partial\phi(x_2)$, then

$$\begin{aligned}
\phi(x_2) - \phi(x_1) &\geq \langle g_1, x_2 - x_1\rangle, \\
\phi(x_1) - \phi(x_2) &\geq \langle g_2, x_1 - x_2\rangle.
\end{aligned}$$

Adding these inequalities gives

$$0 \le \langle g_1 - g_2, x_1 - x_2 \rangle .$$

Since this is true for all $g_1 \in \partial\phi(x_1)$ and $g_2 \in \partial\phi(x_2)$, $\partial\phi$ is a monotone map $X \to \mathcal{P}(X')$. Showing that it is maximal is more difficult. Fortunately this can be done using optimization theory. First, we note that x^* is a global minimizer of a proper convex lower semicontinuous function if and only if $0 \in \partial\phi(x^*)$ (Theorem B.14(6)).

We can now use the characterization of Theorem 2.25 to show that $\partial\phi$ is maximal monotone.

Lemma 2.30. *If X is a Hilbert space and $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is a proper lower semicontinuous convex function, then the subdifferential $\partial\phi\colon X \to \mathcal{P}(X')$ is maximal monotone.*

Proof. We have just seen that $\partial\phi$ is monotone. Now we show that it is maximal. Since ϕ is proper, there is x_0 such that $\phi(x_0) < \infty$. We first need to find a linear lower bound on ϕ. To do this, we consider the point $(x_0, \phi(x_0) - 1) \notin \operatorname{epi}\phi$. Thus by the separating hyperplane theorem there is $(w, \beta) \in X' \times \mathbb{R}$ and $\rho \in \mathbb{R}$ such that $\langle (w,\beta), (x_0, \phi(x_0) - 1)\rangle < \rho$ but $\langle (w,\beta), (x,\alpha)\rangle \ge \rho$ for all $(x,\alpha) \in \operatorname{epi}\phi$. In particular, since $(x_0, \phi(x_0)) \in \operatorname{epi}\phi$, $\langle w, x_0\rangle + \beta\phi(x_0) - \beta < \rho$ but $\langle w, x_0\rangle + \beta\phi(x_0) \ge \rho$; hence $\beta > 0$. Now $(x, \phi(x)) \in \operatorname{epi}\phi$ for all $x \in X$, so $\langle w, x\rangle + \beta\phi(x) \ge \rho$ for all x. Dividing by $\beta > 0$ gives $\phi(x) \ge \rho/\beta + \langle -w/\beta, x\rangle$. Setting $\alpha_0 := \rho/\beta$ and $g_0 := -w/\beta$, we have $\phi(x) \ge \alpha_0 + \langle g_0, x\rangle$ for all $x \in X$.

Now $\psi\colon X \to \mathbb{R} \cup \{\infty\}$ is a proper lower semicontinuous convex function given by $\psi(z) = \frac{1}{2}\|z\|_X^2 + \phi(z) - \langle y, z\rangle$. Also, $\psi(z) \to \infty$ as $\|z\|_X \to \infty$:

$$\begin{aligned}
\psi(z) &\ge \frac{1}{2}\|z\|_X^2 + \alpha_0 + \langle g_0, z - x_0\rangle - \langle y, z\rangle \\
&\ge \frac{1}{2}\|z\|_X^2 + \alpha_0 - \langle g_0, x_0\rangle - \left(\|g_0\|_{X'} + \|y\|_{X'}\right)\|z\|_X \\
&\to \infty \qquad \text{as } \|z\|_X \to \infty.
\end{aligned}$$

We can then apply Theorem B.13 to conclude that there is a global minimizer x^*. By Theorem B.14(6), $0 \in \partial\psi(x^*)$. Since $\partial\psi(x^*) = J_X(x^*) + \partial\phi(x^*) - y$, we have a solution $x = x^*$ of the inclusion

$$y \in J_X(x) + \partial\phi(x).$$

Since this holds for all $y \in X'$, $J_X + \partial\phi$ is surjective and $\partial\phi$ is a maximal monotone operator. □

As a simple example, consider $\phi\colon \mathbb{R} \to \mathbb{R}$ given by $\phi(x) = |x|$; then $\partial\phi(x) = \operatorname{Sgn}(x)$ is a maximal monotone operator. Another example is the *indicator function* for a nonempty closed convex set $K \subseteq X$:

$$I_K(x) = \begin{cases} 0, & x \in K, \\ \infty, & x \notin K. \end{cases} \tag{2.60}$$

Note that I_K is lower semicontinuous (since K is closed), convex (since K is convex), and proper (since $K \neq \emptyset$). Note that minimizing a lower semicontinuous convex function ϕ over a closed convex set K corresponds to minimizing $\phi + I_K$.

An important connection between properties of convex sets and subdifferentials is that $\partial I_K(x) = N_K(x)$, the normal cone for K at x.

Lemma 2.31. *If $K \subseteq X$ is a nonempty closed convex set, $N_K(x) = \partial I_K(x)$ for all $x \in X$.*

Proof. ($\subseteq$) Suppose $y \in N_K(x)$. We show that $y \in \partial I_K(x)$. Note that $N_K(x) \neq \emptyset$, so $x \in K$. From the definition of $N_K(x)$, $\langle y, z-x\rangle \leq 0$ for all $z \in K$. Thus, for any $z \in K$, $I_K(z) = 0 \geq \langle y, z-x\rangle = I_K(x) + \langle y, z-x\rangle$. If $z \notin K$, then $I_K(z) = +\infty$, so $I_K(z) \geq I_K(x) + \langle y, z-x\rangle$ as well. Thus $y \in \partial I_K(x)$.

($\supseteq$) Suppose that $y \in \partial I_K(x)$. Then, $I_K(x) < \infty$, so $x \in K$. Then, for any $z \in K$, $I_K(z) \geq I_K(x) + \langle y, z-x\rangle$, so $0 \geq 0 + \langle y, z-x\rangle$. Thus $y \in N_K(x)$. □

A consequence is that N_K is a maximal monotone operator.

We can compute things like resolvents and Yosida approximations for N_K for $K \subseteq X$: $R_\lambda = (J_X + \lambda N_K)^{-1}$. Then $y = R_\lambda x$ means that $x \in J_X y + \lambda N_K(y)$, or equivalently $x - J_X y \in \lambda N_K(y)$. Since $N_K(y)$ is a cone, we can absorb the factor of $\lambda > 0$ into $N_K(y)$. Then, for any $z \in K$, $0 \geq \langle x - J_X y, z - y\rangle = (J_X^{-1}x - y, z - y)_X$. Thus $y = \Pi_K(J_X^{-1}x)$. The Yosida approximation is then

$$\begin{aligned}(N_K)_\lambda(x) &= \lambda^{-1}(J_X - J_X R_\lambda J_X)(x)\\ &= \lambda^{-1}\left(J_X x - J_X \Pi_K(J_X^{-1} J_X x)\right)\\ &= \lambda^{-1} J_X (x - \Pi_K(x)).\end{aligned} \tag{2.61}$$

2.4.3 Sums of maximal monotone operators

When is the sum of two maximal monotone operators also maximal monotone? The sum of two monotone operators must be monotone: suppose Φ and Ψ are maximal monotone operators $X \to \mathcal{P}(X')$. Then, whenever $y_i \in \Phi(x_i)$ and $z_i \in \Psi(x_i)$ for $i = 1, 2$, we have

$$\begin{aligned}&\langle (y_2 + z_2) - (y_1 + z_1), x_2 - x_1\rangle\\ &\quad = \langle y_2 - y_1, x_2 - x_1\rangle + \langle z_2 - z_1, x_2 - x_1\rangle \geq 0.\end{aligned}$$

But it is not immediately clear that $\Phi + \Psi$ is *maximal* monotone. In fact, it can be false. One of the simplest examples is given by

$$\Phi(x) = \begin{cases} +\mathbb{R}_+ & \text{if } x = -1,\\ \{0\} & \text{if } x < -1,\\ \emptyset & \text{if } x > -1,\end{cases}$$

$$\Psi(x) = \begin{cases} -\mathbb{R}_+ & \text{if } x = +1,\\ \{0\} & \text{if } x > +1,\\ \emptyset & \text{if } x < +1.\end{cases}$$

Note that the graph of $\Phi + \Psi$ is empty!

Another example where $\operatorname{dom}\Phi \cap \operatorname{dom}\Psi \neq \emptyset$ can be constructed in two dimensions using normal cones. Let $\Phi = N_{C_1}$ and $\Psi = N_{C_2}$, where $C_1 = \left\{(x, y) \mid (x-1)^2 + y^2 \leq 1\right\}$ and $C_2 = \left\{(x, y) \mid (x+1)^2 + y^2 \leq 1\right\}$. Then $N_{C_1}(0,0) = \mathbb{R}_+ \mathbf{e}_1$ and $N_{C_2}(0,0) = -\mathbb{R}_+ \mathbf{e}_1$,

where $\mathbf{e}_1 = [1,0]^T$. Furthermore, since $\operatorname{dom}\Phi = C_1$ and $\operatorname{dom}\Psi = C_2$, $\operatorname{dom}(\Phi+\Psi) = \operatorname{dom}\Phi \cap \operatorname{dom}\Psi = C_1 \cap C_2 = \{(0,0)\}$. But $\Phi(0,0)+\Psi(0,0) = \mathbb{R}_+\mathbf{e}_1 - \mathbb{R}_+\mathbf{e}_1 = \mathbb{R}\mathbf{e}_1$. For $\Phi+\Psi$ to be maximal monotone, $I+\Phi+\Psi$ must be surjective; because the domain is just one point, this means that $\Phi+\Psi$ being maximal monotone implies that $\Phi(0,0)+\Psi(0,0) = \mathbb{R}^2$, which is false. Thus $\Phi+\Psi$ is not maximal monotone even though $\operatorname{dom}\Phi \cap \operatorname{dom}\Psi \neq \emptyset$.

Often $\Phi+\Psi$ is maximal monotone for maximal monotone Φ and Ψ; however, the conditions are nontrivial and can cause significant complications. For example, if Ψ is a Lipschitz monotone function $X \to X'$ and Φ is maximal monotone, then $\Phi+\Psi$ is maximal monotone. The simplest general statement along these lines seems to be as follows.

Lemma 2.32. *If Φ and Ψ are maximal monotone $X \to \mathcal{P}(X')$, X a Hilbert space, and*

$$0 \in \operatorname{int}[\operatorname{dom}\Phi - \operatorname{dom}\Psi],$$

then $\Phi+\Psi$ is also maximal monotone.

We can prove this as an easy consequence of Theorem 2.27 as shown in [34, 35]. We follow their approach here.

Proof. Note that $\Phi \times \Psi \colon X \times X \to X' \times X'$ is also maximal monotone since $J_{X\times X} = J_X \times J_X$ and $J_{X\times X} + \Phi \times \Psi = (J_X+\Phi)\times(J_X+\Psi)$ is surjective. Now let $\Delta = \{(x,x) \mid x \in X\}$, the diagonal of $X \times X$, and consider the operator $\Phi \times \Psi + \partial I_\Delta \colon X \times X \to \mathcal{P}\left(X' \times X'\right)$. Now $\operatorname{dom}\Phi \times \Psi - \operatorname{dom}\partial I_\Delta = \operatorname{dom}\Phi \times \operatorname{dom}\Psi - \Delta$, which contains $\frac{1}{2}(\operatorname{dom}\Phi - \operatorname{dom}\Psi) \times \frac{1}{2}(\operatorname{dom}\Psi - \operatorname{dom}\Phi)$, which contains zero in its interior under the assumption that $0 \in \operatorname{int}[\operatorname{dom}\Phi - \operatorname{dom}\Psi]$.

Thus we can apply Theorem 2.27 to show that $J_{X\times X} + \Phi \times \Psi + \partial I_\Delta$ is surjective. For any $\xi, \eta \in X'$ there are $x, y \in X$, where $(\xi,\eta) \in J_X(x) \times J_X(y) + \Phi(x) \times \Psi(y) + \partial I_\Delta(x,y)$. But $\partial I_\Delta(x,y) = \emptyset$ if $x \neq y$, and if $x = y$, then $\partial I_\Delta(x,x) = N_\Delta(x,x) = \Delta^\perp = \left\{(\zeta,-\zeta) \mid \zeta \in X'\right\}$ since Δ is a linear subspace of $X \times X$. That is, for some $\zeta \in X'$,

$$\begin{aligned} \xi &\in J_X(x) + \Phi(x) + \zeta, \\ \eta &\in J_X(x) + \Psi(x) - \zeta. \end{aligned}$$

Adding the two inclusions gives

$$\xi + \eta \in 2J_X(x) + \Phi(x) + \Psi(x).$$

Thus $2J_X + \Phi + \Psi$ is surjective, so $\Phi+\Psi$ is maximal monotone. □

For example, consider $X = H^1(\Omega)$ and

$$K = \left\{ u \in H^1(\Omega) \mid u(x) - \varphi(x) \geq 0 \text{ for all } x \in \Omega \right\}.$$

Then $\Phi = -\nabla^2 \colon H^1(\Omega) \to H^{-1}(\Omega) = H^1(\Omega)'$ is maximal monotone. Now $\operatorname{dom}\Phi = H^1(\Omega) = X$, and since K is a closed convex set in $H^1(\Omega) = X$, $\operatorname{dom} N_K = K$. Now $\operatorname{dom}\Phi - \operatorname{dom}\Psi = X - K = X$, provided $K \neq \emptyset$. Note, for example, that $K \neq \emptyset$ if $\varphi \in H^1(\Omega)$. But X contains zero in its interior, and thus $-\nabla^2 + N_K$ is a maximal monotone operator $X \to \mathcal{P}(X')$.

Another easy consequence of this result is an extension of Theorem B.14(7).

Lemma 2.33. *If $\phi, \psi : X \to \mathbb{R} \cup \{\infty\}$ are convex proper lower semicontinuous functions and $0 \in \operatorname{int}[\operatorname{dom}\partial\phi - \operatorname{dom}\partial\psi]$, we have $\partial(\phi+\psi) = \partial\phi + \partial\psi$.*

Proof. Note that $\partial(\phi+\psi)(z) \supseteq \partial\phi(z) + \partial\psi(z)$ for all z. Now $\phi+\psi$ is proper since $\operatorname{dom}\partial\phi \subseteq \operatorname{dom}\phi$ and $\operatorname{dom}\partial\psi \subseteq \operatorname{dom}\psi$, and our assumption implies that $0 \in \operatorname{dom}\phi - \operatorname{dom}\psi$. That is, there is a point $x^* \in \operatorname{dom}\phi \cap \operatorname{dom}\psi$ so that $x^* \in \operatorname{dom}(\phi+\psi)$.

Each of $\partial\phi$ and $\partial\psi$ is maximal monotone. Since $0 \in \operatorname{int}[\operatorname{dom}\partial\phi - \operatorname{dom}\partial\psi]$, we can apply Lemma 2.32 to show that $\partial\phi + \partial\psi$ is maximal monotone. Since $\partial(\phi+\psi)$ is a maximal monotone operator whose graph contains the graph of $\partial\phi + \partial\psi$, by maximality, $\partial(\phi+\psi) = \partial\phi + \partial\psi$. □

An important special case of the formula $\partial(\phi+\psi) = \partial\phi + \partial\psi$ that does not require a constraint qualification like "$0 \in \operatorname{int}[\operatorname{dom}\partial\phi - \operatorname{dom}\partial\psi]$" is the case where $\phi = I_K$ and $\phi = I_L$ with K and L convex *polyhedral* sets. In this case, $\partial I_K(z) = N_K(z)$ and $\partial I_L(z) = N_L(z)$ are polyhedral cones. We assume that $z \in K \cap L$.

Polyhedral sets can be represented in terms of linear inequalities:

$$K = \bigcap_{j=1}^{m_K} \left\{ x \mid \langle \xi_j, x \rangle \geq \alpha_j \right\},$$
$$L = \bigcap_{j=1}^{m_L} \left\{ x \mid \langle \eta_j, x \rangle \geq \beta_j \right\}.$$

The intersection is therefore also a polyhedral set:

$$K \cap L = \bigcap_{j=1}^{m_K+m_L} \left\{ x \mid \langle \zeta_j, x \rangle \geq \gamma_j \right\},$$

with $\zeta_j = \xi_j$ for $j \leq m_K$ and $\zeta_j = \eta_{j-m_K}$ for $j > m_K$, and similarly for γ_j. If we set $\mathcal{J}(z) = \left\{ j \mid \langle \zeta_j, x \rangle = \gamma_j \right\}$, the tangent cone to $K \cap L$ can be written as

$$T_{K\cap L}(z) = \left\{ x \mid \langle \zeta_j, x \rangle \geq 0,\ j \in \mathcal{J}(z) \right\}.$$

The normal cone $N_{K\cap L}(z) = T_{K\cap L}(z)^\circ = -T_{K\cap L}(x)^*$ can then be easily computed using Lemma 2.34.

Lemma 2.34. *If $P = \{ x \mid \langle d_i, x \rangle \geq 0, i = 1, 2, \ldots, m \}$, then*

$$P^* = \operatorname{cone}\{d_1, d_2, \ldots, d_m\}.$$

Proof. ($\subseteq$) Suppose $\xi \in P^*$. If $\xi \notin \operatorname{cone}\{d_1, d_2, \ldots, d_m\}$, then there is a separating hyperplane

$$\begin{aligned} &\langle z, \xi \rangle < \beta, \\ &\langle z, \eta \rangle \geq \beta \qquad \text{for all } \eta \in \operatorname{cone}\{d_1, d_2, \ldots, d_m\}. \end{aligned}$$

Taking $\eta = 0$ we see that $\beta \leq 0$. Note that $\langle z, \eta \rangle \geq 0$ for all $\eta \in \operatorname{cone}\{d_1, d_2, \ldots, d_m\}$: for all $\alpha \geq 0$ we have $\alpha\eta \in \operatorname{cone}\{d_1, d_2, \ldots, d_m\}$, and so $\langle z, \alpha\eta \rangle \geq \beta$. Dividing by $\alpha > 0$ and taking

$\alpha \to \infty$ we get $\langle z, \eta \rangle \geq 0$. In particular, taking $\eta = d_i$ we see that $\langle z, d_i \rangle \geq 0$, and so $z \in P$. Thus $\langle z, \xi \rangle \geq 0 \geq \beta > \langle z, \xi \rangle$, which is a contradiction. Thus $\xi \in \text{cone}\{d_1, d_2, \ldots, d_m\}$.

($\supseteq$) Suppose $\xi \in \text{cone}\{d_1, d_2, \ldots, d_m\}$. Then $\xi = \sum_{i=1}^{m} \alpha_i d_i$ with $\alpha_i \geq 0$ for all i. Then, if $x \in P$, we have $\langle \xi, x \rangle = \sum_{i=1}^{m} \alpha_i \langle d_i, x \rangle \geq 0$, so $\xi \in P^*$. □

Returning to the matter of normal cones of polyhedral sets, this means that for $z \in K \cap L$,

$$N_{K\cap L}(z) = -\text{cone}\left\{ \zeta_j \mid \langle \zeta_j, z \rangle = \gamma_j \right\}.$$

By using the same arguments for K and L separately,

$$N_K(z) = -\text{cone}\left\{ \xi_j \mid \langle \xi_j, z \rangle = \alpha_j \right\},$$
$$N_L(z) = -\text{cone}\left\{ \eta_j \mid \langle \eta_j, z \rangle = \beta_j \right\}.$$

But $\zeta_j = \xi_j$ if $j \leq m_K$ and $\zeta_j = \eta_{j-m_K}$ otherwise, and comparison of the index sets shows that indeed $N_{K\cap L}(z) = N_K(z) + N_L(z)$ as $\text{cone}(A \cup B) = \text{cone}(A) + \text{cone}(B)$.

Polyhedral sets arise sufficiently often enough that it is useful to realize that no "constraint qualification" type of conditions need to be satisfied in order to set $N_{K\cap L} = N_K + N_L$. This is also useful in the following section, where we show how Lagrange multipliers can be incorporated into VIs.

2.4.4 VIs and Lagrange multipliers

A consequence of the equivalent formulation (2.45), $0 \ni F(z) + N_K(z)$, for VI(F, K) is that we can have VIs with Lagrange multipliers representing certain types of convex constraints. Consider, for example, VI(F, K), where $K = L \cap M$ with L and M being closed convex sets:

$$z \in K \quad \& \quad 0 \leq \langle \widetilde{z} - z, F(z) \rangle \qquad \text{for all } \widetilde{z} \in K.$$

From Lemma 2.32, we note that provided

$$0 \in \text{int}[L - M] = \text{int}[\text{dom}\, \partial I_L - \text{dom}\, \partial I_M],$$

we have

$$\begin{aligned} N_{L\cap M}(z) &= \partial I_{L\cap M}(z) \\ &= \partial (I_L + I_M)(z) \\ &= \partial I_L(z) + \partial I_M(z) \\ &= N_L(z) + N_M(z). \end{aligned}$$

Note that if L and M are polyhedral sets, then by the previous section, we do not need any constraint qualification to ensure that $N_{L\cap M}(z) = N_L(z) + N_M(z)$.

Hence z solves VI(F, K) if and only if $0 \in F(z) + N_K(z) = F(z) + N_L(z) + N_M(z)$. Thus there is a $\mu \in N_M(z)$ where $0 \in F(z) + \mu + N_L(z)$. Then $0 \in F(z) + \mu + N_L(z)$. That is,

$$z \in L \quad \& \quad 0 \leq \langle \widetilde{z} - z, F(z) + \mu \rangle \qquad \text{for all } \widetilde{z} \in L, \tag{2.62}$$
$$\mu \in N_M(z). \tag{2.63}$$

To make the connections with Lagrange multipliers clearer, consider M to be generated by some constraint functions:

$$M = \{ x \mid Ax = b \text{ and } \phi(x) \le 0 \},$$

where $A\colon X \to Y$ is a linear operator and ϕ is a convex proper lower semicontinuous function. Furthermore, we assume that the *Slater constraint qualification* (B.22) holds for ϕ; that is, there is an $\widehat{x}$ where $A\widehat{x} = b$ and $\phi(\widehat{x}) < 0$. We will also suppose that the adjoint operator $A^*\colon Y' \to X'$ has closed range (which is automatically true if X is finite dimensional) and that ϕ is finite in a neighborhood of $\widehat{x}$. Then, for $z \in M$,

$$N_M(z) = N_{\{x \mid Ax = b\}}(z) + N_{\{x \mid \phi(x) \le 0\}}(z).$$

From the Slater constraint qualification (Lemma B.16), $N_{\{x \mid \phi(x) \le 0\}}(z) = \operatorname{cone} \partial\phi(z)$ if $\phi(z) = 0$ and $N_{\{x \mid \phi(x) \le 0\}}(z) = \{0\}$ if $\phi(z) < 0$. On the other hand,

$$\begin{aligned} N_{\{x \mid Ax = b\}}(z) &= T_{\{x \mid Ax = b\}}(z)^\circ \\ &= \{ x \mid Ax = 0 \}^\circ \\ &= \{ \xi \mid \langle \xi, x \rangle \le 0 \text{ for all } x, \text{ where } Ax = 0 \} \\ &= \{ \xi \mid \langle \xi, x \rangle = 0 \text{ for all } x, \text{ where } Ax = 0 \} \\ &= (\ker A)^\perp \;=\; \overline{\operatorname{range} A^*}. \end{aligned}$$

Thus, provided $\operatorname{range} A^*$ is closed, every element of $N_{\{x \mid Ax = b\}}(z)$ can be represented by $A^*\lambda$. Thus we can pick $\mu \ge 0$ and $\zeta \in \partial\phi(z)$, where

$$z \in L \quad \& \quad 0 \le \big\langle \widetilde{z} - z,\, F(z) + A^*\lambda + \mu\,\zeta \big\rangle \qquad \text{for all } \widetilde{z} \in L, \tag{2.64}$$

$$Az = b, \tag{2.65}$$

$$0 \ge \phi(z) \perp \mu \ge 0, \qquad \zeta \in \partial\phi(z). \tag{2.66}$$

The last conditions essentially recover the Karush–Kuhn–Tucker conditions on the Lagrange multiplier for inequality-constrained optimization (B.26). If $\phi(z) = \max_i \phi_i(z)$ with smooth ϕ_i, then we can decompose $\partial\phi(z) = \operatorname{co}\{ \nabla\phi_i(z) \mid \phi_i(z) = \phi(z) \}$ and obtain Lagrange multipliers μ_i satisfying

$$\begin{aligned} z \in L \quad \& \quad 0 &\le \left\langle \widetilde{z} - z,\, F(z) + A^*\lambda + \sum_i \mu_i \nabla\phi_i(z) \right\rangle \\ &\text{for all } \widetilde{z} \in L, \\ Az &= b, \\ 0 \ge \phi_i(z) \perp \mu_i &\ge 0 \qquad \text{for all } i. \end{aligned}$$

In the extreme case, we can make $L = X$, the entire space, and then the conditions reduce to the Karush–Kuhn–Tucker conditions

$$\begin{aligned} 0 &= F(z) + A^*\lambda + \sum_i \mu_i \nabla\phi_i(z), \\ Az &= b, \\ 0 \ge \phi_i(z) \perp \mu_i &\ge 0 \qquad \text{for all } i. \end{aligned}$$

Of course, VI conditions have reentered through the complementarity conditions between $\phi_i(z)$ and μ_i. Note that here $F(z)$ is not necessarily the gradient of any function, so we are not giving necessary conditions for a local minimum, but for a VI.

We can also proceed in the reverse direction, replacing a "VI with Lagrange multipliers" of the form (2.62)–(2.63) with a standard VI over a restricted set. We start with the "VI with Lagrange multipliers":

$$\begin{aligned} z \in L \quad \& \quad & 0 \le \langle \widetilde{z} - z,\, F(z) + \mu \rangle \qquad \text{for all } \widetilde{z} \in L, \\ & \mu \in N_M(z). \end{aligned}$$

The condition that $\mu \in N_M(z)$ implies that $z \in M$. Also, for any $\widetilde{z} \in M$, $\langle \widetilde{z} - z, \mu \rangle \le 0$. Thus if $\widetilde{z} \in L \cap M$, $0 \le \langle \widetilde{z} - z,\, F(z) + \mu \rangle \le \langle \widetilde{z} - z,\, F(z) \rangle$, so

$$z \in L \cap M \quad \& \quad 0 \le \langle \widetilde{z} - z,\, F(z) \rangle \qquad \text{for all } \widetilde{z} \in L \cap M,$$

and z solves $\mathrm{VI}(F, L \cap M) = \mathrm{VI}(F, K)$.

Note that no constraint qualifications are needed for turning a "VI with Lagrange multipliers" into a standard VI, just for the reverse operation.

2.5 Pseudomonotone operators

Pseudomonotonicity has at least two meanings, which we describe here.

There is pseudomonotonicity in the sense of Karamardian [139], which is the following property: $\Phi\colon X \to X'$ is pseudomonotone in the sense of Karamardian if

$$\langle \Phi(y), x - y \rangle \ge 0 \quad \text{implies} \quad \langle \Phi(x), x - y \rangle \ge 0. \tag{2.67}$$

There is also pseudomonotonicity in the sense of Brézis [40] for single-valued functions and Browder [43] and Naniewicz and Panagiotopoulos [188] for set-valued maps: $\Phi\colon X \to \mathcal{P}(X')$. It is pseudomonotonicity in the sense of Brézis et al. that we consider here, which is defined by the following three conditions:

- $\Phi(x)$ is closed, convex, and bounded for each $x \in X$;
- for any finite-dimensional space $F \subset X$, $\Phi|_F$ is upper semicontinuous into X' in the weak topology; and
- if $x_k \rightharpoonup x$ weakly in X as $k \to \infty$, and $y_k \in \Phi(x_k)$ satisfy

$$\limsup_{k \to \infty} \mathrm{Re}\, \langle y_k, x_k - x \rangle \le 0,$$

then for each $z \in X$ there is a $y \in \Phi(x)$ ($y = y(z)$ can depend on z) such that

$$\liminf_{k \to \infty} \mathrm{Re}\, \langle y_k, x_k - z \rangle \ge \mathrm{Re}\, \langle y, x - z \rangle .$$

The single-valued case has an excellent treatment in Zeidler [274, Chap. 27]. Pseudomonotone operators have found application to dynamic problems as well, such as in [153].

The main result for coercive pseudomonotone operators ($\lim_{\|x\|\to\infty} \inf_{\eta\in\Phi(x)} \langle \eta, x\rangle / \|x\| = +\infty$) is the following.

Theorem 2.35. *If $\Phi\colon X \to \mathcal{P}(X')$, with X a reflexive Banach space, is pseudomonotone, is bounded on bounded sets, and is coercive, then Φ is surjective.*

Unlike the Minty–Browder or Rockafellar theorems for maximal monotone operators, there is no uniqueness of x satisfying $\eta \in \Phi(x)$. Thus we do not have resolvents in general for pseudomonotone operators. Also, for single-valued functions in finite dimensions, pseudomonotonicity reduces to continuity. So pseudomonotonicity becomes useful only in infinite dimensions.

Rather than prove Theorem 2.35 directly, we prove a generalization for VIs.

Theorem 2.36. *Suppose $\Phi\colon K \to \mathcal{P}(X')$ is pseudomonotone with nonempty values, where K is a closed convex subset of X, a reflexive Banach space, and Φ is weakly coercive on K in the sense that there are a fixed $z_0 \in K$ and number R where $\langle \eta, x - z_0\rangle > 0$ for all $\eta \in \Phi(x)$, where $x \in K$ and $\|x\| \geq R$. Suppose also that Φ maps bounded sets to bounded sets. Then* $\mathrm{VI}(\Phi, K)$ *has a solution.*

Note that $\mathrm{VI}(\Phi, K)$ for set-valued Φ is the problem of finding z and $\zeta \in \Phi(z)$, where

$$z \in K \quad \& \quad 0 \leq \langle \widetilde{z} - z, \zeta\rangle \qquad \text{for all } \widetilde{z} \in K.$$

If we take $K = X$ in Theorem 2.36, we get Theorem 2.35 as an immediate corollary. The proof below is based on a proof of a related result by de Figueiredo [72].

Proof. Let $\mathcal{F}$ be the set of all finite subsets of K that contain z_0. For each $F \in \mathcal{F}$ let x_F and $\eta_F \in \Phi(x_F)$ solve the VI

$$x_F \in \mathrm{co}(F) \quad \& \quad 0 \leq \langle z - x_F, \eta_F\rangle \qquad \text{for all } z \in \mathrm{co}(F). \tag{2.68}$$

There is a solution to this VI, which we now show. First, $\mathrm{co}(F)$ is finite dimensional and is contained in the finite-dimensional space $\operatorname{span} F$. Since $\operatorname{span} F$ is finite dimensional, we can give it an inner product which generates an equivalent norm.

In the case that Φ is single valued, Φ being pseudomonotone implies that Φ is continuous on $\mathrm{co}(F)$. Then (2.68) is equivalent to

$$0 \in \Phi(x_F) + N_{\mathrm{co}(F)}(x_F)$$

and to

$$x_F = \Pi_{\mathrm{co}(F)}\Big(x_F - J^{-1}_{\operatorname{span} F}(\Phi(x_F))\Big).$$

But $x \mapsto \Pi_{\mathrm{co}(F)}\big(x - J^{-1}_{\operatorname{span} F}(\Phi(x))\big)$ is a continuous function from $\mathrm{co}(F)$ to $\mathrm{co}(F)$, and so by Brouwer's theorem (Proposition A.11) there is a fixed point x_F solving (2.68).

In the case that Φ is multivalued, we use a Galerkin approximation $\Phi_F\colon \mathrm{co}(F) \to (\operatorname{span} F)'$ defined by

$$\Phi_F(x) = \big\{\, \eta' \mid \eta' = \eta|\operatorname{span} F,\ \eta \in \Phi(x) \,\big\}.$$

This is upper semicontinuous with closed convex values. We can approximate Φ_F by a single-valued function $\Phi_{F,\epsilon}$ so that the Hausdorff distance between graph Φ_F and graph $\Phi_{F,\epsilon}$ is less than ϵ for any given $\epsilon > 0$. This can be done, for example, by using a piecewise affine approximation to a selection of Φ_F on a triangulation of F. Then we can solve the VI

$$x_{F,\epsilon} \in \mathrm{co}(F) \quad \& \quad 0 \le \left\langle z - x_{F,\epsilon}, \Phi_{F,\epsilon}(x_{F,\epsilon}) \right\rangle \qquad \text{for all } z \in \mathrm{co}(F). \tag{2.69}$$

Since $\mathrm{co}(F)$ is compact and Φ maps bounded sets to bounded sets, we can pick a convergent subsequence (also denoted $x_{F,\epsilon}$) as $\epsilon \downarrow 0$, so that $x_{F,\epsilon} \to x_F$ and $\Phi_{F,\epsilon}(x_{F,\epsilon}) \rightharpoonup \eta_F \in \Phi(x_F)$ weakly in X'. We know that $\eta_F \in \Phi(x_F)$ since Φ is upper semicontinuous on $\mathrm{co}(F)$. Taking limits of (2.69) as $\epsilon \downarrow 0$ gives (2.68), so we have a solution of $\mathrm{VI}(\Phi, \mathrm{co}(F))$.

Note that all solutions of $\mathrm{VI}(\Phi, \mathrm{co}(F))$ have $\|x\| < R$, as if $\|x\| \ge R$, $\langle \eta, z_0 - x \rangle < 0$ for all $\eta \in \Phi(x)$. Thus $x_F \in K \cap R\overline{B_X}$.

For $G \in \mathcal{F}$, let $V_G = \{ x_F \mid F \in \mathcal{F} \text{ and } G \subseteq F \}$. Now $V_G \subset K \cap R\overline{B_X}$. Let $\overline{A}^w$ denote the weak closure of A. Since $K \cap R\overline{B_X}$ is closed and convex, it is weakly closed. Since X is reflexive, $R\overline{B_X}$ is weakly compact by Alaoglu's theorem. As K is weakly closed (being convex and closed), $K \cap R\overline{B_X}$ is weakly compact. Since $\overline{V_G}^w$ is a weakly closed subset of $K \cap R\overline{B_X}$, $\overline{V_G}^w$ is also weakly compact.

The sets $\overline{V_G}^w$ with $G \in \mathcal{F}$ have the finite intersection property: for any finite collection of such sets $\overline{V_{G_1}}^w, \overline{V_{G_2}}^w, \ldots, \overline{V_{G_m}}^w$,

$$\begin{aligned} \overline{V_{G_1}}^w \cap \overline{V_{G_2}}^w \cap \cdots \cap \overline{V_{G_m}}^w &\supseteq V_{G_1} \cap V_{G_2} \cap \cdots \cap V_{G_m} \\ &= V_{G_1 \cup G_2 \cup \cdots \cup G_m} \ni x_{G_1 \cup G_2 \cup \cdots \cup G_m}, \end{aligned}$$

so $\overline{V_{G_1}}^w \cap \overline{V_{G_2}}^w \cap \cdots \cap \overline{V_{G_m}}^w \ne \emptyset$. Thus $\bigcap_{G \in \mathcal{F}} \overline{V_G}^w \ne \emptyset$ by the finite intersection property (Lemma A.1). Let $\widehat{x} \in \bigcap_{G \in \mathcal{F}} \overline{V_G}^w$.

We now show that $\widehat{x}$ gives a solution of $\mathrm{VI}(\Phi, K)$. First, $\widehat{x} \in K \cap R\overline{B_X} \subseteq K$. So all we need to show is that $\langle z - \widehat{x}, \eta \rangle \ge 0$ for all $z \in K$ with some $\eta \in \Phi(\widehat{x})$.

Pick $F \in \mathcal{F}$, where $\widehat{x}, z \in F$. Since $\widehat{x} \in \overline{V_F}^w$, there must be a sequence $x_k := x_{F_k} \rightharpoonup \widehat{x}$ where $F \subseteq F_k$. If $\eta_k = \eta_{F_k}$, we have

$$0 \ge \langle \eta_k, x_k - \widehat{x} \rangle, \qquad \eta_k \in \Phi(x_k).$$

Taking the limsup of the right-hand side and using pseudomonotonicity of Φ give

$$\begin{aligned} 0 &\ge \liminf_{k \to \infty} \langle \eta_k, x_k - z \rangle \\ &\ge \langle \widehat{\eta}, \widehat{x} - z \rangle, \qquad \widehat{\eta} \in \Phi(\widehat{x}). \end{aligned}$$

Note that the first inequality holds by (2.68) with $F = F_k$, and since $z \in F_k$. Thus we have $\widehat{x}, \widehat{\eta} \in \Phi(\widehat{x})$, which satisfy $\mathrm{VI}(\Phi, K)$. □

2.6 Signorini's problem

Signorini's problem [98, 225, 226] was a landmark problem that led to much of the early work on VIs [160]. The tools we have, along with some multivariable calculus and functional analysis, are sufficient to formulate the problem properly and show the existence of solutions to this problem.

Signorini initially posed his problem in a short paper [225] in 1933, which he much later expanded in more detail [226] in 1959 following a course he gave at his university (Instituto Nazionale di Alta Matematica), where he mentioned this problem to his students and colleagues. Gaetano Fichera and Mauro Picone took up this problem, with Signorini strongly encouraging them. However, finding that this problem with unknown boundary conditions was not part of the literature at that time, Fichera set about trying to develop a way of dealing with the problem. He used the principle of virtual work to gain an entrance to the problem, eventually obtaining a solution late in 1962. By this time, Signorini was in failing health, but was overjoyed that Fichera had been able to find a solution, with a short summary [98] published in 1963 and a complete treatment [99] published in 1964. Fichera describes his involvement in this problem and the development of the theory of VIs in [100].

A more detailed introduction to elasticity is given in Section 6.2. To quickly summarize, the main unknown is the displacement field $\mathbf{u}(\mathbf{x})$ where a point $\mathbf{x}$ in the undeformed body $\Omega \subset \mathbb{R}^d$ is moved to $\mathbf{x}+\mathbf{u}(\mathbf{x})$. Using linearized elasticity, the *strain tensor* $\varepsilon[\mathbf{u}] = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$ and the *stress tensor* σ are related by

$$\sigma_{ij}[\mathbf{u}] = \sum_{k,l=1}^{d} a_{ijkl}\,\varepsilon_{kl}[\mathbf{u}],$$

where a_{ijkl} is a collection of elasticity constants that have certain symmetries: $a_{ijkl} = a_{jikl} = a_{klij}$, etc. Note that both ε and σ are symmetric $d \times d$ matrix-valued functions of position $\mathbf{x}$. The constants a_{ijkl} define the elastic properties of the material of the body. These constants also have some other important properties; most particularly, there is an $\eta > 0$ such that

$$\sum_{i,j,k,l} a_{ijkl}\varepsilon_{ij}\varepsilon_{kl} \geq \eta \sum_{i,j} \varepsilon_{ij}^2.$$

These relationships can be written in short form: $\sigma = \mathcal{A}\varepsilon$ and $\varepsilon : \mathcal{A}\varepsilon \geq \eta\,\varepsilon : \varepsilon$, where $A : B = \sum_{i,j} a_{ij}b_{ij}$ is the standard inner product on matrices.

Physically, Signorini's problem represents an elastic body which makes contact with a frictionless rigid obstacle. The obstacle is represented by the fact that there is a hard limit on the normal displacement of the body on the boundary close to the obstacle. On other parts of the boundary, there are either traction boundary conditions (a known or given force is acting on the boundary) or displacement boundary conditions (where the displacement of the boundary is given). Traction boundary conditions have the form

$$\sigma(\mathbf{x})\,\mathbf{n}(\mathbf{x}) = \mathbf{t}(\mathbf{x}), \qquad \mathbf{x} \in \Gamma_t,$$

while displacement boundary conditions have the form

$$\mathbf{u}(\mathbf{x}) = \mathbf{d}(\mathbf{x}), \qquad \mathbf{x} \in \Gamma_d.$$

On the part of the boundary where contact can occur Γ_c, we have the complementarity condition

$$\begin{aligned} &0 \leq N(\mathbf{x}) \perp \varphi(\mathbf{x}) - \mathbf{n}(\mathbf{x}) \cdot \mathbf{u}(\mathbf{x}) \geq 0, \qquad \mathbf{x} \in \Gamma_c, \\ &\sigma(\mathbf{x})\,\mathbf{n}(\mathbf{x}) = -N(\mathbf{x})\,\mathbf{n}(\mathbf{x}), \qquad \mathbf{x} \in \Gamma_c. \end{aligned}$$

Note that $N(\mathbf{x})$ is the normal contact force that prevents penetration, and $\varphi(\mathbf{x})$ represents the distance from the undeformed body to the obstacle. This is the complementarity representation of the Signorini contact conditions. Within the body we have the standard equations of elasticity, which can be written as

$$0 = \operatorname{div}\sigma + \mathbf{f}(\mathbf{x}), \qquad \mathbf{x} \in \Omega,$$

where $\mathbf{f}(\mathbf{x})$ represents the nonelastic (or external) forces acting on the body. Note that $(\operatorname{div}\sigma)_i = \sum_j \partial\sigma_{ij}/\partial x_j$ so that $\operatorname{div}\sigma$ is a vector-valued function.

So far we have a formulation of Signorini's problem as a CP. Now we will turn it into a VI. To do this we set

$$K = \left\{ \mathbf{u} \in H^1(\Omega) \mid \mathbf{n}\cdot\mathbf{u} \le \varphi \text{ on } \Gamma_c,\ \mathbf{u} = \mathbf{d} \text{ on } \Gamma_d \right\}.$$

Since the restriction of functions in $H^1(\Omega)$ to the boundary or part thereof is a continuous operation, K is a closed and convex subset of $H^1(\Omega)$. For any $\mathbf{w} \in K$,

$$\begin{aligned}
0 &= \int_\Omega (\mathbf{w}-\mathbf{u})\cdot[\operatorname{div}\sigma[\mathbf{u}] + \mathbf{f}]\,dx \\
&= \int_\Omega \sum_i (w_i - u_i)\left[\sum_j \frac{\partial\sigma_{ij}}{\partial x_j} + f_i\right] dx \\
&= \int_\Omega \sum_{i,j}\left[\frac{\partial}{\partial x_j}\left((w_i-u_i)\sigma_{ij}\right) - \left(\frac{\partial w_i}{\partial x_j} - \frac{\partial u_i}{\partial x_j}\right)\sigma_{ij} + (w_i-u_i)\,f_i\right] dx \\
&= \int_{\partial\Omega}\sum_{i,j}(w_i-u_i)\,\sigma_{ij}n_j\,dS - \int_\Omega [(\nabla\mathbf{w}-\nabla\mathbf{u}) : \sigma[\mathbf{u}] - (\mathbf{w}-\mathbf{u})\cdot\mathbf{f}]\,dx.
\end{aligned}$$

The boundary integral term is

$$\int_{\partial\Omega}(\mathbf{w}-\mathbf{u})\cdot\sigma[\mathbf{u}]\mathbf{n}\,dS = \int_{\Gamma_t}(\mathbf{w}-\mathbf{u})\cdot\mathbf{t}\,dS - \int_{\Gamma_c}(\mathbf{w}-\mathbf{u})\cdot\mathbf{n}\,N\,dS$$

since $\sigma[\mathbf{u}]\mathbf{n} = \mathbf{t}$ on Γ_t, $\mathbf{u} = \mathbf{w}$ on Γ_d, and $\sigma[\mathbf{u}]\mathbf{n} = -N\mathbf{n}$ on Γ_c. The complementarity condition between $\mathbf{u}\cdot\mathbf{n} - \varphi$ and N means that the last term becomes

$$\begin{aligned}
-\int_{\Gamma_c}(\mathbf{w}-\mathbf{u})\cdot\mathbf{n}\,N\,dS &= \int_{\Gamma_c}((\varphi - \mathbf{w}\cdot\mathbf{n}) - (\varphi - \mathbf{u}\cdot\mathbf{n}))\,N\,dS \\
&= \int_{\Gamma_t}(\varphi - \mathbf{w}\cdot\mathbf{n})\,N\,dS \ \ge\ 0 \quad \text{since } \varphi - \mathbf{n}\cdot\mathbf{w} \ge 0.
\end{aligned}$$

Thus

$$0 \ge \int_{\Gamma_t}(\mathbf{w}-\mathbf{u})\cdot\mathbf{t}\,dS - \int_\Omega [(\nabla\mathbf{w}-\nabla\mathbf{u}) : \sigma[\mathbf{u}] - (\mathbf{w}-\mathbf{u})\cdot\mathbf{f}]\,dx.$$

Since $\sigma[\mathbf{u}]$ is a symmetric tensor $(\sigma_{ij} = \sigma_{ji})$, $(\nabla\mathbf{u} : \sigma) = \sum_{i,j}\left(\partial u_i/\partial x_j\right)\sigma_{ij} = (\varepsilon[\mathbf{u}] : \sigma[\mathbf{u}])$; similarly $\nabla\mathbf{w} : \sigma[\mathbf{u}] = \varepsilon[\mathbf{w}] : \sigma[\mathbf{u}]$. In addition, if we change signs, we get

$$0 \le \int_\Omega [\varepsilon(\mathbf{w}-\mathbf{u}) : \sigma[\mathbf{u}] - (\mathbf{w}-\mathbf{u})\cdot\mathbf{f}]\,dx - \int_{\Gamma_t}(\mathbf{w}-\mathbf{u})\cdot\mathbf{t}\,dS$$
$$\text{for all } \mathbf{w} \in K.$$

If we write

$$a(\mathbf{u},\mathbf{v}) = \int_\Omega \varepsilon[\mathbf{v}] : \sigma[\mathbf{u}]\, dx,$$
$$\psi(\mathbf{v}) = \int_\Omega \mathbf{v}\cdot\mathbf{f}\, dx + \int_{\Gamma_t} \mathbf{v}\cdot\mathbf{t}\, dS,$$

then we get the VI

$$\mathbf{u}\in K \quad \& \quad 0 \le a(\mathbf{w}-\mathbf{u},\mathbf{u}) - \psi(\mathbf{w}-\mathbf{u}) \qquad \text{for all } \mathbf{w}\in K.$$

We can define $\mathcal{A}\colon H^1(\Omega)\to H^1(\Omega)' = H^{-1}(\Omega)$ by $\langle \mathcal{A}\mathbf{u},\mathbf{v}\rangle = a(\mathbf{u},\mathbf{v})$, and similarly $\psi \in H^1(\Omega)' = H^{-1}(\Omega)$. Note that the boundary integral term $\int_{\Gamma_t}\mathbf{v}\cdot\mathbf{t}\, dS$ is continuous on $H^1(\Omega)$, thanks to the *trace theorem* (Theorem A.9) for Sobolev spaces. The operator $\mathcal{A}$ is elliptic since

$$\begin{aligned}\langle \mathcal{A}\mathbf{u},\mathbf{u}\rangle &= \int_\Omega \varepsilon[\mathbf{u}] : \sigma[\mathbf{u}]\, dx \\ &\ge \int_\Omega \eta\, \|\varepsilon[\mathbf{u}]\|^2\, dx \;\ge\; c\, \|\mathbf{u}\|^2_{H^1}\end{aligned}$$

for a positive constant c, provided that the Lebesgue measure of Γ_d is positive, thanks to Korn's theorem [121]. This is sufficient to apply Theorem 2.22 to show existence, uniqueness, and continuous dependence of solutions.

One thing to note about the VI approach is that it completely removes N from consideration. Questions of the regularity of N do not concern or bother us with this formulation. However, if we *want* information about N, this is not the best approach.

Another way of representing Signorini's problem is as a constrained optimization problem: Minimize the total energy of the system subject to the constraint that the obstacle is not penetrated. That is, we minimize $E[\mathbf{u}]$ over $\mathbf{u}\in K$, where

$$E[\mathbf{u}] = \int_\Omega \frac{1}{2}\varepsilon[\mathbf{u}] : \sigma[\mathbf{u}]\, dx - \int_\Omega \mathbf{u}\cdot\mathbf{f}\, dx - \int_{\Gamma_t}\mathbf{u}\cdot\mathbf{t}\, dS.$$

Note that E is a convex function, continuous and coercive on $K\subset H^1(\Omega)$, and so it is weakly lower semicontinuous (as $H^1(\Omega)$ is a Hilbert space). Thus we can apply Theorem B.13 to show existence of solutions. This can be characterized in terms of subdifferentials if we use the characteristic function I_K for K. Since E is defined over the entirety of $H^1(\Omega)$, it follows that $\partial(E+I_K) = \partial E + \partial I_K = \partial E + N_K$. In other words, the solution must satisfy $0\in\partial(E+I_K)(\mathbf{u})$, which is equivalent to

$$0 = \operatorname{div}\sigma + \mathbf{f} - N\nu\,\mathbf{n} - \nu\,\mathbf{t}, \tag{2.70}$$

where ν is the surface measure on $\Gamma_t\cup\Gamma_c$, N is the normal contact force, and $\mathbf{n}$ is the outward normal unit vector, as usual. To see that (2.70) is correct we should look carefully at $N_K(\mathbf{u})$. Recall that

$$N_K(\mathbf{u}) = \left\{ \mathbf{y}\in H^{-1}(\Omega) \mid \langle \mathbf{y}, \mathbf{w}-\mathbf{u}\rangle \le 0 \text{ for all } \mathbf{w}\in K \right\}.$$

Now $\mathbf{w} \in K$ if and only if $\varphi - \mathbf{w}\cdot\mathbf{n} \geq 0$ on Γ_c. If we choose $\mathbf{w}\cdot\mathbf{n} = \mathbf{u}\cdot\mathbf{n}$ on Γ_c but $\mathbf{w}$ is otherwise arbitrary inside Ω, we see that $\mathbf{y} \in N_K(\mathbf{u})$ must be zero in the interior of Ω. That is, $\mathbf{y}$ is a distribution concentrated on the boundary $\mathbf{y} = \mathbf{v}\,\nu$, where ν is the surface measure of Γ_c. Furthermore, since there is no restriction on the tangential component of $\mathbf{w}$ on Γ_c, the tangential component of $\mathbf{y}$ or $\mathbf{v}$ must be zero. So we write $\mathbf{y} = -N\,\nu\,\mathbf{n}$. Then

$$N_K(\mathbf{u}) = \left\{ \mathbf{y} = -N\,\nu\,\mathbf{n} \mid \int_{\Gamma_c} -N\,\mathbf{n}\cdot(\mathbf{w}-\mathbf{u})\,dS \leq 0 \text{ for all } \mathbf{w} \in K \right\}.$$

If $\varphi(\mathbf{x}) - \mathbf{u}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) > 0$, then we can pick $\mathbf{w} \in K$, where $\mathbf{n}(\mathbf{x})\cdot(\mathbf{w}(\mathbf{x}) - \mathbf{u}(\mathbf{x}))$ can be either positive or negative, so we need to have $N(\mathbf{x}) = 0$. If $\varphi(\mathbf{x}) - \mathbf{u}(\mathbf{x})\cdot\mathbf{n}(\mathbf{x}) = 0$, then no matter how we pick $\mathbf{w} \in K$, we will always have $\mathbf{n}(\mathbf{x})\cdot(\mathbf{w}(\mathbf{x}) - \mathbf{u}(\mathbf{x})) \leq 0$. Then we must have $N(\mathbf{x}) \geq 0$. That is,

$$N_K(\mathbf{u}) = \{ \mathbf{y} = -N\,\nu\,\mathbf{n} \mid \\ 0 \leq N(\mathbf{x}) \perp \varphi(\mathbf{x}) - \mathbf{n}(\mathbf{x})\cdot\mathbf{u}(\mathbf{x}) \geq 0 \text{ for all } \mathbf{x} \in \Gamma_c \}.$$

This brings us back to the complementarity formulation.

This approach actually gives us some information about the regularity of N (something that the VI approach does not): from the Sobolev imbedding theorem (Theorem A.8) and duality we can show that $N \in H^{-1/2}(\Gamma_c)$.

Chapter 3

Formalisms

> There is no branch of mathematics, however abstract, which may not some day be applied to phenomena of the real world.
>
> *Nikolai Lobachevsky*
>
> But the Modern Utopia must be not static but kinetic...
>
> *H.G. Wells*

In this chapter we outline how we can provide a consistent and unified formalism for describing the examples of the previous chapter. The theory behind these formalisms is developed in the following chapter.

3.1 Differential variational inequalities

A differential variational inequality (DVI) is formally defined as the problem of finding a solution pair (u, z) of functions $x\colon [0,T] \to X$ and $z\colon [0,T] \to Z$, where X and Z are Banach spaces ($X = \mathbb{R}^n$ and $Z = \mathbb{R}^m$ for example), such that

$$\frac{dx}{dt}(t) = f(t, x(t), z(t)), \qquad x(t_0) = x_0, \tag{3.1}$$

$$z(t) \in K \qquad \text{for (almost) all } t, \tag{3.2}$$

$$0 \le \langle \widetilde{z} - z(t), F(t, x(t), z(t)) \rangle \qquad \text{for all } \widetilde{z} \in K \text{ and (almost) all } t. \tag{3.3}$$

The function $f\colon [0,T] \times X \times Z \to X$ defines the main dynamics of the system, but $z(t) \in Z$ is determined by the VI (3.1)–(3.3) in terms of $F\colon [0,T] \times X \times Z \to Z'$ and the closed convex set K. In due course we will need to impose additional conditions on f, F, and even K.

We usually interpret a differential equation "$dx/dt(t) = f(t, x(t), z(t))$" as holding if the right-hand side function $t \mapsto f(t, x(t), z(t))$ is an integrable function, the solution $x(\cdot)$ is an absolutely continuous function, and the equation hold for almost all t. In some situations we relax these requirements to handle situations where a weaker or more general solution makes sense.

Often we use an apparently weaker formulation: instead of requiring that (3.3) hold for (almost) all t, we use the integral formulation:

$$0 \le \int_0^T \langle \widetilde{z}(t) - z(t),\, F(t,x(t),z(t))\rangle \, dt \qquad \text{for all continuous } \widetilde{z}\colon [0,T] \to K. \tag{3.4}$$

In fact, it is not even necessary to require this for all *continuous* $\widetilde{z}$; we can simply require that (3.4) hold for all smooth (or $C^\infty[0,T]$) functions $\widetilde{z}(\cdot)$.

Lemma 3.1. *Suppose K is a closed convex set. If* (3.4) *holds for all $\widetilde{z} \in C^\infty[0,T]$, and*

$$t \mapsto (1 + \|z(t)\|_Z)(1 + \|F(t,x(t),z(t))\|_{Z'})$$

is an integrable function, then (3.3) *holds for almost all t.*

Proof. First we show that if (3.4) holds for all $\widetilde{z} \in C^\infty[0,T]$, then (3.4) holds for all bounded and integrable $\widetilde{z}$.

Suppose $\widetilde{z}$ is a bounded integrable function. We extend $\widetilde{z}$ outside $[0,T]$ by setting $\widetilde{z}(t) = z_0$ for some fixed $z_0 \in K$ if $t \notin [0,T]$. Now pick a $C^\infty(\mathbb{R})$ function $\psi : \mathbb{R} \to \mathbb{R}$ which is zero outside the interval $[-1,+1]$ and positive on $(-1,+1)$ and $\int_{-1}^{+1} \psi(s)\,ds = 1$. Let $\psi_\epsilon(s) = \psi(s/\epsilon)/\epsilon$ so that $\int_{-\epsilon}^{+\epsilon} \psi_\epsilon(s)\,ds = 1$ and ψ_ϵ is zero outside $[-\epsilon,+\epsilon]$. The convolution $\widetilde{z}_\epsilon(t) = (\psi_\epsilon * \widetilde{z})(t) = \int_{-\infty}^{+\infty} \psi_\epsilon(t-s)\widetilde{z}(s)\,ds$ is C^∞. Since K is convex, $\widetilde{z}_\epsilon(t) \in K$ for all t. To see this, suppose otherwise: $\widetilde{z}_\epsilon(t) \notin K$. Then by the separating hyperplane theorem there are $w \in X'$ and $\beta \in \mathbb{R}$ such that $\langle x, w\rangle + \beta \ge 0$ for all $x \in K$, but $\langle \widetilde{z}_\epsilon(t), w\rangle + \beta < 0$. On the other hand, for every τ, $\widetilde{z}(\tau) \in K$. So

$$\begin{aligned}
\langle \widetilde{z}_\epsilon(t), w\rangle + \beta &= \langle (\psi_\epsilon * \widetilde{z})(t), w\rangle + \beta \\
&= \int_{-\infty}^{+\infty} \psi_\epsilon(t-s)\,[\langle \widetilde{z}(s), w\rangle + \beta]\,ds \\
&\qquad \left(\text{since } \int \psi_\epsilon = 1\right) \\
&\ge 0
\end{aligned}$$

as $\psi_\epsilon \ge 0$ and $\langle \widetilde{z}(s), w\rangle + \beta \ge 0$ for all s. This contradicts the assumption that $\widetilde{z}_\epsilon(t) \notin K$; therefore $\widetilde{z}_\epsilon(t) \in K$ for all t.

For any integrable function ϕ,

$$\begin{aligned}
\int_{-\infty}^{+\infty} \phi(t) \cdot (\psi_\epsilon * \widetilde{z})(t)\,dt &= \int_{-\infty}^{+\infty} \phi(t) \int_{-\infty}^{+\infty} \psi_\epsilon(t-s)\widetilde{z}(s)\,ds\,dt \\
&= \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} \phi(t)\psi_\epsilon(t-s)\,dt \right) \widetilde{z}(s)\,ds \\
&= \int_{-\infty}^{+\infty} \left(\phi * \check{\psi}_\epsilon\right)(s)\widetilde{z}(s)\,ds,
\end{aligned}$$

where $\check{\psi}_\epsilon(s) = \psi_\epsilon(-s)$. Now $\phi * \check{\psi}_\epsilon \to \phi$ in $L^1(\mathbb{R})$ as $\epsilon \to 0$, and since $\widetilde{z}$ is bounded, $\left(\phi * \check{\psi}_\epsilon\right) \cdot \widetilde{z} \to \phi \cdot \widetilde{z}$ as $\epsilon \to 0$ in $L^1(\mathbb{R})$. Applying this argument to $\phi(t) = F(t,x(t),z(t))$ and $\phi(t) = \langle z(t), F(t,x(t),z(t))\rangle$ shows that if (3.4) holds for all $\widetilde{z}_\epsilon$, then it holds for $\widetilde{z}$.

Now we show that this implies that (3.3) holds for almost all t: for any measurable set $E \subset [0,T]$, $R \geq 0$, and $w \in K$ we can set

$$\widetilde{z}_R(t) = \mathrm{sat}_R\left(\chi_E(t)w + (1-\chi_E(t))z(t)\right),$$

where $\chi_E(t) = 1$ if $t \in E$ and zero otherwise, and $\mathrm{sat}_R(v)$ is the projection of v onto the intersection of K and the closed unit ball of radius R. Then $\widetilde{z}_R(t) \in K$ is bounded and measurable, so (3.4) holds for $\widetilde{z}_R$. Now $\widetilde{z}(t) = \mathrm{sat}_R(w)$ if $t \in E$ and $\widetilde{z}(t) = \mathrm{sat}_R(z(t))$ if $t \notin E$. Taking $R \to \infty$ we note that $\mathrm{sat}_R(z(t)) \to z(t)$. Then, taking limits as $R \to \infty$ and using the dominated convergence theorem, we see that (3.4) holds for $\widetilde{z}(t) := \chi_E(t)w + (1-\chi_E(t))z(t)$. Then $\widetilde{z}(t) - z(t) = \chi_E(t)(w - z(t))$, so

$$\int_E \langle w - z(t),\, F(t,x(t),z(t))\rangle\, dt \geq 0 \qquad \text{for all measurable } E \subset [0,T].$$

Thus the set $\{t \mid \langle w - z(t), F(t,x(t),z(t))\rangle < 0\}$ must be a null set, and so (3.3) must hold for almost all t and all $w \in K$. □

The ability to replace the pointwise condition (3.3) with the integral condition (3.4) is useful for existence proofs, as will be seen later.

If K is a cone, then (3.1)–(3.3) are equivalent to the corresponding *differential complementarity problem* (DCP): Given $x_0 \in \mathbb{R}^n$, $f\colon [0,T]\times\mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}^n$, $F\colon [0,T]\times\mathbb{R}^n\times\mathbb{R}^m \to \mathbb{R}^m$, and $K \subseteq \mathbb{R}^n$ a closed convex cone, find $x\colon [0,T]\to\mathbb{R}^n$ and $z\colon [0,T]\to\mathbb{R}^m$ such that

$$\frac{du}{dt}(t) = f(t,x(t),z(t)), \qquad x(t_0) = x_0, \tag{3.5}$$

$$K \ni z(t) \perp F(t,x(t),z(t)) \in K^* \qquad \text{for almost all } t. \tag{3.6}$$

Since CPs can be cast as VIs, DCPs form a subset of DVIs. Nonetheless, DCPs provide a useful subclass of DVIs.

As with DVIs there is an integral formulation of DCPs in which (3.6) is replaced by

$$\begin{aligned} z(t) &\in K \qquad \text{for almost all } t,\\ F(t,x(t),z(t)) &\in K^* \qquad \text{for almost all } t,\\ 0 &= \int_0^T \langle z(t), F(t,x(t),z(t))\rangle\, dt. \end{aligned}$$

The integral and pointwise formulations of DCPs are equivalent by Lemmas 3.1 and 2.11. By Lemma 2.11, the existence and uniqueness theory for VIs carries over to CPs, although often there is special structure that can be applied to certain CPs that is not apparent in the general formulation of VIs.

3.1.1 A discussion of meanings

As with all formalisms, sometimes we have to stop and ask "What do you *really* mean?" And as we are dealing with a general class of dynamical problems we might be forced to allow solutions that are not regular functions, not even locally integrable functions such as

Dirac-δ functions or more general distributions. So if we have a solution $z(t)$ of a problem which is a distribution (that is a functional on the space of C^∞ functions of compact support), what does it mean to say "$z(t) \in K$ for (almost) all t"? Since the pointwise values of distributions are in general meaningless, this does not make sense at face value. The only operations that make sense for distributions are integrals with smooth functions of bounded support: $\langle z, \phi \rangle = \int_{-\infty}^{+\infty} \phi(t) z(t)\, dt$. For closed convex K we can give an interpretation to "$z(t) \in K$ for (almost) all t" as meaning

$$\frac{\int_{-\infty}^{+\infty} \phi(t) z(t)\, dt}{\int_{-\infty}^{+\infty} \phi(t)\, dt} \in K \qquad \text{for all } 0 \le \phi \in C_0^\infty(\mathbb{R}), \quad \phi \not\equiv 0. \tag{3.7}$$

As a simple test case, consider $K = \mathbb{R}_+$; in other words, when is a distribution nonnegative? The answer is well known and can be found in [127], for example. A real-valued distribution $z(t)$ is nonnegative if and only if $\int_{-\infty}^{+\infty} \phi(t) z(t)\, dt \ge 0$ for all $\phi \ge 0$. As shown in [127], this is equivalent to $z(\cdot)$ being a nonnegative measure. The theory of measure differential inclusions (see Section 4.4.2) can be applied to refine these ideas.

Another case is $K = [0, 1]$. Then (3.7) means that for any nonnegative C^∞ function with compact support ϕ,

$$0 \le \int_{-\infty}^{+\infty} \phi(t) z(t)\, dt \le \int_{-\infty}^{+\infty} \phi(t)\, dt.$$

For a C^∞ function with compact support that can have either sign, $\left| \int_{-\infty}^{+\infty} \phi(t) z(t)\, dt \right| \le \int_{-\infty}^{+\infty} |\phi(t)|\, dt$. (This can be done by splitting any such test function $\phi = \phi_+ - \phi_-$ where $\phi_\pm$ themselves are test functions with $\max(\|\phi_+\|_\infty, \|\phi_-\|_\infty)$ and $\|\phi_+\|_1 + \|\phi_-\|_1$ arbitrarily close to $\|\phi\|_\infty$ and $\|\phi\|_1$, respectively.) Then, if ψ is another C^∞ function of compact support with common support interval $[a, b]$,

$$\begin{aligned} &|\langle z, \phi \rangle - \langle z, \psi \rangle| \\ &\quad \le \left| \int_{-\infty}^{+\infty} (\phi(t) - \psi(t))\, z(t)\, dt \right| \\ &\quad \le \int_{-\infty}^{+\infty} |\phi(t) - \psi(t)|\, dt \\ &\quad \le (b - a)\, \|\phi - \psi\|_\infty, \end{aligned}$$

and so $\phi \mapsto \langle z, \phi \rangle$ is continuous in the supremum norm. This means that we can extend this functional to continuous functions with support in $[a, b]$. Hence we have a measure on $[a, b]$. Then we can apply the Radon–Nikodym theorem to conclude that z can be represented by an integrable function (also denoted by $z(t)$). Since nonnegative C^∞ functions are dense in the space of nonnegative L^1 functions, for any nonnegative L^1 function ϕ, $0 \le \int_{-\infty}^{+\infty} \phi(t) z(t)\, dt \le \int_{-\infty}^{+\infty} \phi(t)\, dt$. In particular, taking $\phi = \chi_{[\sigma,\tau]}$, we get $0 \le \int_\sigma^\tau z(t)\, dt \le \tau - \sigma$ for all $\tau \ge \sigma$. Using the standard results of Lebesgue integration, this implies that for almost all t, $0 \le z(t) \le 1$. By redefining z on a null set, we can assure that $0 \le z(t) \le 1$ for all t. Thus the distributional definition coincides with the usual one.

Here we see that the weak distributional definition for *bounded* sets is equivalent to the ordinary notion. For unbounded sets we have to include the possibility of impulses or other measures which cannot be represented by integrable functions.

3.2 Notion of index

The notion of *index* of a DVI or DCP is very important for the theory that follows. The index is essentially the number of times that the equation $F(t,x,z)=0$ would need to be differentiated with respect to t in order to uniquely specify z in terms of t and x. So, for example, the DCP (which is also called a *linear complementarity system* (LCS))

$$\frac{dx}{dt} = Ax(t) + Bz(t), \tag{3.8}$$

$$w(t) = Cx(t) + Dz(t), \tag{3.9}$$

$$0 \le w(t) \perp z(t) \ge 0 \tag{3.10}$$

has index zero if D is a nonsingular matrix. Note that $F(t,x,z) = Cx + Dz$, so that if $F(t,x,z)=0$, we can put $z = -D^{-1}Cx$ without using any differentiations. On the other hand, if $D=0$, then $(d/dt)F(t,x,z) = (d/dt)(Cx) = C(dx/dt) = C(Ax+Bz) = 0$, so if CB is a nonsingular matrix, then we can write $z = -(CB)^{-1}Ax$. Since one differentiation was sufficient to write z in terms of t and x, this problem has index one.

Impact problems (without friction) for rigid bodies (1.2)–(1.4) with one contact can be put into the form, with $q(t), v(t) \in \mathbb{R}^n$,

$$M(q)\frac{dv}{dt} = k(q,v) - \nabla V(q) + n(q)N, \tag{3.11}$$

$$\frac{dq}{dt} = v, \tag{3.12}$$

$$0 \le \varphi(q) \perp N \ge 0. \tag{3.13}$$

Here we have a DCP with $F(t,q,v,N) = \varphi(q)$. Note that $n(q) = \nabla\varphi(q)$. Here N takes the role of z in the general formulation. Now $\varphi(q)=0$ does not give an equation for N, nor does $0 = (d/dt)\varphi(q) = \nabla\varphi(q)\cdot dq/dt = \nabla\varphi(q)\cdot v$. However, using the notation Hess $f(x)$ for the matrix of second derivatives $\left[\partial^2 f/\partial x_i\,\partial x_j(x)\right]$,

$$\begin{aligned}
0 &= (d/dt)^2\varphi(q) \\
&= \nabla\varphi(q)^T\frac{dv}{dt} + v^T \operatorname{Hess}\varphi(q)\,v \\
&= \nabla\varphi(q)^T M(q)^{-1}\left[k(q,v) - \nabla V(q) + n(q)N\right] + v^T\operatorname{Hess}\varphi(q)\,v \\
&= n(q)^T M(q)^{-1} n(q)N \\
&\qquad + n(q)^T M(q)^{-1}\left[k(q,v) - \nabla V(q)\right] + v^T\operatorname{Hess}\varphi(q)\,v
\end{aligned}$$

does give an equation for N in terms of q and v. Thus this problem has index two.

Coulomb friction (by itself) results in an index-one DVI: for the example of a brick on a ramp as illustrated in Figure 1.4, we have

$$\begin{gathered}
m\frac{dv}{dt} = mg\sin\theta - f, \\
f \in [-\mu N, +\mu N], \quad 0 \le (\widetilde{f} - f)\,v \quad \text{for all } f \in [-\mu N, +\mu N], \\
N = mg\cos\theta.
\end{gathered}$$

Here the friction force is represented by f. The function $F(t,v) = v$. Setting $F(t,v) = 0$ does not give an equation for f. However, $0 = (d/dt)F(t,v) = dv/dt = mg\sin\theta - f$ does give an equation for f. Thus the index for this problem is one.

Index-three problems do not commonly arise in applications, but it is not hard to make one up. Here is an example: find $x\colon [0,T] \to \mathbb{R}^3$ and $z\colon [0,T] \to \mathbb{R}$ satisfying

$$\begin{aligned} \frac{dx_1}{dt} &= x_2, \qquad x_1(0) = 0, \\ \frac{dx_2}{dt} &= x_3, \qquad x_2(0) = -1, \\ \frac{dx_3}{dt} &= z, \qquad x_3(0) = 0, \\ 0 \le x_1(t) &\perp z(t) \ge 0. \end{aligned}$$

Every "solution" has the form $z(t) = \delta'(t) + \alpha\,\delta(t)$, where $\delta(t)$ is the Dirac-δ function, with $\alpha \ge 0$. But $\delta'(t)$ is not a nonnegative distribution as $\int \delta'(t)\phi(t)\,dt = -\phi'(0)$ for any smooth function ϕ, which can be positive or negative even for nonnegative ϕ.

This can be extended to give index-m problems: $d^m x/dt^m = u$, $0 \le x(t) + q(t) \perp z(t) \ge 0$ for almost all t. Problems with index three or higher do not in general have solutions.

3.2.1 Solution behavior

In order to understand how solutions behave, or the different characteristics that solutions have, for different indexes, we consider a simple class of problems of this sort:

$$\frac{d^m x}{dt^m}(t) = z(t) - 1, \tag{3.14}$$

$$x(0) = 1, \quad x^{(j)}(0) = 0, \quad 1 \le j \le m-1,$$

$$0 \le x(t) \perp z(t) \ge 0 \qquad \text{for (almost) all } t. \tag{3.15}$$

This problem has index m.

3.2.2 Index-zero problems

We start with index-zero problems, as they are the simplest to understand. Consider the problem

$$\frac{dx}{dt}(t) = z(t) - 1, \qquad x(0) = 1, \tag{3.16}$$

$$0 \le x(t) + z(t) \perp z(t) \ge 0 \qquad \text{for (almost) all } t. \tag{3.17}$$

The scalar CP

$$0 \le x(t) + z(t) \perp z(t) \ge 0$$

can be solved for $z(t)$ directly. If $x(t) > 0$, then $x(t) + z(t) > 0$ as well (since $z(t) \ge 0$), and so $z(t) = 0$. If $x(t) < 0$, then since $x(t) + z(t) \ge 0$, we must have $z(t) > 0$. So then

$x(t) + z(t) = 0$ by complementarity, and $z(t) = -x(t) > 0$. The case $x(t) = 0$ has the solution $z(t) = 0$ by inspection. Since this is a strongly monotone CP for $z(t)$, there is only one solution.

Thus we can write $z(t)$ in terms of $x(t)$ directly: $z(t) = x(t)_- = \max(0, -x(t))$. We can substitute this into the differential equation for x (3.16):

$$\frac{dx}{dt} = x(t)_- - 1, \qquad x(0) = 1. \tag{3.18}$$

This can be solved in pieces: initially $x(t) > 0$, so we initially have $dx/dt = -1$, and $x(t) = 1 - t$. However, eventually $x(t)$ will reach zero and might then become negative. We reach $x(t) = 0$ at time $t^* = 1$; then we still have $dx/dt < 0$. So for t immediately after $t^* = 1$ we will have $dx/dt = x(t)_- - 1 = -x(t) - 1$, which has solutions $x(t) = -1 + ce^{-t}$. Substituting $x(1) = 0$ we can solve for the constant c: $-1 + ce^{-1} = 0$, so $c = e$. This gives $x(t) = -1 + e^{1-t}$ for $t \geq 1$.

The solution is continuous and smooth, except for the "kink" at $t = 1$. This can be explained by noting that the right-hand side of (3.18) is Lipschitz continuous (but not smooth) in $x(t)$. The standard theory of ordinary differential equations can be applied to (3.18), showing that solutions exist and are unique on any time interval. Note that since $z(t)$ depends in a Lipschitz way on $x(t)$, $z(t)$ is also a continuous (but not necessarily smooth) function of t.

3.2.3 Index-one problems

Index-one problems are more difficult:

$$\begin{aligned} \frac{dx}{dt} &= z(t) - 1, \qquad x(0) = 1, \\ 0 \leq x(t) \perp z(t) &\geq 0 \qquad \text{for almost all } t. \end{aligned}$$

We will assume that $z(\cdot)$ is an integrable function, rather than a general measure. This means that $x(\cdot)$ is absolutely continuous.

Now we require that $x(t) \geq 0$ for all t, which is a condition that must be imposed on the *initial value* $x(0)$. If $x(t) > 0$, then by the complementarity condition, $z(t) = 0$, so $dx/dt = -1$. Eventually we must reach $x(t^*) = 0$ (which happens at $t^* = 1$). But we cannot allow $x(t)$ to become negative. On the other hand, we cannot have $x(t) > 0$ for $t > t^*$ since that would mean there must be a time τ between t^* and t where $x(\tau) > 0$ and $dx/dt(\tau) > 0$, which is impossible. So we must have $x(t) = 0$ for all $t > t^*$. Does this allow us to have a solution? Yes it does, since if $x(t) = 0$ for all $t > t^*$, the complementarity condition allows *any* $z(t) \geq 0$. But for the only possible solution, $dx/dt(t) = 0$ for $t > t^*$. This gives $-1 + z(t) = 0$; that is, $z(t) = +1$ for all $t > t^*$. Note that $z(t)$ is not a continuous function of t. Instead we have a jump discontinuity in $z(t)$ at $t = t^*$.

We can try to write this DCP as a differential equation, but when $x(t) = 0$, *any* $z(t) \geq 0$ satisfies the complementarity condition. Substituting this into the right-hand side for the differential equation gives the *differential inclusion*

$$\frac{dx}{dt}(t) \in \Psi(x(t)) - 1, \qquad x(0) = 1,$$

where $\Psi(x) = \{0\}$ if $x > 0$, $\Psi(0) = \mathbb{R}_+$, and $\Psi(x) = \emptyset$ for $x < 0$. The theory of differential inclusions is quite extensive [19, 73, 103]; existence of solutions can be shown using the theory of *maximal monotone operators* [41].

The sign of the right-hand side of the differential equation and the sign of x in the complementarity condition are crucially important for existence of solutions. For example, if we had the problem

$$\begin{aligned} \frac{dx}{dt}(t) &= -z(t) - 1, \qquad x(0) = 1, \\ 0 \le x(t) \perp z(t) &\ge 0 \qquad \text{for almost all } t, \end{aligned}$$

then we would have $dx/dt(t) = -1$ for $x(t) > 0$. But when we reach $x(t^*) = 0$, we have $dx/dt(t^*) \le -1$ (in fact, $dx/dt(t) \le -1$ for *any* t), so for any $t > t^*$, $x(t) < 0$, which violates the complementarity condition. Thus solutions do not exist in general for this problem.

Similarly, for the problem

$$\begin{aligned} \frac{dx}{dt}(t) &= z(t) - 1, \qquad x(0) = 1, \\ 0 \le x(t) \perp -z(t) &\ge 0 \qquad \text{for almost all } t, \end{aligned}$$

solutions do not exist in general.

3.2.4 Index-two problems

For index-two problems, consider

$$\frac{d^2x}{dt^2}(t) = z(t) - 1, \qquad x(0) = 1, \quad \frac{dx}{dt}(0) = 0, \tag{3.19}$$

$$0 \le x(t) \perp z(t) \ge 0 \qquad \text{for almost all } t. \tag{3.20}$$

It turns out that we have to assume that $z(\cdot)$ can be a measure. Again, for $x(t) > 0$ we have $z(t) = 0$, so until $x(t^*) = 0$ we have $x(t) = 1 - t^2/2$. This means that we have $x(t^*) = 0$ for $t^* = \sqrt{2}$. Note that $dx/dt(t^{*-}) = (d/dt)\left(1 - t^2/2\right)\big|_{t=t^*} = -2t^* < 0$. Since we need $x(t) \ge 0$ for $t > t^*$, this means that the velocity dx/dt has to have a jump discontinuity at t^*. This, in turn, means that $z(t)$ must contain a Dirac-δ function: $z(t) = z^*\,\delta(t - t^*) + z_1(t)$, with $z_1(t)$ a "nicer" function, at least near $t = t^*$. The strength of the impulse z^* can be determined in part from this condition:

$$\begin{aligned} z^* &= \frac{dx}{dt}(t^{*+}) - \frac{dx}{dt}(t^{*-}) \\ &\ge 0 - \left(-2t^*\right) = 2t^* = 2\sqrt{2}. \end{aligned}$$

But this is clearly not sufficient to uniquely specify z^*. In fact, *any* $z^* \ge dx/dt(t^{*-})$ will give a solution: If $z^* > dx/dt(t^{*-})$, then we have the solution

$$x(t) = \left(z^* - dx/dt(t^{*-})\right)\left(t - t^*\right) - \left(t - t^*\right)^2/2$$

and $z(t) = z^*\,\delta(t - t^*)$ for $t^* \le t < t^* + \epsilon$, where $x(t^* + \epsilon) = 0$ again. If we choose $z^* = dx/dt(t^{*-})$, then we have the solution $x(t) = 0$ and $z(t) = z^*\,\delta(t - t^*) + 1$ for $t \ge t^*$.

Since problems of this kind arise in impact mechanics, there has been a strong need to find a way of resolving this nonuniqueness. The usual way in which this is done is to introduce a *coefficient of restitution* e. There are several variations of this idea, but the usual way in which it is formulated, following Newton, is that

$$\frac{dx}{dt}(t^{*+}) = -e\,\frac{dx}{dt}(t^{*-}). \tag{3.21}$$

In the context of general rigid-body dynamics this can be phrased as "the postimpact normal velocity is $-e$ times the preimpact normal velocity." In terms of the rigid-body equations (3.11)–(3.13) given above,

$$n(q(t^*))^T v(t^{*+}) = -e\, n(q(t^*))^T v(t^{*-}).$$

This is *Newton's law of restitution*. For problems with multiple contacts, each contact can have its own coefficient of restitution.

From the requirement that $dx/dt(t^{*+}) \geq 0$ it is clear that $e \geq 0$. Physical principles intervene to limit $e \leq 1$, as $e > 1$ violates the principle that energy must be conserved or dissipated as heat. Macroscopic energy cannot be created spontaneously. There is also the fact that if $e > 1$, $x(t^*) = 0$, $dx/dt(t^*) = 0$, it is still possible for the solution to become nonzero spontaneously. This results in a different kind of nonuniqueness.

To see how this is possible, consider the case where $0 < e < 1$ first. The first impact time is $t_1^* = \sqrt{2}$. Then $x(t_1^*) = 0$, $dx/dt(t_1^{*+}) = +2e\,t_1^*$. So, immediately after $t = t_1^*$, $x(t) = 2et_1^*(t - t_1^*) - (t - t_1^*)^2/2$. The second impact time is then $t_2^* = t_1^* + 4et_1^*$ and $dx/dt(t_2^{*+}) = -e\,dx/dt(t_2^{*-}) = +e\,dx/dt(t_1^{*+}) = 2e^2t_1^*$. Immediately after the second impact time $x(t) = 2e^2t_1^*(t - t_2^*) - (t - t_2^*)^2/2$. Then the third impact time $t_3^* = t_2^* + 4e^2t_1^*$. Continuing in this way we can show that the kth impact time is $t_k^* = t_1^*(1 + 4\sum_{j=1}^{k-1} e^j)$. For $0 < e < 1$ this sequence has a finite limit. Thus we have infinitely many bounces in a finite time. After all these bounces we have $t = \lim_{k\to\infty} t_k^* = t_\infty^*$. Then $x(t_\infty^*) = 0$ and $dx/dt(t_\infty^*)$, and we can continue the solution by setting $z(t) = 1$ and $x(t) = 0$ for $t \geq t_\infty^*$. This is an example of a *Zeno solution* where the set of strict inequalities changes infinitely often in finite time.

Now the differential equation $d^2x/dt^2 = z(t) - 1$ and the complementarity condition $0 \leq x(t) \perp z(t) \geq 0$ are both reversible conditions (that is, replacing t with $T - t$ for both x and z keeps both conditions true). However, reversing time for the restitution law (3.21) results in

$$\frac{dx}{dt}(t^{*+}) = -\frac{1}{e}\,\frac{dx}{dt}(t^{*-}). \tag{3.22}$$

Thus if we time-reverse the solution $x(t)$ and $z(t)$ from the previous paragraph to get $\check{x}(t) = x(T - t)$ and $\check{z}(t) = z(T - t)$, we get a solution of (3.19)–(3.20) with the restitution law (3.22). Note that for $T > t_\infty^*$, the time-reversed solution $\check{x}(\cdot)$ has the initial values $\check{x}(0) = 0$ and $d\check{x}/dt(0) = 0$. Instead of getting only the trivial solution ($\check{x}(t) = 0$ for all t) we also have a solution which spontaneously starts bouncing due to the coefficient of restitution $\check{e} = 1/e > 1$.

There are a number of practical difficulties with coefficients of restitution for real impacts. One is that the coefficient of restitution is far from being a constant for a pair of bodies. The orientation of contact is also very important (see, for example, [250]). Also, there are some theoretical questions as to the appropriateness of using Newton's law of

restitution. Alternatives include Poisson's law of restitution, which is based on splitting the impact interval into a compression phase and an expansion phase and requiring that the ratio of the integral of the normal contact force over the expansion phase be e times the integral of the normal contact force over the compression phase [11, 209]. Other alternatives include energy-based laws of restitution [252, 253].

3.2.5 Index three and higher

Index-three problems have serious questions regarding their existence. Consider, for example, the problem

$$\frac{d^3x}{dt^3}(t) = z(t) - 1, \qquad x(0) = 1, \quad \frac{dx}{dt}(0) = 0, \quad \frac{d^2x}{dt^2}(0) = 0,$$
$$0 \le x(t) \perp z(t) \ge 0 \qquad \text{for almost all } t.$$

Again, for $x(t) > 0$ we have $z(t) = 0$. So until $x(t) = 0$ we have the solution $x(t) = 1 - t^3/3$. The first impact time is at $t^* = 3^{1/3}$. Then we need to change dx/dt instantaneously. If $\eta = dx/dt(t^{*+}) - dx/dt(t^{*-})$, since $dx/dt(t^{*+}) \ge 0$, $\eta \ge 1$. Then

$$\frac{d^3x}{dt^3}(t) = \frac{d^2}{dt^2}\left(\frac{dx}{dt}(t)\right),$$

which is the second derivative of a function with a jump discontinuity at $t = t^*$. This means that $z(t)$ must contain the derivative of a Dirac-δ function at $t = t^*$: $z(t) = z^*\,\delta'(t - t^*) + z_1(t)$, where $z_1(t)$ is a more regular function at $t = t^*$. The trouble with this is that, according to the theory of distributions, a distribution ψ is nonnegative if for every smooth (that is, C^∞) function $\phi \ge 0$ with compact support, $\langle \psi, \phi \rangle \ge 0$. By this definition, δ' cannot be a nonnegative distribution [127], as $\langle \delta', \phi \rangle = -\phi'(0)$, which can be positive, negative, or zero.

Some theories circumvent this difficulty by restricting the class of functions to which they apply. In particular, the theory of linear complementarity systems (see (1.8)–(1.10) above) [124] does this by restricting attention to polygonal cones and functions that are *Bohl distributions*. Bohl distributions locally have the form

$$z(t) = \sum_{j=0}^{m} a_j\,\delta^{(j)}(t - t^*) + w^T e^{Ct} d, \qquad t^* \le t < t^* + \epsilon, \tag{3.23}$$

where C may be a matrix. A Bohl distribution is *initially nonnegative at* t^*, according to [124], if $[a_m, a_{m-1}, \ldots, a_1, a_0]$ is lexicographically positive (that is, the first nonzero in the list is positive), or if all $a_j = 0$ and there is an $\epsilon' > 0$ such that $w^T e^{Ct} d \ge 0$ for all $t^* \le t \le t^* + \epsilon'$.

The restriction to such a narrow class of functions has important implications for other aspects of the theory. Numerical methods that can compute solutions for problems of lower index generally fail for index-three or higher problems. The general principle that "limits of solutions are also solutions" tends to fail. Without this property, not only is it difficult to prove convergence for numerical methods, but the whole concept as a model comes into question. We do not expect any model of the world to be complete, but only an approximation. If an unmodeled disturbance can destroy a given solution, then the model is probably not useful and should be replaced.

3.3 Infinite-dimensional problems

Infinite-dimensional problems often behave differently from finite-dimensional problems; the existence and uniqueness theory are often also quite different, or they must at least deal with some new issues. One of the most important issues is the matter of whether the right-hand side is a bounded or unbounded operator. For partial differential equations, it is usually unbounded. Consider, for example, the heat equation on a bounded open domain $\Omega \subset \mathbb{R}^d$:

$$\frac{\partial u}{\partial t} = \nabla^2 u \qquad \text{in } \Omega, \tag{3.24}$$

$$u(t,x) = 0 \qquad \text{for } x \in \partial\Omega, \tag{3.25}$$

$$u(0,x) = u_0(x) \qquad \text{for } x \in \Omega. \tag{3.26}$$

This has solutions. But the reversed heat equation governed by the partial differential equation

$$\frac{\partial u}{\partial t} = -\nabla^2 u \qquad \text{in } \Omega \tag{3.27}$$

usually does not. The basic reason is that the operator ∇^2 is an unbounded operator. Looking more closely, the eigenvalues λ_k of $-\nabla^2$ go to $+\infty$ as $k \to \infty$. If the eigenfunctions are ϕ_k: $-\nabla^2\phi_k = \lambda_k\phi_k$ with $\phi_k(x) = 0$ for $x \in \partial\Omega$ and $g(x) = 0$ for all $x \in \partial\Omega$, then if we expand the initial conditions as $u_0(x) = \sum_{k=1}^{\infty} u_k^* \phi_k(x)$, then the solution of (3.24)–(3.26) is

$$u(t,x) = \sum_{k=1}^{\infty} u_k^* e^{-\lambda_k t} \phi_k(x).$$

Since $\lambda_k > 0$ for all (sufficiently large) k, $e^{-\lambda_k t} \to 0$ as $t \to +\infty$ for these k. The decaying exponentials indicate that the sum, if it is well defined for $t = 0$, should be well defined for any $t > 0$.

On the other hand, the reversed heat equation has the solution

$$u(t,x) = \sum_{k=1}^{\infty} u_k^* e^{+\lambda_k t} \phi_k(x).$$

Now we have exponential growth in the coefficients of ϕ_k as t increases. For the Laplacian operator $-\nabla^2$ for $\Omega \subset \mathbb{R}^d$, the eigenvalues are asymptotically $\lambda_k \sim \text{const} k^{2/d}$ as $k \to \infty$. Assume that the ϕ_k are orthonormal functions; that is, $\langle \phi_i, \phi_j \rangle_{L^2(\Omega)} = \int_\Omega \phi_i(x)\phi_j(x)\,dx = 0$ if $i \neq j$ and one if $i = j$. Then

$$\|u(t,\cdot)\|^2_{L^2(\Omega)} = \sum_{k=1}^{\infty} \left(u_k^*\right)^2 e^{+2\lambda_k t}$$

for the reversed heat equation. Unless u_k^* decay *very fast*, $\|u(t,\cdot)\|_{L^2(\Omega)} = +\infty$ even for very small $t > 0$. This means that solutions usually *do not exist* for the reversed heat equation.

On the other hand, we will also deal with hyperbolic problems like the wave equation

$$\frac{\partial^2 u}{\partial t^2} \nabla^2 u \qquad \text{in } \Omega, \tag{3.28}$$

$$u(t,x) = 0 \qquad \text{for } x \in \partial\Omega, \tag{3.29}$$

$$u(0,x) = u_0(x) \qquad \text{for } x \in \Omega, \tag{3.30}$$

$$\frac{\partial u}{\partial t}(0,x) = v_0(x) \qquad \text{for } x \in \Omega. \tag{3.31}$$

If we use the eigenfunction decomposition that we used for the heat equation, $u(t,x) = \sum_{k=1}^{\infty} u_k(t)\,\phi_k(x)$, then

$$\frac{d^2 u_k}{dt^2} = -\lambda_k u_k,$$
$$u_k(0) = (u_0)_k, \qquad \frac{du_k}{dt}(0) = (v_0)_k,$$

where $u_0(x) = \sum_{k=1}^{\infty} (u_0)_k\,\phi_k(x)$ and $v_0(x) = \sum_{k=1}^{\infty} (v_0)_k\,\phi_k(x)$. Solving these ordinary differential equations gives

$$u(t,x) = \sum_{k=1}^{\infty} \left[(u_0)_k \cos\left(\lambda_k^{1/2} t\right) + (v_0)_k\,\lambda_k^{-1/2} \sin\left(\lambda_k^{1/2} t\right) \right] \phi_k(x).$$

Instead of having rapid decay in the coefficients of the eigenfunctions, there is rapid oscillation. This makes the behavior of solutions of hyperbolic equations much harder to analyze or control.

3.3.1 Gelfand triples

In order to capture the behavior of these kinds of operators, we usually work in the framework of *Gelfand triples*. A Gelfand triple, also called an evolution triple or rigged Hilbert space, is a pair of Hilbert spaces X and H together with some inclusions

$$X \subseteq H = H' \subseteq X'. \tag{3.32}$$

Here the inclusion $X \subseteq H$ is actually a function $\mathrm{incl}\colon X \to H$ which is one-to-one (that is, $\mathrm{incl}(x) = \mathrm{incl}(z)$ implies $x = z$). Also, we identify H with its dual H'. In practice, this usually means that $H = L^2(A)$ for some domain $A \subset \mathbb{R}^d$ and we identify the function $f \in L^2(A)$ with the functional $g \mapsto \int_A f\,g$. The inclusion $H' \subseteq X'$ is the *adjoint* $\mathrm{incl}^*\colon H' \to X'$ of $\mathrm{incl}\colon X \to H$.

The inclusion function incl should be linear, continuous, one-to-one, and *dense*: the closure of the image $\overline{\mathrm{incl}(X)} = H$. The fact that $\overline{\mathrm{incl}(X)} = H$ is important, as it implies that incl^* is also one-to-one: $\mathrm{incl}^*(u_1) = \mathrm{incl}^*(u_2)$ implies that $\mathrm{incl}^*(u_1 - u_2) = 0$. To see this, note that

$$0 = \left\langle \mathrm{incl}^*(u_1 - u_2),\, v \right\rangle_{X' \times X} = \langle u_1 - u_2,\, \mathrm{incl}(v) \rangle_{H' \times H};$$

by taking limits $\text{incl}(v_k) \to w$ for any $w \in H$, we see that $\langle u_1 - u_2, w\rangle_{H\times H'} = 0$ for all $w \in H$. Thus $u_1 - u_2 = 0$; that is, $u_1 = u_2$. This means that incl^* is one-to-one. In finite dimensions, every vector subspace is closed, so $\text{incl}(X)$ is closed. This means that $\text{incl}(X) = \overline{\text{incl}(X)} = H$, so incl is then one-to-one and onto. This means that we can identify $X = H = H' = X' = \mathbb{R}^n$. But in infinite dimensions dense but not onto inclusions are common. For example, we can take X to be the Sobolev space $H^1(\Omega)$ for an open domain $\Omega \subseteq \mathbb{R}^n$ and $H = L^2(\Omega)$.

Often it is assumed that $\text{incl}\colon X \to H$ is *compact*; that is, if A is a bounded set in X, then $\overline{\text{incl}(A)}$ is compact. In finite dimensions all closed bounded sets are compact, but this is not true for infinite-dimensional Banach spaces. So, in finite dimensions, *any* linear function $X \to H$ is compact. In infinite dimensions some inclusions are compact (such as $H^1(\Omega) \to L^2(\Omega)$), but some are not (such as any identity map on an infinite-dimensional Banach space, or the inclusion $L^p(\Omega) \to L^q(\Omega)$ where $q \le p$). If incl is compact, then incl^* is also compact.

In a Gelfand triple we *do* identify the middle Hilbert space H with its dual H'. This means that we treat the operator $J_H\colon H \to H'$ defined by $\langle J_H(u), v\rangle_{H'\times H} = (u, v)_H$ as the identity operator on $H = H'$. If $H = L^2(\Omega)$, this means that we identify a function $f \in L^2(\Omega)$ with the functional on $L^2(\Omega)$

$$g \mapsto \int_\Omega f(x)\, g(x)\, dx.$$

This is reasonable for most applications involving partial differential equations.

The most crucial aspect of a Gelfand triple is that the duality pairing on X and X' is equivalent to the inner product on H:

$$(u, v)_H = \langle u, v\rangle_{X\times X'} \qquad \text{for all } u \in X \subseteq H,\ v \in H \subseteq X'. \tag{3.33}$$

In the finite-dimensional case, this just means that the duality pairing between $\mathbb{R}^n$ and $(\mathbb{R}^n)' \cong \mathbb{R}^n$ is the same as the inner product on $\mathbb{R}^n$: $(x, y) = x^T y$. In the case where $X = H^1(\Omega)$ and $H = L^2(\Omega)$, this just means that the duality pairing between $u \in H^1(\Omega)$ and $v \in H^{-1}(\Omega) = H^1(\Omega)'$ is given by[3]

$$\langle u, v\rangle_{H^1\times H^{-1}} = \int_\Omega u(x)\, v(x)\, dx,$$

although the integral may need to be understood in the sense of distributions. This integral can also be given a meaning via Plancherel's theorem for Fourier transforms: for $f, g\colon \mathbb{R}^d \to \mathbb{R}$ "sufficiently regular,"

$$\int_{\mathbb{R}^d} f(x)\, g(x)\, dx = (2\pi)^{-d}\, \text{Re} \int_{\mathbb{R}^d} \overline{\mathcal{F}f(\xi)}\, \mathcal{F}g(\xi)\, d\xi. \tag{3.34}$$

Note that here $\overline{(\cdot)}$ means the complex conjugate of $(\cdot)$.

[3]Many authors *define* $H^{-1}(\Omega)$ as $H_0^1(\Omega)'$ and leave the relationship between $H_0^1(\Omega)'$ and $H^1(\Omega)'$ unexamined; here we *define* $H^{-1}(\Omega) = H^1(\Omega)'$. In fact, $H_0^1(\Omega)'$ can be considered as a closed subspace of $H^1(\Omega)'$.

Plancherel's theorem implies that

$$\int_{\mathbb{R}^d} |f(x)|^2\,dx = (2\pi)^{-d} \int_{\mathbb{R}^d} |\mathcal{F}f(\xi)|^2\,d\xi,$$

so Fourier transforms can be used for inner products. In particular, for the H^1 inner product, since $\mathcal{F}[\nabla f](\xi) = i\xi\,\mathcal{F}f(\xi)$,

$$\begin{aligned}(f,g)_{H^1} &= \int_{\mathbb{R}^d} (f(x)\,g(x) + \nabla f(x)\cdot\nabla g(x))\,dx \\ &= (2\pi)^{-d}\,\mathrm{Re}\int_{\mathbb{R}^d} (1+\xi\cdot\xi)\,\overline{\mathcal{F}f(\xi)}\,\mathcal{F}g(\xi)\,d\xi \qquad \text{while} \\ (f,g)_{L^2} &= (2\pi)^{-d}\,\mathrm{Re}\int_{\mathbb{R}^d} \overline{\mathcal{F}f(\xi)}\,\mathcal{F}g(\xi)\,d\xi.\end{aligned}$$

The dual inner product for $H^{-1}(\mathbb{R}^n)$ is given by

$$(f,g)_{H^{-1}} = (2\pi)^{-d}\,\mathrm{Re}\int_{\mathbb{R}^d} (1+\xi\cdot\xi)^{-1}\,\overline{\mathcal{F}f(\xi)}\,\mathcal{F}g(\xi)\,d\xi.$$

Since $f \in H^1(\mathbb{R}^d)$ if and only if $\xi \mapsto (1+\xi\cdot\xi)^{1/2}\,\mathcal{F}f(\xi)$ is in $L^2(\mathbb{R}^d)$, if $\phi\colon H^1(\mathbb{R}^d) \to \mathbb{R}$ is a linear functional, then we can represent ϕ by

$$\phi(f) = \mathrm{Re}\int_{\mathbb{R}^d} \overline{q(\xi)}\,(1+\xi\cdot\xi)^{1/2}\,\mathcal{F}f(\xi)\,d\xi$$

with q a complex-valued function in $L^2(\mathbb{R}^d)$. If we write

$$q(\xi) = (2\pi)^{-d}\,(1+\xi\cdot\xi)^{-1/2}\,\mathcal{F}g(\xi),$$

then $g \in H^{-1}(\mathbb{R}^d)$ and

$$\begin{aligned}\phi(f) &= (2\pi)^{-d}\,\mathrm{Re}\int_{\mathbb{R}^d} \overline{\mathcal{F}g(\xi)}\,\mathcal{F}f(\xi)\,d\xi \\ &= \int_{\mathbb{R}^d} g(x)\,f(x)\,dx \qquad \text{for all } f \in H^1(\mathbb{R}^d).\end{aligned}$$

Thus we have a representation of $H^1(\mathbb{R}^d)'$ as $H^{-1}(\mathbb{R}^d)$ where the duality pairing between $H^1(\mathbb{R}^d)$ and $H^{-1}(\mathbb{R}^d)$ is (formally) equivalent to the usual inner product in $L^2(\mathbb{R}^d)$.

Note that the inclusion $I := \mathrm{incl}^* \circ J_H \circ \mathrm{incl}\colon X \to X'$ of $X \subseteq H = H' \subseteq X'$ is usually very different from the map $J_X\colon X \to X'$ given by $\langle J_X(x), z\rangle_{X'\times X} = (x,z)_X$. For example, if $X = H^1(\mathbb{R}^d)$ and $H = L^2(\mathbb{R}^d)$ as discussed above, then I really is the inclusion $f \mapsto f$ that gives $H^1(\mathbb{R}^d) \to H^{-1}(\mathbb{R}^d)$. On the other hand, $J_X(f) = g$, where $g(x) = \mathcal{F}^{-1}_{\xi\to x}\left[(1+\xi\cdot\xi)\,\mathcal{F}f(\xi)\right] = f(x) - \nabla^2 f(x)$; that is, $J_X = I - \nabla^2$. These are very different operators.

One property that is very important about I is that it is a strictly monotone linear function:

$$\begin{aligned}\langle I(z), z\rangle_{X'\times X} &= \left\langle \mathrm{incl}^*(J_H(\mathrm{incl}(z))), z\right\rangle_{X'\times X} \\ &= \langle J_H(\mathrm{incl}(z)), \mathrm{incl}(z)\rangle_{H'\times H} \\ &= (\mathrm{incl}(z), \mathrm{incl}(z))_H \\ &= \|\mathrm{incl}(z)\|_H^2 = \|z\|_H^2 > 0 \qquad \text{for } z \neq 0.\end{aligned}$$

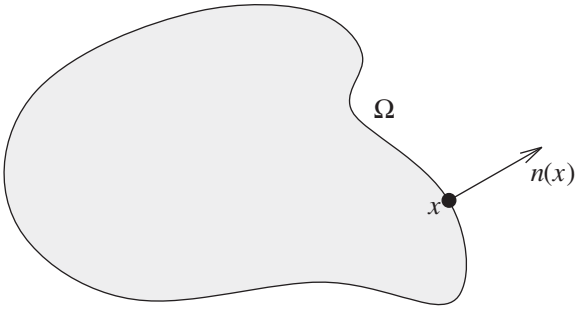


Figure 3.1: Normal direction vector to Ω at $x \in \partial\Omega$.

Note that I is not *strongly* monotone, since there is no guarantee that there is an $\alpha > 0$ where $\|z\|_H \geq \alpha \|z\|_X$ for all $z \in X$. In fact, for infinite-dimensional Banach spaces, if the inclusion $X \subseteq H$ is compact (that is, incl: $X \to H$ is a compact operator), then $\|z\|_H / \|z\|_X$ can be made as close to zero as desired.

This has some important consequences. For example, it would be tempting to apply the theory of maximal monotone operators (as developed in Section 4.2) to situations described by Gelfand triples. However, where we should not identify X with X' (so that J_X cannot be considered to be the identity operator), $A: X \to X'$ maximal monotone does *not* imply that $I + A$ is onto. This is discussed in more detail in Section 4.2.2. In particular, Lemma 4.7 gives a simple condition where a maximal monotone operator $\Phi: X \to \mathcal{P}(X')$ can define a maximal monotone operator $\Phi_H: H \to \mathcal{P}(H') = \mathcal{P}(H)$.

In order to obtain existence of solutions for problems that include, for example, the heat equation (3.24)–(3.26), we restrict our attention to linear *elliptic* operators $A: X \to X'$: there is an $\alpha > 0$ such that

$$\langle A(x), x \rangle_{X' \times X} \geq \alpha \|x\|_X^2 \qquad \text{for all } x \in X. \tag{3.35}$$

For most purposes involving dynamics, this condition can be weakened to what is described here as a *semielliptic* operator: there are $\alpha > 0$ and $\beta \in \mathbb{R}$ such that

$$\langle A(x), x \rangle_{X' \times X} \geq \alpha \|x\|_X^2 - \beta \|x\|_H^2 \qquad \text{for all } x \in X. \tag{3.36}$$

Then $A + \beta I$ is an elliptic operator. As an example, consider the negative Laplacian operator $A = -\nabla^2$ on the space $H^1(\Omega)$ *with Neumann boundary conditions*: $\partial u / \partial n(x) = g(x)$ on the boundary $\partial\Omega$ where $\partial u / \partial n(x)$ is the derivative of u at x in the direction of the outward pointing normal vector to Ω at x. This normal derivative is just $\partial u / \partial n(x) = n(x) \cdot \nabla u(x)$; see Figure 3.1.

With the Neumann boundary conditions, if $f(x) := 1$ for all $x \in \Omega$, then $\partial f / \partial n(x) = 0$ for any $x \in \partial\Omega$ and $-\nabla^2 f = 0$. Thus $\langle f, -\nabla^2 f \rangle_{H^1 \times H^{-1}} = 0$.

3.3.2 Interpolation spaces in Gelfand triples

Often we need to look for intermediate spaces, usually of functions with intermediate levels of smoothness or regularity, other than just X, H, and X' in a Gelfand triple. This can be done using the theory of interpolation spaces, which can be found in many sources such as

[1, 29, 262]. More accessible summaries can be found in, for example, [151, 260, 273]. The various methods for carrying this out are called the complex method, the real method (with either "J" or "K" versions), and the operator method. Here, the simpler operator method will suffice.

We take X to be a Hilbert space in a Gelfand triple $X \subset H = H' \subset X'$ with compact inclusions. Let $A\colon X \to X'$ be an elliptic self-adjoint operator (we can take $A = J_X$, the duality operator). Then $A^{-1} \circ I\colon X \to X' \to X$ is a compact self-adjoint operator: $\big(A^{-1} \circ I(x), y\big)_X = \big(x, A^{-1} \circ I(y)\big)_X$. Then by the spectral theorem for compact self-adjoint operators there is an infinite family of eigenfunctions ϕ_k and eigenvalues $\mu_k > 0$ with $\mu_1 \geq \mu_2 \geq \cdots > 0$ and $\lim_{k\to\infty} \mu_k = 0$, and $\overline{\operatorname{span}\{\phi_1, \phi_2, \ldots\}} = X$. The eigenfunctions can be taken to be orthogonal ($(\phi_i, \phi_j) = 0$ if $i \neq j$), not only in the inner product on X but also in H. We will scale the eigenfunctions ϕ_i so that $\|\phi_i\|_H = 1$. Then each ϕ_i is an eigenfunction of A with $A\phi_i = \lambda_i \phi_i$, with $\lambda_i = \mu_i^{-1}$. (Actually, it should be $A\phi_i = \lambda_i\, I(\phi_i)$, but we identify $\phi_i \in X$ with $I(\phi_i) \in X'$.)

For any $w \in \operatorname{span}\{\phi_1, \phi_2, \ldots\}$ we have the following norm for any given $\theta \in \mathbb{R}$:

$$\left\| \sum_{i=1}^{\infty} \alpha_i \phi_i \right\|_\theta = \left[\sum_{i=1}^{\infty} \lambda_i^\theta \alpha_i^2 \right]^{1/2}. \tag{3.37}$$

This norm is equivalent to the norm on X if $\theta = 1$, the norm on H if $\theta = 0$, and the norm on X' if $\theta = -1$. If we used $A = J_X$ in our construction, the norms would be equal. For each θ, we define the interpolation space X_θ to be the completion of $\operatorname{span}\{\phi_1, \phi_2, \ldots\}$ in the norm $\|\cdot\|_\theta$. If $\rho > \theta$, then $X_\rho \subset X_\theta$, and the imbedding is compact. The fractional power operators A^α defined by $A^\alpha \phi_i = \lambda_i^\alpha \phi_i$ are continuous operators $X_\theta \to X_{\theta-2\alpha}$ with continuous inverses ($(A^\alpha)^{-1} = A^{-\alpha}\colon X_\theta \to X_{\theta+2\alpha}$). Negative θ spaces correspond to dual spaces: $X_\theta \subset H = H' \subset (X_\theta)' = X_{-\theta}$ is a Gelfand triple.

3.4 Differentiation lemmas

Differentiation lemmas are technical results for functions satisfying complementarity conditions or VIs that connects the complementarity condition or VI with properties of the functions concerned. For example, if $a\colon [r,s] \to X$, $b\colon [r,s] \to X'$ with X a Banach space satisfy a generalized complementarity condition

$$K \ni a(t) \perp b(t) \in K^* \qquad \text{for all } t$$

and are smooth, then

$$\begin{aligned} 0 &= \left\langle a(t), \frac{db}{dt}(t) \right\rangle, \\ 0 &\geq \left\langle \frac{da}{dt}(t), \frac{db}{dt}(t) \right\rangle, \\ 0 &\leq \left\langle a(t), \frac{d^2b}{dt^2}(t) \right\rangle \end{aligned}$$

for all t. These results can be generalized to much less regular functions and also applied to VIs. Their use ranges from helping to show existence of solutions to showing energy conservation in certain impact problems.

3.4.1 Differentiation lemmas for CPs

Differentiation lemmas are easier to develop for complementarity conditions, so we start with these. The first result which we prove is a basic result which we can use for a large number of situations.

Lemma 3.2. *Let K be a closed convex cone in a Banach space X which has the Radon–Nikodym property (RNP). Suppose that $K \ni a(t) \perp b(t) \in K^*$ for almost all t and for $t = t_0$, and that b is differentiable at t_0. Then $\langle a(t_0), db/dt(t_0)\rangle = 0$. If a is also differentiable at t_0, then $\langle da/dt(t_0), db/dt(t_0)\rangle \leq 0$. If b is absolutely continuous and is twice differentiable at t_0, then $\left\langle a(t_0), d^2b/dt^2(t_0)\right\rangle \geq 0$.*

Proof. Since $K \ni a(t_0) \perp b(t_0) \in K^*$, for almost all $h > 0$,

$$\langle a(t_0), (b(t_0+h) - b(t_0))/h\rangle \geq 0.$$

Taking limits as $h \downarrow 0$ gives $\langle a(t_0), db/dt(t_0)\rangle \geq 0$. On the other hand, for almost all $h < 0$, $\langle a(t_0), (b(t_0+h) - b(t_0))/h\rangle \leq 0$. Again taking limits $h \uparrow 0$ gives $\langle a(t_0), db/dt(t_0)\rangle \leq 0$. Combining these two inequalities gives $\langle a(t_0), db/dt(t_0)\rangle = 0$.

For the second result, consider the finite difference approximation (for almost all $h \neq 0$)

$$\begin{aligned}
&\left\langle \frac{a(t_0+h) - a(t_0)}{h}, \frac{b(t_0+h) - b(t_0)}{h} \right\rangle \\
&\quad = \frac{1}{h^2} \langle a(t_0+h) - a(t_0), b(t_0+h) - b(t_0)\rangle \\
&\quad = \frac{1}{h^2} (\langle a(t_0+h), b(t_0+h)\rangle + \langle a(t_0), b(t_0)\rangle \\
&\qquad - \langle a(t_0+h), b(t_0)\rangle - \langle a(t_0), b(t_0+h)\rangle) \leq 0.
\end{aligned}$$

Taking limits as $h \to 0$ we see that $\langle da/dt(t_0), db/dt(t_0)\rangle \leq 0$.

For the third result, consider the finite difference approximation

$$\begin{aligned}
&\left\langle a(t_0), \frac{b(t_0+h) - 2b(t_0) + b(t_0-h)}{h^2} \right\rangle \\
&\quad = \frac{1}{h^2} (\langle a(t_0), b(t_0+h)\rangle - 2\langle a(t_0), b(t_0)\rangle + \langle a(t_0), b(t_0-h)\rangle) \geq 0.
\end{aligned}$$

Now we show that $\lim_{h\to 0} (b(t_0+h) - 2b(t_0) + b(t_0-h))/h^2 = b''(t_0)$. Since $b''(t_0)$ exists, $\lim_{r\to 0} \left(b'(t_0+r) - b'(t_0)\right)/r = b''(t_0)$. So for any $\epsilon > 0$ there is a $\delta > 0$ such that

$|r| < \delta$ implies $\left\| \left(b'(t_0+r) - b'(t_0)\right)/r - b''(t_0) \right\| < \epsilon$. So

$$\begin{aligned}
&\left\| \frac{b(t_0+h) - 2b(t_0) + b(t_0-h)}{h^2} - b''(t_0) \right\| \\
&\quad = \left\| \frac{1}{h} \int_0^h \frac{b'(t_0+r) - b'(t_0-r)}{r} \frac{r}{h}\, dr - b''(t_0) \right\| \\
&\quad \le \frac{1}{h} \int_0^h \left\| \frac{b'(t_0+r) - b'(t_0) + b'(t_0) - b'(t_0-r)}{r} - 2b''(t_0) \right\| \frac{r}{h}\, dr \\
&\quad \le \frac{1}{h} \int_0^h 2\epsilon \frac{r}{h}\, dr = \epsilon \qquad (\text{for } |h| < \delta).
\end{aligned}$$

So $\lim_{h\to 0} (b(t_0+h) - 2b(t_0) + b(t_0-h))/h^2 = b''(t_0)$; taking the limit, $\langle a(t_0), b''(t_0)\rangle \ge 0$, as desired. □

Note that it is not even necessary for $(a(t_0+h) - a(t_0))/h \to a'(t_0)$ *strongly* for the first result to hold; this could be weak convergence.

These pointwise results can be extended to prove integrated results. One example is as follows.

Lemma 3.3. *Suppose X is a Banach space with the RNP, K is a closed convex cone in X and*

$$K \ni a(t) \perp b(t) \in K^* \qquad \textit{for almost all } t.$$

If a is absolutely continuous $[0,T] \to X$ and $b \in L^1(0,T;X')$, then $\langle a'(t), b(t)\rangle = 0$ for almost all t. Also, if $a \in H^{1+\alpha}(0,T;X)$ and $b \in H^{-\alpha}(0,T;X)$, $\alpha > -1$, then $\langle a'(t), b(t)\rangle = 0$ in the sense of tempered distributions.

Proof. We start by supposing that $a\colon [0,T] \to X$ is absolutely continuous and that $b \in L^1(0,T;X')$. Then since X has the RNP, $a'(t) = \lim_{h\to 0}(a(t+h) - a(t))/h$ exists almost everywhere. Then, for almost every $t \in [0,T]$, the derivative $a'(t)$ exists *and* $K \ni a(t) \perp b(t) \in K^*$, so that by Lemma 3.2 $\langle a'(t), b(t)\rangle = 0$, as desired.

Now consider $a \in H^{1+\alpha}(0,T;X)$, $b \in H^{-\alpha}(0,T;X)$, and $\alpha > -1$. Note that we do not necessarily require that $\alpha \ge 0$. First we extend $a(t)$ to all real t: $a(t) = a(T)$ if $t > T$, and $a(t) = a(0)$ if $t < 0$; $b(t) = 0$ if $t > T$ or $t < 0$. We wish to show that $\int \phi(t) \langle a'(t), b(t)\rangle\, dt = 0$ for all tempered test functions $\phi \in \mathcal{S}(\mathbb{R})$ with support in $[0,T]$. To do this we first show that this is true for all *nonnegative* $\phi \in \mathcal{S}(\mathbb{R})$.

We can use Fourier transforms to represent

$$\int_0^T \langle (a(t+h) - a(t))/h, \phi(t) b(t)\rangle\, dt$$

via Plancherel's theorem: For $h > 0$,

$$\begin{aligned}
0 &\le \int_{-\infty}^{+\infty} \left\langle \frac{a(t+h)-a(t)}{h}, \phi(t)b(t) \right\rangle dt \\
&= \frac{1}{2\pi} \operatorname{Re} \int_{-\infty}^{+\infty} \frac{e^{i\omega h}-1}{h} \mathcal{F}a(\omega) \overline{\mathcal{F}(\phi b)(\omega)} d\omega \\
&= \frac{1}{2\pi} \operatorname{Re} \int_{-\infty}^{+\infty} \frac{e^{i\omega h}-1}{h} \left(1+\omega^2\right)^{\alpha} \mathcal{F}a(\omega) \left(1+\omega^2\right)^{-\alpha} \overline{\mathcal{F}(\phi b)(\omega)} d\omega.
\end{aligned}$$

Similarly, if $h < 0$, then $0 \ge \int_{-\infty}^{+\infty} \langle (a(t+h)-a(t))/h, \phi(t)b(t) \rangle \, dt$.

Clearly the integrand converges pointwise to

$$i\omega \mathcal{F}a(\omega) \overline{\mathcal{F}(\phi b)(\omega)} = \mathcal{F}\left[a'\right](\omega) \overline{\mathcal{F}(\phi b)(\omega)}.$$

To show that the integral converges we use the dominated convergence theorem. Note that $\omega \mapsto \left(1+\omega^2\right)^{(1+\alpha)/2} \mathcal{F}a(\omega)$ and $\omega \mapsto \left(1+\omega^2\right)^{-\alpha/2} \overline{\mathcal{F}(\phi b)(\omega)}$ are both in $L^2(0,T)$, and so $\omega \mapsto \left(1+\omega^2\right)^{1/2} |\mathcal{F}a(\omega)| \, |\mathcal{F}b(\omega)|$ is integrable.

As $\left|e^{i\omega h}-1\right| / |h| \le |\omega| \le \left(1+\omega^2\right)^{1/2}$, the integrands are bounded uniformly as $h \to 0$ by an integrable function. Thus we can apply the dominated convergence theorem and take $h \to 0$ to obtain

$$\begin{aligned}
0 &= \frac{1}{2\pi} \operatorname{Re} \int_{-\infty}^{+\infty} \left(1+\omega^2\right)^{\alpha} i\omega \mathcal{F}a(\omega) \left(1+\omega^2\right)^{-\alpha} \overline{\mathcal{F}(\phi b)(\omega)} d\omega \\
&= \frac{1}{2\pi} \operatorname{Re} \int_{-\infty}^{+\infty} \left(1+\omega^2\right)^{\alpha} \mathcal{F}\left[a'\right](\omega) \left(1+\omega^2\right)^{-\alpha} \overline{\mathcal{F}(\phi b)(\omega)} d\omega \\
&= \int_0^T \left\langle \frac{da}{dt}(t), \phi(t)b(t) \right\rangle dt.
\end{aligned}$$

Since this is true for any $0 \le \phi \in \mathcal{S}(\mathbb{R})$ with support in $[0,T]$, it can be shown that $\langle a'(t), b(t) \rangle = 0$ in the sense of tempered distributions, so $\langle a'(t), b(t) \rangle = 0$ for almost all t. □

This result can be extended to a having bounded variation $[0,T] \to X$ and b continuous function $[0,T] \to X'$. Note that we cannot allow b to be discontinuous at an atom of a', as the simple example from [243] below shows:

$$a(t) = \begin{cases} 0, & t \le 0, \\ 1, & t > 0, \end{cases}$$

$$b(t) = \begin{cases} 1, & t \le 0, \\ 0, & t > 0. \end{cases}$$

Then $a(t)b(t) = 0$ for all t. However, $a'(t) = \delta(t)$, the Dirac-δ function. Since b is a Borel function, it is bounded, and $b(0) = 1$, we have $\int a'(t)b(t)\,dt = 1$, not zero. Nevertheless, continuity of b is sufficient to obtain the corresponding result for a of bounded variation.

This result turns out to be important in proving energy conservation results for rigid-body dynamics with impact: only when the velocity is discontinuous and the contact force is impulsive can there be work done by a rigid obstacle.

Lemma 3.4. *If $a\colon [0,T] \to X$ has bounded variation, $b\colon [0,T] \to X'$ is continuous, K is a closed convex cone in X, and*

$$K \ni a(t) \perp b(t) \in K^*,$$

then $\langle a'(t), b(t)\rangle = 0$ in the sense of measures.

Proof. Let $\phi\colon [0,T] \to \mathbb{R}$ be continuous. Then we can write $\phi(t) = \phi_+(t) - \phi_-(t)$ with $\phi_+(t) = \max(\phi(t), 0)$ and $\phi_-(t) = \max(-\phi(t), 0)$. Both $\phi_\pm$ are continuous and nonnegative. Let us suppose that ϕ is also nonnegative. Then $\phi(t)b(t) \in K^*$ for all t and ϕb is continuous. We want to show that $\int_0^T \phi(t)\langle a'(t), b(t)\rangle\, dt = 0$. Using the Stieltjes integral, for any $\epsilon > 0$ we can pick a $\delta > 0$ so that for any partition $0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$ with $t_i \le \tau_i \le t_{i+1}$ and $t_{i+1} - t_i < \delta$ for all i

$$\left| \int_0^T \phi(t)\, \langle da(t), b(t)\rangle - \sum_{i=0}^{N-1} \langle a(t_{i+1}) - a(t_i), \phi(\tau_i) b(\tau_i)\rangle \right| < \epsilon.$$

Picking $\tau_i = t_i$, we note that $\langle a(t_{i+1}) - a(t_i), \phi(t_i)b(t_i)\rangle \ge 0$, but if we pick $\tau_i = t_{i+1}$, we get $\langle a(t_{i+1}) - a(t_i), \phi(t_{i+1})b(t_{i+1})\rangle \le 0$. Thus

$$\left| \sum_{i=0}^{N-1} \langle a(t_{i+1}) - a(t_i), \phi(\tau_i) b(\tau_i)\rangle \right| < \epsilon$$

for any choice of $\tau_i \in [t_i, t_{i+1}]$. Thus

$$\left| \int_0^T \phi(t)\, \langle da(t), b(t)\rangle \right| < 2\epsilon.$$

Since $\epsilon > 0$ is arbitrary, we get $\int_0^T \phi(t)\, \langle da(t), b(t)\rangle = 0$. As this is true for all nonnegative continuous ϕ, it must be true for all continuous ϕ. Thus $\langle da(t), b(t)\rangle = 0$ in the sense of measures. □

Lemmas 3.2, 3.3, and 3.4 are the most important and have application to questions of conservation of energy or energy balance for mechanical systems with contact.

Corresponding results for the two-derivative differentiation lemmas follow.

Lemma 3.5. *Suppose that X is a Banach space with the RNP, and that*

$$K \ni a(t) \perp b(t) \in K^* \qquad \text{for almost all } t.$$

If $a \in W^{1,p}(0,T;X)$ and $b \in W^{1,q}(0,T;X')$ with $1/p + 1/q = 1$, then $\langle a'(t), b'(t)\rangle \le 0$ for almost all t. Also, if $a \in H^{1+\alpha}(0,T;X)$ and $b \in H^{1-\alpha}(0,T;X)$ for some $|\alpha| < 1$, then $\langle a'(t), b'(t)\rangle \le 0$ for almost all t.

Proof. For the case where $a \in W^{1,p}(0,T;X)$ and $b \in W^{1,q}(0,T;X')$ with $1/p+1/q=1$, we note that both a and b are differentiable almost everywhere. By Lemma 3.2, $\langle a'(t), b'(t)\rangle \leq 0$ for any point of differentiability t of both a and b. Thus the result holds in this case.

For the second case, we use Fourier transforms applied to

$$0 \geq \left\langle \frac{a(t+h)-a(t)}{h}, \frac{b(t+h)-b(t)}{h} \right\rangle.$$

Let ϕ be a smooth nonnegative function that is zero outside $(\epsilon, T-\epsilon)$ for some $\epsilon > 0$. Then

$$K \ni \phi(t)a(t) \perp \phi(t)b(t) \in K^* \qquad \text{for almost all } t.$$

Extending the functions by zero outside of $(\epsilon, T-\epsilon)$,

$$\begin{aligned}
0 &\geq \int_{-\infty}^{+\infty} \left\langle \frac{\phi(t+h)a(t+h)-\phi(t)a(t)}{h}, \frac{\phi(t+h)b(t+h)-\phi(t)b(t)}{h} \right\rangle dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left\langle \left(\frac{e^{i\omega h}-1}{h}\right) \mathcal{F}[\phi a](\omega), \left(\frac{e^{i\omega h}-1}{h}\right) \mathcal{F}[\phi b](\omega) \right\rangle d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \overline{\left(\frac{e^{i\omega h}-1}{h}\right)} \left(\frac{e^{i\omega h}-1}{h}\right) \langle \mathcal{F}[\phi a](\omega), \mathcal{F}[\phi b](\omega)\rangle \, d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} 2\frac{1-\cos(\omega h)}{\omega^2 h^2} \omega^2 \, \langle \mathcal{F}[\phi a](\omega), \mathcal{F}[\phi b](\omega)\rangle \, d\omega \\
&= \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{1-\cos(\omega h)}{(\omega h)^2} \left\langle \mathcal{F}\left[(\phi a)'\right](\omega), \mathcal{F}\left[(\phi b)'\right](\omega) \right\rangle d\omega.
\end{aligned}$$

Now $\theta \mapsto (1-\cos\theta)/\theta^2$ is a bounded function since it is continuous (except possibly at zero) and $\lim_{\theta\to 0}(1-\cos\theta)/\theta^2 = 1/2$ and $\lim_{\theta\to\pm\infty}(1-\cos\theta)/\theta^2 = 0$. Also $\phi a \in H^{1+\alpha}(\mathbb{R};X)$ and $\phi b \in H^{1-\alpha}(\mathbb{R};X')$, so

$$\omega \mapsto \left\langle \mathcal{F}\left[(\phi a)'\right](\omega), \mathcal{F}\left[(\phi b)'\right](\omega) \right\rangle \qquad \text{is in } L^1(\mathbb{R}).$$

But $(1-\cos(\omega h))/(\omega h)^2 \to 1/2$ as $h \downarrow 0$ pointwise, so by the dominated convergence theorem, the limit as $h \downarrow 0$ of the integral above is

$$\begin{aligned}
0 &\geq \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left\langle \mathcal{F}\left[(\phi a)'\right](\omega), \mathcal{F}\left[(\phi b)'\right](\omega) \right\rangle d\omega \\
&= \int_{-\infty}^{+\infty} \left\langle (\phi a)'(t), (\phi b)'(t) \right\rangle dt.
\end{aligned}$$

But $(\phi a)' = \phi' a + \phi a'$, and so

$$\begin{aligned}
\left\langle (\phi a)'(t), (\phi b)'(t) \right\rangle &= \left\langle \phi'(t)a(t) + \phi(t)a'(t), \phi'(t)b(t) + \phi(t)b'(t) \right\rangle \\
&= \phi'(t)^2 \langle a(t), b(t)\rangle + \phi(t)\phi'(t)\left(\left\langle a'(t), b(t)\right\rangle + \left\langle a(t), b'(t)\right\rangle\right) \\
&\qquad + \phi(t)^2 \left\langle a'(t), b'(t)\right\rangle.
\end{aligned}$$

The first two terms are zero; the first term is because $a(t) \perp b(t)$, and the second is because of the first differentiation lemma. Thus

$$0 \geq \int_0^T \phi(t)^2 \left\langle a'(t), b'(t) \right\rangle dt.$$

Since this is true for all nonnegative continuous ϕ that are zero in a neighborhood of zero and T, we have $\left\langle a'(t), b'(t) \right\rangle \leq 0$ for almost all t. □

We cannot go beyond two derivatives with these differentiation lemmas; simple counterexamples are given in [243]. One of the nice features of these results for DCPs is that they are essentially symmetric in $a(t)$ and $b(t)$. However, this is not so for VIs, as we will see.

3.4.2 Differentiation lemmas for VIs

Some, but not all, of these differentiation lemmas can be transferred to VIs. Consider the following form of parametric VIs:

$$a(t) \in K \qquad \text{for all } t, \tag{3.38}$$

$$0 \leq \langle b(t) - a(t),\, f(t) \rangle \qquad \text{for any } b(t) \in K \text{ for all } t. \tag{3.39}$$

Then, if $a(t)$ and $f(t)$ form a solution of the parametric VI (3.38)–(3.39), we have the following inequalities:

$$0 \leq \left\langle \frac{a(t+h) - a(t)}{h},\, f(t) \right\rangle \qquad \text{for } h > 0,$$

$$0 \geq \left\langle \frac{a(t+h) - a(t)}{h},\, f(t) \right\rangle \qquad \text{for } h < 0.$$

Provided $a'(t)$ exists (even if $(a(t+h) - a(t))/h$ converges only weakly), then

$$0 = \left\langle a'(t),\, f(t) \right\rangle. \tag{3.40}$$

Following the methods of proof of Lemma 3.3, we can show that if $a \in W^{1,p}(0,T;X)$ and $f \in L^q(0,T;X')$ $(1/p + 1/q = 1)$, or $a \in H^{1+\alpha}(0,T;X)$ and $f \in H^{-\alpha}(0,T;X')$, solve the parametric VI (3.38)–(3.39), then $\left\langle a'(t),\, f(t) \right\rangle = 0$ for almost all t.

There are also differentiation lemmas with two derivatives for VIs: If $a(t)$ and $f(t)$ solve (3.38)–(3.39) and are both differentiable at t, we have

$$0 \geq \left\langle \frac{a(t+h) - a(t)}{h},\, \frac{f(t+h) - f(t)}{h} \right\rangle \qquad \text{for } h \neq 0.$$

Taking limits gives

$$0 \geq \left\langle a'(t),\, f'(t) \right\rangle. \tag{3.41}$$

On the other hand, if a is twice differentiable at t and $h \neq 0$,

$$0 \leq \left\langle \frac{a(t+h) - 2a(t) + a(t-h)}{h^2},\, f(t) \right\rangle$$

$$= \frac{1}{h^2} \left[\langle a(t+h) - a(t),\, f(t) \rangle + \langle a(t-h) - a(t),\, f(t) \rangle \right].$$

Taking limits gives

$$0 \leq \langle a''(t),\, f(t) \rangle. \tag{3.42}$$

The formal proofs that these inequalities apply for a and f having the appropriate regularity (for example, that $a \in W^{1,p}(a,b;X)$ and $f \in W^{1,q}(a,b;X')$ satisfying (3.38)–(3.39) implies (3.41) holds for almost all t) follow those of the previous section for CPs.

These differentiation lemmas for VIs are particularly useful for dealing with elastic impact problems, as these are often cast as VIs to avoid dealing with the normal contact forces on the boundary.

Chapter 4

Variations on the Theme

> In most sciences one generation tears down what another has built and what one has established another undoes. In mathematics alone each generation adds a new story to the old structure.
>
> *Hermann Hankel*

There are a number of variations on the theme of DVIs. These include, first and foremost, differential inclusions and maximal monotone differential inclusions. Also mentioned are variants on this approach: projected dynamical systems (PDSs), sweeping processes, parabolic variational inequalities (PVIs), and other approaches based more directly on complementarity, such as linear complementarity systems (LCSs) and convolution complementarity problems (CCPs).

4.1 Differential inclusions

Differential inclusions [19, 73, 228] are a generalization of differential equations of the form

$$\frac{dx}{dt}(t) \in \Phi(t, x(t)), \qquad x(t_0) = x_0. \tag{4.1}$$

The function $\Phi\colon [0,T] \times X \to \mathcal{P}(X)$ is a set-valued function. In full generality, differential inclusions do not have solutions, just as differential equations in their full generality do not have solutions. Usually for differential equations, we require that the right-hand side function $\Phi(t,x)$ be Lipschitz, or even just continuous, in x. Carathéodory's existence theorem (Theorem C.5) requires just continuity in x and integrability in t. Some of these restrictions are not necessary for differential inclusions. Filippov [102, 103] developed the theory of differential inclusions for dealing with discontinuous ordinary differential equations, such as arise with Coulomb friction. Consider the problem of a brick on a ramp considered in Section 1.2:

$$m\frac{dv}{dt} = mg\,\sin\theta - \mu\, mg\,\cos\theta\,\operatorname{sgn}(v), \qquad v(0) = v_0, \tag{4.2}$$

where we take $\operatorname{sgn}(v) = +1$ if $v > 0$, -1 if $v < 0$, and 0 if $v = 0$. As given, the differential equation (4.2) has no solution beyond the time when $v(t^*) = 0$, which will happen in finite time for $0 < \theta < \tan^{-1}\mu$. The reason is that if $v(t^*) = 0$, then to have $v(t) > 0$ for some time $t > t^*$, there must be a time τ between t^* and t where $dv/dt(\tau) > 0$ and $v(\tau) > 0$, which contradicts the formula in (4.2). Similarly, it is not possible to have $v(t) < 0$ for $t > t^*$. So our only possible solution is to have $v(t) = 0$ for $t > t^*$. However, this means that $dv/dt(t) = 0$ for $t > t^*$. This means that

$$0 = mg\,\sin\theta - \mu\, mg\,\cos\theta\,\operatorname{sgn}(0) = mg\,\sin\theta \neq 0,$$

a contradiction! Thus there are no solutions after $v(t^*) = 0$.

The remedy for this lack of existence is to extend the set of values of the right-hand side of (4.2) to allow $v(t) = 0$ for $t > t^*$. The idea is to replace

$$\frac{dx}{dt} = f(t, x(t)),$$

where $f(t, x)$ is a discontinuous function of x, with

$$\frac{dx}{dt} \in \Phi(t, x(t)),$$

where

$$\Phi(t, x) = \bigcap_{\delta > 0} \overline{\operatorname{co}}\, f(t, x + \delta B),$$

where B is the unit ball centered on the origin: $B = \{\, y \mid \|y\| < 1 \,\}$, and $\overline{\operatorname{co}}\, A = \overline{\operatorname{co} A}$ is the closure of the convex hull of $A \subseteq X$. The *convex hull* of a set A is

$$\operatorname{co} A = \left\{ \sum_{i=1}^{m} \theta_i x_i \mid \sum_{i=1}^{m} \theta_i = 1, \theta_i \geq 0 \ \text{ for all } i, x_i \in A \ \text{ for all } i \right\},$$

the set of convex combination of elements of A. This ensures that in the resulting differential inclusion $dx/dt \in \Phi(t, x)$, the value of $\Phi(t, x)$ always has closed convex values. To understand the need for the sets $\Phi(t, x)$ to be convex, we need to understand set-valued integrals.

4.1.1 Set-valued integrals

For a set-valued function $\Phi\colon [0, T] \to \mathcal{P}(\mathbb{R}^n)$ we would like to define the set-valued integrals for $0 \leq a \leq b \leq T$ by

$$\int_a^b \Phi(t)\,dt = \left\{ \int_a^b \phi(t)\,dt \mid \phi(t) \in \Phi(t) \text{ for all } t \right\}. \tag{4.3}$$

The functions $\phi\colon [0, T] \to \mathbb{R}^n$ satisfying $\phi(t) \in \Phi(t)$ for all (or almost all) t is called a selection of Φ. However, there are some obstacles to using (4.3) as a definition, as the integrals $\int_a^b \phi(t)\,dt$ require that ϕ be integrable functions, which implies that the ϕ are measurable functions in the sense of Lebesgue measure. This rather technical condition imposes some easy-to-satisfy restrictions on Φ.

A set-valued function $\Phi\colon [0,T] \to \mathcal{P}(\mathbb{R}^n)$ is said to be *measurable* if for each ball $x + r\,B$, the set $\{\, t \in [0,T] \mid \Phi(t) \cap (x + r\,B) \neq \emptyset \,\}$ is a Lebesgue measurable subset of $[0,T]$ (see Section 2.1.4). Equivalently, Φ is measurable if for every $x \in \mathbb{R}^n$, the distance function $t \mapsto d\,(x, \Phi(t)) := \inf_{y \in \Phi(t)} \|x - y\|$ is a measurable function $[0,T] \to \mathbb{R}$ [21]. Since functions $[0,T] \to \mathbb{R}$ that are *not* measurable are very difficult to construct (and the construction usually requires the axiom of choice), the requirement of measurability of Φ is a technical necessity but not a practical difficulty.

The measurability of $\Phi\colon [0,T] \to \mathcal{P}(\mathbb{R}^n)$ implies that Φ has measurable selections: $\phi\colon [0,T] \to \mathbb{R}^n$, where ϕ is both a measurable function and a selection of Φ. We introduce the notation $\|A\|$ for *sets* A:

$$\|A\| = \begin{cases} \sup_{a \in A} \|a\|, & A \neq \emptyset, \\ +\infty, & A = \emptyset. \end{cases}$$

The set-valued function Φ is *integrable* if it is measurable and the function $t \mapsto \|\Phi(t)\|$ is a Lebesgue integrable function $[0,T] \to \mathbb{R}$. Note that integrability implies that the Lebesgue measure of $\{\, t \in [0,T] \mid \Phi(t) = \emptyset \,\}$ is zero.

The importance of integrability of Φ is that this implies the existence of integrable selections ϕ, so that we then *can* use (4.3) as the definition of $\int_a^b \Phi(t)\,dt$.

The connection with convexity comes through Aumann's theorem [22].

Theorem 4.1. *For any integrable set-valued function* $\Phi\colon [0,T] \to \mathcal{P}(\mathbb{R}^n)$ *with closed values and* $0 \le a \le b \le T$,

$$\overline{\int_a^b \Phi(t)\,dt} = \int_a^b \overline{co}\,\Phi(t)\,dt.$$

Proof. ($\subseteq$) First note that $\Phi(t) \subseteq \overline{\mathrm{co}}\,\Phi(t)$ for all t, so $\int_a^b \Phi(t)\,dt \subseteq \int_a^b \overline{\mathrm{co}}\,\Phi(t)\,dt$. Now we show that $\int_a^b \overline{\mathrm{co}}\,\Phi(t)\,dt$ is a closed set. Suppose that we have $y_m = \int_a^b \varphi_m(t)\,dt$ with $\varphi_m(t) \in \overline{\mathrm{co}}\,\Phi(t)$ for almost all t and that $y_m \to y$ in $\mathbb{R}^n$.

The sequence φ_m satisfies $\|\varphi_m(t)\| \le \|\Phi(t)\|$ for almost all t, so the φ_m are equi-integrable. Then by the Dunford–Pettis theorem there is a weakly convergent subsequence which we denote by $\varphi_m \rightharpoonup \varphi$. Since the dual space $L^1(a,b;\mathbb{R}^n)'$ is $L^\infty(a,b;\mathbb{R}^n)$, this means that for any bounded integrable function $\xi\colon [a,b] \to \mathbb{R}^n$, $\int_a^b \langle \xi(t), \varphi_m \rangle(t)\,dt \to \int_a^b \langle \xi(t), \varphi(t) \rangle\,dt$ as $n \to \infty$ in the subsequence. Let $E = \{\, t \in [a,b] \mid \varphi(t) \notin \overline{\mathrm{co}}\,\Phi(t) \,\}$. Then, for each $t \in E$, by the separating hyperplane theorem, there are a $w(t) \in \mathbb{R}^n$ and a $\beta(t) \in \mathbb{R}$ such that $\langle w(t), \varphi(t) \rangle < \beta(t)$ and $\langle w(t), z \rangle \ge \beta(t)$ for all $z \in \overline{\mathrm{co}}\,\Phi(t)$. Without loss of generality, we can ensure that $\|w(t)\| = 1$ and $|\beta(t)| \le \sup_{z \in \Phi(t)} \|z\| + \|\varphi(t)\|$. By Filippov's lemma (Lemma 2.8), we can ensure that both $w(t)$ and $\beta(t)$ are measurable functions of $t \in E$. From the bounds on w and β we see that $\langle w(t), \varphi(t) \rangle - \beta(t)$ is an integrable function of t, and $\int_E (\langle w(t), \varphi_m(t) \rangle - \beta(t))\,dt \to \int_E \big(\langle w(t)^T, \varphi(t) \rangle - \beta(t)\big)\,dt$. Now $\varphi_m(t) \in \overline{\mathrm{co}}\,\Phi(t)$ for almost all $t \in E$, so $\int_E \big(\langle w(t)^T, \varphi_m(t) \rangle - \beta(t)\big)\,dt \ge 0$. On the other hand, $\langle w(t), \varphi(t) \rangle - \beta(t) < 0$ for almost all $t \in E$. The only way the two integrals can be equal is if the Lebesgue measure of E is zero. Thus $\varphi(t) \in \overline{\mathrm{co}}\,\Phi(t)$ for almost all $t \in [a,b]$, and so φ is a selection of $\overline{\mathrm{co}}\,\Phi(t)$, and $y = \int_a^b \varphi(t)\,dt$. Hence $\int_a^b \overline{\mathrm{co}}\,\Phi(t)\,dt$ is closed, and therefore $\overline{\int_a^b \Phi(t)\,dt} \subseteq \int_a^b \overline{\mathrm{co}}\,\Phi(t)\,dt$.

To show the reverse inclusion, suppose we have φ, a selection of $\overline{\text{co}}\,\Phi$. We wish to find a sequence φ_m of selections of Φ such that $\int_a^b \varphi_m(t)\,dt \to \int_a^b \varphi(t)\,dt$. For each integer $m \geq 1$, let $\epsilon_m = 1/m$. Then, since $\varphi(t) \in \overline{\text{co}}\,\Phi(t)$, we can find $n+1$ points $\varphi_m^{(k)}(t) \in \Phi(t)$, $k = 1, 2, \ldots, n+1$, and coefficients $\theta_k(t) \geq 0$ with $\sum_{k=1}^{n+1} \theta_k(t) = 1$ such that

$$\left\| \varphi(t) - \sum_{k=1}^{n+1} \theta_k(t)\varphi_m^{(k)}(t) \right\| < \epsilon_m.$$

By Filippov's lemma we can assume that $\varphi_m^{(k)}$ and θ_k are measurable for all k. We want to replace the sum $\sum_{k=1}^{n+1} \theta_k(t)\varphi_m^{(k)}(t)$ by a single function $\varphi_n(t) \in \Phi(t)$ for almost all t. We do this by rapidly switching between the $\varphi_m^{(k)}$ functions, allowing each $\varphi_m^{(k)}$ to be "on" for a fraction of the time that approaches θ_k in a suitable sense. Let $\omega > 0$ be given; this will be interpreted as the speed of cycling through the $\varphi_m^{(k)}$. For each interval $[r/\omega, (r+1)/\omega)$, $r \in \mathbb{Z}$, we set $\rho_{k,r,\omega} = \omega \int_{r/\omega}^{(r+1)/\omega} \theta_k(t)\,dt$, the average value of θ_k on this interval. Note that $\sum_{j=1}^{n+1} \rho_{j,r,\omega} = 1$ for all r. Let $F_{k,r,\omega} = [r + \sum_{j=1}^{k-1} \rho_{j,r,\omega}, r + \sum_{j=1}^{k} \rho_{j,r,\omega})/\omega$, and let $F_{k,\omega} = \bigcup_{r\in\mathbb{Z}} F_{k,r,\omega}$; these are disjoint sets whose union is $\mathbb{R}$. Now let $\varphi_{m,\omega}(t) = \varphi_m^{(k)}(t)$ whenever $t \in F_{k,\omega}$, or equivalently

$$\varphi_{m,\omega}(t) = \sum_{k=1}^{n+1} \chi_{F_{k,\omega}}(t)\varphi_m^{(k)}(t).$$

By Alaoglu's theorem, there is a weak* convergent subsequence $\chi_{F_{k,\omega}} \overset{*}{\rightharpoonup} \widehat{\theta}_k$ as $\omega \to \infty$ in $L^\infty(a,b;\mathbb{R})$. It is easily shown that for any interval $[c,d]$, the integrals $\int_c^d \chi_{F_{k,\omega}}(t)\,dt \to \int_c^d \theta_k(t)\,dt$ as $\omega \to \infty$, so the weak* limit of any convergent subsequence is θ_k. Since $\varphi_m^{(k)} \in L^1(a,b;\mathbb{R}^n)$ and $L^\infty(a,b;\mathbb{R}) = L^1(a,b;\mathbb{R})'$, $\int_a^b \chi_{F_{k,\omega}}(t)\varphi_m^{(k)}(t)\,dt \to \int_a^b \theta_k(t)\varphi_m^{(k)}(t)\,dt$ as $\omega \to \infty$.

Choose $\omega = \omega_m$ sufficiently large so that $\|\int_a^b \chi_{F_{k,\omega}}(t)\varphi_m^{(k)}(t)\,dt - \int_a^b \theta_k(t)\varphi_m^{(k)}(t)\,dt\| < \epsilon_m$. This gives the selection $\varphi_m = \varphi_{m,\omega_m}$ of Φ. Finally,

$$\begin{aligned}
&\left\| \int_a^b \varphi(t)\,dt - \int_a^b \varphi_m(t)\,dt \right\| \\
&\quad \leq \left\| \int_a^b \varphi(t)\,dt - \int_a^b \sum_{k=1}^{n+1} \theta_k(t)\varphi_m^{(k)}(t)\,dt \right\| \\
&\qquad + \sum_{k=1}^{n+1} \left\| \int_a^b \theta_k(t)\varphi_m^{(k)}(t)\,dt - \int_a^b \chi_{F_{k,\omega_m}}(t)\varphi_m^{(k)}(t)\,dt \right\| \\
&\quad \leq \epsilon_m(b-a) + (n+1)\epsilon_m.
\end{aligned}$$

Taking $m \to \infty$ we see that $\int_a^b \varphi_m(t)\,dt \to \int_a^b \varphi(t)\,dt$, so that $\int_a^b \overline{\text{co}}\,\Phi(t)\,dt \subseteq \overline{\int_a^b \Phi(t)\,dt}$.

With both inclusions shown, $\int_a^b \overline{\text{co}}\,\Phi(t)\,dt = \overline{\int_a^b \Phi(t)\,dt}$. $\square$

4.1.2 Integral and differential definitions of solutions to differential inclusions

This definition of integrals of set-valued functions enables us to give a pair of equivalent conditions for solutions of differential inclusions $dx/dt \in \Phi(t, x)$.

Definition 4.2. *A solution of the differential inclusion on* $[0, T]$*,*

$$\frac{dx}{dt} \in \Phi(t, x) \subseteq \mathbb{R}^n,$$

is an absolutely continuous function $x\colon [0, T] \to X$ *where* x *is an absolutely continuous function where* $dx/dt(t) \in \Phi(t, x(t))$ *for almost all* t*. Equivalently, for all* $t_1 < t_2$ *in* $[0, T]$*,*

$$x(t_2) \in x(t_1) + \int_{t_1}^{t_2} \Phi(\tau, x(\tau))\, d\tau. \tag{4.4}$$

Note that we cannot use

$$x(t) \in x(0) + \int_0^t \Phi(\tau, x(\tau))\, d\tau \qquad \text{for all } t \tag{4.5}$$

in the definition. The inclusion (4.5) is implied by (4.4), but the converse is false. For example, take $\Phi(t, x) = [-1, +1]$ so that solutions are simply functions $\mathbb{R} \to \mathbb{R}$ with Lipschitz constant one. But the functions satisfying (4.5) are all functions satisfying $|x(t) - x(0)| \le t$.

Definition 4.2 can be extended to differential inclusions in a Banach space X, provided X has the RNP. This property ensures that absolute continuity implies differentiability almost everywhere. Note that reflexive spaces such as Hilbert spaces automatically have the RNP.

4.1.3 Existence of solutions to differential inclusions

There is another condition that is needed for solutions to exist for a differential inclusion: a "no-blow-up" condition that prevents solutions going to infinity in finite time. One way of ensuring this for ordinary differential equations $dx/dt = f(t, x) \in \mathbb{R}^n$ is to impose the condition that $\langle x,\, f(t,x)\rangle \le C\left(1 + \|x\|^2\right)$ for all x. Then

$$\begin{aligned}\frac{d}{dt}\left(\|x(t)\|^2\right) &= 2\left\langle \frac{dx}{dt}(t), x(t)\right\rangle \\ &= 2\,\langle f(t, x(t)), x(t)\rangle \\ &\le 2C\left(1 + \|x(t)\|^2\right).\end{aligned}$$

Thus $\|x(t)\| \le \sqrt{e^{2Ct}\left(\|x(0)\|^2 + 1\right) - 1} \le 1 + e^{Ct}\,\|x(0)\|$ for $t \ge 0$.

The same idea can be used to prevent blowup for differential inclusions:

$$\langle y, x\rangle \le C\left(1 + \|x\|^2\right) \quad \text{for all } x \text{ and } y \in \Phi(t, x). \tag{4.6}$$

We will use a stronger condition for the following proof:

$$\|\Phi(t,x)\| \le C(t)(1+\|x\|) \qquad \text{for all } t, x \tag{4.7}$$

with $C(\cdot)$ an integrable function. This condition along with upper semicontinuity of Φ and the values $\Phi(t,x)$ having closed convex values is enough to show existence of solutions to the differential inclusion $dx/dt \in \Phi(t,x)$, $x(t_0) = x_0$, at least in finite dimensions.

Theorem 4.3. *Suppose that $\Phi\colon [0,T] \times \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n)$ has the following properties:*

1. *$\Phi(t,\cdot)$ is upper semicontinuous for all t;*
2. *$\|\Phi(t,x)\| \le C(t)(1+\|x\|)$ for all x, where C is an integrable function;*
3. *the values $\Phi(t,x)$ are closed convex sets for all t and x.*

Then the differential inclusion $dx/dt \in \Phi(t,x)$, $x(t_0) = x_0$ has a solution $x(t)$ for all $t \ge t_0$.

Proof. The method of proof is via a variation of Euler's method for numerical solution of ordinary differential equations and is close to the proof of convergence by Taubert [258]. Choose $h > 0$ as the "step size" of the method. We can then construct a sequence of functions $x^h(t)$ which we will show has a limit point; any limit point of this sequence turns out to be a solution of the differential inclusion. Note that assumptions 1 and 2 together imply that $\Phi(\cdot,x)$ is an integrable set-valued function for all x.

The construction of $x^h(t)$ proceeds inductively. We have $x(t_0) = x_0$ as the initial condition. Then we can pick an integrable selection φ_0 of $\Phi(\cdot, x(t_0))$ on $[t_0, t_1)$ where $t_1 = t_0 + h$. Let

$$x^h(t) = x(t_0) + \int_{t_0}^{t} \varphi_0(\tau)\,d\tau, \qquad t_0 \le t \le t_1.$$

This gives $x^h(t_1)$. In general, let $t_k = t_0 + kh$. Suppose that we have constructed $x^h(t)$ for $t_0 \le t \le t_k$. We want to then construct $x^h(t)$ for $t_k \le t \le t_{k+1}$ consistently with the previous construction. Pick an integrable selection φ_k of $t \mapsto \Phi(t, x^h(t_k))$ and set

$$x^h(t) = x^h(t_k) + \int_{t_k}^{t} \varphi_k(\tau)\,d\tau, \qquad t_k \le t \le t_{k+1}.$$

We can bound $\|x^h(t)\|$ since $dx^h/dt(t) \in \Phi(t, x^h(t_k))$ for $t_k \le t \le t_{k+1}$. Thus $\|dx^h/dt(t)\| \le C(t)\left(1 + \|x^h(t_k)\|\right)$ for $t_k \le t \le t_{k+1}$ with $C(t)$ an integrable function. Thus

$$\left\|x^h(t_{k+1})\right\| \le \left\|x^h(t_k)\right\| + \int_{t_k}^{t_{k+1}} C(\tau)\,d\tau \left(1 + \left\|x^h(t_k)\right\|\right).$$

If we let $\eta_k = 1 + \|x^h(t_k)\|$, then $\eta_{k+1} \le \left(1 + \int_{t_k}^{t_{k+1}} C(\tau)\,d\tau\right)\eta_k$ for all k. Thus

$$\begin{aligned}\eta_k &\le \eta_0 \prod_{j=0}^{k-1}\left(1 + \int_{t_j}^{t_{j+1}} C(\tau)\,d\tau\right)\\ &\le \eta_0 \exp\left(\sum_{j=0}^{k-1}\int_{t_j}^{t_{j+1}} C(\tau)\,d\tau\right)\\ &\le (1 + \|x(t_0)\|)\exp\left(\int_{t_0}^{t_k} C(\tau)\,d\tau\right).\end{aligned}$$

Since $\|x^h(t)\| \le \eta_{k+1}$ for all $t_k \le t \le t_{k+1}$, we have a bound on $x^h(t)$ for all $t \ge t_0$ that is independent of $h > 0$. On any finite interval $[t_0, t^*]$ we have

$$\|x^h(t)\| \le (1 + \|x(t_0)\|)\exp\left(\int_{t_0}^{t^*} C(\tau)\,d\tau\right).$$

Furthermore, the x^h are uniformly absolutely continuous on $[t_0, t^*]$ since

$$\|dx^h/dt(t)\| \le C(t)\left(1 + \max_{t_0\le t\le t^*}\|x^h(t)\|\right).$$

Thus the family of functions is equicontinuous on $[t_0, t^*]$. Then we can apply the theorem of Arzela and Ascoli to show that there is a uniformly convergent subsequence $x^h \to \widehat{x}$ as $h \downarrow 0$ in this subsequence.

We wish to show that $\widehat{x}$ is a solution of the differential inclusion. This amounts to showing that

$$\widehat{x}(t) \in \widehat{x}(s) + \int_s^t \Phi(\tau, \widehat{x}(\tau))\,d\tau \qquad \text{for all } t_0 \le s < t$$

and that $\widehat{x}(t_0) = x_0$. Let $\widetilde{x}^h(t) = x^h(t_k)$ for $t_k \le t < t_{k+1}$, so that

$$x^h(t) \in x^h(s) + \int_s^t \Phi(\tau, \widetilde{x}^h(\tau))\,d\tau \qquad \text{for all } s < t.$$

As $h \downarrow 0$, $\max_{t_0\le t\le t^*}\|\widetilde{x}^h(t) - x^h(t)\| \to 0$ in the subsequence since x^h are uniformly absolutely continuous and hence uniformly continuous. Also, $\widehat{x}$ is absolutely continuous since it is the uniform limit of a sequence of functions which are uniformly absolutely continuous as $h \downarrow 0$. Now dx^h/dt is a selection of $t \mapsto \Phi(t, \widetilde{x}^h(t))$. For every $\epsilon > 0$ and $t \in [t_0, t^*]$ there is a $\rho(t) > 0$ such that $\Phi(t, \widehat{x}(t) + \rho(t)B) \subseteq \Phi(t, \widehat{x}(t)) + \epsilon B$. By Filippov's lemma we can assume that δ is a measurable function. Thus, for almost all t, $\operatorname{dist}(dx^h/dt(t), \Phi(t, \widehat{x}(t)) + \epsilon B) \to 0$ as $h \downarrow 0$ in the subsequence. Since the functions dx^h/dt are equi-integrable, by the Dunford–Pettis theorem, there is a further subsequence in which dx^h/dt converges weakly in $L^1(t_0, t^*; \mathbb{R}^n)$ to a limit $\widehat{\varphi}$ as $h \downarrow 0$. Since $\Phi(t, x)$ is a closed convex set for all t and x, we can use the separating hyperplane theorem to show

that $\widehat{\varphi}(t) \in \Phi(t, \widehat{x}(t)) + \epsilon \overline{B}$ for almost all t. Also $\int_s^t dx^h/dt(\tau)\,d\tau \to \int_s^t \widehat{\varphi}(\tau)\,d\tau$ as $h \downarrow 0$ in the further subsequence. Thus

$$\widehat{x}(t) \in \widehat{x}(s) + \int_s^t \left[\Phi(\tau, \widehat{x}(\tau)) + \epsilon \overline{B}\right] d\tau.$$

Since this is true for every $\epsilon > 0$ and $0 \le s < t$, we see that $\widehat{x}$ solves the differential inclusion, as desired. □

Condition 2 of Theorem 4.3 can be replaced by the following two weaker conditions: for any $R > 0$, suppose that

- $\|\Phi(t, x)\| \le k_R(t)$ for all t and $\|x\| \le R$, where k_R is integrable, and
- $\langle y, x\rangle \le C(t)\left(1 + \|x\|^2\right)$ for all $y \in \Phi(t, x)$, where $C(\cdot)$ is integrable.

The existence of solutions can be shown for these weaker conditions by applying Theorem 4.3 to the modified differential inclusion

$$\frac{dx_R}{dt} \in \Phi(t, \Pi_{R\overline{B}}(x_R)), \qquad x(t_0) = x_0.$$

This is possible because $\|\Phi(t, x)\| \le k_R(t)$ for all t and $\|x\| \le R$ with k_R integrable. Because of the bound $\langle y, x\rangle \le C(t)\left(1 + \|x\|^2\right)$ for all $y \in \Phi(t, x)$, it can be shown that $\|x_R(t)\| \le 1 + \exp\left(\int_{t_0}^t C(\tau)\,d\tau\right)\|x_0\|$, independently of R. Thus, for sufficiently large $R > 0$, $\|x_R(t)\| < R$ for all $t \in [t_0, t^*]$, and so $dx_R/dt(t) \in \Phi(t, x_R(t))$, and we have solved the differential inclusion on $[t_0, t^*]$.

Proving uniqueness of solutions requires some extra conditions. Even for ordinary differential equations, continuity of the right-hand side does not guarantee uniqueness:

$$\frac{dx}{dt} = \sqrt{\max(x, 0)}, \qquad x(0) = 0 \tag{4.8}$$

has the solutions $x(t) = 0$ for all t, $x(t) = \frac{1}{4}t^2$ for all t, and for any $t^* > 0$, $x(t) = 0$ for $t \le t^*$ and $x(t) = \frac{1}{2}(t - t^*)^2$ for $t \ge t^*$. Failure of uniqueness is usually blamed on the lack of Lipschitz continuity of the function $x \mapsto \sqrt{\max(x, 0)}$ at zero. This effect, though, is the result of an extreme instability of the differential equation (4.8) at $x = 0$: $d/dx\left(\sqrt{\max(x, 0)}\right) = \frac{1}{2}x^{-1/2}$ for $x > 0$, so that as $x \downarrow 0$, $d/dx\left(\sqrt{\max(x, 0)}\right) \to +\infty$, indicating extreme instability. On the other hand, the differential equation $dx/dt = -\sqrt{\max(x, 0)}$ *does* have unique solutions: here $d/dx\left(-\sqrt{\max(x, 0)}\right) \to -\infty$ as $x \downarrow 0$. Again, Lipschitz continuity fails at $x = 0$, but the differential equation is extremely stable. It is so stable, in fact, that solutions reach $x(t) = 0$ in finite time, rather than just $\lim_{t\to\infty} x(t) = 0$.

This distinction between stability and instability is the key idea behind the idea of *one-sided Lipschitz continuity* for differential inclusions: there is a one-sided Lipschitz constant L such that

$$\begin{aligned} &\langle y_1 - y_2, x_1 - x_2\rangle \le L\,\|x_1 - x_2\|^2 \\ &\qquad \text{for all } y_1 \in \Phi(t, x_1),\ y_2 \in \Phi(t, x_2). \end{aligned} \tag{4.9}$$

Then, if we have two solutions $x_1(t)$ and $x_2(t)$ of $dx/dt \in \Phi(t, x)$,

$$\frac{d}{dt}\|x_1(t) - x_2(t)\|^2 = 2\left\langle \frac{dx_1}{dt}(t) - \frac{dx_2(t)}{dt}, x_1(t) - x_2(t)\right\rangle \\ \leq 2L\,\|x_1(t) - x_2(t)\|^2.$$

Then, by means of Gronwall's lemma,

$$\|x_1(t) - x_2(t)\| \leq e^{L(t-t_0)}\,\|x_1(t_0) - x_2(t_0)\|.$$

If $x_1(t_0) = x_2(t_0) = x_0$, then $x_1(t) = x_2(t)$ for all $t \geq t_0$. Note that one-sided Lipschitz continuity not only ensures uniqueness of solutions but also ensures Lipschitz continuity of the solutions in terms of the initial conditions x_0.

Examples of one-sided Lipschitz continuous set-valued functions include $s \mapsto -\mathrm{Sgn}(s)$ as defined in (2.18), with $L = 0$. Thus the differential inclusion for a brick on a ramp with Coulomb friction (4.2) has unique solutions.

Note that uniqueness and Lipschitz continuity of the solution in terms of the initial conditions still hold if the one-sided condition is modified to allow the one-sided Lipschitz constant to depend on t:

$$\langle y_1 - y_2, x_1 - x_2\rangle \leq L(t)\,\|x_1 - x_2\|^2 \\ \text{for all } y_1 \in \Phi(t, x_1),\ y_2 \in \Phi(t, x_2),$$

where $L(t)$ is an integrable function of t. The bound on $\|x_1(t) - x_2(t)\|$ should then read as

$$\|x_1(t) - x_2(t)\| \leq \exp\left(\int_{t_0}^{t} L(\tau)\,d\tau\right)\|x_1(t_0) - x_2(t_0)\|.$$

This idea of one-sided Lipschitz continuity can be extended to infinite dimensions and unbounded operators by the theory of maximal monotone operators.

4.1.4 Comparison with DVIs

DVIs can be represented as differential inclusions. A simple way is to define the set of solutions of the VI

$$u \in K \quad \& \quad 0 \leq \langle \widetilde{u} - u, F(t, x(t), u)\rangle \qquad \text{for all } \widetilde{u} \in K$$

as $\mathrm{sol}(F(t, x(t), \cdot), K)$. Then we can write the DVI (3.1)–(3.3) as the differential inclusion

$$\frac{dx}{dt}(t) \in f(t, x(t), \mathrm{sol}(F(t, x(t), \cdot), K)), \qquad x(t_0) = x_0. \tag{4.10}$$

Often the reverse is true, but this usually requires some structure on the differential inclusion. In practice this is almost always available, but it is also possible to construct "wild" differential inclusions which cannot be reformulated as DVIs.

For example, consider the class of piecewise smooth problems [235, 236]

$$\frac{dx}{dt}(t) = f_i(t, x(t)), \\ \text{where } h_i(x(t)) < \min_{j:j\neq i} h_j(x(t)). \tag{4.11}$$

We assume that all functions involved are smooth and that i ranges over $\{1, 2, \ldots, m\}$. Let the index set $I(x) := \{ i \mid h_i(x) = \min_j h_j(x) \} \neq \emptyset$. We also assume $\{ \nabla h_i(x) \mid i \in I(x) \}$ is an affinely independent set for all x; that is, the affine space generated by $\{ \nabla h_i(x) \mid i \in I(x) \}$ is *not* generated by any strict subset. This is equivalent to saying that

$$\{ \nabla h_i(x) - \nabla h_p(x) \mid i \in I(x) \backslash \{p\} \}$$

is a linearly independent set for some $p \in I(t, x)$. The choice of $p \in I(x)$ makes no difference to the condition: any p is as good as any other. Let $R_i = \{ x \mid h_i(x) < \min_{j:j \neq i} h_j(x) \}$; then $I(x) = \{ i = 1, 2, \ldots, m \mid x \in \overline{R_i} \}$.

We use the Filippov reformulation as a differential inclusion:

$$\frac{dx}{dt}(t) \in \text{co} \{ f_i(t, x(t)) \mid i \in I(x) \}. \tag{4.12}$$

This can be represented as a DVI: Let $\Sigma_m \subset \mathbb{R}^m$ be the unit simplex

$$\Sigma_m := \left\{ \theta \in \mathbb{R}^m \mid \theta_i \geq 0 \text{ for } i = 1, 2, \ldots, m, \sum_{i=1}^m \theta_i = 1 \right\}.$$

Then (4.12) can be represented as

$$\frac{dx}{dt}(t) = \sum_{i=1}^m \theta_i(t) f_i(t, x(t)), \tag{4.13}$$

$$\theta(t) \in \Sigma_m, \tag{4.14}$$

$$0 \leq \left(\widetilde{\theta} - \theta(t)\right)^T h(x(t)) \qquad \text{for all } \widetilde{\theta} \in \Sigma_m, \tag{4.15}$$

where $h(x) = [h_1(x), h_2(x), \ldots, h_m(x)]^T$.

To see that these are equivalent, note that (4.14)–(4.15) is equivalent to $\theta = \theta(t)$ minimizing $\sum_{i=1}^m \theta_i h_i(x(t))$ over $\theta \in \Sigma_m$: thus (4.14)–(4.15) implies $\theta_i(t) = 0$ if $h_i(x(t)) > \min_j h_j(x(t))$, and so

$$\sum_{i=1}^m \theta_i(t) f_i(t, x(t)) \in \text{co} \{ f_i(t, x(t)) \mid i \in I(x) \}.$$

Conversely, any element of co $\{ f_i(t, x(t)) \mid i \in I(x) \}$ can be represented as

$$\sum_{i=1}^m \theta_i(t) f_i(t, x(t)),$$

where $\theta(t)$ satisfies (4.14)–(4.15).

Other kinds of differential inclusions can arise naturally, such as in modeling Coulomb friction forces in three dimensions. If the normal contact force $N(t)$ is known (or can be computed in terms of the state vector $x(t)$), then Coulomb's friction law means that the friction force $F(t)$ must be in the direction $-v_{slip}(t)/\left\|v_{slip}(t)\right\|$ with magnitude $\mu N(t)$ if the slip velocity $v_{slip}(t) \neq 0$. If the slip velocity $v_{slip}(t) = 0$, then $F(t)$ can be *any* vector inside

a closed ball centered at the origin with radius $\mu N(t)$. Writing $x(t) = [q(t)^T, v(t)^T]^T$ for the state vector, with $q(t)$ the configuration of the system and $v(t)$ a representation of the velocities, and the slip velocity given by $v_{slip}(t) = H(q(t))v(t)$, the differential equation for a mechanical system has the form

$$M(q(t))\frac{dv}{dt}(t) \in k(q(t), v(t)) - \mu N(t) H(q(t))^T \partial\phi(H(q(t))v(t)),$$
$$\frac{dq}{dt}(t) = G(q(t))v(t),$$

where $M(q)$ is the mass matrix, $k(q,v)$ contains all noncontact forces, and $\phi(v_{slip}) = \|v_{slip}\|$. Since ϕ is convex, "$w \in \partial\phi(z)$" is equivalent to the VI of the second kind

$$0 \le \phi(\widetilde{z}) - \phi(z) - \langle w, \widetilde{z} - z\rangle \qquad \text{for all } \widetilde{z}.$$

This in turn is equivalent to a VI of the first kind, and so we can represent the differential inclusion for this friction problem as a DVI.

4.2 Maximal monotone operators and differential inclusions

The negative of a maximal monotone operator can be used as the right-hand side for differential inclusions. This not only extends Theorem 4.3 on the existence of solutions to differential inclusions, but the solutions have some important properties, such as uniqueness for given initial conditions.

Since the domain of a maximal monotone operator does not have to be the whole Hilbert space, this approach can be used to show the existence of solutions for differential equations involving unbounded operators such as the heat equation

$$\frac{\partial u}{\partial t} = \nabla^2 u. \tag{4.16}$$

This approach can also allow us to incorporate other conditions such as $u(t,x) \ge \varphi(x)$ for all t and $x \in \Omega$, although there are some complications in doing this.

In the particular case of the negative Laplacian operator $-\nabla^2$, we take $\phi\colon H^1(\Omega) \to \mathbb{R}$ given by $\phi(u) = \frac{1}{2}\int_\Omega \|\nabla u(x)\|^2\, dx$ as the convex lower semicontinuous and proper function: $-\nabla^2 u = \nabla\phi(u)$ and $\partial\phi(u) = \{\nabla\phi(u)\}$ in $H^1(\Omega)'$.

4.2.1 Theory of maximal monotone differential inclusions

We first consider the differential inclusion

$$0 \in \frac{du}{dt} + \Phi(u), \qquad u(0) = u_0, \tag{4.17}$$

where Φ is a maximal monotone operator $X \to \mathcal{P}(X')$ for a Hilbert space *where we can identify X and X'*; that is, we take $J_X = I$. In the context of a Gelfand triple $V \subset H = H' \subset V'$, we need $\Phi\colon H \to \mathcal{P}(H)$. (If X' is *not* identified with X, these results can be used to show the existence and properties of the differential inclusion $0 \in J_V(du/dt) + \Phi(u)$.)

Theorem 4.4. *If $\Phi\colon X \to \mathcal{P}(X') = \mathcal{P}(X)$ is a maximal monotone operator for a Hilbert space X identified with X', then solutions exist for the differential inclusion*

$$0 \in \frac{du}{dt} + \Phi(u), \qquad u(0) = u_0 \in \operatorname{dom}\Phi.$$

Furthermore, the solution satisfies the following properties:

1. *$u(t) \in \operatorname{dom}\Phi$ for all $t > 0$;*
2. *u is Lipschitz on $[0,\infty)$ with $\|du/dt(t)\| \le \left\|\Phi^0(u_0)\right\|$ for almost all $t > 0$;*
3. *$d^+u/dt(t) + \Phi^0(u(t)) = 0$ for all $t > 0$;*
4. *the map $t \mapsto \Phi^0(u(t))$ is continuous from the right, and $\left\|\Phi^0(u(t))\right\|$ is a nonincreasing function of t; hence also $\|du/dt(t)\|$ is a decreasing function of t;*
5. *if $u(t)$ and $\widehat{u}(t)$ are solutions of $0 \in du/dt + \Phi(u)$ with (possibly) different initial conditions, then $\|u(t) - \widehat{u}(t)\| \le \|u(0) - \widehat{u}(0)\|$.*

Before we give the proof of this, consider a differential inclusion with a maximal monotone operator:

$$0 \in \frac{du}{dt} + \operatorname{Sgn}(u) + c, \qquad u(0) = u_0.$$

If $u_0 > 0$, then for an initial interval $du/dt = -1 - c$. If $|c| < 1$, then $u(t) = u_0 - (1+c)t$. In a finite time $t^* = u_0/(1+c)$ we have $u(t^*) = 0$. But for $t > t^*$ with $|c| < 1$ the only possible solution is $u(t) = 0$. Clearly $du/dt(t)$ exists only for *almost all* t, but the one-sided derivative $d^+u/dt(t)$ does exist for *all* t.

Many of the properties mentioned in the theorem above are clearly on display: du/dt is a nonincreasing function of t; the minimum norm point $\operatorname{Sgn}(u(t)) + c$ is continuous from the right in t.

Proof. We show item 5 first: if u and $\widehat{u}$ are any two solutions of $0 \in du/dt + \Psi(u)$ for a monotone Ψ,

$$\frac{d}{dt}\|u(t) - \widehat{u}(t)\|^2 = 2\left\langle u(t) - \widehat{u}(t), \frac{du}{dt}(t) - \frac{d\widehat{u}}{dt}(t)\right\rangle \le 0,$$

so $\|u(t) - \widehat{u}(t)\| \le \|u(0) - \widehat{u}(0)\|$.

Now we show existence of solutions via Yosida approximations (2.57): let

$$0 = \frac{du_\lambda}{dt} + \Phi_\lambda(u_\lambda), \qquad u_\lambda(0) = u_0.$$

Since Φ_λ are Lipschitz, these differential equations have solutions $u_\lambda(t)$ for $t \ge 0$. We show that the u_λ form a Cauchy sequence as $\lambda \downarrow 0$. First we prove that $\|du_\lambda/dt(t)\|$ is a decreasing function of t. Recall that $\Phi_\lambda(u) = (u - R_\lambda(u))/\lambda$, and R_λ is Lipschitz of constant one. Then

$$\frac{du_\lambda}{dt}(t) = -\frac{1}{\lambda}(u_\lambda - R_\lambda(u_\lambda)).$$

Using the variation-of-parameters formula for $t > s$,

$$u_\lambda(t) = e^{-(t-s)/\lambda} u_\lambda(s) + \frac{1}{\lambda}\int_s^t e^{-(t-\tau)/\lambda} R_\lambda(u_\lambda(\tau))\,d\tau,$$

$$u_\lambda(t+h) = e^{-(t-s)/\lambda} u_\lambda(s+h) + \frac{1}{\lambda}\int_s^t e^{-(t-\tau)/\lambda} R_\lambda(u_\lambda(\tau+h))\,d\tau$$

since the equation is autonomous. Subtracting and taking norms give

$$\begin{aligned}\|u_\lambda(t+h) - u_\lambda(t)\| &\le e^{-(t-s)/\lambda}\,\|u_\lambda(s+h) - u_\lambda(s)\| \\ &\quad + \frac{1}{\lambda}\int_s^t e^{-(t-\tau)/\lambda}\,\|u_\lambda(\tau+h) - u_\lambda(\tau)\|\,d\tau\end{aligned}$$

since R_λ is Lipschitz with constant one. By means of a Gronwall lemma (e.g., Lemma C.3), $\|u_\lambda(t+h) - u_\lambda(t)\| \le \eta(t)$, where

$$\begin{aligned}\eta(t) &= e^{-(t-s)/\lambda}\eta(s) + \frac{1}{\lambda}\int_s^t e^{-(t-\tau)/\lambda}\eta(\tau)\,d\tau, \\ \eta(s) &= \|u_\lambda(s+h) - u_\lambda(s)\|.\end{aligned}$$

Simple calculations show that $\eta(t) = \eta(s)$ for all $t > s$. Thus we have $\|u_\lambda(t+h) - u_\lambda(t)\| \le \|u_\lambda(s+h) - u_\lambda(s)\|$ for all $t > 0$. Dividing by $h > 0$ and taking $h \to 0$ give $\|du_\lambda/dt(t)\| \le \|du_\lambda/dt(s)\|$ whenever $t > s \ge 0$ and both derivatives exist. But since u_λ is the solution of a Lipschitz differential equation, du_λ/dt exists everywhere. So $\|\Phi_\lambda(u_\lambda(t))\| = \|du_\lambda/dt(t)\| \le \|du_\lambda/dt(0)\| = \|\Phi_\lambda(u_0)\| \le \left\|\Phi^0(u_0)\right\|$.

Now to show that the sequence is a Cauchy sequence consider $\lambda, \mu > 0$ and the difference between the corresponding differential equations:

$$0 = \frac{du_\lambda}{dt} - \frac{du_\mu}{dt} + \Phi_\lambda(u_\lambda) - \Phi_\mu(u_\mu).$$

Taking inner products with $u_\lambda - u_\mu$ gives

$$0 = \frac{1}{2}\frac{d}{dt}\left\|u_\lambda - u_\mu\right\|^2 + \left\langle \Phi_\lambda(u_\lambda) - \Phi_\mu(u_\mu),\, u_\lambda - u_\mu\right\rangle.$$

Note that

$$\begin{aligned}u_\lambda - u_\mu &= \big(u_\lambda - R_\lambda(u_\lambda)\big) + \big(R_\lambda(u_\lambda) - R_\mu(u_\mu)\big) + \big(R_\mu(u_\mu) - u_\mu\big) \\ &= \lambda\Phi_\lambda(u_\lambda) + R_\lambda(u_\lambda) - R_\mu(u_\mu) - \mu\Phi_\mu(u_\mu).\end{aligned}$$

Since $\Phi_\lambda(u_\lambda) \in \Phi(R_\lambda(u_\lambda))$ and $\Phi_\mu(u_\mu) \in \Phi(R_\mu(u_\mu))$, by monotonicity of Φ we have $\left\langle \Phi_\lambda(u_\lambda) - \Phi_\mu(u_\mu),\, R_\lambda(u_\lambda) - R_\mu(u_\mu)\right\rangle \ge 0$. So

$$\begin{aligned}&\left\langle \Phi_\lambda(u_\lambda) - \Phi_\mu(u_\mu),\, u_\lambda - u_\mu\right\rangle \\ &\quad \ge \left\langle \Phi_\lambda(u_\lambda) - \Phi_\mu(u_\mu),\, \lambda\Phi_\lambda(u_\lambda) - \mu\Phi_\mu(u_\mu)\right\rangle \\ &\quad \ge \lambda\,\|\Phi_\lambda(u_\lambda)\|^2 + \mu\left\|\Phi_\mu(u_\mu)\right\|^2 - (\lambda+\mu)\,\|\Phi_\lambda(u_\lambda)\|\left\|\Phi_\mu(u_\mu)\right\| \\ &\quad \ge \lambda\,\|\Phi_\lambda(u_\lambda)\|^2 + \mu\left\|\Phi_\mu(u_\mu)\right\|^2 \\ &\qquad\quad - \frac{\lambda+\mu}{2}\left(\|\Phi_\lambda(u_\lambda)\|^2 + \left\|\Phi_\mu(u_\mu)\right\|^2\right) \\ &\quad = \frac{\lambda-\mu}{2}\,\|\Phi_\lambda(u_\lambda)\|^2 + \frac{\mu-\lambda}{2}\left\|\Phi_\mu(u_\mu)\right\|^2 \ge -\frac{|\lambda-\mu|}{2}\left\|\Phi^0(u_0)\right\|^2.\end{aligned}$$

Hence

$$\frac{d}{dt}\left\|u_\lambda - u_\mu\right\|^2 \le |\lambda-\mu|\left\|\Phi^0(u_0)\right\|^2,$$

and since $u_\lambda(0) = u_\mu(0) = u_0$, $\left\|u_\lambda(t) - u_\mu(t)\right\| \le (|\lambda-\mu|t)^{1/2}\left\|\Phi^0(u_0)\right\|$. So the family u_λ converges uniformly on bounded sets with limit $u(t)$, where

$$\|u_\lambda(t) - u(t)\| \le (\lambda t)^{1/2}\left\|\Phi^0(u_0)\right\|.$$

Also $R_\lambda(u_\lambda) \to u$ uniformly as $\lambda \downarrow 0$ since

$$\|R_\lambda(u_\lambda(t)) - u_\lambda(t)\| \le \lambda\,\|\Phi_\lambda(u_\lambda(t))\| \le \lambda\left\|\Phi^0(u_0)\right\|.$$

Since $\|\Phi_\lambda(u_\lambda(t))\| \le \left\|\Phi^0(u_0)\right\|$, taking $\lambda \downarrow 0$, $u(t) \in \operatorname{dom}\Phi$ (which shows item 1), as the graph of Φ is closed. Since $\Phi_\lambda(u_\lambda(t)) \in \Phi(R_\lambda(u_\lambda(t))$ and $R_\lambda(u_\lambda(t)) \to u(t)$, then the limit of $\Phi_\lambda(u_\lambda(t))$ as $\lambda \downarrow 0$ is in $\Phi(u(t))$. The bounds on $\|\Phi_\lambda(u_\lambda(t))\|$ then imply $\left\|\Phi^0(u(t))\right\| \le \left\|\Phi^0(u_0)\right\|$. Repeating this argument with initial condition $\widehat{u}(0) = u(t_0)$ shows that $\left\|\Phi^0(u(t))\right\| \le \left\|\Phi^0(u(t_0))\right\|$ for all $t > t_0$. Thus $t \mapsto \left\|\Phi^0(u(t))\right\|$ is a nonincreasing function (which is the second part of item 4).

Since the u_λ are uniformly Lipschitz as $\lambda \downarrow 0$ so that the limit is also Lipschitz, with constant $\left\|\Phi^0(u_0)\right\|$, then $\|du_\lambda/dt\|_{L^\infty} \le \left\|\Phi^0(u_0)\right\|$, which is item 2. Thus there is a weakly converging subsequence $du_\lambda/dt \rightharpoonup w$ in $L^2(0,T;X)$; by standard methods $w = du/dt$. Since $\lim_{\lambda\downarrow 0}\Phi_\lambda(u_\lambda(t)) = \Phi(u(t))^0 \in \Phi(u(t))$, by item 1 in Lemma 2.24 $0 \in du/dt(t) + \Phi(u(t))$ for almost all t.

To show continuity from the right of $t \mapsto \Phi^0(u(t))$, we show this holds at $t = 0$. So consider a sequence $t_n \downarrow 0$ as $n \to \infty$. Since $\Phi^0(u(t_n))$ is uniformly bounded, by Alaoglu's theorem there is a weakly convergent subsequence (also denoted by $\Phi^0(u(t_n))$) with weak limit y. By Mazur's lemma, $\|y\| \le \left\|\Phi^0(u(0))\right\|$. Because Φ has a strong $\times$ weak closed graph, $y \in \Phi(u(0))$ so $y = \Phi^0(u(0))$. Since this is the only possible limit, $\Phi^0(u(t_n)) \rightharpoonup \Phi^0(u(0))$ as $n \to \infty$. Since $\left\|\Phi^0(u(t_n))\right\| \to \left\|\Phi^0(u(0))\right\|$ as $n \to \infty$, combined with weak convergence, we have strong convergence: $\Phi^0(u(t_n)) \to \Phi^0(u(0))$ as $n \to \infty$. This shows the first part of item 4.

Finally, continuity from the right for $t \mapsto \Phi^0(u(t))$ and $0 = du/dt(t) + \Phi^0(u(t))$ for almost all t (from $0 \in du/dt(t) + \Phi(u(t))$ and $\|du/dt(t)\| \le \left\|\Phi^0(u(t))\right\|$ whenever $du/dt(t)$ exists) shows that $(u(t+h) - u(t))/h = -(1/h)\int_t^{t+h}\Phi^0(u(\tau))\,d\tau$, and taking $h \downarrow 0$ gives $0 = d^+u/dt(t) + \Phi^0(u(t))$ for *all* $t \ge 0$. This shows item 3. □

These results, though remarkable, are still somewhat restrictive. However, they are easily extended to handle combinations of maximal monotone and external functions. First we show that solutions exist for $f(t) \in du/dt + \Phi(u)$, $u(0) = u_0$ for $f \in L^1(0,T;X)$, and not just for $u_0 \in \operatorname{dom}\Phi$ but also for $u_0 \in \overline{\operatorname{dom}\Phi}$. We start with a lemma following Brézis' path.

Lemma 4.5. *Suppose that u and v are solutions of the following differential inclusions, where $\Phi\colon X \to \mathcal{P}(X)$ is maximal monotone with X' identified with X:*

$$f(t) \in \frac{du}{dt} + \Phi(u),$$
$$g(t) \in \frac{dv}{dt} + \Phi(v).$$

Assume that f and g are in $L^1(0,T;X)$. Then for $0 \le s \le t \le T$

$$\|u(t) - v(t)\| \le \|u(s) - v(s)\| + \int_s^t \|f(\tau) - g(\tau)\|\, d\tau.$$

Proof. We use the basic inequality for almost all t,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u(t)-v(t)\|^2 &= \left\langle \frac{du}{dt}(t) - \frac{dv}{dt}(t), u(t)-v(t)\right\rangle \\ &\le \langle f(t)-g(t), u(t)-v(t)\rangle \\ &\le \|f(t)-g(t)\|\,\|u(t)-v(t)\|,\end{aligned}$$

so that $(d/dt)\|u(t)-v(t)\| \le \|f(t)-g(t)\|$. Integrating gives our result. □

To show the existence of solutions for $f(t) \in du/dt + \Phi(u)$, $u(0) = u_0$, we take limits from functions f, where solutions do exist (by Theorem 4.4) to any given $f \in L^1(0,T;X)$, and use Lemma 4.5 to show that the convergence is uniform.

Corollary 4.6. *If $f \in L^1(0,T;X)$ and $\Phi\colon X \to \mathcal{P}(X)$ is maximal monotone with X' identified with X, then solutions exist and are unique for*

$$f(t) \in \frac{du}{dt} + \Phi(u), \qquad u(0) = u_0 \in \overline{\operatorname{dom}\Phi}.$$

Proof. First solutions exist for piecewise constant f and $u_0 \in \operatorname{dom}\Phi$. Consider a sequence $0 = t_0 < t_1 < \cdots$ and $f(t) = f_i$ for $t \in [t_i, t_{i+1})$. On each interval $[t_i, t_{i+1}]$ the differential inclusion becomes

$$0 \in \frac{du}{dt} + \Phi(u) - f_i, \qquad u(t_i) = u_i,$$

where $u_i = u(t_i)$ is obtained from the solution on the previous interval $[t_{i-1}, t_i]$. So, for a given $f \in L^1(0,T;X)$ and $u_0 \in \overline{\operatorname{dom}\Phi}$, consider a sequence f_k of piecewise constant functions that converges to f in $L^1(0,T;X)$ and $u_{0,k} \to u_0$ with $u_{0,k} \in \operatorname{dom}\Phi$. By Lemma 4.5, the solutions u_k to

$$f_k(t) \in \frac{du_k}{dt} + \Phi(u_k), \qquad u_k(0) = u_{k,0}$$

satisfy

$$\|u_k(t) - u_l(t)\| \le \|u_{k,0} - u_{l,0}\| + \int_0^t \|f_k(\tau) - f_l(\tau)\|\, d\tau,$$

so that u_k is a Cauchy sequence in space of continuous functions $C(0,T;X)$. Thus $u_k \to u$ uniformly as $k \to \infty$. Then, as the graph of Φ is closed, for every $s < t$,

$$u(t) - u(s) \in \int_s^t \big[f(\tau) - \Phi(u(\tau))\big]\, d\tau,$$

and so u is absolutely continuous and

$$f(t) \in \frac{du}{dt}(t) + \Phi(u(t))$$

for almost all t, and $u(0) = u_0$. Uniqueness (and continuous dependence on $u_0 \in \overline{\operatorname{dom}\Phi}$) follows from Lemma 4.5. □

The stronger properties about the one-sided derivatives $d^+u/dt + \Phi^0(u) = 0$ for the differential inclusion without f do not hold for *all* t, but the modification

$$0 = \frac{d^+u}{dt}(t) + \big(\Phi(u(t)) - f^+(t)\big)^0$$

holds for all t, where $f^+(t) = \lim_{h\downarrow 0}(1/h)\int_t^{t+h} f(\tau)\,d\tau$ exists. If X is a separable Hilbert space, then almost all t is a Lebesgue point for f.

We can extend the above theory to allow $f(t,u)$ as long as $f(t,u)$ is Lipschitz in u. Suppose that $f\colon [0,T]\times X \to X$ is a function where $u \mapsto f(t,u)$ is a Lipschitz function with constant $L(t)$ for all t with L an integrable function, $t \mapsto f(t,u)$ measurable, and $f(t,u)$ bounded by $\|f(t,u)\| \le k(t) + L(t)\|u\|$ with both k and L integrable. Then the differential inclusion

$$f(t,u(t)) \in \frac{du}{dt} + \Phi(u), \qquad u(0) = u_0 \in \operatorname{dom}\Phi \tag{4.18}$$

has unique solutions for maximal monotone Φ. To see this, we consider a Picard-type iteration: given $u_k \in C(0,T;X)$, let u_{k+1} be the solution of

$$f(t,u_k(t)) \in \frac{du_{k+1}}{dt} + \Phi(u_{k+1}), \qquad u_{k+1}(0) = u_0.$$

This can be thought of in terms of the operator $\mathcal{G}\colon C(0,T;X) \to C(0,T;X)$ given by $\mathcal{G}v$, which is the solution of

$$f(t,v(t)) \in \frac{du}{dt} + \Phi(u), \qquad u(0) = u_0.$$

Now $\mathcal{G}$ is a Lipschitz operator with Lipschitz constant $\int_0^T L(\tau)\,d\tau$ since $\|\mathcal{G}v - \mathcal{G}w\|_{C(0,T;X)}$ $= \max_{0\le t\le T}\|\mathcal{G}v(t) - \mathcal{G}w(t)\|$, and by Lemma 4.5,

$$\begin{aligned}\|\mathcal{G}v(t) - \mathcal{G}w(t)\| &\le \int_0^t \|f(\tau,v(\tau)) - f(\tau,w(\tau))\|\,d\tau\\ &\le \int_0^t L(\tau)\,\|v(\tau) - w(\tau)\|\,d\tau\\ &\le \left(\int_0^t L(\tau)\,d\tau\right)\|v - w\|_{C(0,T;X)}.\end{aligned}$$

Thus if $T > 0$ is chosen sufficiently small so that $\int_0^T L(\tau)\,d\tau < 1$, $\mathcal{G}$ is a contraction map, and so by the contraction mapping theorem there is a unique fixed point $u \in C(0,T;X)$, which solves

$$f(t,u(t)) \in \frac{du}{dt} + \Phi(u), \qquad u(0) = u_0.$$

Once a solution is obtained on $[0,T]$, the above result can be used to show the existence of a unique solution on $[T,2T]$, and then on $[2T,3T]$, etc. Thus a unique solution exists for this variant for all $t \geq 0$.

4.2.2 Maximal monotone operators and Gelfand triples

Often we have a situation in which we have a Gelfand triple of Hilbert spaces $X \subseteq H = H' \subseteq X'$ with a maximal monotone operator $\Phi\colon X \to \mathcal{P}(X')$. To apply the above theory we need a maximal monotone operator $\Psi\colon H \to H = H'$. It is tempting to simply define

$$\Phi_H(u) = \begin{cases} \Phi(u) \cap H & \text{if } u \in X, \\ \emptyset & \text{if } u \notin X. \end{cases} \tag{4.19}$$

This might or might not be a maximal monotone operator.

As a simple example, consider the operator $\Phi\colon H^1(-1,+1) \to \mathcal{P}(H^{-1}(-1,+1))$ given by $\Phi(f) = \{f(0)\delta\}$, where δ is the Dirac-δ function. This is well defined, since by the Sobolev imbedding theorem every function in $H^1(-1,+1)$ is continuous. It is maximal monotone, as it is the gradient of the smooth convex function $\phi\colon H^1(-1,+1) \to \mathbb{R}$ given by $\phi(f) = \frac{1}{2}f(0)^2$. However, if we take $H = L^2(-1,+1)$, then, as Dirac-δ functions do not belong to H, $\Phi(f) \cap H$ is empty unless $f(0) = 0$. Thus $\Phi(f) = \{0\}$ if $f \in H^1(-1,+1)$ and $f(0) = 0$, and $\Phi(f) = \emptyset$ otherwise. This means that Φ is not maximal monotone: it can be strictly extended to form the zero function on H.

However, there are easily checked cases in which Φ_H is also maximal monotone.

Lemma 4.7. *Suppose that $\Phi\colon X \to \mathcal{P}(X')$ is a maximal monotone operator in a Gelfand triple of Hilbert spaces $X \subseteq H = H' \subseteq X'$, and that there is an $\alpha > 0$ such that $\Phi - \alpha J_X$ is monotone, where $J_X\colon X \to X'$ is the usual duality operator. Then Φ_H is maximal monotone $H \to \mathcal{P}(H)$.*

First we need an extra lemma.

Lemma 4.8. *If $\Psi\colon X \to \mathcal{P}(X')$ is maximal monotone and strongly monotone in that there is an $\eta > 0$ such that if $\zeta \in \Psi(z)$, $\xi \in \Psi(x)$, then*

$$\langle \zeta - \xi, z - x \rangle \geq \eta \, \|z - x\|_X^2 ,$$

then for any $\Upsilon\colon X \to X'$ Lipschitz with Lipschitz constant less than or equal to η, $\Psi + \Upsilon\colon X \to \mathcal{P}(X')$ is also maximal monotone.

Proof. We wish to show that for any $\beta > 0$, the operator $\beta J_X + \Psi + \Upsilon$ is onto X'. So consider the problem of solving $\beta J_X(x) + \Psi(x) + \Upsilon(x) \ni y$ for any given $y \in X'$ and

$\beta > 0$. This can be rewritten as $x + (\beta J_X + \Psi)^{-1} \Upsilon(x) = (\beta J_X + \Psi)^{-1}(y)$. This can be solved by an iterative method:

$$x_{k+1} = (\beta J_X + \Psi)^{-1}(y) - (\beta J_X + \Psi)^{-1} \Upsilon(x_k).$$

Now $(\beta J_X + \Psi)^{-1} \Upsilon$ is a single-valued Lipschitz operator with Lipschitz constant $L_\Upsilon / (\beta + \eta) < 1$ where L_Υ is the Lipschitz constant for Υ. Applying the contraction mapping principle shows that there is indeed a unique solution. □

Now we can continue with the proof of Lemma 4.7.

Proof. From Lemma 4.8 it can be shown that $\Phi - \alpha J_X$ is maximal monotone. Our task is to show that for any $\gamma > 0$, $\gamma I + \Phi_H$ is surjective, where $I\colon H \to H$ is the identity map; this map can be identified with the inclusion map $X \subseteq H = H' \subseteq X'$ on X. Now, for any $\gamma > 0$, $\gamma I + \Phi$ is maximal monotone, since I is both Lipschitz and monotone. (We can repeatedly apply Lemma 4.8.) Furthermore, $\gamma I + \Phi\colon X \to \mathcal{P}(X')$ is strongly monotone with constant greater than or equal to α. Thus $\gamma I - \alpha J_X + \Phi\colon X \to \mathcal{P}(X')$ is also maximal monotone. Thus $\gamma I + \Phi = \alpha J_X + (\gamma I - \alpha J_X + \Phi)$ is onto X'. For any $y \in H = H' \subseteq X'$, there is an $x \in X$ such that $y \in \gamma x + \Phi(x)$. Since $x \in X \subseteq H$, $y - \gamma x \in \Phi(x)$ and since $y - \gamma x \in H$, we have $y - \gamma x \in \Phi_H(x)$. Thus $y \in (\gamma I + \Phi_H)(x)$ for some $x \in X$. This means that $\gamma I + \Phi_H$ is onto H, and so $\Phi_H\colon H \to \mathcal{P}(H)$ is a maximal monotone operator on H. □

As we see in the next section, maximal monotone operators in Gelfand triples can be used to show the existence and uniqueness of solutions to obstacle problems.

4.2.3 Application to the heat equation and obstacle problems

To see how we can use this for partial differential equations and related problems, consider again the heat equation with Dirichlet boundary conditions:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla^2 u, \qquad u(t,x) = u_0(x) \qquad \text{for all } x \in \Omega, \\ u(t,x) &= 0 \qquad \text{for all } x \in \partial\Omega, \end{aligned}$$

where Ω is a bounded open set in $\mathbb{R}^d$. Finding the right function spaces in which to set up this problem is not a trivial issue. For u in practically any function space worth considering, $\nabla^2 u$ can be defined in the sense of distributions. However, this does not allow us to infer the existence of solutions. A natural modern approach is to use Gelfand triples: we look for solutions in the Sobolev space $H_0^1(\Omega)$, the space of functions $u \in L^2(\Omega)$ where the distributional gradient $\nabla u \in L^2(\Omega)$, and the restriction (or, more accurately, the *trace* of u) to the boundary $\partial\Omega$ is zero. Thanks to the divergence theorem,

$$\begin{aligned} \int_\Omega v(x) \left(-\nabla^2 u(x)\right) dx &= -\int_{\partial\Omega} v(x) \frac{\partial u}{\partial n}(x)\, dS(x) + \int_\Omega \nabla v(x) \cdot \nabla u(x)\, dx \\ &= \int_\Omega \nabla v(x) \cdot \nabla u(x)\, dx \qquad (\text{if } v = 0 \text{ on } \partial\Omega) \end{aligned}$$

is defined whenever $u, v \in H_0^1(\Omega)$. By duality, then, we can think of $-\nabla^2 \colon H_0^1(\Omega) \to H_0^1(\Omega)'$. The Gelfand triple we can use is then

$$H_0^1(\Omega) \subset L^2(\Omega) = L^2(\Omega)' \subset H_0^1(\Omega)'.$$

This identifies a function $f \in L^2(\Omega)$ with the functional $g \mapsto \int_\Omega f(x)\,g(x)\,dx$.

However, this does not allow us to identify $H_0^1(\Omega)$ with $H_0^1(\Omega)'$, as the natural isomorphism $J \colon H_0^1(\Omega) \to H_0^1(\Omega)'$ is given by $w \mapsto w - \nabla^2 w$.

If we wish to deal with the heat equation within the framework of maximal monotone operators, we need to use $L^2(\Omega)$ as our Hilbert space. In *this* space, $-\nabla^2$ is the subdifferential of the proper lower semicontinuous convex function

$$\phi(f) = \begin{cases} \int_\Omega \|\nabla f(x)\|^2\,dx, & f \in H_0^1(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

Then we have unique solutions to the differential inclusion

$$0 \in \frac{\partial u}{\partial t} + \partial\phi(u), \qquad u(0,x) = u_0(x)$$

not only for $u_0 \in \operatorname{dom}\partial\phi = \left\{ w \in L^2(\Omega) \mid \nabla^2 w \in L^2(\Omega) \right\}$ but also for $u_0 \in \overline{\operatorname{dom}\partial\phi} = L^2(\Omega)$! That's right; we do not even need $u_0 \in H_0^1(\Omega)$ for solutions to exist. It is enough for u_0 to be in $L^2(\Omega)$.[4]

Now let us consider incorporating a constraint that $u(t,x)$ satisfy $u(t,x) \geq \varphi(x)$ for all $x \in \Omega$. This kind of problem is known as an *obstacle problem* in the partial differential equations community. What should happen if $u(t,x) = \varphi(x)$ to prevent $u(t,x) < \varphi(x)$ from happening? We will assume that there should be some restoring quantity in the differential equation that prevents $u(t,x) < \varphi(x)$: call it $w(t,x)$. If $u(t,x) > \varphi(x)$, we should take $w(t,x) = 0$ so that the heat equation applies. If $u(t,x) = \varphi(x)$, we should take $w(t,x) \geq 0$ so as to "push" the solution away from $u(t,x) < \varphi(x)$. So our system of conditions becomes

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla^2 u - w(t,x), \qquad u(0,x) = u_0(x) \qquad \text{for } x \in \Omega, \\ 0 \geq w(t,x) &\perp u(t,x) - \varphi(x) \geq 0 \qquad \text{for } x \in \Omega. \end{aligned}$$

This is a complementarity formulation of a parabolic obstacle problem. To turn this into a maximal monotone differential inclusion, we need to construct a closed convex set $K = \left\{ z \in H_0^1(\Omega) \mid z(x) \geq \varphi(x) \text{ for all } x \in \Omega \right\}$. Provided $\varphi(x) \leq 0$ on $\partial\Omega$ and $\varphi \in H^1(\Omega)$, K is a nonempty closed convex set in $H_0^1(\Omega)$. Now, considering $K \subset H_0^1(\Omega)$,

$$N_K(u) = \left\{ w \in H_0^1(\Omega)' \mid w \leq 0 \text{ and } w(x)(u(x) - \varphi(x)) = 0 \text{ for all } x \in \Omega \right\}.$$

Now $N_K(u) = \partial I_K(u)$, where $I_K \colon H_0^1(\Omega) \to \mathbb{R} \cup \{\infty\}$ is the indicator function for K. Then $\phi + I_K$ is a proper convex lower semicontinuous function on $H_0^1(\Omega)$. However, what we

[4] Harmonic analysts go beyond even this level of regularity to consider $u_0 \in L^1(\Omega)$ or even measures. Much of this work requires the maximum principle for the heat equation, while the maximal monotone operator approach does not require it, making the maximal monotone approach more appropriate for *systems* of partial differential equations.

really need is a proper convex lower semicontinuous function on $L^2(\Omega)$. The obvious way to do this is to set

$$\psi(u) = \begin{cases} \phi(u) + I_K(u) & \text{if } u \in H_0^1(\Omega), \\ +\infty & \text{otherwise.} \end{cases}$$

This is clearly a convex function (since ϕ, I_K are convex, and $H_0^1(\Omega)$ is a convex subset of $L^2(\Omega)$) and proper (since $\psi(\varphi_+) = \phi(\varphi_+) < \infty$, where $\varphi_+(x) = \max(\varphi(x), 0)$ for $\varphi \in H_0^1(\Omega)$). The harder part is to show that ψ is lower semicontinuous. So suppose that $u_k \to u$ in $L^2(\Omega)$, and that $\limsup_{k\to\infty} \psi(u_k) < \infty$. (If $\limsup_{k\to\infty} \psi(u_k) = \infty$, there is nothing to prove.) This means that for sufficiently large k, u_k are bounded in $H_0^1(\Omega)$. Ignoring the finite set of $u_k \notin H_0^1(\Omega)$, let $\|u_k\|_{H_0^1(\Omega)} \le C$ for all k. Since the u_k are bounded in $H_0^1(\Omega)$, there is a weakly convergent subsequence $u_k \rightharpoonup \widehat{u}$ in $H_0^1(\Omega)$. Since $H_0^1(\Omega)$ is compactly embedded into $L^2(\Omega)$, $u_k \to \widehat{u}$ in $L^2(\Omega)$ in a suitable subsequence; therefore $\widehat{u} = u$. Thus $u \in H_0^1(\Omega)$. Also, since $u_k \rightharpoonup u$ weakly in $H_0^1(\Omega)$, by Mazur's lemma $\psi(u) = \phi(u) + I_K(u) \le \limsup_{k\to\infty} \phi(u_k) + I_K(u_k) = \limsup_{k\to\infty} \psi(u_k)$. Thus ψ is lower semicontinuous.

This means that our obstacle problem can be treated as a maximal monotone differential inclusion

$$0 \in \frac{\partial u}{\partial t} + \partial\psi(u), \qquad u(0) = u_0 \in \overline{\operatorname{dom} \partial\psi}.$$

Note that the closure of the $\operatorname{dom}\psi$ is taken *in* $L^2(\Omega)$. This means that we can take u_0 to be any function in $L^2(\Omega)$ where $u_0 \ge \varphi$. Thus solutions exist and are unique for this problem.

A word of warning though: we should be careful about identifying $\partial\psi$ with $-\nabla^2 + N_K$. The reason is that $\partial(\phi_1 + \phi_2) = \partial\phi_1 + \partial\phi_2$ does not always hold.

An alternative approach to this is to note that $-\nabla^2 \colon X = H_0^1(\Omega) \to H_0^1(\Omega)' = X'$ is a Lipschitz (between these spaces), maximal monotone operator. Then, since $K \subseteq X = H_0^1(\Omega)$ is a nonempty closed convex set, N_K is also a maximal monotone operator $X \to \mathcal{P}(X')$. By Lemma 2.32, since $\operatorname{dom}(-\nabla^2) = X$ and $\operatorname{dom}(N_K) = K \ne \emptyset$, we have (in X) $\operatorname{interior}(\operatorname{dom}(-\nabla^2)) \cap \operatorname{dom}(N_K) = K \ne \emptyset$, so $-\nabla^2 + N_K$ is a maximal monotone operator $X \to X'$. Now we take $H = L^2(\Omega)$, which we identify with H' to form a Gelfand triple $X \subset H = H' \subset X'$. Now $J_X = I - \nabla^2$. For any $0 < \alpha < 1$, $\left(-\nabla^2 + N_K + \alpha I\right) - \alpha J_X$ is monotone. So, by Lemma 4.7, $-\nabla^2 + N_K + \alpha I$ is maximal monotone. So our obstacle problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla^2 u - N_K(u) \\ &= \alpha u - \left(-\nabla^2 u + N_K(u) + \alpha u\right) \end{aligned}$$

has a right-hand side of the form "Lipschitz−Maximal monotone," and so it has unique solutions for given initial values. In fact, if $u_{1,0}$ is one initial value for solution $u_1(\cdot)$, and $u_{2,0}$ is another initial value for solution $u_2(\cdot)$, then

$$\|u_1(t) - u_2(t)\|_{L^2(\Omega)} \le e^{\alpha t} \left\|u_{1,0} - u_{2,0}\right\|_{L^2(\Omega)}.$$

Since this is true for all $0 < \alpha < 1$, we can take α as small as we please so that in the limit $\alpha \downarrow 0$,

$$\|u_1(t) - u_2(t)\|_{L^2(\Omega)} \le \left\|u_{1,0} - u_{2,0}\right\|_{L^2(\Omega)}.$$

4.2.4 Uniqueness of solutions and maximal monotone operators

While having a differential inclusion of the form

$$\frac{dx}{dt}(t) \in f(t, x(t)) - \Phi(x(t)), \qquad x(t_0) = x_0$$

with $f(t, x)$ Lipschitz in x and Φ maximal monotone is sufficient to ensure uniqueness of solutions, it is far from being necessary. In fact, maximal monotone operators are tightly constrained in certain respects. Consider, for example, the differential inclusion

$$\frac{dx}{dt}(t) \in f(t, x(t)) - g(x(t))\,\mathrm{Sgn}(\phi(x(t))), \qquad x(t_0) = x_0, \tag{4.20}$$

where $\phi\colon \mathbb{R}^n \to \mathbb{R}$, and $\mathrm{Sgn}(u) = \{+1\}$ if $u > 0$, $\{-1\}$ if $u < 0$, and $[-1, +1]$ if $u = 0$. We assume that if $\phi(x) = 0$, then $\nabla\phi(x) \neq 0$, and that f and g are Lipschitz.

If the set-valued function $\Psi(x) := -g(x)\,\mathrm{Sgn}(\phi(x))$ is "Lipschitz − maximal monotone," then this set-valued function must satisfy the following one-sided Lipschitz condition: there is an L such that

$$\langle y - w, x - z\rangle \leq L\,\|x - z\|^2 \qquad \text{for all } x, z, y \in \Psi(x), w \in \Psi(z).$$

Pick a point x^* such that $\phi(x^*) = 0$. Let $n^* = \nabla\phi(x^*)$.

First we show that $\nabla\phi(x^*)\,g(x^*) \geq 0$. For any $\eta > 0$ sufficiently small, $\phi(x^* + \eta\, n^*) > 0$ and $\phi(x^* - \eta\, n^*) < 0$. Then the one-sided Lipschitz condition implies that

$$\left(-g(x^* + \eta n^*) - g(x^* - \eta n^*)\right)^T \left(2\,\eta\, n^*\right) \leq L\,\left\|2\eta\, n^*\right\|^2.$$

Dividing by $\eta > 0$ and taking $\eta \downarrow 0$ give

$$-g(x^*)^T n^* \leq 0,$$

as desired. It turns out that the condition $g(x^*)^T n^* = \nabla\phi(x^*)\,g(x^*) > 0$ is *sufficient* to guarantee uniqueness for (4.20); see Section 5.2.2.

We will now see that the one-sided Lipschitz condition implies that $\nabla\phi(x^*)$ and $g(x^*)$ must also be *parallel*.

Let $d \in \mathbb{R}^n$ be a nonzero direction perpendicular to n^*: $d^T n^* = 0$. Let β_H be a bound for the Hessian matrix $\mathrm{Hess}\,\phi(x)$ for x in a neighborhood of x^*. For $\eta > 0$ sufficiently small,

$$\begin{aligned}\phi(x^* + \eta\, d + C\,\eta^2\, n^*) &\geq \phi(x^*) + \eta\,\nabla\phi(x^*)\,d + C\,\eta^2\,\nabla\phi(x^*)\,n^* \\ &\quad - \frac{1}{2}\beta_H\,\left\|\eta\, d + C\eta^2 n^*\right\|^2 \\ &= \left(C\,\left\|n^*\right\|^2 - \frac{1}{2}\beta_H\,\|d\|^2\right)\eta^2 + \mathcal{O}(\eta^3).\end{aligned}$$

So if we choose $C > \frac{1}{2}\beta_H\,\|d\|^2 / \|n^*\|^2$, we have $\phi(x^* + \eta\, d + C\,\eta^2\, n^*) > 0$ for all $\eta > 0$ sufficiently small. Pick another vector $d' \neq 0$ such that $\left(d'\right)^T n^* = 0$. In the same way as above, if we choose C sufficiently large and positive, we can ensure that $\phi(x^* + \eta\, d' - C\eta^2\, n^*) < 0$

for all $\eta > 0$ sufficiently small as well. Then, if the one-sided Lipschitz condition is satisfied, we have

$$\begin{aligned}&\left(-g(x^* + \eta d + C\eta^2 n^*) - g(x^* + \eta d' - C\eta^2 n^*)\right)^T \left(\eta\left(d - d'\right) + 2C\eta^2 n^*\right)\\ &\qquad \le L \left\| \eta\left(d - d'\right) + 2C\eta^2 n^* \right\|^2.\end{aligned}$$

Dividing by $\eta > 0$ gives

$$\begin{aligned}&\left(-g(x^* + \eta d + C\eta^2 n^*) - g(x^* + \eta d' - C\eta^2 n^*)\right)^T \left(\left(d - d'\right) + 2C\eta n^*\right)\\ &\qquad \le L\eta \left\| \left(d - d'\right) + 2C\eta n^* \right\|^2.\end{aligned}$$

Taking $\eta \downarrow 0$ then gives

$$-2\, g(x^*)^T \left(d - d'\right) \le 0$$

for any d, d' perpendicular to n^*. So $g(x^*)$ must be perpendicular to every vector perpendicular to n^*; in other words, $g(x^*)$ must be parallel to $n^* = \nabla\phi(x^*)$.

Functions of the form $x \mapsto g(x)\,\mathrm{Sgn}(\phi(x))$ are thus maximal monotone (or "Lipschitz + maximal monotone") only under fairly restrictive assumptions on g and ϕ: whenever $\phi(x) = 0$, $\nabla\phi(x)\, g(x) \ge 0$ with $\nabla\phi(x)$ and $g(x)$ parallel. Thus arbitrarily small perturbations to $g(x)$ on the surface $\phi(x) = 0$ can destroy this property.

Nevertheless, uniqueness of solutions to (4.20) can be shown if $\nabla\phi(x)\, g(x) > 0$ on the surface $\phi(x) = 0$. Away from this surface, uniqueness is clear by the Lipschitz properties of f and g. Uniqueness can be shown via the uniqueness theorem for DVIs (Theorem 5.3 in Section 5.2.2). The DVI theorem can be applied to

$$\begin{aligned}&\frac{dx}{dt}(t) = f(t, x(t)) + g(x(t))\, z(t), \qquad x(t_0) = x_0,\\ z(t) \in [-1, +1], \qquad &0 \le (\widetilde{z} - z(t))\, \phi(x(t)) \qquad \text{for all } \widetilde{z} \in [-1, +1].\end{aligned}$$

In the borderline case of $\nabla\phi(x)\, g(x) = 0$ where $g(x)$ is not necessarily parallel to $\nabla\phi(x)$, solutions are not necessarily unique. Take, for example, $f(t, x) \equiv 0$, $\phi\colon \mathbb{R}^2 \to \mathbb{R}$ given by $\phi(x_1, x_2) = x_1$ and $g(x_1, x_2)^T = [0, 1]$. Any absolutely continuous function $x(\cdot)$ with $x_1(t) \equiv 0$ and $x_2(\cdot)$ nondecreasing and Lipschitz with constant one is then a solution of the differential inclusion. Thus solutions are not unique in this case.

4.3 Projected dynamical systems

Projected dynamical systems (PDSs) are a class of dynamical systems that have been investigated as a class by Nagurney, Dupuis, Cojocaru, Daniele, and Jonker (see [62, 63, 83, 187]); however, the first mathematical investigations of these concepts go back at least to Henry [125] and extended by Cornet [65] to nonconvex sets which are nevertheless "regular" in the sense of Clarke [55]. The initial applications they had in mind were for the dynamics of economic networks. The basic idea and theory come from maximal monotone differential equations where the maximal monotone operator used is the normal cone operator N_K for a closed convex set K.

Suppose that $f\colon [0,T]\times X\to X$, where X is a Hilbert space (where we identify X and X'), $x\mapsto f(t,x)$ is Lipschitz continuous with constant L for all t, and $t\mapsto f(t,x)$ is continuous for all $x\in K$. Then consider the differential inclusion

$$0\in\frac{dx}{dt}+N_K(x)+f(t,x),\qquad x(0)=x_0\in K. \tag{4.21}$$

This represents a dynamical system which "usually" satisfies the differential equation $dx/dt(t)=-f(t,x(t))$, provided $x(t)\in \text{interior}\, K$. But when $x(t)$ reaches the boundary ∂K the trajectory is prevented from leaving K because $N_K(x)$ is part of the differential inclusion. Here it is important that $N_K(x)$ can be unbounded since otherwise $f(t,x)$ could be sufficiently large to overcome its effect and $x(t)$ could leave K.

Since N_K is a maximal monotone operator, solutions exist and are unique for this differential inclusion. But, by the theory of the previous section on maximal monotone differential inclusions, any solution satisfies

$$\frac{d^+x}{dt}(t)=(-N_K(x(t))-f(t,x(t)))^0\qquad \text{for all } t\geq 0,$$

where C^0 is the minimal norm point of a closed convex set C: $C^0=\Pi_C(0)=\arg\min_{z\in C}\|z\|$. So, if we set $g=f(t,x(t))$ and $C=N_K(x(t))$, we have

$$\begin{aligned}\frac{d^+x}{dt}(t)&=\Pi_{-C-g}(0)=-\Pi_{C+g}(0)\\&=-\Pi_C(-g)-g.\end{aligned}$$

However, the projections onto a closed convex cone C and its polar cone $C^\circ=-C^*$ satisfy $z=\Pi_C(z)+\Pi_{C^\circ}(z)$ for all z in a Hilbert space X identified with X' by Moreau's decomposition theorem (Lemma B.7). Thus $-\Pi_C(-g)-g=(-g)-\Pi_C(-g)=\Pi_{C^\circ}(-g)$. But $C=N_K(x(t))$, so $C^\circ=T_K(x(t))$. Thus

$$\frac{d^+x}{dt}(t)=\Pi_{T_K(x(t))}(-f(t,x(t))). \tag{4.22}$$

That is, we replace $-f(t,x(t))$ with its projection on the tangent cone $T_K(x(t))$ at $x(t)$. This keeps the solution from leaving K. Another way of formulating this is to note that

$$\Pi_{T_K(x)}(v)=\Pi'_K(x;v),$$

the directional derivative of Π_K at x in the direction v. Thus a PDS can be formulated as

$$\frac{d^+u}{dt}(t)=\Pi'_K(x(t);-f(t,x(t))). \tag{4.23}$$

The characterizations (4.22) and (4.23) are useful, but the fundamental theory behind PDSs comes from maximal monotone differential inclusions.

In terms of DVIs, suppose that the set $K\subseteq\mathbb{R}^n$ is represented in terms of convex functions $K=\{x\mid g_i(x)\leq 0, i=1,2,\dots,m\}$, where each g_i is a convex function. Provided the *Slater constraint qualification* (B.22) holds, the tangent cone $T_K(x)$ is given by the linearization of the constraints:

$$T_K(x)=\left\{z\mid \nabla g_i(x)^T z\leq 0, \text{ where } g_i(x)=0, i=1,2,\dots,m\right\},$$

so the normal cone is

$$N_K(x) = \text{co}\,\{\nabla g_i(x) \mid g_i(x) = 0, i = 1, 2, \ldots, m\}.$$

The maximal monotone differential inclusion can be represented by the DCP

$$0 = \frac{dx}{dt}(t) + \sum_{i=1}^{m} \lambda_i(t)\nabla g_i(x(t)),$$
$$0 \le \lambda_i(t) \perp g_i(x(t)) \ge 0 \qquad \text{for all } i \text{ and } t \ge 0.$$

Provided the matrix $[\nabla g_i(x) \mid g_i(x) = 0, i = 1, 2, \ldots, m]$ has full rank for all x, this DCP has index-one. To go beyond index-one problems in general requires a weaker and more general notion of solution. This is needed for impact problems (which are index two in general), for example. In the next section we investigate an approach which has been used to provide a mathematical foundation for impact problems.

4.4 Sweeping processes

Sweeping processes are a true generalization of maximal monotone differential inclusions rather than a subclass. These were invented by Moreau [178, 180], and the concept was extended and applied by Moreau [179, 182] and others such as Castaing, Monteiro Marques, Valadier, and Kunze (see [49, 166, 174, 130, 149]). Sweeping processes provide a way to introduce discontinuous changes, as we will see.

4.4.1 Pure sweeping processes

The basic idea is that there is a time-dependent state vector $x(t)$ which stays within a moving closed convex set $C(t)$ but otherwise tries not to move. The set $C(t)$ "sweeps" the state vector along with it as it moves. If $C(t)$ changes continuously, then the state vector $x(t)$ should also change continuously (but not necessarily smoothly, as it can transition from being still to suddenly being swept along). But if $C(t)$ changes discontinuously (particularly when $C(t)$ becomes smaller), then $x(t)$ can jump.

Recall that we can use the Hausdorff metric d_H (see (2.10)) to define a distance between closed and bounded sets.

To measure how much of an "excess" one set has over another, we can use the one-sided "metric":

$$\delta_H(A, B) = \sup_{a \in A} d(a, B),$$

so that the Hausdorff metric is $d_H(A, B) = \max(\delta_H(A, B), \delta_H(B, A))$. This one-sided metric satisfies $\delta_H(A, A) = 0$ and $\delta_H(A, C) \le \delta_H(A, B) + \delta_H(B, C)$, but usually $\delta_H(A, B) \ne \delta_H(B, A)$.

The basic assumptions needed for sweeping processes are the following:

1. $C(t)$ is a closed and bounded convex set for all t.
2. There is a function $r(t)$ such that $\delta_H(C(s), C(t)) \le r(t) - r(s)$ for all $t \ge s$.
3. $0 \in dx/dt(t) + N_{C(t)}(x(t))$.

The nature of the function r determines the regularity of the solution. Since r is a nondecreasing function, it is at worst a function of *bounded variation*. If r is absolutely continuous, then the solution $x(t)$ is absolutely continuous and $0 \in dx/dt + N_{C(t)}(x(t))$ can be interpreted as a differential inclusion. But if r is not absolutely continuous, then dx/dt cannot be interpreted as an ordinary derivative; if r has a jump, then x can also jump.

One way of establishing the existence of solutions for these problems is by means of a time discretization: $x_n \approx x(t_n)$, $t_n = t_0 + n\,h$, and

$$x_{n+1} = \Pi_{C(t_{n+1})}(x_n).$$

This is called the *catching-up algorithm* [174]. This will be used to establish existence of solutions in Section 4.4.4. Before we deal with that, we describe in what sense the differential inclusion is understood, as measures do not have pointwise values.

4.4.2 Measure differential inclusions

Measure differential inclusions (MDIs) are a variant on the concept of differential inclusions which allows the solutions to have discontinuities. It cannot be an arbitrary (discontinuous) function, but rather the solution must be a function $x(\cdot)$ with bounded variation, so that the differential measure $dx(\cdot)$ is a measure. See Section A.4 for a definition of differential measures. MDIs were first formally named by Moreau [179, 180], although they were previously used (but not named) by Schatzman [219, 220]. The theory was further developed by Monteiro Marques [174] in the context of sweeping processes and later by Stewart [240, 241].

An MDI has the form

$$\frac{dx}{dt}(t) \in \Phi(t, x(t)), \tag{4.24}$$

where $x(\cdot)$ is a function of bounded variation and $\Phi(t, x(t))$ is closed and convex and can be an unbounded set. What is different about MDIs is how to interpret the inclusion. This is necessary because measures do not have pointwise values, and we cannot assume that $x(\cdot)$ is an absolutely continuous function.

There is an additional issue: Since $x(\cdot)$ can be discontinuous, what should we use for the value $x(t)$ in the right-hand side? Usually we take the limit from above $x(t^+) = \lim_{s \downarrow t} x(s)$ which exists for any function of bounded variation. Using $x(t^+)$ makes the formulation consistent with the theory of maximal monotone operators: if $\Phi(t, x) = -\Psi(x)$, where $\Psi\colon X \to \mathcal{P}(X) = \mathcal{P}(X')$ is a maximal monotone, then a solution of the MDI $dx/dt \in -\Psi(x(t^+))$ is necessarily of bounded variation, while solutions of $dx/dt \in -\Psi(x(t^-))$ are not. Later we will see an example of this.

Consider the basic MDI

$$\frac{dx}{dt}(t) \in \Phi(t) \subseteq X. \tag{4.25}$$

As for ordinary differential inclusions, we will assume that

- Φ is a measurable function $[a, b] \to \mathcal{P}(X)$, and
- $\Phi(t)$ is a closed convex set for all t.

For $x(\cdot)$ of bounded variation, the differential measure $dx(\cdot)$ is a measure whose variation measure $|dx|(\cdot)$ is finite. That is, $|dx|(E) < \infty$ for any Borel set $E \subseteq [a,b]$. Moreau, Monteiro Marques, and Schatzman all assumed that $\Phi(t)$ is a closed convex *cone*. The formulation of what (4.25) means can then be given like this: if $\mu = dx$, the Radon–Nikodym derivative (see Section A.4)

$$\frac{d\mu}{d|\mu|}(t) \in \Phi(t) \qquad \text{for } |dx|\text{-almost all } t. \tag{4.26}$$

Requiring that $\Phi(t)$ is always a cone is rather restrictive: ordinary differential equations cannot be represented as MDIs of this kind. In the theory of ordinary differential equations, solutions $x(\cdot)$ are absolutely continuous. In the language of measures, this amounts to saying the dx is an absolutely continuous measure with respect to the Lebesgue measure (which we can represent as dt). The ordinary derivative (existing almost everywhere in the Lebesgue measure) is then $dx/dt(t) = d(dx)/d(dt)(t)$, where the right-hand side is a Radon–Nikodym derivative. But $d(dx)/d(dt)$ is ugly notation, so we use dx/dt instead.

To incorporate ordinary differential equations, we must give a special place to the Lebesgue measure, and we do this by using the *Lebesgue decomposition* $dx = \mu_s + \mu_{ac}$ where μ_{ac} is absolutely continuous with respect to the Lebesgue measure, and μ_s is singular with respect to the Lebesgue measure. That is, there is a Lebesgue integrable function h where $\mu_{ac}(E) = \int_E h(t)dt$ for all Borel E, and there is a Lebesgue null set F where $\mu_s(E) = \mu_s(E \cap F)$ for all Borel E. The singular part μ_s can contain things like Dirac-δ functions as well as more exotic measures.

The absolutely continuous part is the "nice" part, and we can identify the absolutely continuous part of dx/dt with $h(t)$. Thus the absolutely continuous part of the MDI can be understood as

$$h(t) = \frac{d\mu_{ac}}{d\lambda}(t) \in \Phi(t) \qquad \text{for Lebesgue almost all } t, \tag{4.27}$$

where λ is the Lebesgue measure.

Since the singular part involves things like momentarily infinite values, we should think of μ_s as belonging to the vectors in $\Phi(t)$ "at infinity." For convex sets $K \subseteq X$ there is a natural "limit" at infinity: the *recession cone*

$$\begin{aligned} K_\infty &= \left\{ \lim_{k\to\infty} t_k y_k \mid t_k \downarrow 0 \text{ as } k \to \infty,\ y_k \in K \right\} \\ &= \bigcap_{s>0} s(K - y) \qquad \text{for any } y \in K. \end{aligned}$$

For the absolutely continuous part we can think of the measure differential μ_{ac} as having pointwise values: those of $h(t)$. For the singular part we use the idea of Moreau and others but with $\Phi(t)$ replaced by its recession cone:

$$\frac{d\mu_s}{d|\mu_s|}(t) \in \Phi(t)_\infty \qquad \text{for } |\mu_s| \text{ almost all } t. \tag{4.28}$$

Combining (4.27) and (4.28) we have a definition of what "$dx/dt \in \Phi(t)$" means for $x(\cdot)$ having bounded variation. It should be noted that if $\Phi(t)$ is a *cone* for all t, then (4.27) and (4.28) together are equivalent to (4.26).

While it is very useful to have a suitable *definition* of a concept, it is even better if we can use the concept to prove results, especially about convergence. In the case of measures, what we usually have is weak* convergence, treating the space of (signed) measures on, say, the interval $[a,b]$ as the dual space to $C[a,b]$, the space of continuous functions on $[a,b]$. Unfortunately, the Lebesgue decomposition of a measure is not stable under weak* convergence. Consider, for example, a "bed of nails" measure on $[0,1]$:

$$\mu_m(t) = \frac{1}{m}\sum_{j=0}^{m-1} \delta\left(t - \frac{j}{m}\right). \tag{4.29}$$

(Think of m nails pointing up with spacing $1/m$ between them.) The weak* limit of this sequence of measures as $m \to \infty$ is simply the Lebesgue measure. (If the nails are close enough together, it feels like a flat board; I have actually felt this at a science museum!) However, μ_m is purely singular; that is, its absolutely continuous part is zero. Nevertheless, the weak* limit has no singular part at all. Conversely, consider a standard "approximation" to the Dirac-δ function:

$$\psi_h(t) = \begin{cases} 1/h, & 0 \le t \le h, \\ 0 & \text{otherwise.} \end{cases}$$

Each ψ_h gives a measure $\psi_h \lambda$ that is absolutely continuous. But its weak* limit is the Dirac-δ measure. Thus the weak* limit of a purely absolutely continuous measure can be purely singular.

An alternative definition is given in [240, 241]: We say "$d\mu/dt \in \Phi(t)$" in the sense of MDIs if for every continuous $\phi\colon [a,b] \to \mathbb{R}_+$ not identically zero,

$$\frac{\int_{[a,b]} \phi \, d\mu}{\int_{[a,b]} \phi \, dt} \in \overline{\text{co}} \bigcup_{t:\phi(t)>0} \Phi(t). \tag{4.30}$$

This is called the *weak definition*. On the other hand, (4.27)–(4.28) is called the *strong definition*. Clearly if $\mu_k \rightharpoonup^* \mu$ and μ_k satisfy (4.30), then μ also satisfies (4.30). But when can we tell if the two concepts are equivalent? Fortunately, they are equivalent under mild conditions, especially if $X = \mathbb{R}^n$. If $X = \mathbb{R}^n$, then (4.27)–(4.28) is equivalent to (4.30), provided

(MDI-H1) $\min\{\|y\| \mid y \in \Phi(t)\}$ is a locally bounded function of t;

(MDI-H2) $\Phi(t)$ is a closed convex set for all t, and $\operatorname{graph}\Phi$ is closed;

(MDI-H3) the recession cone $\Phi(t)_\infty$ is a *pointed* cone for all t.

A simple example can show why the pointedness condition (MDI-H3) is necessary. Let $\mathbf{d}(t) = [\cos(t), \sin(t)]^T \in \mathbb{R}^2$, and set $\Phi(t) = \left\{\mathbf{z} \in \mathbb{R}^2 \mid \mathbf{d}(t)^T\mathbf{z} \ge 0\right\}$. If $\phi\colon [a,b] \to \mathbb{R}_+$ is not identically zero, then it is nonzero on an open interval (c,d). But each $\Phi(t)$ is a half-space, and provided $t \not\equiv s \pmod{2\pi}$, $\text{co}\,(\Phi(s) \cup \Phi(t)) = \mathbb{R}^2$—the entire plane. This means that the right-hand side of (4.30) in this case is always $\mathbb{R}^2$, and any measure μ with values in $\mathbb{R}^2$ is a solution of "$d\mu/dt \in \Phi(t)$" according to the weak definition. However, this is definitely not the case with the strong definition. (You could try $\mu = -\mathbf{d}(t)\nu$, where ν is any nonzero and nonnegative measure.)

In practice, conditions (MDI-H1), (MDI-H2), and (MDI-H3) for equivalence of the weak and strong definitions for MDIs hold for $X = \mathbb{R}^n$. However, the situation for infinite-dimensional problems is more complex: According to [240] the Banach space X should be a separable reflexive space, $\Phi(t) \subseteq L + R\overline{B_X}$, where L is a *strongly pointed cone* (that is, L^* should have a nonempty interior), $R > 0$, and $\overline{B_X}$ is the closed unit ball in X. That is, in infinite dimensions, to prove equivalence we assume that X is a separable reflexive space, and replace condition (MDI-H3) with the requirement that

(MDI-H3b) $\Phi(t) \subset R\overline{B} + L$, where L is *strongly pointed*; that is, L^* has nonempty interior.

Theorem 4.9. *Suppose that $\Phi(t) \subseteq \mathbb{R}^n$ satisfies conditions* (MDI-H1)–(MDI-H3) *above. A function of bounded variation $x\colon [0,T] \to \mathbb{R}^n$ is a solution of the MDI*

$$\frac{dx}{dt}(t) \in \Phi(t) \tag{4.31}$$

in the strong sense if and only if it is a solution to (4.31) *in the weak sense.*

In finite dimensions, note that it is sufficient to have $\Phi(t)_\infty$ pointed for all t instead of (MDI-H3) $\Phi(\tau) \subseteq L + R\overline{B_X}$ for all τ in a neighborhood of t.

Monteiro Marques [174] and Moreau [180] both use a simpler version of the strong solution condition, which is applicable for closed convex *cone*-valued functions $\Phi\colon [a,b] \to \mathcal{P}(X)$: writing $\mu = dx$ for the differential measure, they simply require that the Radon–Nikodym derivative

$$\frac{d\mu}{d(\lambda + |\mu|)}(t) \in \Phi(t) \qquad \text{for } \lambda + |\mu| \text{ almost all } t. \tag{4.32}$$

The equivalence of weak and strong solution concepts for MDIs avoids some of the complexity of dealing with the strong solution concept that Monteiro Marques and Moreau had to deal with. In particular, in showing that discrete-time approximations do indeed converge to solutions, weak* convergence of the discrete-time measures is all that can usually be shown. The weak solution concept can usually be shown to hold in the limit easily, while the Lebesgue decomposition is not continuous under weak* limits of measures and Radon–Nikodym derivatives are also not very well behaved with respect to weak* convergence.

The proof of equivalence of the weak and strong solution concepts involves taking sequences $\phi_k \to \chi_E$ pointwise as $k \to \infty$ for E an open set to obtain $\mu(E)/\lambda(E) \in \overline{\text{co}}\bigcup_{t\in E}\Phi(t)$, where λ is the Lebesgue measure. Then by taking nested unions and intersections we can show that for any Borel $E' \subset E$, E open, $\mu(E')/\lambda(E') \in \overline{\text{co}}\bigcup_{t\in E}\Phi(t)$ if $\lambda(E') > 0$ and $\mu(E') \in \left[\overline{\text{co}}\bigcup_{t\in E}\Phi(t)\right]_\infty$ if $\lambda(E') = 0$. Taking Radon–Nikodym derivatives then gives $d\mu_{ac}/dt(t) \in \overline{\text{co}}\bigcup_{t\in E}\Phi(t)$ and $d\mu_{sing}/d\left|\mu_{sing}\right|(t) \in \left[\overline{\text{co}}\bigcup_{t\in E}\Phi(t)\right]_\infty$ whenever E is an open set containing t. The pointedness property of $\Phi(t)_\infty$ combined with the closed graph of Φ and $t \mapsto \Phi(t)_\infty$ is then used to remove the unions, giving the strong formulation.

4.4.3 Moreau's product rule

Moreau's product rule is a rule for obtaining the measure differential of the product of two functions of bounded variation. This generalizes the usual rule for absolutely continuous

functions:

$$\frac{d}{dt}(u\,v) = \frac{du}{dt}\,v + u\,\frac{dv}{dt}.$$

However, if u and v are functions of bounded variation, then "du/dt" and "dv/dt" are measures with impulses at the jumps of u and v, respectively. The trouble then is that the "$u(t)$" may not be defined at a particular time t, where there is a jump in u. However, one-sided limits $u(t^{\pm})$ and $v(t^{\pm})$ are well defined. The following rule applies to a wide variety of situations in both finite and infinite dimensions.

Lemma 4.10. *Suppose $\beta\colon X \times Y \to Z$ is a continuous bilinear form with X, Y, and Z Banach spaces, and with $u\colon [a,b] \to X$, $v\colon [a,b] \to Y$ functions of bounded variation; then $z(t) = \beta(u(t), v(t))$ is a function of bounded variation $[a,b] \to Z$ and*

$$dz = \beta(u^+, dv) + \beta(du, v^-)$$

in the sense of differential measures.

Proof. Consider a partition $P : a = t_0 < t_1 < \cdots < t_{N-1} < t_N = b$. Then

$$\begin{aligned} z(t_{i+1}) - z(t_i) &= \beta(u(t_{i+1}), v(t_{i+1})) - \beta(u(t_i), v(t_i)) \\ &= \beta(u(t_{i+1}), v(t_{i+1})) - \beta(u(t_{i+1}), v(t_i)) \\ &\qquad + \beta(u(t_{i+1}), v(t_i)) - \beta(u(t_i), v(t_i)) \\ &= \beta(u(t_{i+1}) - u(t_i), v(t_i)) + \beta(u(t_{i+1}), v(t_{i+1}) - v(t_i)), \end{aligned}$$

so

$$\|z(t_{i+1}) - z(t_i)\| \le \|\beta\| \big[\|u(t_{i+1}) - u(t_i)\|\, \|v(t_i)\| + \|u(t_{i+1})\|\, \|v(t_{i+1}) - v(t_i)\| \big].$$

Adding over $i = 0, 1, \ldots, N-1$ and taking the supremum over all such partitions P give

$$\bigvee_a^b z \le \|\beta\| \left[\|v\|_\infty \bigvee_a^b u + \|u\|_\infty \bigvee_a^b v \right],$$

so $z(\cdot)$ has bounded variation, and thus dz is a differential measure.

To show the product rule, consider a continuous function $\zeta\colon [a,b] \to Z'$; we will show that

$$\int_{[a,b]} \langle \zeta, dz \rangle = \int_{[a,b]} \big\langle \zeta, \beta(u^+, dv) + \beta(du, v^-) \big\rangle.$$

Such integrals can be approximated by Riemann–Stieltjes integrals: for any $\epsilon > 0$ there is a $\delta > 0$ where for any partition P with

$$|P| := \max_{i=0,1,\ldots,N-1} |t_{i+1} - t_i| \le \delta$$

we have

$$\left| \int_{[a,b]} \langle \zeta, dz \rangle - \sum_{i=0}^{N-1} \langle \zeta(t_i), z(t_{i+1}) - z(t_i) \rangle \right| \le \epsilon.$$

Now if we let $u_P(t) = u(t_{i+1})$ for $t_i \le t < t_{i+1}$, and $v_P(t) = v(t_i)$ for $t_i < t \le t_{i+1}$, then as $|P| \to 0$, $u_P(t) \to \lim_{s \downarrow t} u(t) = u(t^+)$; similarly $v_P(t) \to v(t^-)$ as $|P| \to 0$. Let $\zeta_P(t) = \zeta(t_i)$ for $t_i \le t < t_{i+1}$. Clearly $\zeta_P \to \zeta$ uniformly as $|P| \to 0$ by uniform continuity of ζ. On the other hand,

$$\sum_{i=0}^{N-1} \langle \zeta(t_i), z(t_{i+1}) - z(t_i) \rangle = \int_{[a,b]} \langle \zeta_P(t), \beta(u_P(t), dv(t)) + \beta(du(t), v_P(t)) \rangle .$$

By the dominated convergence theorem for general measures and pointwise convergence of

$$\begin{aligned} \langle \zeta_P(t), \beta(u_P(t), \cdot) \rangle &\to \big\langle \zeta(t), \beta(u^+(t), \cdot) \big\rangle \qquad \text{in } Y', \\ \langle \zeta_P(t), \beta(\cdot, v_P(t)) \rangle &\to \big\langle \zeta(t), \beta(\cdot, v^-(t)) \big\rangle \qquad \text{in } X', \end{aligned}$$

we have the limit

$$\left| \int_{[a,b]} \langle \zeta, dz \rangle - \int_{[a,b]} \big\langle \zeta, \beta(u^+, dv) + \beta(du, v^-) \big\rangle \right| \le \epsilon.$$

As $\epsilon > 0$ is arbitrary, we have the equality

$$\int_{[a,b]} \langle \zeta, dz \rangle = \int_{[a,b]} \big\langle \zeta, \beta(u^+, dv) + \beta(du, v^-) \big\rangle$$

for any continuous ζ. Thus

$$dz = \beta(u^+, dv) + \beta(du, v^-),$$

as desired. □

Note that we can reverse the roles of u and v to obtain the equivalent formula

$$dz = \beta(u^-, dv) + \beta(du, v^+).$$

In the case where $u = v$ and β is symmetric, we have

$$dz = \beta(u^- + u^+, du).$$

These equalities have a number of applications to impulsive differential equations, just as the standard product rule has many applications to smooth differential equations. We will see one application in the next section and another in the section on the existence of solutions for rigid-body dynamics with Coulomb friction.

4.4.4 MDIs and discontinuous sweeping processes

Consider the sweeping process governed by the differential inclusion

$$\frac{dx}{dt}(t) \in -N_{C(t)}(x(t^+)) + f(t, x(t)). \tag{4.33}$$

We deal with these problems in several steps. The first is to treat the basic problem

$$\frac{dx}{dt}(t) \in -N_{C(t)}(x(t^+)). \tag{4.34}$$

From this we can use a shifting technique to reduce a problem of the form

$$\frac{du}{dt}(t) \in -N_{C(t)}(u(t^+)) + f(t) \tag{4.35}$$

to a problem of the form (4.34). Perturbation bounds combined with a Picard-type iteration will bring us to the general problem (4.33).

A point about the formulation is that solutions can be discontinuous, as $C(t)$ can be discontinuous in t, often forcing the solution to jump. It is important that the right-hand side depends on the *postjump* state $u(t^+)$.

The crucial assumption is that

$$\delta_H(C(s), C(t)) \le r(t) - r(s) \qquad \text{for all } s < t,$$

where r is a nondecreasing function, which therefore has bounded variation. We assume that $r(\cdot)$ is right continuous, so that $r(s^+) = \lim_{t \downarrow s} r(t) = r(s)$ for all s. We define $C^+(s) = \{\lim_{k\to\infty} x_k \mid x_k \in C(t_k),\ t_k \downarrow s\}$ and assume that $C(\cdot)$ is also right continuous in the sense that

$$C(s) = C^+(s) \qquad \text{for all } s. \tag{4.36}$$

There are two main ways of analyzing systems like this. One is to use a time discretization (e.g., the "catching-up" algorithm), and the other is to use a regularization (e.g., the Yosida approximation). If we "freeze" $C(t)$, then this system is covered by the section on maximal monotone differential inclusions. However, since $C(t)$ changes, and discontinuously, this no longer applies. However, we can use this as a starting point for our analysis. Let $P : 0 = t_0 < t_1 < \cdots < t_N = T$ be a partition of the interval $[0, T]$. We define the piecewise constant $C_P(t) = C(t_i)$ for $t_i \le t < t_{i+1}$. The solution for this piecewise constant problem

$$\frac{du_P}{dt}(t) \in -N_{C_P(t)}(u(t^+)), \qquad u(t_0) = u_0 \in C(t_0)$$

is easily shown to be the result of the catching-up algorithm [174, 178]: $u_P(t) = u_i$ for $t_i \le t < t_{i+1}$, where

$$u_{i+1} = \Pi_{C(t_{i+1})}(u_i).$$

In Moreau [178], existence of a solution is shown by showing convergence of the approximate solutions u_P as P becomes more refined. Unlike the approach in the previous section, we need to make sure that P contains most of the jumps of $r(\cdot)$; that is, the catching-up algorithm needs to know when to jump. This is needed to ensure *uniform* convergence of the $u_P(\cdot)$, rather than just pointwise convergence. A basic geometric tool to analyze the results of refining the partition is the following lemma.

Lemma 4.11. *If C is a closed convex set in a Hilbert space X, then for any $x, y \in X$,*

$$\|x - \Pi_C(y)\|^2 \le \|x - y\|^2 + 2\,d(x, C)\,d(y, C).$$

Proof. Let $u = \Pi_C(x)$ and $v = \Pi_C(y)$. Then

$$\begin{aligned}
&\|x - \Pi_C(y)\|^2 - \|x - y\|^2 \\
&\quad = \langle x - v, x - v\rangle - \langle x - y, x - y\rangle \\
&\quad = \langle 2x - v - y, y - v\rangle \\
&\quad = 2\langle x - u, y - v\rangle + 2\langle u - v, y - v\rangle - \|y - v\|^2 \\
&\quad \le 2\langle x - u, y - v\rangle - \|y - v\|^2 \qquad (\text{as } u \in C) \\
&\quad \le 2\,\|x - u\|\,\|y - v\| \;=\; 2\,d(x, C)\,d(y, C),
\end{aligned}$$

as desired. □

Now if P' is a refinement of P (that is, $P \subset P'$), then we can bound the difference between u_P and $u_{P'}$ at certain times according to the following lemma from [174].

Lemma 4.12. *Let $P : 0 = t_0 < t_1 < \cdots < t_N = T$ and let P' be a refinement of P. For an interval $I_i = [t_i, t_{i+1})$, denote the intervals of P' contained in I_i by $J_1 = [t'_1, t'_2)$, $J_2 = [t'_2, t'_3)$, …, $J_m = [t'_m, t'_{m+1})$ and let $J_{m+1} = [t'_{m+1}, t'_{m+2})$ be the following interval of P' (if $i = N - 1$, set $J_{m+1} = \emptyset$). Then whenever $\sigma \in J_1$ and $\tau \in I_i \cup J_{m+1}$ with $\sigma \le \tau$ we have*

$$\|u_P(\tau) - u_{P'}(\tau)\|^2 - \|u_P(\sigma) - u_{P'}(\sigma)\|^2 \le 2\left(r(t_{i+1}^-) - r(t_i)\right)^2.$$

Proof. Let $x_i = u_P(t_i)$ and $y_j = u_{P'}(t)$ for $t \in J_j$. Then $y_{j+1} = \Pi_{C(t'_{j+1})}(y_j)$, and so by Lemma 4.11 we have

$$\|x_i - y_{j+1}\|^2 - \|x_i - y_j\|^2 \le 2\,d(x_i, C(t'_{j+1}))\,d(y_j, C(t'_{j+1})).$$

Adding over $j = 0, 1, \ldots, k-1$ ($k \le m$) gives

$$\|x_i - y_{k+1}\|^2 - \|x_i - y_1\|^2 \le 2\sum_{j=1}^{k-1} d(x_i, C(t'_{j+1}))\,d(y_j, C(t'_{j+1})).$$

But $x_i \in C(t_i) = C(t'_1)$, so $d(x_i, C(t'_{j+1})) \le \delta_H(C(t'_1), C(t'_{j+1})) \le r(t'_{j+1}) - r(t'_1) \le r(t_{i+1}) - r(t_i)$. On the other hand, since $y_j \in C(t'_j)$, we have

$$d(y_j, C(t'_{j+1})) \le \delta_H\left(C(t'_j), C(t'_{j+1})\right) \;\le\; r(t'_{j+1}) - r(t'_j).$$

Summing over $j = 1, 2, \ldots, k-1$ ($k \le m$) gives

$$\begin{aligned}
\|x_i - y_{k+1}\|^2 - \|x_i - y_1\|^2 &\le 2\sum_{j=1}^{k-1} d(x_i, C(t'_{j+1}))\,d(y_j, C(t'_{j+1})) \\
&\le 2\sum_{j=1}^{k-1} (r(t_{i+1}) - r(t_i))\left(r(t'_{j+1}) - r(t'_j)\right) \\
&\le 2\,(r(t_{i+1}) - r(t_i))^2.
\end{aligned}$$

Thus, for any $\sigma \in I_i$ and $\tau \in I_i$ with $\sigma \le \tau$, we have

$$\begin{aligned}\|u_P(\tau) - u_{P'}(\tau)\|^2 &\le \|u_P(\sigma) - u_{P'}(\sigma)\|^2 + 2\left(r(t'_m) - r(t_i)\right)^2 \\ &\le \|u_P(\sigma) - u_{P'}(\sigma)\|^2 + 2\left(r(t_{i+1}^-) - r(t_i)\right)^2.\end{aligned}$$

We now consider the case with $\tau \in J_{m+1}$. Now $x_{i+1} = \Pi_{C(t_{i+1})}(x_i)$ and $y_{m+1} = \Pi_{C(t'_{m+1})}(y_m)$, but $t'_{m+1} = t_{i+1}$, and projection onto a convex set is nonexpansive, so $\|x_{i+1} - y_{m+1}\| \le \|x_i - y_m\|$. Thus, for $\tau \in J_{m+1}$,

$$\|u_P(\tau) - u_{P'}(\tau)\|^2 \le \|u_P(\sigma) - u_{P'}(\sigma)\|^2 + 2\left(r(t_{i+1}^-) - r(t_i)\right)^2,$$

as desired. □

With this lemma done, we can show uniform convergence of the approximations u_P for suitable partitions P, and we can show that the limits are solutions.

Theorem 4.13. *Suppose $C\colon [0,T] \to \mathcal{P}(X)$, X is a Hilbert space with closed convex values that is continuous from the right in the sense of* (4.36)*, and*

$$\delta_H(C(s), C(t)) \le r(t) - r(s) \qquad \text{for all } t \ge s.$$

Then solutions exist for the sweeping process

$$\frac{du}{dt}(t) \in -N_{C(t)}(u(t^+)), \qquad u(0) = u_0$$

in the sense of MDIs. Furthermore, such a solution can be constructed by limits of approximate trajectories using the catching-up algorithm.

Proof. To complete the existence proof, we construct partitions P such that

$$\sum_{i=0}^{N-1} \left(r(t_{i+1}^-) - r(t_i)\right)^2$$

is arbitrarily small. Since

$$\sum_{i=0}^{N-1} \left(r(t_{i+1}^-) - r(t_i)\right) \le r(t_N) - r(t_0) \;=\; r(T) - r(0)$$

is bounded, it suffices to ensure that $\max_i \left(r(t_{i+1}^-) - r(t_i)\right)$ is sufficiently small. Choosing $\epsilon > 0$, we start with all jumps t with $r(t^+) - r(t^-) \ge \epsilon/2$ in P. Since $r(\cdot)$ has bounded variation, there is only a finite number of such points. Then it is possible to add points so that the total variation on each open interval (t_i, t_{i+1}) is less than $\epsilon/2$. Then $r(t_{i+1}) - r(t_i^-) \le \epsilon$ for all i. Then, from the bounds in Lemma 4.12, if P' is a refinement of P,

$$\|u_P(t) - u_{P'}(t)\| \le 2\,(r(T) - r(0))\,\epsilon.$$

Thus we can choose a uniformly convergent subsequence $u_k \to u$ where $u_k = u_{P_k}$ and P_{k+1} is a refinement of P_k for all k.

We now need to show that the limit $u(\cdot)$ indeed solves the sweeping process. First $u_P(t) \in C_P(t)$ for all partitions P in the sequence used to construct $u(\cdot)$. We can without loss of generality assume that $|P_k| \to 0$ as $k \to \infty$, where $|P| = \max_{i=0,1,\dots,N-1} t_{i+1} - t_i$. Taking pointwise limits, if $t \in \bigcup_k P_k$, we have $u(t) \in C(t)$. Then using right continuity of $u(\cdot)$ (being the uniform limit of right continuous functions) we have $u(t) \in C(t)$ for all t.

Now we wish to show that the MDI

$$\frac{du}{dt}(t) \in -N_{C(t)}(u(t^+))$$

holds. From the catching-up process, $u_P(\cdot)$ satisfies the MDI

$$\frac{du_P}{dt}(t) \in -N_{C_P(t)}(u_P(t^+)).$$

That is, if $\widetilde{u}_P(t) \in C_P(t)$ for all t, we have

$$0 \le \int_{[0,T]} \langle \widetilde{u}_P(t) - u_P(t), du_P(t) \rangle .$$

Now suppose $\widetilde{u}(t) \in C(t)$ for all t. Set $\widetilde{u}_P(t) = \Pi_{C_P(t)}(\widetilde{u}(t))$. For $t_i \le t < t_{i+1}$ in the partition P, we have

$$\begin{aligned} \|\widetilde{u}_P(t) - \widetilde{u}(t)\| = d(\widetilde{u}(t), C_P(t)) \;&=\; d(\widetilde{u}(t), C(t_i)) \\ &\le \delta_H(C(t_i), C(t)) \;\le\; r(t) - r(t_i) \;\le\; \epsilon \end{aligned}$$

for given $\epsilon > 0$ and $P = P_k$ with k sufficiently large. Thus $\widetilde{u}_P(\cdot) \to \widetilde{u}(t)$ *uniformly* as $P = P_k$ and $k \to \infty$.

Note that in a partition P, $\|u_P(t_{i+1}) - u_P(t_i)\| \le r(t_{i+1}) - r(t_i)$, and u_P is constant over intervals (t_i, t_{i+1}). Thus the variation of u_P is

$$\begin{aligned} \sum_{i=0}^{N-1} \|u_P(t_{i+1}) - u_P(t_i)\| &\le \sum_{i=0}^{N-1} \big[r(t_{i+1}) - r(t_i) \big] \\ &= r(t_N) - r(t_0) \;=\; r(T) - r(0). \end{aligned}$$

Thus the differential measures du_P are uniformly bounded.

Taking $P = P_k$ and $k \to \infty$ gives $\widetilde{u}_P - u_P \to \widetilde{u} - u$ uniformly. On the other hand, $du_P \rightharpoonup^* du$ weak* as measures because $u_P \to u$ pointwise. (Alternatively we could use Alaoglu's theorem to produce a weak* convergent subsequence.) Then

$$0 \le \int_{[0,T]} \langle \widetilde{u}_P(t) - u_P(t), du_P(t) \rangle \to \int_{[0,T]} \langle \widetilde{u}(t) - u(t), du(t) \rangle$$

for all $\widetilde{u}\colon [0,T] \to X$ with $\widetilde{u}(t) \in C(t)$ for all t. Thus the Radon–Nikodym derivative of du with respect to its variation $|du|$, which we denote by $du/|du|$, satisfies $du/|du|\,(t) \in -N_{C(t)}(u(t^+))$, and $u(\cdot)$ satisfies the MDI. □

Uniqueness can be shown via Moreau's product rule (Lemma 4.10): suppose that

$$\frac{du}{dt}(t) \in -N_{C(t)}(u(t^+)),$$
$$\frac{dv}{dt}(t) \in -N_{C(t)}(v(t^+)).$$

Let $w = u - v$. Then

$$\begin{aligned} d\,\|w\|^2 = d\,\langle w, w\rangle &= \left\langle w^+ + w^-, dw\right\rangle \\ &= 2\left\langle w^+, dw\right\rangle - \left\langle w^+ - w^-, dw\right\rangle. \end{aligned}$$

Note that $\left\langle w^+ - w^-, dw\right\rangle$ is a purely atomic measure, which can be written as

$$\left\langle w^+(t) - w^-(t), dw(t)\right\rangle = \sum_s \left\| w(s^+) - w(s^-)\right\|^2 \delta(t-s) \;\geq\; 0,$$

where the sum is taken over all s, where $w(s^+) \neq w(s^-)$. Thus

$$\begin{aligned} d\,\|w\|^2 &\leq 2\left\langle w^+, dw\right\rangle \\ &= 2\left\langle u^+ - v^+, du - dv\right\rangle \\ &= 2\left\langle u^+ - v^+, \frac{du}{|du|+|dv|} - \frac{dv}{|du|+|dv|}\right\rangle (|du|+|dv|), \end{aligned}$$

where $|du|$ is the variation measure of du, and $du/(|du|+|dv|)$ is the Radon–Nikodym derivative of the measure du with respect to $|du|+|dv|$. Now

$$du/(|du|+|dv|)(t) \in -N_{C(t)}(u(t^+)),$$

and similarly for v, so from the monotonicity of $x \mapsto N_{C(t)}(x)$, we have

$$d\,\|w\|^2 \leq 0.$$

That is, for $t \geq s$, $\|w(t)\| \leq \|w(s)\|$. If $u(0) = v(0)$, then $w(0) = 0$, and so $w(t) = 0$, and $u(t) = v(t)$, for all $t \geq 0$. Thus, given the initial conditions, the solution is unique. Furthermore, we have the fact that $u(0) \mapsto u(t)$ is a nonexpansive map for $t \geq 0$.

Sweeping processes of the form

$$\frac{du}{dt}(t) \in -N_{C(t)}(u(t^+)) + \psi(t), \qquad u(t_0) = u_0$$

have solutions for all $\psi \in L^1(0,T;X)$. These problems can be reduced to basic sweeping processes by setting

$$v(t) = u(t) - \int_{t_0}^{t} \psi(\tau)\,d\tau,$$
$$D(t) = C(t) - \int_{t_0}^{t} \psi(\tau)\,d\tau,$$
$$\rho(t) = r(t) + \int_{t_0}^{t} \|\psi(\tau)\|\,d\tau.$$

Then v solves the basic sweeping process

$$\frac{dv}{dt}(t) \in -N_{D(t)}(v(t^+)), \qquad v(t_0) = u_0,$$

and $\delta_H(D(s), D(t)) \le \rho(t) - \rho(s)$ for all $t \ge s$. Furthermore, we can construct $u(\cdot)$ from $v(\cdot)$.

This can be used to "bootstrap" the general problem (4.33):

$$\frac{du}{dt}(t) \in -N_{C(t)}(u(t^+)) + f(t, u(t)), \qquad u(t_0) = u_0$$

with $f(t,\cdot)$ Lipschitz with Lipschitz constant $k(t)$, $k(\cdot) \in L^1(0,T)$, and $f(\cdot,0) \in L^1(0,T;X)$. Set $u^{(0)}(t) = u_0$ for all t, and from this we can start a Picard-type iteration: Let $u^{(j+1)}$ be the solution of

$$\frac{du^{(j+1)}}{dt}(t) \in -N_{C(t)}(u^{(j+1)}(t^+)) + f(t, u^{(j)}(t)), \qquad u^{(j+1)}(t_0) = u_0.$$

The iteration map $u^{(j)} \mapsto u^{(j+1)}$ is a contraction map on a sufficiently small interval $[t_0, T]$. To show this, consider

$$\begin{aligned}
\frac{du}{dt}(t) &\in -N_{C(t)}(u(t^+)) + \phi(t), \qquad u(t_0) = u_0,\\
\frac{dv}{dt}(t) &\in -N_{C(t)}(v(t^+)) + \psi(t), \qquad u(t_0) = u_0.
\end{aligned}$$

Using the same techniques as used above to show uniqueness for the basic sweeping process, for $w = v - u$,

$$\begin{aligned}
d\,\|w\|^2(t) &\le 2\left\langle w(t^+), \psi(t) - \phi(t)\right\rangle dt \qquad \text{or}\\
\left(\left\|w(t^+)\right\| + \left\|w(t^-)\right\|\right) d\,\|w\|(t) &\le 2\left\|w(t^+)\right\|\,\|\phi(t) - \psi(t)\|\,dt.
\end{aligned}$$

Since $\left\|w(t^+)\right\| \le \left\|w(t^-)\right\|$ for all t, $d\,\|w\|(t) \le \|\phi(t) - \psi(t)\|\,dt$, and so

$$\|u(t) - v(t)\| \le \|u(t_0) - v(t_0)\| + \int_{t_0}^{t} \|\phi(t) - \psi(t)\|\,dt.$$

From this we can show that the Picard iteration above has Lipschitz constant $\int_{t_0}^{T} k(\tau)\,d\tau < 1$ for $T - t_0 > 0$ sufficiently small. By the contraction mapping principle on a sufficiently small interval, there is one and only one solution of (4.33). By continuation arguments it can be shown that there is one and only one solution on the interval $[t_0, \infty)$.

There are a number of generalizations of the idea of sweeping processes. Some of these deal with nonconvex $C(t)$, but they satisfy a bound on the nonconvexity so that the nearest point projection map $\Pi_{C(t)}$ is well defined and Lipschitz in a suitably small neighborhood of $C(t)$. Another approach is to allow operators more general than the normal cone operator, such as

$$\frac{du}{dt}(t) \in -\Phi(t,u) + f(t,u), \qquad u(t_0) = u_0,$$

where $\Phi(t,\cdot)$ is a maximal monotone operator. Care must be taken to make sure that $\Phi(t,u)$ does not vary "too much" with changes in t. Systems of this kind and generalizations are considered in [170, 234].

4.5 Linear complementarity systems

Linear complementarity systems (LCSs) have already been discussed (3.8)–(3.10). We first review their formulation:

$$\frac{dx}{dt}(t) = Ax(t) + Bz(t) + Ef(t), \qquad x(0) = x_0, \tag{4.37}$$

$$w(t) = Cx(t) + Dz(t) + Ff(t), \tag{4.38}$$

$$0 \le w(t) \perp z(t) \ge 0 \qquad \text{for almost all } t. \tag{4.39}$$

The function $f(t)$ is an external input. LCSs were introduced by Çamlıbel, Heemels, Schumacher, Weiland, and van der Schaft in a series of papers [48, 123, 124, 264]. Their theory is mainly based on the use of *Laplace transforms*:

$$\mathcal{L}f(s) = \int_0^\infty e^{-st} f(t)\,dt. \tag{4.40}$$

If this is applied to (4.37)–(4.38), we get

$$\begin{aligned} s\,\mathcal{L}x(s) - x_0 &= A\,\mathcal{L}x(s) + B\,\mathcal{L}z(s) + E\,\mathcal{L}f(s), \\ \mathcal{L}w(s) &= C\,\mathcal{L}x(s) + D\,\mathcal{L}z(s) + F\,\mathcal{L}f(s). \end{aligned}$$

Solving for $\mathcal{L}x(s)$ we now have

$$\begin{aligned} \mathcal{L}w(s) = \Big[D + C\,(s\,I - A)^{-1}\,B\Big]\mathcal{L}z(s) \\ + C\Big[(s\,I - A)^{-1}\,x_0 + \Big(F + (s\,I - A)^{-1}\,E\Big)\,\mathcal{L}f(s)\Big]. \end{aligned}$$

If we consider the complementarity condition (4.39) componentwise, we see that $0 \le w_i(t) \perp z_i(t) \ge 0$ for almost all t and all i. If we seek solutions that are piecewise analytic, then on any piece $t_1 \le t \le t_2$ we have either $y_i(t) \equiv 0$ or $u_i(t) \equiv 0$. The conditions "$z_i(t) \ge 0$" and "$w_i(t) \ge 0$" are the most problematic from this point of view. However, the small t behavior of $z_i(t)$ and $w_i(t)$ is closely related to the large s behavior of $\mathcal{L}z_i(s)$ and $\mathcal{L}w_i(s)$, respectively. If $w_i(t) > 0$ for an interval $[0, \epsilon)$ with $\epsilon > 0$ and grows no faster than exponentially, then for sufficiently large $s > 0$, $\mathcal{L}w_i(s) > 0$. This idea leads to the following definition: $f : \mathbb{R}_+ \to \mathbb{R}$ is *initially positive* if and only if $\mathcal{L}f(s) > 0$ for all sufficiently large $s > 0$. We say $f : \mathbb{R}_+ \to \mathbb{R}$ is *initially nonnegative* if f is initially positive, or it is *initially zero* (that is, $f(t) = 0$ for all $t \in [0, \epsilon')$ for some $\epsilon' > 0$). A better justification for the definition of "initially positive" can be found in the following lemma.

Lemma 4.14. *If $f : \mathbb{R}_+ \to \mathbb{R}$ grows no faster than exponentially and it is analytic on $[0, \epsilon)$ for $\epsilon > 0$, and is initially positive (in the sense just described), then there is an $\epsilon' > 0$ such that $f(t) > 0$ for $0 < t < \epsilon'$.*

Proof. First we introduce a definition: a vector $[a_1, a_2, a_3, \ldots, a_p]$ is *lexicographically positive* if there is a $1 \le j \le p$ such that $a_j > 0$ and $a_i = 0$ for all $i < j$. In other words, a vector is lexicographically positive if the first nonzero entry is positive.

For f analytic on $[0, \epsilon)$, there is a $0 < \rho < \epsilon$ such that f has a Taylor series expansion:

$$f(t) = f(0) + f'(0)\,t + f''(0)\,\frac{t^2}{2!} + f'''(0)\,\frac{t^3}{3!} + \cdots$$

for $0 \le t < \rho$. The Laplace transform is asymptotically

$$\mathcal{L}f(s) = f(0)s^{-1} + f'(0)s^{-2} + f''(0)s^{-3} + f'''(0)s^{-4} + \cdots + \mathcal{O}(s^{-m-1})$$

as $s \to +\infty$ on the real axis. Now $\mathcal{L}f(s) > 0$ for sufficiently large $s > 0$ if and only if the vector $[f(0), f'(0), f''(0), f'''(0), \ldots]$ is lexicographically positive. Suppose the first nonzero entry is $f^{(k)}(0)$. Then, for $0 \le t < \rho$,

$$\begin{aligned} f(t) &= f^{(k)}(0)\frac{t^k}{k!} + f^{(k+1)}(0)\frac{t^{k+1}}{(k+1)!} + f^{(k+2)}(0)\frac{t^{k+2}}{(k+2)!} + \cdots \\ &= f^{(k)}(0)\frac{t^k}{k!} + \mathcal{O}(t^{k+1}). \end{aligned}$$

If $f^{(k)}(0) > 0$, then $f(t) > 0$ for all $0 < t < \epsilon'$ for some $\epsilon' > 0$. □

The idea now is to look for solutions of the Laplace transformed CP:

$$\begin{aligned} \mathcal{L}w(s) = &\Big[D + C\,(s\,I - A)^{-1}\,B\Big]\mathcal{L}z(s) \\ &+ C\Big[(s\,I - A)^{-1}\,x_0 + \Big(F + (s\,I - A)^{-1}\,E\Big)\,\mathcal{L}f(s)\Big], \\ 0 \le \mathcal{L}w(s) \perp \mathcal{L}z(s) \ge 0 \qquad &\text{for sufficiently large } s > 0. \end{aligned}$$

Provided f is a Bohl distribution (3.23), we assume that all the Laplace transforms $\mathcal{L}w(s)$, $\mathcal{L}z(s)$, and $\mathcal{L}f(s)$ are rational functions of s, at least for considering the short-time behavior of the solutions. This essentially assumes that there is a time interval $[0, \epsilon)$, $\epsilon > 0$, on which the active sets $\{i = 1, \ldots, n \mid w_i(t) = 0\}$ and $\{i = 1, \ldots, n \mid z_i(t) = 0\}$ do not change. To simplify the expressions, let

$$\begin{aligned} G(s) &= D + C\,(s\,I - A)^{-1}\,B, \\ \mathcal{L}q(s) &= C\left[(s\,I - A)^{-1}\,x_0 + \left(F + (s\,I - A)^{-1}\,E\right)\mathcal{L}f(s)\right]. \end{aligned}$$

This gives us the *rational complementarity problem* (RCP) [123]: Given $G(s)$ and $\mathcal{L}q(s)$ rational functions, find $\mathcal{L}z(s)$ and $\mathcal{L}w(s)$ rational functions of s so that

$$\mathcal{L}w(s) = G(s)\mathcal{L}z(s) + \mathcal{L}q(s), \tag{4.41}$$

$$0 \le \mathcal{L}w(s) \perp \mathcal{L}z(s) \ge 0 \qquad \text{for sufficiently large } s > 0. \tag{4.42}$$

Since we are interested in the behavior of $G(s)$ for large $s > 0$, we use a Laurent series expansion

$$G(s) = G_0 + G_1 s^{-1} + G_2 s^{-2} + G_3 s^{-3} + \cdots.$$

The matrices G_i can be computed explicitly in terms of A, B, C, and D using the formula for $G(s)$:

$$\begin{aligned} G_0 &= D, \\ G_1 &= CB, \\ G_2 &= CAB, \\ G_3 &= CA^2B, \text{ etc.} \end{aligned}$$

The *index* of the LCS (4.37)–(4.39) is the smallest value of k such that $G_0 + G_1 s^{-1} + \cdots + G_k s^{-k}$ is nonsingular for sufficiently large $s > 0$. Thus, if D is nonsingular, then the index is zero. If $D + CBs^{-1}$ is nonsingular for sufficiently large $s > 0$, then the index is one.

The existence theory of the RCP (4.41)–(4.42) can be developed using standard linear complementarity theory.

Theorem 4.15. *Suppose that $G(s)$ is a P-matrix for all sufficiently large $s > 0$ and that $G(s)$, $\mathcal{L}f(s)$ are real rational functions of s. Then there is a solution $\mathcal{L}w(s)$ and $\mathcal{L}z(s)$ to* (4.41)–(4.42) *with both $\mathcal{L}w(s)$ and $\mathcal{L}z(s)$ rational functions of s.*

Proof. Consider a sequence of real numbers $s_k \to \infty$ as $k \to \infty$. Since $G(s)$ is a P-matrix for sufficiently large $s > 0$, for sufficiently large k there is a unique solution to LCP($\mathcal{L}q(s_k)$, $G(s_k)$); denote it by $\widehat{z}^{(k)}$ so that

$$\begin{aligned} \widehat{w}^{(k)} &= G(s_k)\widehat{z}^{(k)} + \mathcal{L}q(s_k), \\ 0 \le \widehat{w}^{(k)} &\perp \widehat{z}^{(k)} \ge 0. \end{aligned}$$

Let the active set $I_k = \{\, i \mid \widehat{z}_i^{(k)} > 0 \,\}$. Since every active set is a subset of $\{1, 2, \ldots, n\}$, there are only finitely many possible values for I_k. At least one will be repeated infinitely often: Suppose $I_k = J$ for infinitely many k. Let $\overline{J}$ be the complement of J: $\overline{J} = \{1, 2, \ldots, n\} \backslash J$. Splitting the equation $\widehat{w}^{(k)} = G(s_k)\widehat{z}^{(k)} + \mathcal{L}q(s_k)$ into components in J and in $\overline{J}$, we get

$$\begin{aligned} 0 &= G(s_k)_{JJ}\widehat{z}_J^{(k)} + \mathcal{L}q(s_k)_J, \qquad & \widehat{z}_{\overline{J}}^{(k)} &= 0, \\ \widehat{w}_{\overline{J}}^{(k)} &= G(s_k)_{\overline{J}J}\widehat{z}_J^{(k)} + \mathcal{L}q(s_k)_{\overline{J}}, \qquad & \widehat{w}_J^{(k)} &= 0. \end{aligned}$$

Since $G(s_k)$ is a P-matrix, all principal submatrices are invertible, so we have

$$\begin{aligned} \widehat{z}_J^{(k)} &= -G(s_k)_{JJ}^{-1}\,\mathcal{L}q(s_k)_J \ge 0, \\ \widehat{w}_J^{(k)} &= \mathcal{L}q(s_k)_{\overline{J}} - G(s_k)_{\overline{J}J}G(s_k)_{JJ}^{-1}\,\mathcal{L}q(s_k)_J \ge 0. \end{aligned}$$

If we write $\widehat{z}_J(s) = -G(s)_{JJ}^{-1}\,\mathcal{L}q(s)_J$ and $\widehat{w}_{\overline{J}}(s) = \mathcal{L}q(s)_{\overline{J}} - G(s)_{\overline{J}J}G(s)_{JJ}^{-1}\,\mathcal{L}q(s)_J$, we have two rational functions of s that are nonnegative for infinitely many $s_k \to \infty$. Since $G(s)$ and $\mathcal{L}q(s)$ are rational functions of s, so are $\widehat{u}_J(s)$ and $\widehat{y}_{\overline{J}}(s)$. Combining these facts shows that for all sufficiently large $s > 0$, $\widehat{z}_J(s) \ge 0$ and $\widehat{w}_J(s) \ge 0$. Since for sufficiently large $s > 0$, $G(s)$ is a P-matrix, LCP($\mathcal{L}q(s)$, $G(s)$) has a unique solution, which must therefore be given by $\widehat{z}_J(s)$ and $\widehat{w}_{\overline{J}}(s)$. This choice of active set gives the solution of the RCP (4.41)–(4.42). □

Note that the fact that $G(s)$ and $\mathcal{L}q(s)$ are rational functions of s is used in showing that $\widehat{z}_J(s_k) \ge 0$, $\widehat{w}_{\overline{J}}(s_k) \ge 0$ for infinitely many $s_k \to +\infty$ implies that $\widehat{z}_J(s) \ge 0$, $\widehat{w}_{\overline{J}}(s) \ge 0$ for all sufficiently large $s > 0$. This is also true if $G(s)$ and $\mathcal{L}q(s)$ are "meromorphic at infinity"; that is, $G(1/z)$ and $\mathcal{L}q(1/z)$ are analytic functions in a neighborhood of $z = 0$ and have a finite order pole at $z = 0$.

Another fact about this result is that not only is the solution unique, but the active set $I(s) = \{\, i \mid z_i(s) > 0 \,\}$ is also constant for sufficiently large $s > 0$.

With the solution of the RCP, we can identify the active set and hence find the active set for a suitably small time interval, at least provided $\widehat{z}(s) = \mathcal{L}z(s) = \mathcal{O}(1/s)$ as $s \to \infty$.

Slower decay indicates that the solution includes Dirac-δ functions or its derivatives, and so it cannot be analytic on $[0,\epsilon)$ for any $\epsilon > 0$. If $\mathcal{L}z(s) = \mathcal{O}(1)$ as $s \to +\infty$, then no derivatives of Dirac-δ functions can appear in the solution. Since $\widehat{z}(s) = \mathcal{L}z(s) \geq 0$, the strength of a Dirac-δ function in the solution must be greater than or equal to 0. If the strength of the δ-function is zero, then $\mathcal{L}z(s) = \mathcal{O}(1/s)$ and everything works out. So let's consider what happens if the strength of the δ-function is positive. It is then possible that immediately after the δ-function, the solution might go negative. For example, if $z(t) = \delta(t) - 1$, $\mathcal{L}z(s) = 1 - 1/s > 0$ for $s > 1$. Now consider the LCS

$$\begin{aligned} \frac{d^2x}{dt^2} &= z(t) + 1, \qquad x(0) = 0, \qquad \frac{dx}{dt}(0) = -1, \\ w(t) &= x(t), \\ 0 \leq w(t) \perp z(t) &\geq 0 \qquad \text{for all } t. \end{aligned}$$

This gives the RCP

$$0 \leq \mathcal{L}w(s) = \frac{1}{s^2}\mathcal{L}z(s) - \frac{1}{s^2} + \frac{1}{s^3} \perp \mathcal{L}z(s) \geq 0.$$

The solution to this is $\mathcal{L}z(s) = 1 - 1/s$, so that $z(t) = \delta(t) - 1$. Immediately after the δ-function, we need to formulate and solve a new RCP for different initial conditions for $x(\cdot)$: $x(0^+) = 0$, $dx/dt(0^+) = 0$. This gives the RCP

$$0 \leq \mathcal{L}y^+(s) = \frac{1}{s^2}\mathcal{L}u^+(s) + \frac{1}{s^3} \perp \mathcal{L}u^+(s) \geq 0,$$

which has the solution $\mathcal{L}z^+(s) = 0$, $\mathcal{L}w^+(s) = 1/s^3$. That is, for immediately after the δ-function, we have $z(t) \equiv 0$. In fact, the solution for all times is $z(t) = \delta(t)$, and $w(t) = x(t) = \frac{1}{2}t^2$ for $t > 0$.

This process of restarting the problem with new initial conditions when the RCP indicates the presence of a δ-function or one of its derivatives can be repeated if necessary until an analytic function is obtained.

This RCP approach to LCSs can be extended to problems with infinite-dimensional dynamics with A a bounded linear operator $X \to X$ (X a Banach space), as long as the complementarity conditions apply in finite dimensions. That is, we require that $B\colon \mathbb{R}^m \to X$ and $C\colon X \to \mathbb{R}^m$. Then $G(s) = D + C(sI - A)^{-1}B$ is analytic for $|s| > \rho(A)$, where $\rho(A)$ is the *spectral radius* of A.[5] For $|s| > \|A\|$ we have

$$\begin{aligned} (sI - A)^{-1} &= s^{-1}(I - A/s)^{-1} \\ &= s^{-1}I + s^{-2}A + s^{-3}A^2 + \cdots, \end{aligned}$$

so $G(s)$ is analytic at $s = \infty$. Consider the following example:

$$\begin{aligned} \frac{dx_i}{dt}(t) &= x_{i+1}(t) + b_i z(t), \qquad x_i(0) = x_i^0, \\ w(t) &= \sum_{i=1}^{\infty} c_i x_i(t) + d z(t), \\ 0 \leq w(t) \perp z(t) &\geq 0 \qquad \text{for all } t. \end{aligned}$$

[5]The spectral radius $\rho(A)$ of a linear operator $X \to X$ is the supremum of $|\lambda|$ over λ in the spectrum of A: the spectrum of A is the set of all $\lambda \in \mathbb{C}$ where $\lambda I - A$ is not invertible. For $X = \mathbb{R}^n$, $\rho(A)$ is simply the maximum absolute value of the eigenvalues of A.

We assume that $X = \ell^2$, the space of square summable sequences:

$$\ell^2 = \left\{ (x_1, x_2, x_3, \ldots) \mid \sum_{i=1}^{\infty} |x_i|^2 < \infty \right\}.$$

We will assume that $b = (b_1, b_2, b_3, \ldots)$ and $c = (c_1, c_2, c_3, \ldots) \in \ell^2$. The matrix for A is

$$A = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & 0 & 1 & \\ & & & 0 & \ddots \\ & & & & \ddots \end{bmatrix},$$

which is the well-known shift operator on ℓ^2. The resolvent $(sI - A)^{-1}$ can be computed explicitly: solving $(sI - A)z = w$ for $z \in \ell^2$ and $w \in \ell^2$ can be done as follows. For each $i \geq 1$ we have $s\,z_i - z_{i+1} = w_i$. If we write z_{i+1} in terms of z_i, we expect to see exponential growth for $|s| > 1$. So we work in the reverse direction: $z_i = (z_{i+1} + w_i)/s$. This means we can write z_i as an *infinite* sum $z_i = \sum_{k=0}^{\infty} s^{-k} w_{i+k}$. This gives

$$(sI - A)^{-1} = \begin{bmatrix} s^{-1} & s^{-2} & s^{-3} & s^{-4} & \cdots \\ & s^{-1} & s^{-2} & s^{-3} & \cdots \\ & & s^{-1} & s^{-2} & \cdots \\ & & & s^{-1} & \cdots \\ & & & & \ddots \end{bmatrix},$$

which is a bounded operator $\ell^2 \to \ell^2$ for $|s| > 1$.Then we can compute

$$G(s) = d + \sum_{k=1}^{\infty} s^{-k} \sum_{i=1}^{\infty} c_i b_{i+k},$$

which gives us the Laurent series for $G(s)$ at infinity. Solutions thus exist if for $s > 0$ sufficiently large, $d + s^{-1} \sum_{i=1}^{\infty} c_i b_{i+1} > 0$.

4.6 Convolution complementarity problems

Convolution complementarity problems (CCPs) introduce a different kind of dynamics and have the following form: Given $m\colon [0,T] \to \mathbb{R}^{n\times n}$, $q\colon [0,T] \to \mathbb{R}^n$, and a closed convex cone $K \subseteq \mathbb{R}^n$ where $q(0) \in K^*$, find $z\colon [0,T] \to \mathbb{R}^n$ such that

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^* \qquad \text{for almost all } t, \tag{4.43}$$

$$(k * z)(t) = \int_0^T k(t-\tau)\, z(\tau)\, d\tau. \tag{4.44}$$

This can be used to represent the LCS (3.8)–(3.10) by setting $k(t) = D\,\delta(t) + C\,e^{At}\,B$ and $q(t) = C\,e^{At}\,x_0$ for $t \geq 0$. The index of the CCP is the smallest index r such that the

distributional derivative $m^{(r)}(t)$ has a Dirac-δ function at $t = 0$. This definition is equivalent to the index described above for DVIs. However, r need not be an integer; we can also use *fractional derivatives* [145] (see Chapter 7).

CCPs can be used to solve certain problems involving partial differential equations with complementarity conditions, as long as the dynamics are linear and time invariant, and the complementarity conditions are finite dimensional, such as impact of a rod at one end.

4.6.1 Index-zero CCPs

First, we will consider index-zero problems. In particular, suppose that $m(t) = m_0\,\delta(t) + m_1(t)$, where m_1 is an integrable function. Then (4.43) becomes

$$K \ni z(t) \perp m_0 z(t) + (m_1 * z)(t) + q(t) \in K^* \qquad \text{for almost all } t.$$

This can be solved by means of a Picard iteration: Given $z^{(0)}$, compute $z^{(1)}$, $z^{(2)}$, ... by the iteration

$$K \ni z^{(k+1)}(t) \perp m_0 z^{(k+1)}(t) + \left(m_1 * z^{(k)}\right)(t) + q(t) \in K^* \tag{4.45}$$

for almost all t. Then we can transform the problem if m_0 has the uniqueness property for the GCP

$$K \ni z \perp m_0 z + q \in K^*. \tag{4.46}$$

Let us assume that m_0 is a strongly monotone matrix (that is, $m_0 + m_0^T$ is positive definite). Let $q \mapsto z = \mathrm{sol}_{K,m_0}(q)$ be the solution operator for the static problem (4.46). If m_0 is a strictly monotone matrix, then sol_{K,m_0} is a Lipschitz function with Lipschitz constant $1/\lambda_{\min}(\frac{1}{2}(m_0 + m_0^T))$. Then the Picard iteration (4.45) leads to

$$z^{(k+1)}(t) = \mathrm{sol}_{K,m_0}\left(\left(m_1 * z^{(k)}\right)(t) + q(t)\right) \qquad \text{for almost all } t.$$

The operator $z(\cdot) \mapsto y(\cdot)$, where $y(t) = \mathrm{sol}_{K,m_0}((m_1 * z)(t) + q(t))$, is a Lipschitz operator $C[0,T] \to C[0,T]$ with Lipschitz constant

$$\int_0^T \|m_1(t)\|\, dt / \lambda_{\min}\left(\frac{1}{2}(m_0 + m_0^T)\right).$$

For sufficiently small $T > 0$, $\int_0^T \|m_1(t)\|\, dt/\lambda_{\min}(\frac{1}{2}(m_0 + m_0^T)) < 1$, and so we have a contraction map. (We will choose $T > 0$ so that $\int_0^T \|m_1(t)\|\, dt \le \frac{1}{2}\lambda_{\min}(\frac{1}{2}(m_0 + m_0^T))$.) Then by the contraction mapping theorem there is a unique fixed point; thus there is a unique fixed point in $C[0,T]$ and one and only one solution in $C[0,T]$.

The solution can be extended to $[T, 2T]$ by means of a "shift" technique: let $s = t - T$ for $T \le t \le 2T$; then put $q_1(s) = q(T+s) + \int_0^T m_1(T + s - \tau) z(\tau)\, d\tau$. Then using the techniques of the previous paragraph we have a solution of

$$K \ni z_1(s) \perp m_0 z_1(s) + (m_1 * z_1)(s) + q_1(s) \in K^* \qquad \text{for } 0 \le s \le T.$$

Then we set $z(t) = z_1(t - T)$ for $T \le t \le 2T$. The process can be repeated indefinitely, so there is a unique solution $z(t)$ for $t \ge 0$.

There is a point that should be understood about the connection between the matrix m_0 and the cone K. If m_0 is strongly monotone, then there are no restrictions on what the closed convex cone K can be. However, if we restrict K, then more general matrices m_0 are allowed. The crucial condition is that the solution operator sol_{K,m_0} for the static problem (4.46) needs to be Lipschitz. This is the case if, for example, $K = \mathbb{R}^n_+$ and m_0 is a P-matrix, which can be very far from being strongly monotone.

4.6.2 Index-one CCPs

Index-one CCPs have the following form: Given $m(\cdot)$ and $q(\cdot)$, find $z(\cdot)$, where

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^* \qquad \text{for almost all } t,$$

where $m(\cdot)$ is an integrable function and $m(0^+)$ is a suitable nonsingular matrix. It turns out that we need significantly stronger conditions for existence and uniqueness of solutions to index-one problems than for index-zero problems. The question of uniqueness has some subtleties to it that has led to some apparently simple open questions [245].

The simplest example of an index-one CCP is to have $m(t) = m_0$ for all $t > 0$. Then $(m * z)(t) = m_0 \int_0^t z(\tau)\,d\tau$. Let $w(t) = m_0 \int_0^t z(\tau)\,d\tau + q(t)$ so that $dw/dt(t) = m_0 z(t) + q'(t)$. Then we can turn this problem into a DCP:

$$\begin{aligned} \frac{dw}{dt} &= m_0 z(t) + q'(t), \qquad w(0) = q(0), \\ K \ni z(t) &\perp w(t) \in K^* \qquad \text{for almost all } t. \end{aligned}$$

Typically we require that q' be integrable (so that q is absolutely continuous), and of course, $q(0) \in K^*$. In general, we can show that if $q \in W^{1,p}(0,T;\mathbb{R}^n)$ and $q(0) \in K^*$, then the solution $z \in L^p(0,T;\mathbb{R}^n)$. We show existence and uniqueness of solutions to index-one CCPs in Section 4.6.2 under suitable conditions.

4.6.3 Index-two and higher-index CCPs

Index-two CCPs involve kernel functions $m(t)$ which are asymptotically $m(t) \sim m_0 t$ as $t \downarrow 0$. Such problems can be considered as a starting point for understanding simple impact problems such as

$$\begin{aligned} \frac{d^2x}{dt^2}(t) &= f(t) + N(t), \qquad x(0),\ \frac{dx}{dt}(0) \quad \text{given} \\ 0 \le x(t) &\perp N(t) \ge 0 \qquad \text{for all } t. \end{aligned}$$

As might be expected for such a problem, the solution $N(t)$ is typically a measure, as it can contain impulses. Furthermore, solutions might exist, but they cannot be expected to have unique solutions when impact occurs. After all, for rigid-body dynamics we need a *coefficient of restitution* to determine the velocities immediately after impact. Indeed, existence can be proved for the CCP: Given $m\colon [0,T] \to \mathbb{R}^{n\times n}$ and $q\colon [0,T] \to \mathbb{R}^n$, find $z\colon [0,T] \to \mathbb{R}^n$ satisfying

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^* \qquad \text{for all } t,$$

provided $m(0^+) = 0$, $m'(0^+) = m_0$, m'' in $L^\infty(0,T;\mathbb{R}^{n\times n})$, and q'' is in $L^p(0,T;\mathbb{R}^n)$ and $q(0) \in K^*$.

4.6.4 Fractional index problems

Fractional index problems do exist, but the dynamics are not generated by ordinary differential equations. These are most easily described in terms of CCPs. These are CCPs where $m(t) \sim t^{\alpha-1} m_0/\Gamma(\alpha)$ as $t \downarrow 0$, where m_0 is, say, strongly monotone, and α is not an integer. Existence of solutions can be shown for $0 < \alpha < 1$ and $1 < \alpha < 2$ by means of index reduction and the differentiation lemmas (see Section 3.4). Uniqueness can be shown for $0 < \alpha < 1$ if m_0 is symmetric positive definite, but not for $1 < \alpha < 2$ at the time of this writing.

The first CCP to appear as such in the literature is of this kind and is due to Petrov and Schatzman [207]. They obtained a CCP of this kind with $\alpha = 1\frac{1}{2}$ from studying the impact problem for a viscoelastic rod (impact occurs at $x = 0$):

$$\begin{aligned} u_{tt} &= u_{xx} + \beta u_{txx}, \qquad t > 0,\ 0 < x < L, \\ 0 &= u_x(t,L) + \beta u_{tx}(t,L), \\ N(t) &= u_x(t,0) + \beta u_{tx}(t,0), \\ 0 \le N(t) &\perp u(t,0) \ge 0. \end{aligned}$$

By means of constructing a Green's function or fundamental solution for this problem, they found

$$u(t,0) = \int_0^t m(t-\tau) N(\tau) d\tau + q(\tau),$$

where $m(t) \sim \text{const} t^{1/2}$ as $t \downarrow 0$ and const is a positive constant. The function $q(\cdot)$ is obtained from the initial values $u(0,x)$ and $u_t(0,x)$ for $0 < x < L$. Existence of solutions can be shown either by the original techniques of [207] or by more recent techniques [243]. Furthermore, it can be shown that the contact forces do no work, and therefore the energy loss can be accounted for from the viscous term βu_{txx} alone.

Other fractional index CPs can be found in [249], which also develops the theory of such problems for $0 < \alpha < 1$.

4.7 Parabolic variational inequalities

Parabolic variational inequalities (PVIs) are VIs involving the first derivative of the unknown functions with respect to time which have the form

$$u(t) \in K, \tag{4.47}$$

$$0 \le \left\langle \widetilde{u} - u(t), \frac{du}{dt}(t) - f(u(t)) \right\rangle \qquad \text{for all } \widetilde{u} \in K, \tag{4.48}$$

where K is a closed convex set in a Banach space X. Discussions of PVIs can be found in [25, 189]. PVIs can be represented as differential inclusions or as DVIs.

An example of a PVI is the oxygen uptake in a biological tissue, described in Section 1.4.3. Oxygen diffuses through a tissue which absorbs the oxygen at a fixed rate as long as it is available. But if the oxygen concentration drops to zero, the cells in the tissue are assumed to go into hibernation and resume normal uptake when the oxygen concentration becomes positive again. The main variable is $u(t,\mathbf{x})$ being the concentration of oxygen

at point $\mathbf{x}$ at time t. Where the oxygen concentration is positive, the following reaction-diffusion equation holds:

$$\frac{\partial u}{\partial t} = \nabla \cdot (D \nabla u) - r_{\max}.$$

The set $K = \left\{ u \in H^1(\Omega) \mid u \geq 0 \right\}$ represents nonnegative oxygen concentrations. When $u(t,\mathbf{x}) = 0$, this partial differential equation becomes inoperative. Instead we simply ensure that $u(t,\mathbf{x}) \geq 0$; that is, the oxygen concentration does not become negative. This can be represented by the VI

$$u(t,\mathbf{x}) \geq 0 \quad \& \quad 0 \leq (\widetilde{u} - u(t,\mathbf{x})) \cdot \left[\frac{\partial u}{\partial t} - \nabla \cdot (D \nabla u) + r_{\max} \right] \qquad \text{for all } \widetilde{u} \geq 0$$

for all t and $\mathbf{x}$. Alternatively, this can be represented by the complementarity formulation

$$\begin{gathered} \frac{\partial u}{\partial t} = \nabla \cdot (D \nabla u) - r_{\max} + s(t,\mathbf{x}), \\ 0 \leq s(t,\mathbf{x}) \perp u(t,\mathbf{x}) \geq 0 \qquad \text{for all } t \text{ and } \mathbf{x}. \end{gathered}$$

4.7.1 Comparison with maximal monotone differential inclusions

To represent (4.47)–(4.48) as a differential inclusion, note that (4.48) is equivalent to

$$0 \in \frac{du}{dt} - f(u(t)) + N_K(u(t)) \tag{4.49}$$

by (2.41). The representation (4.49) leads directly to existence and uniqueness results via the theory of maximal monotone differential inclusions (Section 4.2) for Lipschitz f since N_K is a maximal monotone operator. We just need $u(0) \in K$.

In Gelfand triples, we need to be a little careful in how we apply the theory of maximal monotone differential inclusions, as we need to ensure that the operator $-f + N_K \colon X \to \mathcal{P}(X')$ in a Gelfand triple $X \subset H = H' \subset X'$ can be turned into a maximal monotone plus Lipschitz operator $(-f + N_K)_H \colon H \to \mathcal{P}(H)$ as we identify $H = H'$. Recall that for $\Phi \colon X \to \mathcal{P}(X')$, $\Phi_H(u) = \Phi(u) \cap H$ for $u \in X$ and $\Phi(u) = \emptyset$ otherwise (4.19). For example, if we take $X = H^1(\Omega)$ and $H = L^2(\Omega)$ for a domain $\Omega \subset \mathbb{R}^d$, we can set $K = \{ w \in X \mid w(x) \geq \varphi(x), \text{ for all } x \in \Gamma \}$ for a suitable subset $\Gamma \subset \Omega$ and $\varphi \colon \Gamma \to \mathbb{R}$. Then we can take $f \colon H^1(\Omega) \to H^{-1}(\Omega)$ to be $f(u) = \pm \nabla^2 u$. Then $f(u) = +\nabla^2 u$ leads to maximal monotone $L^2(\Omega) \to \mathcal{P}(L^2(\Omega))$ by Lemma 4.7, while $f(u) = -\nabla^2 u$ does not.

4.7.2 Comparison with DVIs

To turn (4.49) into a DVI we note that $z(t) \in N_K(u(t))$ just means $z(t) \in \partial I_K(u(t))$, and by the Fenchel duality theorem (Theorem B.15), this is equivalent to $u(t) \in \partial \left(I_K^* \right)(z(t)) = \partial \sigma_K(z(t))$. That is, $z(t)$ minimizes the convex lower semicontinuous function $v \mapsto \sigma_K(v) - \langle u(t), v \rangle$. This is a VI of the second kind:

$$z(t) \in X \quad \& \quad 0 \leq \sigma_K(\widetilde{z}) - \sigma_K(z(t)) + \langle \widetilde{z} - z(t), -u(t) \rangle \qquad \text{for all } \widetilde{z} \in X.$$

Thus we can represent a PVI as a DVI with index one.

The converse is only partly true; DVIs are in fact a larger class of problems. Consider the index-one DVI

$$\begin{aligned} \frac{dx}{dt} &= f(x(t)) + B(x(t))\, z(t), \\ z(t) \in K \;\;\&\;\; 0 &\le \langle \widetilde{z} - z(t),\, G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K. \end{aligned}$$

Then the VI is equivalent to $-G(x(t)) \in N_K(z(t)) = \partial I_K(z(t))$. Again using Fenchel duality, this is equivalent to

$$z(t) \in \partial \left(I_K^*\right)(-G(x(t))) = \partial \sigma_K(-G(x(t))).$$

If G is affine, then the feasible set for $x(t)$ is convex, which is always the case for PVIs. For general nonlinear G this is not so. Assume for now that K is a closed convex *cone*. Then $\sigma_K = I_{K^*}$, and

$$\frac{dx}{dt} \in f(x(t)) + B(x(t))\, \partial I_{K^*}(-G(x(t))).$$

For G affine and B constant with $B = \nabla G^*$,

$$\frac{dx}{dt} \in f(x(t)) - N_{G^{-1}(K^*)}(x(t)),$$

or equivalently,

$$x(t) \in G^{-1}(K^*) \quad \&\quad 0 \le \left\langle \widetilde{x} - x(t),\, \frac{dx}{dt} - f(x(t)) \right\rangle \quad \text{for all } \widetilde{x} \in G^{-1}(K^*).$$

That is, under these conditions the DVI is also a PVI.

Chapter 5

Index Zero and Index One

In this chapter we will consider index-zero and index-one DVIs and various special cases and generalizations. Index-zero inequalities are the easiest kind of DVI to solve since the "algebraic" part of the solution (the part where derivatives do not appear) can be found in terms of the "differential" part of the solution. Substituting this into the differential equation gives an ordinary differential equation without an unknown "algebraic" variable. Thus we can reduce these problems to the study of ordinary differential equations.

Index-zero problems can typically be reduced to Lipschitz differential equations. Index-one DVIs, on the other hand, are considerably more interesting. However, index-zero problems can be used as a starting point for many different approximations and analyses of index-one and other problems.

5.1 Index-zero problems

5.1.1 Existence and uniqueness

Consider the index-zero DVI in the general form:

$$\frac{dx}{dt}(t) = f(t, x(t), z(t)), \qquad u(t_0) = u_0, \tag{5.1}$$

$$z(t) \in K \qquad \text{for all } t, \tag{5.2}$$

$$0 \le \langle \widetilde{z} - z(t), F(t, x(t), z(t)) \rangle \qquad \text{for all } \widetilde{z} \in K \text{ and almost all } t. \tag{5.3}$$

The VI part (5.2)–(5.3) has a unique solution $z(t)$, given t and $x(t)$, provided $F(t, x(t), \cdot)$ is, say, strongly monotone. The existence and uniqueness results for index-zero problems are usually obtained by showing equivalence with a Lipschitz ordinary differential equation. Consider the parametrized variational inequality VI$(F(t, x, \cdot), K)$:

$$z \in K \quad \& \quad 0 \le \langle \widetilde{z} - z, F(t, x, z) \rangle \qquad \text{for all } \widetilde{z} \in K.$$

The solution of this VI gives us a map $(t, x) \mapsto z = \psi(t, x)$; $\psi(t, x)$ is a singleton if, for example, $F(t, x, \cdot)$ is uniformly strongly monotone:

$$\langle F(t, x, z_1) - F(t, x, z_2), z_1 - z_2 \rangle \geq \eta \, \|z_1 - z_2\|^2$$

for some $\eta > 0$ independent of t and x. We also assume that $F(t, x, z)$ is Lipschitz in x:

$$\|F(t, x_1, z) - F(t, x_2, z)\| \leq L_F(t) \, \|x_1 - x_2\|$$

for all x_1, x_2, and $z \in K$, where $L_F(\cdot)$ is an integrable function on $[t_0, t_1]$. We also suppose that $F(t, x, z^*)$ is integrable in t for some fixed $z^* \in K$:

$$\left\| F(t, 0, z^*) \right\| \leq \beta_F(t),$$

where $\beta_F(\cdot)$ is an integrable function on $[t_0, t_1]$. Now suppose z_1 solves VI$(F(t, x_1, \cdot), K)$ and z_2 solves VI$(F(t, x_2, \cdot), K)$; that is, $z_1 = \psi(t, x_1)$ and $z_2 = \psi(t, x_2)$. Then $z_1, z_2 \in K$, so

$$\begin{aligned} 0 &\leq \langle z_2 - z_1, F(t, x_1, z_1) \rangle, \\ 0 &\leq \langle z_1 - z_2, F(t, x_2, z_2) \rangle. \end{aligned}$$

Adding gives

$$\begin{aligned} 0 &\leq -\langle z_2 - z_1, F(t, x_2, z_2) - F(t, x_1, z_1) \rangle \\ &\leq -\langle z_2 - z_1, F(t, x_1, z_2) - F(t, x_1, z_1) \rangle \\ &\qquad + \langle z_2 - z_1, F(t, x_1, z_2) - F(t, x_2, z_2) \rangle \\ &\leq -\eta \, \|z_2 - z_1\|^2 + \|z_2 - z_1\| \, \|F(t, x_1, z_2) - F(t, x_2, z_2)\| \\ &\leq -\eta \, \|z_2 - z_1\|^2 + \|z_2 - z_1\| \, L_F(t) \, \|x_1 - x_2\|. \end{aligned}$$

Rearranging and dividing by $\|z_2 - z_1\|$ give

$$\|z_2 - z_1\| \leq \frac{L_F(t)}{\eta} \, \|x_1 - x_2\|.$$

We also need a bound on the solutions z of VI$(F(t, x, \cdot), K)$: To bound $\|z\|$, note that if z^* is a fixed element of K; then

$$\begin{aligned} 0 &\leq \left\langle z^* - z, F(t, x\, z) \right\rangle \\ &\leq -\left\langle z^* - z, F(t, x, z^*) - F(t, x, z) \right\rangle \\ &\qquad + \left\langle z^* - z, F(t, x, z^*) \right\rangle \\ &\leq -\eta \, \left\| z^* - z \right\|^2 + \left\| z^* - z \right\| \, \left\| F(t, x, z^*) \right\|, \end{aligned}$$

and so rearranging and dividing by $\|z^* - z\|$ give

$$\begin{aligned} \left\| z^* - z \right\| &\leq \frac{1}{\eta} \left\| F(t, x, z^*) \right\| \\ &\leq \frac{1}{\eta} \left(\beta_F(t) + \|x\| \right), \end{aligned}$$

and so $\|z\| \leq \|z^*\| + (\beta_F(t) + \|x\|)/\eta$.

We need our dynamics to be governed by a Lipschitz function

$$\|f(t, x_1, z_1) - f(t, x_2, z_2)\| \leq L_x(t)\,\|x_1 - x_2\| + L_z\,\|z_1 - z_2\|$$

(L_x integrable on $[t_0, t_1]$) and also with a bound

$$\|f(t, 0, 0)\| \leq \beta_f(t),$$

where β_f is an integrable function on $[t_0, t_1]$.

Combining these results shows first that $\psi(t, \cdot)$ is Lipschitz with constant $L_F(t)/\eta$; this in turn shows that $\widetilde{f}(t, x) := f(t, x, \psi(t, x))$ is Lipschitz in x with constant $L_x(t) + L_z\,L_F(t)/\eta$, which is an integrable function of t. Furthermore, we can bound $\left\|\widetilde{f}(t, 0)\right\|$ by an integrable function. Then our DVI (5.1)–(5.3) is equivalent to the ordinary differential equation

$$\frac{dx}{dt}(t) = \widetilde{f}(t, x(t)), \qquad x(t_0) = x_0,$$

which has a unique solution on $[t_0, t_1]$ by the well-known theorem on existence and uniqueness of Lipschitz ordinary differential equations (Theorem C.1).

Modifications to the assumptions can be made: L_z can be made a function of time as well $L_z(t)$, and so can η to give $\eta(t)$. The crucial issue is whether $L_x(t) + L_z(t)\,L_F(t)/\eta(t)$ is an integrable function of t.

Note that the solution $x(t)$ is absolutely continuous in t. If $f(t, x, z)$ is continuous in t as well, then $\widetilde{f}(t, x)$ is Lipschitz in x and continuous in t, so $x(\cdot)$ is then a C^1 function. However, since $\psi(t, x)$ is unlikely to be differentiable in x, we do not expect that $x(\cdot)$ will be a C^2 function.

5.1.2 Index-zero CPs

Now we suppose that K is a closed convex *cone*. Then the problem becomes

$$\begin{aligned} &\frac{dx}{dt}(t) = f(t, x(t), z(t)), \qquad x(t_0) = x_0, \\ &K \ni z(t) \perp F(t, x(t), z(t)) \in K^*. \end{aligned}$$

Since it is often possible, especially for certain cones K such as $K = \mathbb{R}^n_+$, to extend the existence and uniqueness results for VIs, we can extend the existence and uniqueness results for DVIs for certain cones K. Existence and uniqueness results for CP(Φ, $\mathbb{R}^n_+$) have been developed (see [169] and [95, Section 3.5.2]). A particularly useful concept in this context is that of a *uniform P-function* for $K = \mathbb{R}^n_+$ or more generally a *uniform P(K)-function* where $K = \prod_{i=1}^m K_i$.

Suppose that $K = \prod_{i=1}^m K_i$ and each K_i is a closed convex cone in X_i where $X = \prod_{i=1}^m X_i$. A function $\Phi\colon X \to X'$ is a uniform P(K)-function if there is an $\eta > 0$ where

$$\max_{i=1,2,\ldots,m} \langle \Phi_i(y) - \Phi_i(z), y_i - z_i \rangle \geq \eta\,\|y - z\|_X^2 \tag{5.4}$$

for all $y, z \in X$. Note that z_i is the component or projection of $z \in X = \prod_{i=1}^m X_i$ in X_i. The solution of CP(Φ, K) is unique if Φ is a uniform P(K)-function. First $\left(\prod_{i=1}^m K_i\right)^* =$

$\prod_{i=1}^m K_i^*$. Also, if y and z are two solutions, then

$$\begin{aligned} y_i &\in K_i, \qquad 0 \le \langle \Phi(y)_i, z_i - y_i \rangle \qquad \text{for all } i, \\ z_i &\in K_i, \qquad 0 \le \langle \Phi(z)_i, y_i - z_i \rangle \qquad \text{for all } i. \end{aligned}$$

Thus

$$0 \ge \langle \Phi(y)_i - \Phi(z)_i, y_i - z_i \rangle \qquad \text{for all } i.$$

Taking the maximum over i gives

$$0 \ge \max_i \langle \Phi(y)_i - \Phi(z)_i, y_i - z_i \rangle \ge \eta \, \|y - z\|_X^2 ,$$

and so $y = z$. Thus CP(Φ, K) has at most one solution.

Now suppose that $F(t, x, \cdot)$ is a uniform P(K)-function with the same parameter $\eta > 0$ independent of t or x, and that $F(t, \cdot, z)$ is Lipschitz with a Lipschitz constant L_F independent of t or z. Then the solution map $z = \psi(t, x)$, where z satisfies that

$$K \ni z \perp F(t, x, z) \in K^*,$$

is Lipschitz in x: if $z_1 = \psi(t, x_1)$ and $z_2 = \psi(t, x_2)$, we have for each $i = 1, 2, \ldots, m$,

$$\begin{aligned} 0 &\ge \big\langle F(t, x_1, z_1)_i - F(t, x_2, z_2)_i, (z_1)_i - (z_2)_i \big\rangle \\ &= \big\langle F(t, x_1, z_1)_i - F(t, x_1, z_2)_i, (z_1)_i - (z_2)_i \big\rangle \\ &\qquad + \big\langle F(t, x_1, z_2)_i - F(t, x_2, z_2)_i, (z_1)_i - (z_2)_i \big\rangle \\ &\ge \big\langle F(t, x_1, z_1)_i - F(t, x_1, z_2)_i, (z_1)_i - (z_2)_i \big\rangle - L_F \, \|x_1 - x_2\| \, \|z_1 - z_2\| . \end{aligned}$$

Taking the maximum over $i = 1, 2, , \ldots, m$ and using the uniform P(K) property give

$$0 \ge \eta \, \|z_1 - z_2\|^2 - L_F \, \|x_1 - x_2\| \, \|z_1 - z_2\| .$$

Rearranging and dividing by $\|z_1 - z_2\|$ give

$$\|z_1 - z_2\| \le \frac{L_F}{\eta} \, \|x_1 - x_2\| ,$$

and $\psi(t, \cdot)$ is Lipschitz with constant L_F/η. Substituting into the differential equation gives an ordinary differential equation with Lipschitz right-hand side:

$$\frac{dx}{dt}(t) = f(t, x(t), \psi(t, x(t))), \qquad x(t_0) = x_0.$$

5.1.3 Normal compliance for mechanical contact

For a number of reasons the Signorini contact conditions are not preferred by some investigators interested in mechanical contact and impact problems. Instead, a common approach is to use normal compliance, which involves representing contact by a stiff spring which applies no force when there is no interpenetration. But when there is interpenetration, the spring force is (for example) proportional to the depth of interpenetration. This can be

represented or approximated by an index-zero DVI. For a state vector q and velocity vector v, we assume that there is a function $\varphi(q)$ so that there is no penetration if $\varphi(q) \geq 0$. For normal compliance we typically have the normal contact force $N = k\,\nabla\varphi(q)\left[\varphi(q)\right]_-$ where $[u]_- := \max(0, -u)$ and k is the stiffness of the normal compliance "spring." If the admissible set of states is $\{\, q \mid \varphi_i(q) \geq 0,\ i = 1, 2, \ldots, m \,\}$ and the normal contact force for constraint $\varphi_i(q) \geq 0$ is λ_i, then (1.3)–(1.4) can be modified to allow normal compliance:

$$M(q)\frac{dv}{dt} = -\nabla V(q) + k(q,v) - \sum_{i=1}^{m} \lambda_i\, \nabla\varphi_i(q), \tag{5.5}$$

$$\frac{dq}{dt} = v, \tag{5.6}$$

$$0 \leq \lambda_i \perp \varphi_i(q) + \frac{1}{k}\lambda_i \geq 0 \qquad \text{for all } i \text{ and } t. \tag{5.7}$$

Then, if $\varphi_i(q) \geq 0$, we have $\lambda_i = 0$, and if $\varphi_i(q) \leq 0$, we have $\lambda_i = k\left[\varphi_i(q)\right]_-$. Here λ takes the role of z in the above theory. The crucial function is $F(q, v, \lambda) := \varphi(q) + \lambda/k$, which is clearly strongly monotone in λ with $\eta = 1/k$.

The main advantage of the normal compliance approach is that the equations of motion are just ordinary differential equations and not some more complex type of problem. However, most bodies are stiff, which means that k is large. This naturally leads us to the problem of what happens as $k \to \infty$, which is a main topic of Section 6.1.

5.2 Index-one problems

For index-one problems, the VI cannot be solved from knowing the time t and state $x(t)$. In these cases, one differentiation of $F(t, x, z)$ in time will give a function which is invertible in z. These are index-one problems. In these problems there is less regularity in $z(t)$ than in $x(t)$, and this means that we will need some more structure in the problems. Specifically, in this section we will concentrate on DVIs of the form

$$\frac{dx}{dt}(t) = f(t, x(t)) + B(x(t))z(t), \qquad x(t_0) = x_0, \tag{5.8}$$

$$z(t) \in K \qquad \text{for all } t, \tag{5.9}$$

$$0 \leq \langle \widetilde{z} - z(t), G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K \text{ and } t. \tag{5.10}$$

We will call these *pure index-one DVIs*. Pure index-one DVIs can be used to represent resource limit problems, projected dynamical systems (PDSs), Coulomb friction problems with known normal contact force, and many other kinds of systems.

Sometimes we have VIs which *can* be solved for *some* components of $z(t)$; these are *mixed-index* problems. The theory is somewhat complex, as the lack of regularity of the "index-one" components should not be allowed to affect the "index zero" components. Typically mixed-index problems have the form

$$\frac{dx}{dt}(t) = f(t, x(t), z(t)) + B(x(t))z(t), \qquad x(t_0) = x_0, \tag{5.11}$$

$$z(t) \in K, \tag{5.12}$$

$$0 \leq \langle \widetilde{z} - z(t), F(t, z(t)) + G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K, \tag{5.13}$$

for all t. In addition we need conditions like

- $F(t,\cdot)$ is a monotone function for all t, and
- $\langle F(t,z_1)-F(t,z_2), z_1-z_2\rangle = 0$ implies $f(t,x(t),z_1) = f(t,x(t),z_2)$.

Nevertheless, such a theory can be developed and usefully applied. Examples of mixed-index problems include electrical circuits with ideal diodes.

In this section we will concentrate on problems of the form (5.8)–(5.10). Much of the material in this section is based on [199].

5.2.1 Pure index-one DVIs

Pure index-one DVIs do not always have solutions, particularly if K is unbounded, as is the case for DCPs (differential complementarity problems), where K is a cone. For example, if K is a closed convex cone, then there are no solutions $z(t)$ to the VI (5.9)–(5.10) if $G(x(t)) \notin K^*$. Thus the state must often be constrained, and the initial value must also satisfy $G(x_0) \in K^*$ if K is a cone.

Of crucial importance are the matrices $\nabla G(x(t))\, B(x(t))$ where $\nabla G(x)$ is the Jacobian matrix of $G(x)$ with respect to x. In finite dimensions, solutions exist if $\nabla G(x)\, B(x)$ is uniformly positive definite, but uniqueness can be hard to establish without also assuming symmetry of $\nabla G(x(t))\, B(x(t))$. These results can be extended in special cases (for example, if $\nabla G(x)\, B(x)$ is a P(K)-matrix uniformly over t and x, with $K = \prod_{i=1}^{m} K_i$ to obtain existence).

The existence proofs are based on a time discretization. Since, for K a cone, we must have $G(x(t)) \in K^*$ for all t, an implicit time discretization is needed. For our first existence result, we will work in $\mathbb{R}^n$ and suppose that f, B, and G have suitable Lipschitz properties, and that $\nabla G(x)\, B(x)$ is uniformly strongly L-copositive:

$$\langle w, \nabla G(x)\, B(x)\, w\rangle \geq \eta_{\nabla GB}\, \|w\|^2 \qquad \text{for all } w \in L \tag{5.14}$$

for all t, x, and w with $\eta_{\nabla GB} > 0$.

Theorem 5.1. *Suppose that $X = \mathbb{R}^n$ and*

(DVI-A1) *$x \mapsto f(t,x) \in X$ is measurable in t, Lipschitz in x with constant L_f, and $\|f(t,0)\| \leq \beta_{f,0}$ for all t,*

(DVI-A2) *$K = C + L \subset X$, where C is closed, convex, and bounded, and L is a closed convex cone,*

(DVI-A3) *$G(x_0) \in L^*$ (compatibility condition),*

(DVI-A4) *$\nabla G(x)\, B(x)$ is uniformly strongly L-copositive for all x, (5.14), with constant $\eta_{\nabla GB} > 0$,*

(DVI-A5) *B and ∇G are bounded and Lipschitz functions with bounds β_B and $\beta_{\nabla G}$ and Lipschitz constants L_B and $L_{\nabla G}$, respectively.*

Then there is a solution to the DVI (5.8)–(5.10).

Proof. The proof here uses a simple implicit time-stepping approach based on the Euler method. It is, however, explicit for $f(t,x)$ and implicit for much of the VI part. The crucial bounds for the solutions of the VI come from (2.47).

At each time step we solve the following VI: Given x^ℓ, find z^ℓ such that

$$x^{\ell+1} = x^\ell + \int_{t_\ell}^{t_{\ell+1}} f(\tau, x^\ell)\, d\tau + h\, B(x^\ell) z^\ell, \tag{5.15}$$

$$z^\ell \in K \;\;\&\;\; 0 \le \left\langle \widetilde{z} - z^\ell,\, G(x^\ell) + \nabla G(x^\ell)\left(x^{\ell+1} - x^\ell\right)\right\rangle \tag{5.16}$$

for all $\widetilde{z} \in K$. This has a solution since $z^\ell \mapsto G(x^\ell) + \nabla G(x^\ell)\left(x^{\ell+1} - x^\ell\right)$ is an affine map $z^\ell \mapsto b^\ell + h\,\nabla G(x^\ell)\, B(x^\ell) z^\ell$ with $h\,\nabla G(x^\ell)\, B(x^\ell)$ strongly L-copositive and $K = C + L$. From Lemma 2.16 we have the bound based on (2.47)

$$\left\| z^\ell \right\| \le \gamma \left(1 + d(b^\ell/h,\, L^*)\right),$$

where γ depends on C, L, and $\eta_{\nabla G\, B}$. Note that $G(x^0) = G(x_0) \in L^*$, so $d(G(x^0), L^*) = 0$. In general, $b^\ell = G(x^\ell) + \nabla G(x^\ell) \int_{t_\ell}^{t_{\ell+1}} f(\tau, x^\ell)\, d\tau$, so

$$\begin{aligned} d(b^\ell,\, L^*) &\le d(G(x^\ell),\, L^*) + \left\| \nabla G(x^\ell) \right\| \int_{t_\ell}^{t_{\ell+1}} \left\| f(\tau, x^\ell) \right\| d\tau \\ &\le d(G(x^\ell),\, L^*) + h\, \beta_{\nabla G} \left[\beta_{f,0} + L_f \left\| x^\ell \right\| \right]. \end{aligned}$$

Now

$$\begin{aligned} G(x^{\ell+1}) &= G(x^\ell) + \int_0^1 \nabla G\left(x^\ell + s\left(x^{\ell+1} - x^\ell\right)\right)\left(x^{\ell+1} - x^\ell\right) ds \\ &= G(x^\ell) + \nabla G(x^\ell)\left(x^{\ell+1} - x^\ell\right) \\ &\qquad + \int_0^1 \left[\nabla G\left(x^\ell + s\left(x^{\ell+1} - x^\ell\right)\right) - \nabla G(x^\ell)\right] ds \\ &\qquad\qquad \times \left(x^{\ell+1} - x^\ell\right). \end{aligned}$$

With $L_{\nabla G}$ the Lipschitz constant of ∇G,

$$\begin{aligned} G(x^{\ell+1}) &= G(x^\ell) + \nabla G(x^\ell) \left[\int_{t_\ell}^{t_\ell+1} f(\tau, x^\ell)\, d\tau + h\, B(x^\ell) z^\ell \right] + \eta^\ell, \\ \left\| \eta^\ell \right\| &\le \frac{1}{2} L_{\nabla G} \left\| x^{\ell+1} - x^\ell \right\|^2. \end{aligned}$$

Now

$$G(x^\ell) + \nabla G(x^\ell) \left[\int_{t_\ell}^{t_\ell+1} f(\tau, x^\ell)\, d\tau + h\, B(x^\ell) z^\ell \right] \in L^*,$$

since it appears on the right-hand side of the inner product in the VI (5.16). Thus $d(G(x^{\ell+1}), L^*) \le \left\| \eta^\ell \right\| \le \frac{1}{2} L_{\nabla G} \left\| x^{\ell+1} - x^\ell \right\|^2$. For $f(t,0)$ bounded by $\beta_{f,0}$, from the Lipschitz

constant we have

$$\begin{aligned}
\left\|x^{\ell+1}-x^{\ell}\right\| &\le h\left[\beta_{f,0}+L_f\left\|x^{\ell}\right\|+\beta_B\left\|z^{\ell}\right\|\right], \qquad \text{so}\\
d(G(x^{\ell+1}),L^*)/h &\le \frac{1}{2}L_{\nabla G}\,h\left[\beta_{f,0}+L_f\left\|x^{\ell}\right\|+\beta_B\left\|z^{\ell}\right\|\right]^2,\\
\left\|z^{\ell}\right\| &\le \gamma\left[1+\beta_{\nabla G}\left(\beta_{f,0}+L_f\left\|x^{\ell}\right\|+d(G(x^{\ell}),L^*)/h\right)\right],\\
\left\|x^{\ell+1}\right\| &\le \left\|x^{\ell}\right\|+h\left[\beta_{f,0}+L_f\left\|x^{\ell}\right\|+\beta_B\left\|z^{\ell}\right\|\right].
\end{aligned}$$

Substituting the bound on $z^{\ell+1}$ and $d(G(x^{\ell+1}),L^*)$ into the other inequalities above gives

$$\begin{aligned}
\left\|x^{\ell+1}\right\| &\le \left\|x^{\ell}\right\|+h\left[\beta_{f,0}+L_f\left\|x^{\ell}\right\|+\beta_B\left\|z^{\ell}\right\|\right],\\
\left\|z^{\ell+1}\right\| &\le \gamma\left[1+\beta_{\nabla G}\left(\beta_{f,0}+L_f\left\|x^{\ell}\right\|+d(G(x^{\ell}),L^*)/h\right)\right]\\
&\le \gamma\left[1+\beta_{\nabla G}\beta_{f,0}+\beta_{\nabla G}L_f\left(\left\|x^{\ell}\right\|+h\left[\beta_{f,0}+L_f\left\|x^{\ell}\right\|+\beta_B\left\|z^{\ell}\right\|\right]\right)\right.\\
&\qquad \left.+\frac{1}{2}h\,\beta_{\nabla G}L_G\left[\beta_{f,0}+L_f\left\|x^{\ell}\right\|+\beta_B\left\|z^{\ell}\right\|\right]^2\right].
\end{aligned}$$

Setting $\theta^{\ell;h}=\left\|x^{\ell}\right\|$ and $\psi^{\ell;h}=\left\|z^{\ell}\right\|$ we can apply the nonlinear discrete Gronwall-type Lemma 5.2 (which follows this proof), with

$$\begin{aligned}
\rho_0(\theta,\psi) &= \beta_{f,0}+L_f\theta+\beta_B\psi,\\
\sigma_0(\theta) &= \gamma\left[1+\beta_{\nabla G}\beta_{f,0}+\beta_{\nabla G}L_f\theta\right].
\end{aligned}$$

Thus, for any finite $T>t_0$ and $\epsilon>0$, we have for sufficiently small $h>0$,

$$\begin{aligned}
\left\|x^{\ell}\right\| &\le \widehat{\theta}(t_\ell)+\epsilon,\\
\left\|z^{\ell}\right\| &\le \gamma\left[1+\beta_{\nabla G}\beta_{f,0}+\beta_{\nabla G}L_f\widehat{\theta}(t_\ell)\right]+\epsilon,
\end{aligned}$$

where $\widehat{\theta}(t)$ grows exponentially in t (with growth rate $(1+\beta_B\gamma\beta_{\nabla G})\,L_f$). Thus, on any finite interval $[t_0,T]$, $\left\|x^{\ell}\right\|$ and $\left\|z^{\ell}\right\|$ are uniformly bounded, independently of $h>0$.

Set x_h to be the interpolant

$$x_h(t)=x^{\ell}+\int_{t_\ell}^{t} f(\tau,x^{\ell})\,d\tau+(t-t_\ell)\;B(x^{\ell})z^{\ell}$$

for $t_\ell\le t\le t_{\ell+1}$, $z_h(t)=z^{\ell}$ for $t_\ell<t<t_{\ell+1}$, and $\widetilde{x}_h(t)=x^{\ell}$ for $t_\ell<t<t_{\ell+1}$. Since the functions $x_h(\cdot)$ are uniformly Lipschitz and uniformly bounded in $\mathbb{R}^n$, by the Arzela–Ascoli theorem, there is a uniformly convergent subsequence. Let $\widehat{x}(\cdot)$ be the limit of such a sequence. Note that we also have uniform convergence of $\widetilde{x}_h(\cdot)\to\widehat{x}(\cdot)$. Since the functions $z_h(\cdot)$ are uniformly bounded in $L^\infty(0,T;\mathbb{R}^m)$, there is a further weak* convergent subsequence $z_h(\cdot)\rightharpoonup^*\widehat{z}(\cdot)$. We wish to show that $\widehat{x}(\cdot)$ and $\widehat{z}(\cdot)$ together satisfy the conditions of the DVI (5.8)–(5.10).

First note that $dx_h/dt(t) = f(t,\widetilde{x}_h(t)) + B(\widetilde{x}_h(t))z_h(t)$ for almost all t, and that $x_h(t_0) = x_0$. Thus for $s < t$ in the interval $[t_0, T]$ we have

$$x_h(t) - x_h(s) = \int_s^t \left[f(\tau,\widetilde{x}_h(\tau)) + B(\widetilde{x}_h(\tau))z_h(\tau) \right] d\tau.$$

Taking limits in the appropriate subsequence then gives

$$\widehat{x}(t) - \widehat{x}(s) = \int_s^t \left[f(\tau,\widehat{x}(\tau)) + B(\widehat{x}(\tau))\widehat{z}(\tau) \right] d\tau.$$

Note that because $z_h(\cdot) \rightharpoonup^* \widehat{z}(\cdot)$ weak*, it is important that the right-hand side of (5.8) be linear in z. Note that the set of functions

$$\left\{ \zeta \in L^2(t_0, T; \mathbb{R}^m) \mid \zeta(t) \in K \text{ for all } t \right\}$$

is a closed convex set, and it is also weakly closed by Mazur's lemma. Since $z_h(\cdot) \rightharpoonup^* \widehat{z}(\cdot)$ in $L^\infty(t_0, T; \mathbb{R}^m)$ we have $z_h(\cdot) \rightharpoonup^* \widehat{z}(\cdot)$ in $L^2(t_0, T; \mathbb{R}^m)$, but weak and weak* convergence in $L^2(t_0, T; \mathbb{R}^m)$ are identical since it is a reflexive space. Thus the weak limit $\widehat{z}(\cdot)$ has the property that $\widehat{z}(t) \in K$ for all t.

Finally, we need to show that for any continuous $\widetilde{z}\colon\ [t_0, T] \to K$, we have

$$0 \le \int_{t_0}^T \langle \widetilde{z}(t) - \widehat{z}(t),\, G(\widehat{x}(t)) \rangle \, dt.$$

Now for each ℓ we have

$$0 \le \left\langle \frac{1}{h} \int_{t_\ell}^{t_{\ell+1}} \widetilde{z}(\tau)\, d\tau - z^\ell,\, G(x^\ell) + \int_{t_\ell}^{t_{\ell+1}} f(\tau, x^\ell)\, d\tau + h \nabla G(x^\ell) B(x^\ell) z^\ell \right\rangle.$$

Let $\eta_h(t) = \int_{t_\ell}^{t_{\ell+1}} f(\tau, x^\ell)\, d\tau + h \nabla G(x^\ell) B(x^\ell) z^\ell$ for $t_\ell \le t < t_{\ell+1}$. Summing over ℓ from zero to r gives

$$0 \le \int_{t_0}^{t_{r+1}} \langle \widetilde{z}(\tau) - z_h(\tau),\, G(\widetilde{x}_h(\tau)) + \eta_h(\tau) \rangle \, d\tau.$$

Note that $\|\eta_h(t)\| \le h \left[\beta_{f,0} + L_f \|\widetilde{x}_h(t)\| + \beta_B \beta_{\nabla G} \|z_h(t)\| \right]$. Since $\widetilde{x}_h(\cdot)$ and $z_h(\cdot)$ are uniformly bounded on $[t_0, T]$, $\|\eta_h\|_{L^\infty} \to 0$ as $h \downarrow 0$. Also $G(\widetilde{x}_h(\cdot))$ converges uniformly to $G(\widehat{x}(\cdot))$ by continuity of G. Then weak convergence $z_h(\cdot) \rightharpoonup \widehat{z}(\cdot)$ is enough to obtain convergence of the integrals. Setting $r = \lfloor (T - t_0)/h \rfloor$, by boundedness of the integrand, we have

$$0 \le \int_{t_0}^T \langle \widetilde{z}(\tau) - z_h(\tau),\, G(\widetilde{x}_h(\tau)) + \eta_h(\tau) \rangle \, d\tau + \mathcal{O}(h).$$

Taking limits in the appropriate subsequence then gives

$$0 \le \int_{t_0}^T \langle \widetilde{z}(\tau) - \widehat{z}(\tau),\, G(\widehat{x}(\tau)) \rangle \, d\tau.$$

Since this holds for all continuous $\widetilde{z}\colon [t_0, T] \to K$, by standard arguments,

$$0 \le \langle \widetilde{z} - \widehat{z}(t), G(\widehat{x}(t)) \rangle$$

for all $\widetilde{z} \in K$ for almost all t, as desired. By redefining $\widehat{z}(t)$ on a null set, we can make this hold for all t. Thus $\widehat{x}(\cdot)$ and $\widehat{z}(\cdot)$ form a solution of (5.8)–(5.10), as desired. □

Note that a side effect of this proof is that it is clear that $G(\widehat{x}(t)) \in L^*$ for all t.

An important element of this proof is the following nonlinear discrete Gronwall lemma.

Lemma 5.2. *Suppose that*

$$\begin{aligned} \theta^{\ell+1;h} &\le \theta^{\ell;h} + h\,\rho(\theta^{\ell;h}, \psi^{\ell;h}; h), \qquad \theta^{0;h} = \theta_0, \\ \psi^{\ell+1;h} &\le \sigma(\theta^{\ell;h}, \psi^{\ell;h}; h), \qquad 0 \le \psi^{0;h} \le \sigma(\theta_0, 0; h), \end{aligned}$$

where ρ and σ satisfy the following:

1. *$\rho(\theta, \psi; h)$ and $\sigma(\theta, \psi; h)$ have nonnegative values and are nondecreasing in θ and ψ;*
2. *$\rho(\theta, \psi; h)$ and $\sigma(\theta, \psi; h)$ are locally Lipschitz in (θ, ψ) with Lipschitz constant independent of $h > 0$;*
3. *$\rho(\theta, \psi; h) \to \rho_0(\theta, \psi)$ and $\sigma(\theta, \psi; h) \to \sigma_0(\theta)$ as $h \downarrow 0$ uniformly in (θ, ψ) over bounded sets.*

Then, if the solution $\widehat{\theta}$ of

$$\frac{d\widehat{\theta}}{dt}(t) = \rho_0(\widehat{\theta}(t), \sigma_0(\widehat{\theta}(t))), \qquad \widehat{\theta}(0) = \theta_0 \tag{5.17}$$

is finite on $[0, T]$, then for every $\epsilon > 0$ there is an $h_0^* > 0$ such that

$$\begin{aligned} \theta^{\ell;h} &\le \widehat{\theta}(t_\ell) + \epsilon, \\ \psi^{\ell;h} &\le \sigma_0(\widehat{\theta}(t_\ell)) + \epsilon, \end{aligned}$$

where $t_\ell := \ell\, h \in [0, T]$ whenever $0 < h \le h_0^*$.

Proof. Note first that given ρ_0, σ_0, and θ_0, $\widehat{\theta}$ is the unique solution of the differential equation (5.17) since $\theta \mapsto \rho_0(\theta, \sigma_0(\theta))$ is locally Lipschitz. For $\eta > 0$ let

$$\frac{d\widehat{\theta}_\eta}{dt}(t) = \rho_0(\widehat{\theta}_\eta(t), \sigma_0(\widehat{\theta}_\eta(t)) + \eta) + \eta, \qquad \widehat{\theta}_\eta(0) = \theta_0.$$

By continuous dependence of solutions on parameters, $\widehat{\theta}_\eta(t) \to \widehat{\theta}(t)$ as $\eta \downarrow 0$ uniformly in $t \in [0, T]$. Thus, for $\eta > 0$ sufficiently small, $\left|\widehat{\theta}_\eta(t) - \widehat{\theta}(t)\right| < \epsilon$ for all $t \in [0, T]$. We now need to show that for sufficiently small $h > 0$, if $\ell\, h \le T$,

$$\theta^{\ell;h} \le \widehat{\theta}_\eta(t_\ell) \quad \text{and} \quad \psi^{\ell;h} \le \sigma_0(\widehat{\theta}_\eta(t)) + \eta.$$

Given $\eta > 0$, let $R(\eta) = \widehat{\theta}_\eta(T) + \sigma_0(\widehat{\theta}_\eta(T)) + \eta$. Note that $R(\eta)$ is a nondecreasing function of $\eta > 0$.

Show true for $\ell = 0$: Clearly, from the initial values, $\theta^{0;h} = \theta_0 = \widehat{\theta}_\eta(t_0)$ and $\psi^{0;h} \le \sigma(\theta_0, 0; h) \le \sigma_0(\theta_0) + \eta = \sigma_0(\widehat{\theta}_\eta(t_0)) + \eta$, as desired.

Suppose true for $k = \ell$; show true for $k = \ell + 1$: Suppose that $\theta^{\ell;h} \le \widehat{\theta}_\eta(t_\ell)$ and $\psi^{\ell;h} \le \sigma_0(\widehat{\theta}_\eta(t_\ell)) + \eta$ and that $(\ell+1)h \le T$. Then

$$\begin{aligned}\theta^{\ell+1;h} &\le \theta^{\ell;h} + h\,\rho(\theta^{\ell;h}, \psi^{\ell;h}; h)\\ &\le \widehat{\theta}_\eta(t_\ell) + h\,\rho(\widehat{\theta}_\eta(t_\ell), \sigma_0(\widehat{\theta}_\eta(t_\ell)) + \eta)\end{aligned}$$

since $\rho(\theta, \psi; h)$ is nondecreasing in θ and ψ. But since ρ has nonnegative values, $\widehat{\theta}_\eta$ is also nondecreasing, and combined with the fact that $\rho(\theta, \psi; h)$ and $\sigma(\theta, \psi; h)$ are also nondecreasing in θ and ψ, we have

$$\begin{aligned}h\,\rho(\widehat{\theta}_\eta(t_\ell), \sigma_0(\widehat{\theta}_\eta(t_\ell)) + \eta) &\le \int_{t_\ell}^{t_{\ell+1}} \rho(\widehat{\theta}_\eta(\tau), \sigma_0(\widehat{\theta}_\eta(\tau); h) + \eta)\, d\tau\\ &\le \int_{t_\ell}^{t_{\ell+1}} \left[\rho_0(\widehat{\theta}_\eta(\tau), \sigma_0(\widehat{\theta}_\eta(\tau)) + \eta) + \eta\right] d\tau\\ &= \int_{t_\ell}^{t_{\ell+1}} d\widehat{\theta}_\eta/dt(\tau)\, d\tau \;=\; \widehat{\theta}_\eta(t_{\ell+1}) - \widehat{\theta}_\eta(t_\ell).\end{aligned}$$

Therefore,

$$\theta^{\ell+1;h} \le \widehat{\theta}_\eta(t_{\ell+1}) \le R_0(\eta).$$

On the other hand,

$$\begin{aligned}\psi^{\ell+1;h} &\le \sigma(\theta^{\ell;h}, \psi^{\ell;h}; h)\\ &\le \sigma_0(\theta^{\ell;h}) + \eta\\ &\le \sigma_0(\widehat{\theta}_\eta(t_{\ell+1})) + \eta \;\le\; R_0(\eta),\end{aligned}$$

also as desired.

Thus by induction, the result holds for all ℓ, where $\ell h \le T$.

Taking $\eta > 0$ sufficiently small, we have the result that for given $\epsilon > 0$ and sufficiently small $h > 0$,

$$\begin{aligned}\theta^{\ell;h} &\le \widehat{\theta}(t_\ell) + \epsilon,\\ \psi^{\ell;h} &\le \sigma_0(\widehat{\theta}(t_\ell)) + \epsilon,\end{aligned}$$

provided $0 \le \ell h \le T$. □

For infinite-dimensional problems we have some additional complications due to the lack of compactness. One method of overcoming these problems is to use (or assume) pseudomonotonicity (see Section 2.5). For example, a crucial part of the proof of Theorem 5.1 is that there is a solution z^ℓ of

$$z^\ell \in K \;\&\; 0 \le \left\langle \widetilde{z} - z, b^\ell + \nabla G(x^\ell) B(x^\ell) z^\ell \right\rangle \qquad \text{for all } \widetilde{z} \in K.$$

Strong L-copositivity of $\nabla G(x^\ell)B(x^\ell)$ is sufficient in finite dimensions to show existence of solutions, but in infinite dimensions we need to make an additional assumption, such as pseudomonotonicity. If we have pseudomonotonicity of $\nabla G(x^\ell)B(x^\ell)$, then $z \mapsto b^\ell + \nabla G(x^\ell)B(x^\ell)z$ is also coercive since $\nabla G(x^\ell)B(x^\ell)$ is strongly L-copositive, and so by Theorem 2.36 the VI has a solution.

Thinking in terms of finite-dimensional approximations to an infinite-dimensional problem, consider the DVI

$$\begin{aligned} \frac{dx}{dt}(t) &= f(t,x(t)) + B(x(t))z(t), \qquad x(t_0) = x_0, \\ z(t) \in K \ \ \& \ \ 0 &\le \langle \widetilde{z} - z(t), G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K. \end{aligned}$$

Consider $x(t) \in X$ and $z(t) \in Z$, with X and Z reflexive Banach spaces. Then $G\colon X \to Z'$, $f\colon [t_0, T] \times X \to X$, and $B\colon X \to \mathcal{L}(Z, X)$.

Pick finite subsets $\{u_1, u_2, \ldots, u_m\} \subset C$ and $\{v_1, v_2, \ldots, v_m\} \subset L$, and let $Z_m = \operatorname{span}\{u_1, v_1, u_2, v_2, \ldots, u_m, v_m\}$ and $K_m = C \cap Z_m + L \cap Z_m$. Then there is an approximation $(x_m(t), z_m(t))$ that satisfies the approximate DVI

$$\begin{aligned} \frac{dx_m}{dt}(t) &= f(t,x_m(t)) + B(x_m(t))z_m(t), \qquad x(t_0) = x_0, \\ z_m(t) \in K_m \ \ \& \ \ 0 &\le \langle \widetilde{z} - z_m(t), G(x_m(t)) \rangle \qquad \text{for all } \widetilde{z} \in K_m. \end{aligned}$$

We can prove that solutions exist for this approximate DVI (even though $x_m(t)$ may be in an infinite-dimensional space) since precompactness of the set of values $z_m(t)$ gives precompactness of the values $x_m(t)$. The problem is that if $m \to \infty$, we have *weak* convergence of a subsequence of the $z_m(\cdot)$ and $x_m(\cdot)$, which is not enough to prove that the DVI is satisfied in the limit without some more assumptions about the nature of f, B, and G.

We make the following additional assumptions.

(DVI-A6) The functions $\nabla G(x) = G_0 + \nabla G_1(x)$, $B(x) = B_0 + B_1(x)$, where $\nabla G_1(x)$ and $B_1(x)$ are collectively compact; that is, $\bigcup_x \left(\nabla G_1(x)\overline{B_X}\right)$ is precompact in Z' and $\bigcup_x \left(B_1(x)\overline{B_X}\right)$ is precompact in X.

(DVI-A7) The function $f(t,x) = f(x)$ for all t and x, and f being compact and $u_m \rightharpoonup u$ weakly in X implies $f(u_m) \to f(u)$ strongly in X, $\nabla G(u_m) \to \nabla G(u)$ in $\mathcal{L}(X, Z')$, and $B(u_m) \to B(u)$ in $\mathcal{L}(Z, X)$.

(DVI-A8) The operator $G_0 B_0\colon Z \to Z'$ is monotone and self-adjoint.

Assumptions (DVI-A6) and (DVI-A7) essentially require that ∇G and B be not too far from being constant and $\nabla G(x)\, B(x)$ not too far from being monotone and self-adjoint.

These assumptions are fairly strong, and weaker conditions can be found for existence of solutions to DVIs. However, we will see how these conditions give existence of solutions for infinite-dimensional problems. Let us suppose that we have a subsequence in which $z_m(\cdot) \rightharpoonup z(\cdot)$ weakly in $L^2(0,T;Z)$ and $x_m(\cdot) \rightharpoonup x(\cdot)$ in $L^2(0,T;X)$. Then $z(t) \in K$ for almost all t. Also suppose that for sufficiently large m in the subsequence, $\widetilde{z}(t) \in K_m$. (If not, then since $K_m \subseteq K_{m+1} \subseteq \cdots$, we can project $\widetilde{z}(t)$ onto K_{m^*} for some m^* in the subsequence.)

Then, for all m sufficiently large in the subsequence,

$$\begin{aligned}
0 &\le \langle \widetilde{z}(t) - z_m(t),\, G(x_m(t))\rangle \\
&= \left\langle \widetilde{z}(t) - z_m(t),\, G(x_0) + \int_{t_0}^{t} \nabla G(x_m(\tau))\, \frac{dx_m}{d\tau}(\tau)\, d\tau \right\rangle \\
&= \langle \widetilde{z}(t) - z_m(t),\, G(x_0)\rangle + \int_{t_0}^{t} \langle \widetilde{z}(t) - z_m(t),\, \nabla G(x_m(\tau))\, f(x_m(\tau))\rangle\, d\tau \\
&\qquad + \int_{t_0}^{t} \langle \widetilde{z}(t) - z_m(t),\, \nabla G(x_m(\tau))\, B(x_m(\tau))\, z_m(\tau)\rangle\, d\tau .
\end{aligned}$$

Note that f is a continuous compact function $X \to X$, and by the Arzela–Ascoli theorem there is a uniformly convergent subsequence of m, where $f(x_m(\cdot)) \to f(x(t))$ by (DVI-A7). Integrating over $t_0 \le t \le T$, we get

$$\begin{aligned}
0 &\le \int_{t_0}^{T} \langle \widetilde{z}(t) - z_m(t),\, G(x_m(t))\rangle\, dt \\
&= \int_{t_0}^{T} \langle \widetilde{z}(t) - z_m(t),\, G(x_0)\rangle\, dt \\
&\qquad + \int_{t_0}^{T} \int_{t_0}^{t} \langle \widetilde{z}(t) - z_m(t),\, \nabla G(x_m(\tau))\, f(x_m(\tau))\rangle\, d\tau\, dt \\
&\qquad + \int_{t_0}^{T} \int_{t_0}^{t} \langle \widetilde{z}(t),\, \nabla G(x_m(\tau))\, B(x_m(\tau))\, z_m(\tau)\rangle\, d\tau\, dt \\
&\qquad - \int_{t_0}^{T} \int_{t_0}^{t} \langle z_m(t),\, \nabla G(x_m(\tau))\, B(x_m(\tau))\, z_m(\tau)\rangle\, d\tau\, dt .
\end{aligned}$$

The first term converges to $\int_{t_0}^{T} \langle \widetilde{z}(t) - z(t),\, G(x_0)\rangle\, dt$; the second term converges to

$$\int_{t_0}^{T} \int_{t_0}^{t} \langle \widetilde{z}(t) - z(t),\, \nabla G(x(\tau))\, f(x(\tau))\rangle\, d\tau\, dt$$

using (DVI-A7); the third term converges to $\int_{t_0}^{T} \int_{t_0}^{t} \langle \widetilde{z}(t),\, \nabla G(x(\tau))\, B(x(\tau))\, z(\tau)\rangle\, d\tau\, dt$ again using (DVI-A7); for the fourth term, note that $\nabla G(x_m(t))\, B(x_m(t)) \to \nabla G(x(t))\, B(x(t))$ uniformly in t since the $x_m(\cdot)$ are uniformly Lipschitz. Now $\nabla G(x(t))\, B(x(t)) - G_0 B_0$ is a compact operator for all t, and by continuity in t and compactness of the interval $[t_0, T]$, in a (further) subsequence

$$\begin{aligned}
&(\nabla G(x(t))\, B(x(t)) - G_0 B_0)\, z_m(t) \\
&\qquad \to (\nabla G(x(t))\, B(x(t)) - G_0 B_0)\, z(t) \qquad \text{as } m \to \infty.
\end{aligned}$$

Thus

$$\begin{aligned}
&(\nabla G(x_m(t))\, B(x_m(t)) - G_0 B_0)\, z_m(t) \\
&\qquad \to (\nabla G(x(t))\, B(x(t)) - G_0 B_0)\, z(t) \qquad \text{as } m \to \infty,
\end{aligned}$$

and so

$$\int_{t_0}^{T}\int_{t_0}^{t} \langle z_m(t), (\nabla G(x_m(t))\, B(x_m(t)) - G_0\, B_0)\, z_m(\tau)\rangle \, d\tau\, dt$$

converges to

$$\int_{t_0}^{T}\int_{t_0}^{t} \langle z(t), (\nabla G(x(t))\, B(x(t)) - G_0\, B_0)\, z(\tau)\rangle \, d\tau\, dt.$$

The remainder is

$$\int_{t_0}^{T}\int_{t_0}^{t} \langle z_m(t), G_0\, B_0\, z_m(\tau)\rangle \, d\tau\, dt.$$

But $G_0\, B_0$ is self-adjoint by (DVI-A8), so

$$\langle z_m(t), G_0\, B_0\, z_m(\tau)\rangle = \langle z_m(\tau), G_0\, B_0\, z_m(t)\rangle\,;$$

thus the remaining term is

$$\begin{aligned}
&\frac{1}{2}\int_{t_0}^{T}\int_{t_0}^{T} \langle z_m(t), G_0\, B_0\, z_m(\tau)\rangle \, d\tau\, dt \\
&\qquad = \frac{1}{2}\left\langle \int_{t_0}^{T} z_m(t)\, dt,\, G_0\, B_0 \int_{t_0}^{T} z_m(\tau)\, d\tau \right\rangle,
\end{aligned}$$

which is a convex quadratic function of $\int_{t_0}^{T} z_m(t)\, dt$, which converges weakly to $\int_{t_0}^{T} z(t)\, dt$. Hence, by Mazur's lemma,

$$\begin{aligned}
&\liminf_{m\to\infty} \frac{1}{2}\int_{t_0}^{T}\int_{t_0}^{T} \langle z_m(t), G_0\, B_0\, z_m(\tau)\rangle \, d\tau\, dt \\
&\qquad \geq \frac{1}{2}\int_{t_0}^{T}\int_{t_0}^{T} \langle z(t), G_0\, B_0\, z(\tau)\rangle \, d\tau\, dt.
\end{aligned}$$

Combining all these results, we have

$$\begin{aligned}
0 \leq \int_{t_0}^{T} &\langle \widetilde{z}(t) - z(t), G(x_0)\rangle \, dt \\
&+ \int_{t_0}^{T}\int_{t_0}^{t} \langle \widetilde{z}(t) - z(t), \nabla G(x(\tau))\, f(x(\tau))\rangle \, d\tau\, dt \\
&+ \int_{t_0}^{T}\int_{t_0}^{t} \langle \widetilde{z}(t), \nabla G(x(\tau))\, B(x(\tau))\, z(\tau)\rangle \, d\tau\, dt \\
&- \int_{t_0}^{T}\int_{t_0}^{t} \langle z(t), \nabla G(x(\tau))\, B(x(\tau))\, z(\tau)\rangle \, d\tau\, dt \\
= \int_{t_0}^{T} &\langle \widetilde{z}(t) - z(t), G(x(t))\rangle \, dt,
\end{aligned}$$

and the weak limit is indeed a solution of the original problem in infinite dimensions.

This approach cannot deal with unbounded operators unless there are some special conditions included, as the bounds on $z_m(t)$ depend on $f(x_m(t))$. If $f(x)$ is an unbounded operator, then we lose the bounds on $z_m(t)$. If, for example, $f(x)$ generates a semigroup that leaves K invariant, then by alternating time steps for $dx/dt = f(x)$ and for the DVI without $f(x)$, bounds on $z_m(t)$ can be found independent of m.

An alternative approach for infinite-dimensional problems is to use PVIs as discussed in Section 4.7, although then $G(x)$ must be affine and $B = \nabla G^*$. To turn the DVI into something like a PVI, the VI part

$$z(t) \in K \;\&\; 0 \leq \langle \widetilde{z} - z(t), G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K$$

can be represented as

$$\begin{aligned} 0 &\in G(x(t)) + N_K(z(t)) \\ &= G(x(t)) + \partial I_K(z(t)). \end{aligned}$$

Using Fenchel duality, $v \in \partial\phi(u)$ if and only if $u \in \partial\phi^*(v)$, and we can rewrite the VI as

$$z(t) \in \partial I_K^*(-G(x(t))).$$

Substituting this into the differential equation gives the differential inclusion

$$\frac{dx}{dt}(t) \in f(x(t)) + B(x(t))\,\partial I_K^*(-G(x(t))), \qquad x(t_0) = x_0.$$

If G is affine and $B = \nabla G^*$, then under a constraint qualification,

$$\frac{dx}{dt}(t) \in f(x(t)) - \partial\left(I_K^* \circ -G\right)(x(t)), \qquad x(t_0) = x_0,$$

which is a maximal monotone differential equation, and all of the associated theory applies. However, if $G(x)$ is nonlinear, then $I_K^* \circ -G$ is usually neither convex nor convex plus Lipschitz. In such cases, it is necessary to fall back on theory such as described earlier in this section.

5.2.2 Uniqueness of solutions of index-one DVIs

While $\nabla G(x)\,B(x)$ positive definite is (apart from technical conditions) sufficient to guarantee existence of solutions to index-one DVIs, it is certainly not sufficient to guarantee uniqueness. If K is a closed convex cone, then the DVI is a DCP. A natural assumption might be that if the associated LCP($K, q, \nabla G(x)\,B(x)$) has a unique solution for all x and q, then the DCP also has unique solutions. However, this is not true, even if $\nabla G(x)$ and $B(x)$ are constant.

Examples of nonuniqueness for DCPs with $\nabla G(x)$ and $B(x)$, and even constant $f(t, x, z)$, were found by Bernard and el Kharroubi [30], where $\nabla G(x)\,B(x)$ is a P-matrix and $K = \mathbb{R}^3_+$. The nontrivial solutions for the DCP of Bernard and el Kharroubi look like Figure 5.1. Mandelbaum [164] went further and showed nonuniqueness for a system with $\nabla G(x)\,B(x)$ positive definite but not symmetric, $K = \mathbb{R}^2_+$, and $f(t, x, z)$ nonconstant. In fact, $f(t, x, z)$ in Mandelbaum's counterexample is a complicated function that can be C^∞ but not analytic.

Both of these examples involve solutions that are *reverse Zeno*. That is, for one of the solutions the *active set* $I(t) := \{i \mid z_i(t) = 0\}$ changes infinitely often in an interval $[t^*, t^* + \epsilon)$ for any $\epsilon > 0$.

This appears to contradict the theory of LCSs [124], which says that solutions to

$$\begin{aligned} \frac{dx}{dt} &= A\,x(t) + B\,z(t), \qquad x(t_0) = x_0, \\ w(t) &= C\,x(t) + D\,z(t), \\ 0 \le w(t) \perp z(t) &\ge 0 \qquad \text{for all } t \end{aligned}$$

are unique, provided $D + s^{-1}CB$ is a P-matrix for sufficiently large positive s. However, these results are not contradictory: the theory of LCSs considers only *Bohl distributions* as solutions, which immediately rules out reverse Zeno solutions.

Uniqueness in general can be proved in finite-dimensional problems provided that $\nabla G(x)\,B(x)$ is *symmetric* positive definite.

Theorem 5.3. *Consider the DVI*

$$\frac{dx}{dt} = f(t,x) + B(x)z(t), \qquad x(t_0) = x_0, \tag{5.18}$$

$$z(t) \in K, \tag{5.19}$$

$$0 \le (\widetilde{z} - z(t))^T\,[F(z(t)) + G(x(t))] \qquad \textit{for all } \widetilde{z} \in K, \tag{5.20}$$

where f, B, G, and F are all Lipschitz with ∇G also continuous. Assume that $F\colon \mathbb{R}^n \to \mathbb{R}^n$ is monotone and that $\nabla G(x)\,B(x)$ is symmetric positive definite. Then the solution (x, z) is unique.

The proof here follows [244]. Before we go on to the proof of this result, we first begin with a lemma in linear algebra inspired by the BFGS optimization method.

Lemma 5.4. *Let $C = \{(U,V) \mid V^T U$ is symmetric positive definite$\}$. There is a map $Q\colon C \to \mathbb{R}^{n\times n}$ where whenever $V^T U$ is symmetric positive definite, $U = Q(U,V)\,V$ and $Q(U,V)$ is symmetric positive definite. Furthermore, this map is locally Lipschitz.*

Proof. Suppose that U is $n \times m$; then $P := V^T U$ is $m \times m$ and V is also $n \times m$. Note that the rank of both U and V must be m. Fix Q_0 to be a symmetric positive definite $n \times n$ matrix. (We could put $Q_0 = I$, for example.) Then put

$$Q = (I - V P^{-1} U^T)^T\, Q_0 (I - V P^{-1} U^T) + U P^{-1} U^T.$$

This formula is inspired by the DFP and BFGS quasi-Newton update formulas (see [192, pp. 196–198]). Clearly Q is symmetric and positive semidefinite. Since Q_0 and P^{-1} are positive definite, if $x^T Q x = 0$, then $(I - V P^{-1} U^T)x = 0$ and $U^T x = 0$. The latter equation implies that $(I - V P^{-1} U^T)x = x - V P^{-1} 0 = x$, so the only x where $x^T Q x = 0$ is $x = 0$. Thus Q is positive definite.

To show $QV = U$ we carry out a simple calculation, noting that $V^T U = P = P^T = U^T V$:

$$\begin{aligned} QV &= (I - VP^{-1}U^T)^T Q_0 (I - VP^{-1}U^T)V + UP^{-1}U^T V \\ &= (I - VP^{-1}U^T)^T Q_0 (V - VP^{-1}P) + UP^{-1}P \\ &= (I - VP^{-1}U^T)^T Q_0 (V - V) + U = U. \end{aligned}$$

For the final part, we note that $P \mapsto P^{-1}$ is locally Lipschitz for positive definite P, so from the formula we see that $(U, V) \mapsto Q$ is a locally Lipschitz map. □

We can now continue with a proof of our main uniqueness result.

***Proof of Theorem* 5.3.** Suppose there are two solutions (x_1, z_1) and (x_2, z_2). Let

$$t^* = \sup\{t \geq t_0 \mid x_1(t) = x_2(t),\ z_1(s) = z_2(s) \text{ for almost all } t_0 \leq s \leq t\}.$$

By shifting the initial time we can assume without loss of generality that $t^* = +\infty$ or $t^* = t_0$. If there are indeed two distinct solutions, then $t^* < +\infty$, so we can assume that $t^* = t_0$. We will show that this leads to a contradiction.

Let $P(x) = \nabla G(x)\, B(x)$, which is a symmetric positive definite matrix for all x. Thus by Lemma 5.4 there is a locally Lipschitz function $Q(x)$ where $Q(x)$ is symmetric positive definite for all x and $\nabla G(x) = B(x)^T Q(x)$ for all x.

Let

$$\begin{aligned} \beta_x &= \max\{\|x_1(t)\|, \|x_2(t)\| \mid t_0 \leq t \leq t_0 + 1\} + 1, \\ \beta_z &= \max\{\|z_1(t)\|, \|z_2(t)\|\} + 1. \end{aligned}$$

Note that $R_z \in L^1(t_0, t_0 + 1)$. On the closed ball $\{x \mid \|x\| \leq \beta_x\}$ the functions f, B, ∇G, and Q have Lipschitz constants denoted, respectively, by L_f, L_B, $L_{\nabla G}$, and L_Q, and the functions are bounded, respectively, by β_f, β_B, $\beta_{\nabla G}$, and β_Q. We restrict our attention to $t \in [t_0, t_0 + 1]$. Finally, let $\eta_Q = \min\{\lambda_{\min}(Q(x)) \mid \|x\| \leq \beta_x\}$.

Now, since $z_1(t), z_2(t) \in K$,

$$\begin{aligned} \langle z_2(t) - z_1(t),\, G(x_1(t)) + F(z_1(t))\rangle \geq 0, \\ \langle z_1(t) - z_2(t),\, G(x_2(t)) + F(z_2(t))\rangle \geq 0. \end{aligned}$$

Adding gives (after some rearrangement)

$$\langle z_2(t) - z_1(t),\, G(x_2(t)) + F(z_2(t)) - G(x_1(t)) - F(z_1(t))\rangle \leq 0.$$

Since F is monotone, $\langle z_2(t) - z_1(t),\, F(z_2(t)) - F(z_1(t))\rangle \geq 0$, so

$$\langle u_2(t) - u_1(t),\, G(x_2(t)) - G(x_1(t))\rangle \leq 0. \tag{5.21}$$

Now

$$G(x_2(t)) - G(x_1(t)) = \nabla G(x_1(t))(x_2(t) - x_1(t)) + h(t),$$

where $\|h(t)\| \leq L_{\nabla G}\|x_2(t) - x_1(t)\|^2$. But $P(x) := \nabla G(x)\, B(x)$ is symmetric positive definite and locally Lipschitz continuous. Thus both $\nabla G(x)$ and $B(x)$ must have full rank.

Then (5.21) can be rewritten as

$$\Big\langle u_2(t)-u_1(t),\, B(x_1(t))^T\, Q(x_1(t))(x_2(t)-x_1(t))\Big\rangle \le \beta_z(t)\, L_{\nabla G}\|x_2(t)-x_1(t)\|^2.$$

Thus

$$\langle B(x_1(t))u_2(t)-B(x_1(t))u_1(t),\, Q(x_1(t))(x_2(t)-x_1(t))\rangle \le \beta_z(t)\, L_{\nabla G}\|x_2(t)-x_1(t)\|^2,$$

which implies that

$$\begin{aligned}&\langle B(x_2(t))u_2(t)-B(x_1(t))u_1(t),\, Q(x_1(t))(x_2(t)-x_1(t))\rangle\\ &\quad\le (\beta_z(t)\, L_{\nabla G}+\beta_Q L_B)\|x_2(t)-x_1(t)\|^2.\end{aligned}$$

Using the differential equation (5.18), we get

$$\begin{aligned}&\Big\langle x_2'(t)-f(x_2(t))-x_1'(t)+f(x_1(t)),\, Q(x_1(t))(x_2(t)-x_1(t))\Big\rangle\\ &\quad\le (\beta_z(t)\, L_{\nabla G}+\beta_Q\,\beta_z(t)\, L_B)\|x_2(t)-x_1(t)\|^2.\end{aligned}$$

Since f is locally Lipschitz, we get

$$\begin{aligned}&\Big\langle x_2'(t)-x_1'(t),\, Q(x_1(t))(x_2(t)-x_1(t))\Big\rangle\\ &\quad\le (\beta_z(t)\, L_{\nabla G}+\beta_Q(\beta_z(t)\, L_B+L_f))\|x_2(t)-x_1(t)\|^2.\end{aligned}$$

Since Q is Lipschitz and symmetric,

$$\begin{aligned}&\frac{d}{dt}\left[\frac{1}{2}\,\langle x_2(t)-x_1(t),\, Q(x_1(t))(x_2(t)-x_1(t))\rangle\right]\\ &\quad\le (\beta_z(t)\, L_{\nabla G}+\beta_Q(\beta_z(t)\, L_B+L_f)+L_Q(\beta_f+\beta_B\beta_z(t)))\|x_2(t)-x_1(t)\|^2\\ &\quad\le 2\frac{\beta_z(t)\, L_{\nabla G}+\beta_Q(\beta_z(t)\, L_B+L_f)+L_Q(\beta_f+\beta_B\,\beta_z(t))}{\eta_Q}\\ &\qquad\times\frac{1}{2}\,\langle x_2(t)-x_1(t),\, Q(x_1(t))(x_2(t)-x_1(t))\rangle\,.\end{aligned}$$

A Gronwall lemma can then give the result that

$$\begin{aligned}&\langle x_2(t)-x_1(t),\, Q(x_1(t))(x_2(t)-x_1(t))\rangle\\ &\quad\le e^{C(t)}\Big\langle x_2(t^*)-x_1(t^*),\, Q(x_1(t^*))(x_2(t^*)-x_1(t^*))\Big\rangle=0,\end{aligned}$$

where

$$C(t):=\int_{t^*}^{t} 2(\beta_z(s)L_{\nabla G}+\beta_Q(\beta_z(s)L_B+L_f)+L_Q(\beta_f+\beta_B\beta_z(s)))ds/\eta_Q.$$

As $x_1(t^*)=x_2(t^*)$, for some $\epsilon>0$, $x_2(t)=x_1(t)$ for all $t\in[t^*,t^*+\epsilon]$, contradicting the assertion that $t^*<+\infty$. Hence the solution of the DVI is unique. □

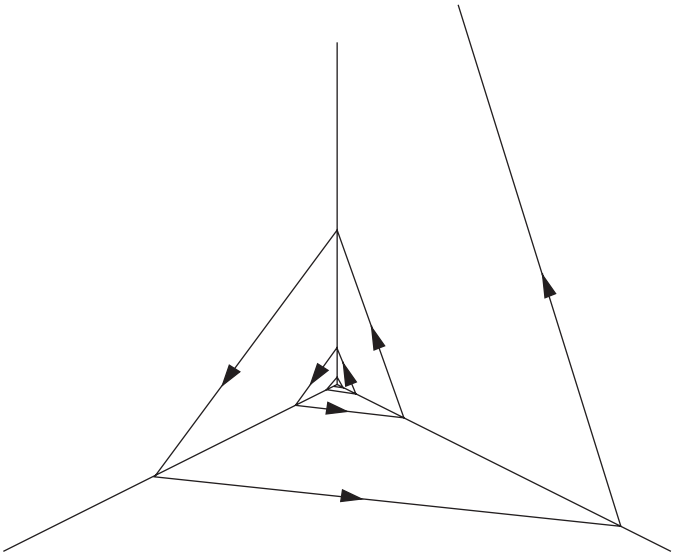


Figure 5.1: Nontrivial solutions for the DCP of Bernard and el Kharroubi. Reprinted with permission.

This result not only establishes uniqueness, but the proof can also be used to show that the solution map $x(t_0) \mapsto x(t)$ for fixed $t > t_0$ is also locally Lipschitz. This proof makes no use of the structure of K. Simpler problems for which sharper conditions can be found include linear G and constant B.

A question arises here: How important is *symmetry* to uniqueness? In certain situations we can show that there are plenty of matrices ∇G and B which give uniqueness without $\nabla G\, B$ being symmetric [245]. Understanding this requires a deeper understanding of how the matrix $\nabla G\, B$ relates to the solutions of the DVI. To go deeper into these issues, [245] considers DCPs of the form

$$\frac{dw}{dt}(t) = M z(t) + q(t), \qquad w(0) = w_0, \tag{5.22}$$

$$K^* \ni w(t) \perp z(t) \in K, \tag{5.23}$$

which were essentially those studied by Mandelbaum [164] and Bernard and el Kharroubi [30]. For existence we need $w_0 \in K^*$. The example of nonuniqueness of Bernard and el Kharroubi was for $K = \mathbb{R}^3_+$ and

$$M = \begin{bmatrix} 1 & 3 & 0 \\ 0 & 1 & 3 \\ 3 & 0 & 1 \end{bmatrix},$$

which is a P-matrix, but is not positive definite. Then, with $q(t) \equiv -[1, 1, 1]^T$ and $w_0 = 0$, there are both the zero solution $w(t) \equiv 0$ and a "cobweb" solution that spirals out from the origin in finite time. A similar solution is in Stewart [236, Appendix C]. See Figure 5.1.

Mandelbaum was able to construct an example with $K = \mathbb{R}^2_+$ and

$$M = \begin{bmatrix} 1 & -2 \\ 1 & 1 \end{bmatrix},$$

with a nonanalytic $q(\cdot)$ giving nonunique solutions. Note that this matrix M is not only a P-matrix (which gives uniqueness for the static problem) but is also positive definite. The basis of this nonuniqueness result is the following result of Mandelbaum [164].

Lemma 5.5. *The DCP* (5.22)–(5.23) *with* $K = \mathbb{R}^n_+$ *has unique solutions for all* w_0 *and* $q(\cdot)$ *if and only if the only pair* (ζ, ω) *satisfying*

$$\frac{d\omega}{dt} = M\zeta(t), \qquad \omega(0) = 0, \tag{5.24}$$

$$\omega(t) \circ \zeta(t) \leq 0 \qquad \text{for all } t \tag{5.25}$$

with the inequality in (5.25) *understood componentwise is* $\omega \equiv 0$ *and* $\zeta \equiv 0$ *(almost everywhere).*

Note that "$\circ$" is the componentwise or Hadamard product: $(a \circ b)_i = a_i b_i$. While a natural generalization to Mandelbaum's result would be to replace "$\circ$" by a general Jordan algebra, this does not in fact hold. At the time of this writing, it is not known exactly which matrices M ensure uniqueness of solutions of (5.22)–(5.23), except for 1×1 and 2×2 matrices. For the 1×1 case, M simply needs to be a positive number. For the 2×2 case,

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

gives unique solutions for (5.22)–(5.23) with $K = \mathbb{R}^2_+$ if and only if $a > 0$, $d > 0$, $ad - bc > 0$, and $ad + bc \geq 0$ [245]. This set strictly includes all symmetric positive definite matrices, but it is definitely smaller than the set of all strictly monotone matrices (whose symmetric part is positive definite).

Proving nonuniqueness via Mandelbaum's theorem for specific examples involves finding ζ and ω not identically zero satisfying (5.24)–(5.25). In the case of $M = \left[\begin{smallmatrix} 1 & -2 \\ 1 & 1 \end{smallmatrix}\right]$, this can be done geometrically; see Figure 5.2.

Since we must have $\zeta(t) \circ \omega(t) \leq 0$ for all t, we must have $\zeta_1(t)\omega_1(t) \leq 0$ and $\zeta_2(t)\omega_2(t) \leq 0$. If $\omega(t)$ is in the first quadrant where $\omega_1(t), \omega_2(t) > 0$, we must have $\zeta_1(t), \zeta_2(t) \leq 0$. Similarly we obtain constraints on the signs of $\zeta_1(t)$ and $\zeta_2(t)$ depending on the quadrant that $\omega(t)$ belongs to. In Figure 5.2, the dashed lines show the directions of $d\omega/dt$ for $\zeta_1(t) > 0$, $\zeta_2(t) = 0$ and for $\zeta_1(t) = 0$, $\zeta_2(t) > 0$. Naturally we pick the admissible direction for $d\omega/dt$ which moves away from the origin as fast as possible, in the hopes of leaving the origin. This gives the cobweb dynamics apparent in Figure 5.2.

To show the opposite, that it is not possible to leave the origin, we can use a Lyapunov function argument. That is, we seek a function $V(\omega)$ such that we can guarantee that V has a global minimum at zero, and that $(d/dt)\,V(\omega(t)) \leq 0$ for all t. Typically, the function V is Lipschitz but not smooth, and care must be taken when crossing discontinuities of ∇V. For more details, see [164, 245].

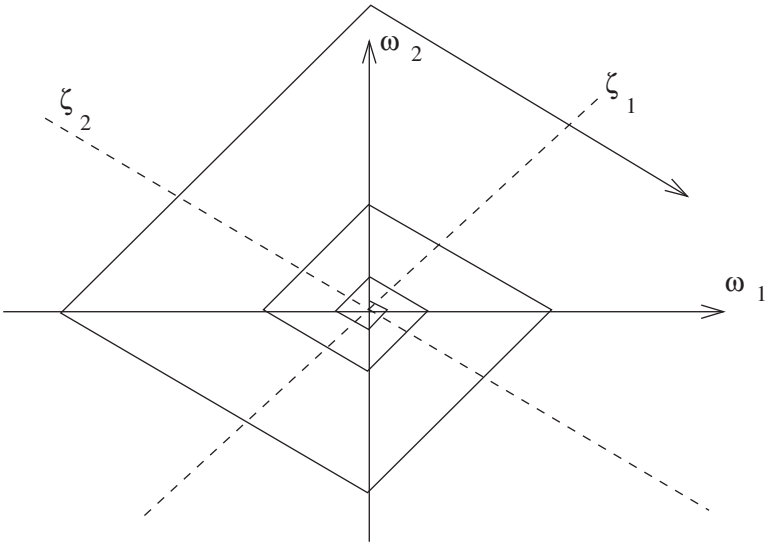


Figure 5.2: Geometry of Mandelbaum's counterexample to uniqueness of solutions. Reprinted with permission.

5.3 Convolution complementarity problems

In this section we return to the topic of CCPs first mentioned in Section 4.6. In Section 4.6.1, index-zero CCPs are shown to have solutions which are unique under mild conditions. While Section 4.6.2 treats index-one CCPs, it does not establish their main properties, particularly existence and uniqueness of solutions. This we do in this section.

Recall from (4.43)–(4.44) that a CCP in finite dimensions has the following form: Given $m\colon [0,T] \to \mathbb{R}^{n\times n}$ and $q\colon [0,T] \to \mathbb{R}^n$ and a closed convex set $K \subset \mathbb{R}^n$, find $z\colon [0,T] \to \mathbb{R}^n$ satisfying

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^*, \qquad \text{where}$$
$$(m * z)(t) = \int_0^t m(t-\tau) z(\tau)\, d\tau.$$

As an example of how this can be applied in practice, consider a rod impacting a table as discussed in Section 6.2.4. This problem can be represented as a CCP for the normal contact force $N(t)$ with the kernel function

$$m(t) = c^{-1}\left[H(ct) + \sum_{k=1}^{\infty} 2\, H(ct - 2k\ell) \right]. \tag{5.26}$$

In (5.26), $H(s)$ is the Heaviside function $H(s) = +1$ for $s > 0$, $H(s) = 0$ for $s < 0$, and $H(0)$ is either undefined or $1/2$. Also c is the wave speed in the rod, and ℓ is the length of the rod. Despite the fact that this is a mechanical impact problem formally with index two, we see that the CCP representing it is in fact an index-one CCP: $m(0^+) = c^{-1} > 0$.

Furthermore, even though $m(\cdot)$ is a discontinuous function, it has bounded variation on any finite interval, and on a sufficiently small interval $(0, T^*)$ it is constant.

CCPs of index one turn out to be useful in studying, for example, certain mechanical impact problems, even though these problems are formally of index two. Another application of CCPs that cannot be easily treated by other means is the example of a diode at the end of a transmission line in (1.13)–(1.15); see also Figure 1.10.

5.3.1 Existence of solutions to CCPs

First we give an existence result for a general class of index-one CCPs. The following theorem was first proved in [242] for $K = \mathbb{R}^n_+$ and $m(0^+)$ a P-matrix using a time-discretization argument. The argument used here is a shorter one using a differentiation lemma.

Theorem 5.6. *Suppose that M_0 is a symmetric positive semidefinite matrix and $m : [0, T] \to \mathbb{R}^{n\times n}$ has bounded variation with $M_0 + \rho\, m\left(0^+\right)$ strongly K-copositive for a closed convex cone K and $\rho > 0$. Then, provided $q(0) \in K^*$ and $q \in W^{1,2}(0, T; \mathbb{R}^n)$, there is a solution to the CCP*

$$K \ni z(t) \perp M_0 z(t) + (m * z)(t) + q(t) \in K^* \qquad \text{for all } t. \tag{5.27}$$

The solution $z \in L^2(0, T; \mathbb{R}^n)$.

If $M_0 = 0$ and instead of $q \in W^{1,2}(0, T; \mathbb{R})$ we have $q \in W^{1,p}(0, T; \mathbb{R})$, $1 \le p \le \infty$, then a solution z exists for (5.27) *with $z \in L^p(0, T; \mathbb{R})$.*

Proof. Suppose for now that $q \in W^{1,2}(0, T; \mathbb{R}^n)$. Consider the approximate problem

$$K \ni z_\epsilon(t) \perp (M_0 + \epsilon I) z_\epsilon(t) + (m * z_\epsilon)(t) + q(t) \in K \tag{5.28}$$

for all t. Since $M_0 + \epsilon I$ is positive definite, there is a well-defined Lipschitz solution operator $\mathrm{sol}_\epsilon(b) = z$ for the LCP

$$K \ni z \perp (M_0 + \epsilon I) z + b \ni K^*.$$

Thus (5.28) can be represented as

$$z_\epsilon(t) = \mathrm{sol}_\epsilon\left((m * z_\epsilon)(t) + q(t)\right).$$

Standard Picard iterations

$$z_\epsilon^{(k+1)}(t) = \mathrm{sol}_\epsilon\left(\left(m * z_\epsilon^{(k)}\right)(t) + q(t)\right)$$

converge by a contraction mapping argument over $C([0, T]; \mathbb{R}^n)$, at least on sufficiently small intervals $\left[0, T^*\right]$, $T^* > 0$. Standard continuation arguments allow us to extend this

existence result to intervals $[0,T]$ for any $T>0$. The solutions z_ϵ are absolutely continuous since for $t>s$,

$$\begin{aligned}\|z_\epsilon(t)-z_\epsilon(s)\| &\le L_\epsilon \|(m*z_\epsilon)(t)+q(t)-(m*z_\epsilon)(s)-q(s)\| \\ &\le L_\epsilon\left[\left(\|m\|_{L^\infty}+\bigvee_0^s m\right)\|z_\epsilon\|_{L^\infty}(t-s)+\|q(t)-q(s)\|\right],\end{aligned} \tag{5.29}$$

and q is absolutely continuous. Note that m' does exist in the sense of distributions and is a measure.

We now wish to obtain bounds on the solutions z_ϵ that are independent of $\epsilon>0$. We do this using the differentiation lemma (Lemma 3.2) applied to (5.28):

$$\left\langle z_\epsilon(t), \frac{d}{dt}\left((M_0+\epsilon I)z_\epsilon(t)+(m*z_\epsilon)(t)+q(t)\right)\right\rangle = 0$$

for almost all t. That is,

$$\begin{aligned}0 &= \left\langle z_\epsilon(t), \frac{d}{dt}\left[(M_0+\epsilon I)z_\epsilon(t)\right]\right\rangle + \left\langle z_\epsilon(t), \frac{d}{dt}(m*z_\epsilon)(t)\right\rangle \\ &\qquad + \left\langle z_\epsilon(t), q'(t)\right\rangle \\ &= \frac{d}{dt}\frac{1}{2}\left\langle z_\epsilon(t),(M_0+\epsilon I)z_\epsilon(t)\right\rangle + \left\langle z_\epsilon(t), \frac{d}{dt}(m*z_\epsilon)(t)\right\rangle \\ &\qquad + \left\langle z_\epsilon(t), q'(t)\right\rangle.\end{aligned} \tag{5.30}$$

Note that whenever the derivative exists,

$$\begin{aligned}&\frac{d}{dt}(m*z_\epsilon)(t) \\ &\quad = \lim_{h\downarrow 0}\frac{1}{h}\left[\int_0^{t+h} m(t+h-\tau)z_\epsilon(\tau)\,d\tau - \int_0^t m(t-\tau)z_\epsilon(\tau)\,d\tau\right] \\ &\quad = \lim_{h\downarrow 0}\frac{1}{h}\left[\int_0^t \left(m(t+h-\tau)-m(t-\tau)\right)z_\epsilon(\tau)\,d\tau \right. \\ &\qquad \left. + \int_t^{t+h} m(t+h-\tau)z_\epsilon(\tau)\,d\tau\right].\end{aligned}$$

Now

$$\begin{aligned}&h^{-1}\int_0^t \left(m(t+h-\tau)-m(t-\tau)\right)z_\epsilon(\tau)\,d\tau \\ &\qquad = \int_0^t h^{-1}\int_{[0,h)} m'(t+s-\tau)\,ds\,z_\epsilon(\tau)\,d\tau.\end{aligned}$$

Now $h^{-1}\int_{[0,h)} m'(t+s-\tau)\,ds \to m'(t-\tau)$ weak* as measures over τ. Furthermore, $\lim_{s\downarrow 0} m(s) = m\left(0^+\right)$. So

$$\frac{d}{dt}(m * z_\epsilon)(t) = m\left(0^+\right) z_\epsilon(t) + \int_{[0,t)} m'(t-\tau)\, z_\epsilon(\tau)\, d\tau.$$

Integrating (5.30) over $[0,T]$ gives

$$\begin{aligned} 0 = {} & \frac{1}{2}\langle z_\epsilon(T), (M_0+\epsilon I) z_\epsilon(T)\rangle - \frac{1}{2}\langle z_\epsilon(0), (M_0+\epsilon I) z_\epsilon(0)\rangle \\ & + \int_0^T \left\langle z_\epsilon(t), m\left(0^+\right) z_\epsilon(t)\right\rangle dt \\ & + \int_0^T \int_{[0,t)} \left\langle z_\epsilon(t), m'(t-\tau)\, z_\epsilon(\tau)\right\rangle d\tau\, dt \\ & + \int_0^T \left\langle z_\epsilon(t), q'(t)\right\rangle dt. \end{aligned}$$

Since $z_\epsilon(0) = \mathrm{sol}_\epsilon(q(0)) = 0$, we can remove the terms involving $z_\epsilon(0)$, and so we obtain the inequality

$$\begin{aligned} \int_0^T \left\langle z_\epsilon(t), -q'(t)\right\rangle dt &\geq \frac{1}{2}\langle z_\epsilon(T), M_0 z_\epsilon(T)\rangle + \int_0^T \left\langle z_\epsilon(t), m\left(0^+\right) z_\epsilon(t)\right\rangle dt \\ &\quad + \int_0^T \int_{[0,t)} \left\langle z_\epsilon(t)\, m'(t-\tau)\, z_\epsilon(\tau)\right\rangle d\tau\, dt \\ &\geq \frac{1}{2}\langle z_\epsilon(T), M_0 z_\epsilon(T)\rangle + \int_0^T \left\langle z_\epsilon(t), m\left(0^+\right) z_\epsilon(t)\right\rangle dt \\ &\quad - \|z_\epsilon\|^2_{L^2(0,T)} \bigvee_{0^+}^T m, \end{aligned}$$

where $\bigvee_{a^+}^b m = \lim_{s\downarrow a} \bigvee_s^b m$. That is,

$$\begin{aligned} \|z_\epsilon\|_{L^2(0,T)} \left\|q'\right\|_{L^2(0,T)} &\geq \frac{1}{2}\left\|M_0^{1/2} z_\epsilon(T)\right\|^2 + \int_0^T \left\langle z_\epsilon(t), m\left(0^+\right) z_\epsilon(t)\right\rangle dt \\ &\quad - \|z_\epsilon\|^2_{L^2(0,T)} \bigvee_{0^+}^T m. \end{aligned}$$

To bound $\int_0^T \left\langle z_\epsilon(t), m\left(0^+\right) z_\epsilon(t)\right\rangle dt$,

$$\begin{aligned} & \int_0^T \left\langle z_\epsilon(t), m\left(0^+\right) z_\epsilon(t)\right\rangle dt \\ & \quad = \int_0^T \left\langle z_\epsilon(t), \left(m\left(0^+\right) + \rho M_0\right) z_\epsilon(t)\right\rangle dt - \rho \int_0^T \langle z_\epsilon(t), M_0 z_\epsilon(t)\rangle\, dt \\ & \quad \geq \eta \|z_\epsilon\|^2_{L^2(0,T)} - \rho T \left\|M_0^{1/2} z_\epsilon\right\|^2_{L^\infty(0,T)}, \end{aligned}$$

where $\eta > 0$ is the constant for strong K-copositivity. If we pick $T^* \geq T > 0$, then

$$\|z_\epsilon\|_{L^2(0,T^*)} \left\|q'\right\|_{L^2(0,T^*)} \geq \frac{1}{2}\left\|M_0^{1/2} z_\epsilon(T)\right\|^2 - \rho T \left\|M_0^{1/2} z_\epsilon\right\|^2_{L^\infty(0,T^*)} + \eta \|z_\epsilon\|^2_{L^2(0,T^*)} - \|z_\epsilon\|^2_{L^2(0,T^*)} \bigvee_{0^+}^{T} m.$$

Taking the supremum over $T \in (0, T^*]$ gives

$$\|z_\epsilon\|_{L^2(0,T^*)} \left\|q'\right\|_{L^2(0,T^*)} \geq \left(\frac{1}{2} - \rho T\right)\left\|M_0^{1/2} z_\epsilon\right\|^2_{L^\infty(0,T^*)} + \left(\eta - \bigvee_{0^+}^{T} m\right)\|z_\epsilon\|^2_{L^2(0,T^*)}.$$

Note that $\bigvee_{0^+}^{T} m \to 0$ as $T \downarrow 0$. So, for sufficiently small $T > 0$, we can make $\bigvee_{0^+}^{T} m \leq \frac{1}{2}\eta$ and $\rho T \leq 1/4$, so that

$$\|z_\epsilon\|_{L^2(0,T^*)} \left\|q'\right\|_{L^2(0,T^*)} \geq \frac{1}{4}\left\|M_0^{1/2} z_\epsilon\right\|^2_{L^\infty(0,T^*)} + \frac{\eta}{2}\|z_\epsilon\|^2_{L^2(0,T^*)}.$$

This shows first, uniform boundedness of $\|z_\epsilon\|_{L^2(0,T^*)}$ and, second, uniform pointwise boundedness of $M_0 z_\epsilon$.

By Alaoglu's theorem, there is a weakly convergent subsequence (also denoted by z_ϵ) where $z_\epsilon \rightharpoonup \widehat{z}$. We wish to show that $\widehat{z}$ is a (weak) solution of the CCP (5.27). Note that $z_\epsilon(t) \in K$ for all t, and since $L^2(0,T;K)$ is closed convex set, it is also weakly closed. Thus the limit satisfies $\widehat{z}(t) \in K$ for almost all t. Similarly,

$$(M_0 + \epsilon I) z_\epsilon + m * z_\epsilon + q \rightharpoonup M_0 \widehat{z} + m * \widehat{z} + q,$$

and so by the same arguments $M_0\widehat{z}(t) + (m * \widehat{z})(t) + q(t) \in K^*$ for almost all t. Finally, to show the orthogonality condition, note that since $z(\cdot) \mapsto \int_0^T \langle z(t), M_0 z(t)\rangle\, dt$ is a continuous convex function on $L^2(0,T)$, by Mazur's lemma,

$$\begin{aligned} 0 &= \limsup_{\epsilon \downarrow 0} \int_0^T \langle z_\epsilon(t), (M_0 + \epsilon I) z_\epsilon(t) + (m * z_\epsilon)(t) + q(t)\rangle\, dt \\ &\geq \int_0^T \langle \widehat{z}(t), M_0\widehat{z}(t) + (m * \widehat{z})(t) + q(t)\rangle\, dt. \end{aligned}$$

Since the final integral can never be negative, as $\widehat{z}(t) \in K$ and $M_0\widehat{z}(t) + (m * \widehat{z})(t) + q(t) \in K^*$ for almost all t, we conclude that $\widehat{z}$ does indeed solve the CCP (5.27).

The results so far show just that there is a solution $\widehat{z} \in L^2(0,T;\mathbb{R}^n)$, provided $q' \in L^2(0,T;\mathbb{R}^n)$. To extend this result to the case $M_0 = 0$ and $q \in W^{1,p}(0,T;\mathbb{R}^n)$, consider a

sequence $q_\epsilon \to q$ in $W^{1,p}(0,T;\mathbb{R}^n)$ with each $q_\epsilon \in L^2(0,T;\mathbb{R}^n)$ and $q_\epsilon(0) = q(0) \in K^*$. In particular, take

$$q_\epsilon(t) = q(0) + \int_0^t \max\left(1, \epsilon \left\|q'(t)\right\|\right)^{-1} q'(\tau)\, d\tau,$$

so that $\left\|q'_\epsilon(t)\right\| \le \left\|q'(t)\right\|$ for almost all t. There exists a corresponding solution $z_\epsilon \in L^2(0,T;\mathbb{R}^n)$. Note that if $M_0 = 0$, then $m(0^+)$ is a strongly K-copositive matrix and there is a constant $\eta_0 > 0$ such that $\langle z, m_0 z\rangle \ge \eta_0 \|z\|^2$ for all $z \in K$. Now $m * z_\epsilon$ is in $W^{1,2}(0,T;\mathbb{R}^n)$, and so it is absolutely continuous and differentiable almost everywhere. Thus, by the differentiation lemma (Lemma 3.3),

$$0 = \left\langle z_\epsilon(t), \frac{d}{dt}\left((m * z_\epsilon)(t) + q_\epsilon(t)\right)\right\rangle$$

for almost all t. Writing $m(t) = m(0^+)\,H(t) + m_1(t)$, where H is the Heaviside function $H(t) = 0$ if $t \le 0$ and $H(t) = 1$ if $t > 0$, we note that

$$(m * z_\epsilon)' = m' * z_\epsilon = m(0^+) z_\epsilon + m'_1 * z_\epsilon.$$

Thus the differentiation lemma implies

$$\begin{aligned} 0 = \left\langle z_\epsilon(t), m(0^+) z_\epsilon(t)\right\rangle &+ \left\langle z_\epsilon(t), \left(m'_1 * z_\epsilon\right)(t)\right\rangle \\ &+ \left\langle z_\epsilon(t), q'_\epsilon(t)\right\rangle, \end{aligned}$$

and so

$$\eta_0 \|z_\epsilon(t)\|^2 \le \|z_\epsilon(t)\| \left[\left\|\left(m'_1 * z_\epsilon\right)(t)\right\| + \left\|q'_\epsilon(t)\right\|\right].$$

Dividing by $\|z_\epsilon(t)\|$ for $\|z_\epsilon(t)\| \ne 0$ (the inequality is obviously true otherwise) we get

$$\eta_0 \|z_\epsilon(t)\| \le \left\|\left(m'_1 * z_\epsilon\right)(t)\right\| + \left\|q'_\epsilon(t)\right\|.$$

Now m'_1 is a measure (perhaps better written as a differential measure dm_1) which has no atom at zero: $dm(t) = m(0^+)\,\delta(t) + dm_1(t)$ in the sense of differential measures. Then we have the bound

$$\begin{aligned} \left\|\left(m'_1 * z_\epsilon\right)(t)\right\| &= \left\|\int_{(0,t]} dm_1(\tau)\, z_\epsilon(t-\tau)\, d\tau\right\| \\ &\le \int_{(0,t]} |dm_1(\tau)|\, \|z_\epsilon(t-\tau)\|\, d\tau \\ &= (|dm_1| * \zeta_\epsilon)(t), \end{aligned}$$

where $\zeta_\epsilon(t) = \|z_\epsilon(t)\|$. Note that $|dm_1|$ is the variation measure of dm_1. Thus

$$\eta_0\, \zeta_\epsilon(t) \le (|dm_1| * \zeta_\epsilon)(t) + \left\|q'(t)\right\|$$

since $\|q'_\epsilon(t)\| \le \|q'(t)\|$ for all $\epsilon > 0$. By Lemma C.3, letting $\mu(\tau) = \int_{(0,\tau]} |dm_1|$, we have

$$\zeta_\epsilon(t) \le \|q'(t)\| + \int_{(0,t]} \|q'(t-\tau)\| \, e^{\mu(t)-\mu(\tau)} \, d\mu(\tau)$$

for all $\epsilon > 0$ and almost all t. Note that the right-hand side is in $L^p(0,T)$ since $t \mapsto \|q'(t)\|$ is in $L^p(0,T)$ and by Young's lemma for convolutions with measures (A.20). Taking weak limits of the $z_\epsilon \rightharpoonup z$, z is a solution of the DVI which is in $L^p(0,T;\mathbb{R}^n)$, as required. □

Note that solving an index-one CCP behaves like a differentiation of $q(t)$. The method of proof (especially for the case $q \in W^{1,p}(0,T;\mathbb{R}^n)$) also demonstrates the power of the differentiation lemmas of Section 3.4.

These results can be extended to infinite dimensions under some additional restrictions. For example, if $M_0 = 0$, we have existence of solutions if m_0 is strongly K-copositive and can be written as the sum of a compact linear operator and a monotone linear operator.

5.3.2 Uniqueness for CCPs

Uniqueness of solutions requires both a little more regularity and some restrictions on the structure of the problem. If $m(t) = m_0$ for all $t > 0$, then we can easily show uniqueness if m_0 is *symmetric* and positive definite: If we have two solutions

$$\begin{aligned} K \ni z_1(t) \perp (m * z_1)(t) + q(t) \in K^*, \\ K \ni z_2(t) \perp (m * z_2)(t) + q(t) \in K^*, \end{aligned}$$

then we set $w_i = (m * z_i)(t) + q(t)$, $\zeta(t) = z_1(t) - z_2(t)$, and $\omega(t) = w_1(t) - w_2(t)$. Since $K \ni z_i(t) \perp w_i(t) \in K^*$ for almost all t, we have

$$\begin{aligned} \langle \zeta(t), \omega(t) \rangle &= \langle z_1(t) - z_2(t), w_1(t) - w_2(t) \rangle \\ &= \langle z_1(t), w_1(t) \rangle - \langle z_1(t), w_2(t) \rangle \\ &\qquad - \langle z_2(t), w_1(t) \rangle + \langle z_2(t), w_2(t) \rangle \le 0 \end{aligned}$$

for almost all t. Now $\omega(t) = (m * \zeta)(t)$ and we are assuming $m(t) = m_0$ for all $t > 0$, so this means that for any $\tau > 0$,

$$\begin{aligned} 0 \ge \int_0^\tau \langle \omega(t), \zeta(t) \rangle \, dt &= \int_0^\tau \langle (m * \zeta)(t), \zeta(t) \rangle \, dt \\ &= \int_0^\tau \left\langle \int_0^t m_0 \zeta(s), \zeta(t) \right\rangle ds \, dt \\ &= \int_0^\tau \int_0^t \langle m_0 \zeta(s), \zeta(t) \rangle \, ds \, dt. \end{aligned}$$

If m_0 is symmetric, this is half of the integral over the square $[0,\tau]\times[0,\tau]$:

$$\begin{aligned} 0 &\geq \frac{1}{2}\int_0^\tau \int_0^\tau \langle m_0\zeta(s),\zeta(t)\rangle\, ds\, dt \\ &= \frac{1}{2}\left\langle m_0 \int_0^\tau \zeta(s)\, ds, \int_0^\tau \zeta(t)\, dt \right\rangle. \end{aligned}$$

Thus $\int_0^\tau \zeta(s)ds = 0$ for all $\tau > 0$; thus $\zeta(s) = 0$ for almost all $s > 0$. This means that $z_1(s) = z_2(s)$ for almost all $s > 0$, and so the solution is (essentially) unique.

Dealing with a nonconstant kernel function $m(t)$ requires using integration by parts twice to obtain a suitable inequality which shows uniqueness. This method can also show a weak kind of continuity of the solution map $q(\cdot)\mapsto z(\cdot)$.

Theorem 5.7. *Suppose that M_0 is a symmetric positive semidefinite matrix and that $m(\cdot)$ and $m_0 = m(0^+)$ satisfy the following assumptions:*

- *m_0 is symmetric positive definite on range$(M_0)^\perp$. That is, if $u, v \neq 0$ are in range$(M_0)^\perp$, then $\langle u, m_0 v\rangle = \langle v, m_0 u\rangle$ and $\langle u, m_0 u\rangle > 0$.*
- *$\left(m_0 - m_0^T\right)$ range $M_0 \subseteq (\text{range}\, M_0)^\perp$, $\left(m_0 - m_0^T\right)(\text{range}\, M_0)^\perp \subseteq \text{range}\, M_0$.*
- *$m\colon [0,T]\to\mathbb{R}^{n\times n}$ has bounded variation with m' also a function of bounded variation on an interval $(0,T^*)$, $T^* > 0$.*

Then the CCP

$$K \ni z(t) \perp M_0 z(t) + (m * z)(t) + q(t) \in K^* \qquad \text{for all } t \tag{5.31}$$

has a unique solution.

This result extends the result in [242] by allowing $M_0 \neq 0$; that is, there can be an index-zero part as well as an index-one part. A pure index-zero problem requires M_0 only to be strongly monotone in order to show existence and uniqueness. A pure index-one problem has $M_0 = 0$, in which case we need $m_0 = m(0^+)$ to be symmetric as well as positive definite for this proof to hold.

Proof. Suppose that z_1 and z_2 are two solutions of (5.31). Let $\zeta(t) = z_1(t) - z_2(t)$. If we let $w_1 = M_0 z_1 + m * z_1 + q$, $w_2 = M_0 z_2 + m * z_2 + q$, and $\omega(t) = w_1(t) - w_2(t)$, by complementarity we have $\langle \omega(t), \zeta(t)\rangle \leq 0$ for all t. If $z_1 \neq z_2$, we let $t^* = \min\{t \geq 0 \mid z_1(t) \neq z_2(t)\}$. By shifting the initial time, if $t^* < \infty$, we can make $t^* = 0$. Our task then is to show uniqueness of solutions on some sufficiently small interval $\left[0, \widehat{T}\,\right]$, $\widehat{T} > 0$.

Integrating $\langle \omega(t), \zeta(t)\rangle \leq 0$ over an interval $\left[0, \widehat{T}\,\right]$ we get

$$0 \geq \int_0^{\widehat{T}} \langle \zeta(t), M_0\zeta(t) + (m * \zeta)(t)\rangle\, dt. \tag{5.32}$$

We look carefully at

$$\int_0^{\widehat{T}} \langle \zeta(t), (m * \zeta)(t) \rangle \, dt = \int_0^{\widehat{T}} \int_0^t \langle \zeta(t), m(t-\tau)\zeta(\tau) \rangle \, d\tau \, dt$$
$$= \frac{1}{2} \int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle \zeta(t), \widetilde{m}(t-\tau)\zeta(\tau) \rangle \, d\tau \, dt,$$

where

$$\widetilde{m}(t) = \begin{cases} m(t), & t > 0, \\ m(-t)^T, & t < 0. \end{cases}$$

Note that $\widetilde{m}$ may not be continuous at zero if $m(0^+)$ is not symmetric. Let $m_0 = m(0^+)$. Also, let $Z(t) = \int_0^t \zeta(\tau) d\tau$. Then we can write, using $\nu(t) = \widetilde{m}'(t)$ and $\sigma(t) = \nu'(t)$ for $t \neq 0$ and integration by parts,

$$\begin{aligned}
&\int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle \zeta(t), \widetilde{m}(t-\tau)\zeta(\tau) \rangle \, d\tau \, dt \\
&\quad = \int_0^{\widehat{T}} \int_0^{\widehat{T}} \left\langle Z'(t), \widetilde{m}(t-\tau) Z'(\tau) \right\rangle dt \, d\tau \\
&\quad = \int_0^{\widehat{T}} \left[\left\langle Z(t), \widetilde{m}(t-\tau) Z'(\tau) \right\rangle \Big|_{t=0}^{t=\tau^-} + \left\langle Z(t), \widetilde{m}(t-\tau) Z'(\tau) \right\rangle \Big|_{t=\tau^+}^{t=\widehat{T}} \right. \\
&\qquad \left. - \int_0^{\widehat{T}} \left\langle Z(t), \nu(t-\tau) Z'(\tau) \right\rangle dt \right] d\tau \\
&\quad = \int_0^{\widehat{T}} \left[\left\langle Z(\tau), \widetilde{m}(0^-) Z'(\tau) \right\rangle - \left\langle Z(\tau), \widetilde{m}(0^+) Z'(\tau) \right\rangle \right] d\tau \\
&\qquad + \int_0^{\widehat{T}} \left\langle Z(\widehat{T}), \widetilde{m}(\widehat{T}-\tau) Z'(\tau) \right\rangle d\tau \\
&\qquad - \int_0^{\widehat{T}} \int_0^{\widehat{T}} \left\langle Z(t), \nu(t-\tau) Z'(\tau) \right\rangle d\tau \, dt \\
&\quad = \int_0^{\widehat{T}} \left\langle Z(\tau), \left(m_0^T - m_0 \right) Z'(\tau) \right\rangle d\tau \\
&\qquad + \int_0^{\widehat{T}} \left\langle Z(\widehat{T}), \widetilde{m}(\widehat{T}-\tau) Z'(\tau) \right\rangle d\tau \\
&\qquad - \int_0^{\widehat{T}} \langle Z(t), \nu(t-\tau) Z(\tau) \rangle \Big|_{t=0}^{t=\tau^-} d\tau \\
&\qquad - \int_0^{\widehat{T}} \langle Z(t), \nu(t-\tau) Z(\tau) \rangle \Big|_{t=\tau^+}^{t=\widehat{T}} d\tau \\
&\qquad - \int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle Z(t), \sigma(t-\tau) Z(\tau) \rangle \, d\tau \, dt
\end{aligned}$$

$$
\begin{aligned}
&= \int_0^{\widehat{T}} \left\langle Z(\tau), \left(m_0^T - m_0\right) Z'(\tau) \right\rangle d\tau \\
&\quad + \left\langle Z(\widehat{T}), \widetilde{m}(\widehat{T}-\tau)\, Z(\tau) \right\rangle \Big|_{\tau=0}^{\tau=\widehat{T}^-} + \int_0^{\widehat{T}} \left\langle Z(\widehat{T}), \nu(\widehat{T}-\tau)\, Z(\tau) \right\rangle d\tau \\
&\quad - \int_0^{\widehat{T}} \left\langle Z(\tau), \left(\nu(0^-) - \nu(0^+)\right) Z(\tau) \right\rangle d\tau \\
&\quad - \int_0^{\widehat{T}} \left\langle Z(\widehat{T}), \nu(\widehat{T}-\tau)\, Z(\tau) \right\rangle d\tau \\
&\quad - \int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle Z(t), \sigma(t-\tau)\, Z(\tau) \rangle \, d\tau \, dt \\
&= \left\langle Z(\widehat{T}), m_0\, Z(\widehat{T}) \right\rangle + \int_0^{\widehat{T}} \left\langle Z(\tau), \left(m_0^T - m_0\right) Z'(\tau) \right\rangle d\tau \\
&\quad + \int_0^{\widehat{T}} \left\langle Z(\tau), \left(\nu(0^+) - \nu(0^-)\right) Z(\tau) \right\rangle d\tau \\
&\quad - \int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle Z(t), \sigma(t-\tau)\, Z(\tau) \rangle \, d\tau \, dt. \qquad (5.33)
\end{aligned}
$$

We split $\zeta(t) = \zeta_1(t) + \zeta_2(t)$, where $\zeta_1(t) \in \operatorname{range} M_0$ and $\zeta_2(t) \perp \operatorname{range} M_0$, and let $Z_1(t) = \int_0^t \zeta_1(\tau)\, d\tau$, $Z_2(t) = \int_0^t \zeta_2(\tau)\, d\tau$. Let λ_M be the smallest nonzero eigenvalue of M_0; $\lambda_M > 0$. Note that $\|Z_i(\tau)\| \le \tau^{1/2} \|\zeta_i\|_{L^2(0,\tau)}$.

Returning to (5.32), substituting (5.33) gives

$$
\begin{aligned}
0 \ge &\int_0^{\widehat{T}} \langle \zeta(t)\, M_0 \zeta(t) \rangle \, dt + \left\langle Z(\widehat{T}), m_0\, Z(\widehat{T}) \right\rangle \\
&+ \int_0^{\widehat{T}} \left\langle Z(\tau), \left(m_0^T - m_0\right) \zeta(\tau) \right\rangle d\tau \\
&+ \int_0^{\widehat{T}} \left\langle Z(\tau), \left(\nu(0^+) - \nu(0^-)\right) Z(\tau) \right\rangle d\tau \\
&- \int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle Z(t), \sigma(t-\tau)\, Z(\tau) \rangle \, d\tau \, dt.
\end{aligned}
$$

The biggest difficulty is with the antisymmetric part of m_0:

$$
\int_0^{\widehat{T}} \left\langle Z(\tau), \left(m_0^T - m_0\right) \zeta(\tau) \right\rangle d\tau.
$$

Here we need to use the splitting and the properties that $\left(m_0 - m_0^T\right) \operatorname{range} M_0 \subseteq (\operatorname{range} M_0)^\perp$ and $\left(m_0 - m_0^T\right) (\operatorname{range} M_0)^\perp \subseteq \operatorname{range} M_0$. From the splitting $Z(t) = Z_1(t) + Z_2(t)$, $Z_1(t) \in \operatorname{range} M_0$, $Z_2(t) \in (\operatorname{range} M_0)^\perp$; then

$$\begin{aligned}
&\int_0^{\widehat{T}} \left\langle Z(\tau), \left(m_0^T - m_0\right) Z'(\tau) \right\rangle d\tau \\
&\quad = \int_0^{\widehat{T}} \left\langle Z_1(\tau), \left(m_0^T - m_0\right) Z_2'(\tau) \right\rangle d\tau \\
&\qquad + \int_0^{\widehat{T}} \left\langle Z_2(\tau), \left(m_0^T - m_0\right) Z_1'(\tau) \right\rangle d\tau \\
&\quad = \left\langle Z_1(\tau), \left(m_0^T - m_0\right) Z_2(\tau) \right\rangle \Big|_{\tau=0}^{\tau=\widehat{T}} \\
&\qquad - \int_0^{\widehat{T}} \left\langle Z_1'(\tau), \left(m_0^T - m_0\right) Z_2(\tau) \right\rangle d\tau \\
&\qquad + \left\langle Z_2(\tau), \left(m_0^T - m_0\right) Z_1'(\tau) \right\rangle d\tau \\
&\quad = \left\langle Z_1(\widehat{T}), \left(m_0^T - m_0\right) Z_2(\widehat{T}) \right\rangle \\
&\qquad + 2 \int_0^{\widehat{T}} \left\langle Z_2(\tau), \left(m_0^T - m_0\right) \zeta_1(\tau) \right\rangle d\tau .
\end{aligned}$$

Note the following bounds: for suitable constants α and λ_m,

$$\begin{aligned}
&\int_0^{\widehat{T}} \langle \zeta(t)\, M_0 \zeta(t) \rangle \, dt \geq \lambda_m \, \|\zeta_1\|^2_{L^2(0,\widehat{T})}, \\
&\quad \left\langle Z(\widehat{T}), m_0\, Z(\widehat{T}) \right\rangle \geq \lambda_m \left\| Z_2(\widehat{T}) \right\|^2 - \alpha \left\| Z_1(\widehat{T}) \right\| \left\| Z(\widehat{T}) \right\| .
\end{aligned}$$

Also note that

$$\begin{aligned}
&\left| \int_0^{\widehat{T}} \left\langle Z_2(\tau), \left(m_0^T - m_0\right) \zeta_1(\tau) \right\rangle d\tau \right| \leq \widehat{T}^{1/2} \left\| m_0^T - m_0 \right\| \, \|Z_2\|_{L^\infty(0,\widehat{T})} \|\zeta_1\|_{L^2(0,\widehat{T})}, \\
&\left| \int_0^{\widehat{T}} \left\langle Z(\widehat{T}), \nu(\widehat{T} - \tau)\, Z(\tau) \right\rangle d\tau \right| \leq \widehat{T}^{1/2} \, \|\nu\|_{L^\infty(0,\widehat{T})} \left\| Z(\widehat{T}) \right\| \, \|Z\|_{L^\infty(0,\widehat{T})}, \\
&\left| \int_0^{\widehat{T}} \left\langle Z(\tau), \left(\nu(0^+) - \nu(0^-)\right) Z(\tau) \right\rangle d\tau \right| \leq \widehat{T} \left\| \nu(0^+) - \nu(0^-) \right\| \, \|Z\|^2_{L^\infty(0,\widehat{T})}, \\
&\left| \int_0^{\widehat{T}} \int_0^{\widehat{T}} \langle Z(t), \sigma(t - \tau)\, Z(\tau) \rangle \, d\tau \, dt \right| \leq \widehat{T} \, \|\sigma\|_{\mathcal{M}(0,\widehat{T})} \, \|Z\|^2_{L^\infty(0,\widehat{T})}.
\end{aligned}$$

As $\widehat{T} \downarrow 0$, $\|\sigma\|_{\mathcal{M}(0,\widehat{T})} \to 0$. With these bounds we obtain

$$\begin{aligned}
0 \geq \lambda_M \|\zeta_1\|^2_{L^2(0,\widehat{T})} + \lambda_m \left\|Z_2(\widehat{T})\right\|^2 - \alpha \left\|Z_1(\widehat{T})\right\| \left\|Z(\widehat{T})\right\| \\
- \left\|m_0^T - m_0\right\| \left\|Z_1(\widehat{T})\right\| \left\|Z_2(\widehat{T})\right\| \\
- 2\widehat{T}^{1/2} \left\|m_0^T - m_0\right\| \|Z_2\|_{L^\infty(0,\widehat{T})} \|\zeta_1\|_{L^2(0,\widehat{T})} \\
- \widehat{T} \left(\left\|\nu(0^+) - \nu(0^-)\right\| + \|\sigma\|_{\mathcal{M}(0,\widehat{T})}\right) \|Z\|^2_{L^\infty(0,\widehat{T})}.
\end{aligned}$$

Using $\|Z(\tau)\| \leq \|Z_1(\tau)\| + \|Z_2(\tau)\| \leq \tau^{1/2} \|\zeta_1\|_{L_2(0,\tau)} + \|Z_2(\tau)\|$, we obtain for $0 < \widehat{T} \leq \widetilde{T} \leq \min(1, T^*)$,

$$\begin{aligned}
0 &\geq \lambda_M \|\zeta_1\|^2_{L^2(0,\widehat{T})} + \lambda_m \left\|Z_2(\widehat{T})\right\|^2 \\
&\quad - \alpha \widetilde{T}^{1/2} \|\zeta_1\|_{L^2(0,\widetilde{T})} \left(\widetilde{T}^{1/2} \|\zeta_1\|_{L^2(0,\widetilde{T})} + \|Z_2\|_{L^\infty(0,\widetilde{T})}\right) \\
&\quad - 3\widetilde{T}^{1/2} \left\|m_0^T - m_0\right\| \|\zeta_1\|_{L^2(0,\widetilde{T})} \|Z_2\|_{L^\infty(0,\widetilde{T})} \\
&\quad - \widetilde{T} \left(\left\|\nu(0^+) - \nu(0^-)\right\| + \|\sigma\|_{\mathcal{M}(0,\widetilde{T})}\right) \\
&\qquad \times \left(\widetilde{T}^{1/2} \|\zeta_1\|_{L^2(0,\widetilde{T})} + \|Z_2\|_{L^\infty(0,\widetilde{T})}\right)^2 \\
&\geq \lambda_M \|\zeta_1\|^2_{L^2(0,\widehat{T})} + \lambda_m \left\|Z_2(\widehat{T})\right\|^2 \\
&\quad - C\, \widetilde{T}^{1/2} \left(\|\zeta_1\|^2_{L^2(0,\widetilde{T})} + \|Z_2\|^2_{L^\infty(0,\widetilde{T})}\right)
\end{aligned}$$

for a suitable positive constant C. (Note that this uses the inequality $ab \leq \frac{1}{2}(a^2 + b^2)$.) Taking the supremum over $0 < \widehat{T} \leq \widetilde{T}$ gives

$$0 \geq \lambda_M \|\zeta_1\|^2_{L^2(0,\widehat{T})} + \lambda_m \left\|Z_2(\widehat{T})\right\|^2 - C\, \widetilde{T}^{1/2} \left(\|\zeta_1\|^2_{L^2(0,\widetilde{T})} + \|Z_2\|^2_{L^\infty(0,\widetilde{T})}\right).$$

Therefore, provided $\lambda_M, \lambda_m > C\, \widetilde{T}^{1/2}$, or equivalently if $\widetilde{T} < \min(\lambda_M/C, \lambda_m/C)^2$, we have $\|\zeta_1\|_{L^2(0,\widetilde{T})} = \|Z_2\|_{L^\infty(0,\widetilde{T})} = 0$, and we must have $\zeta_1 \equiv 0$ and $Z_2 \equiv 0$. Thus $\zeta_1 \equiv \zeta_2 \equiv 0$ and $z_1 \equiv z_2$ on $(0, \widetilde{T})$. Thus we have uniqueness. □

The role of symmetry here is quite important. The antisymmetry of the index-one matrix $m_0 - m_0^T$ has to be controlled by the index-zero matrix M_0. If $M_0 = 0$, for a pure index-one problem, this proof requires m_0 to be symmetric, which is essentially equivalent to the requirement that $\nabla G(x)\, B(x)$ be symmetric in the DVI case.

5.4 Application: Circuits with diodes

Diodes are electrical devices that are intended to allow current to pass in only one direction. There is, of course, some nonideal behavior, but unless we wish to model these nonideal characteristics, we are led to consider DCPs. While a diode itself has a static voltage-current relationship, many other circuit elements such as capacitors and inductors do not (see Figure 5.3).

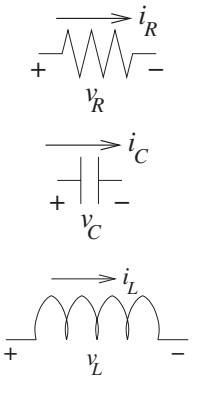


Figure 5.3: Circuit elements and their behavior.

5.4.1 Obtaining differential equations from circuits

The method described below is not the only method for obtaining differential equations from circuits. For more information, see, for example, [136].

Circuits without diodes are modeled by differential equations or by differential algebraic equations. To generate these models, the circuit needs to be analyzed. Circuits, such as those shown in Figure 1.6 in Section 1.3, are given as networks or graphs with nodes (specific points in the circuit) and edges (circuit elements). Each node has an associated voltage and each edge has an associated current. Since it is only the voltage differences across each circuit element that results in currents, there is one node that is usually assigned zero voltage. In engineering terms, this is the "earth" node of the circuit.

The equations connecting the different components are the two Kirchhoff laws:

- the total current flowing into any node is zero, and
- the sum of voltage differences around any loop is zero.

To ensure that the total current flowing into any node is zero, we create new current variables. This we do by creating a set of loops so that any consistent set of currents can be represented in terms of currents around each loop. We then require that the sum of voltage drops around each loop be zero.

To handle these problems, we need a more formal way of describing graphs or networks. For other sources for graph theory, see, for example, [33, 79, 117, 263, 266]. A *graph* is then a collection of nodes together with a set of edges that join these nodes:[6] the graph $G = (V, E)$, where V is the collection of nodes, and E is a set of edges e with two functions start, end: $E \to V$, where start(e) is the starting vertex of e, and end(e) is the end vertex of e. We denote the set of vertices of a graph G by $V(G)$ and the set of edges of G by $E(G)$. Each edge $e \in E$ may have information other than just the start and end nodes, such as the circuit element for that edge. A *subgraph* $H = (W, F)$ of $G = (V, E)$ consists of a subset of the nodes $W \subseteq V$ and a subset of the edges $F \subseteq E$ of G with start(f), end(f) $\in W$ for every $f \in F$. We say that an edge e is incident to a node x if

[6] Technically, this is the description of a *hypergraph*.

$x = \text{start}(e)$ or $x = \text{end}(e)$. A graph is *connected* if for any two nodes $x \neq y \in V$ there is a sequence $x = s_0, e_1, s_1, e_2, s_2, \ldots, e_m, s_m = y$ where edge e_i is incident to nodes s_{i-1} and s_i. This path can be written as

$$x = s_0 \xrightarrow{e_1} s_1 \xrightarrow{e_2} s_2 \xrightarrow{e_3} \cdots \xrightarrow{e_m} s_m = y.$$

A sequence like this is called a *path*. If a path has the same starting and ending points ($x = y$), it is called a *cycle*. A cycle with one edge is called a *loop*. Note that we are ignoring the orientation of the edges in defining paths. We need undirected paths in order to define trees and spanning trees below. On the other hand, orientations will be needed to define the direction of current flow in a network, and some circuit elements (such as diodes) have a definite direction.

The tool we need from graph or network theory is the *minimal spanning tree* (MST). A *tree* is a connected graph or network with no cycles. The MST T of G is a connected subgraph of G that includes all nodes of the original graph, but this is not true of any strict subgraph of T. Note that T is a tree since, if it were not, it would contain a cycle, and an edge can be removed from a cycle without making the graph disconnected. Each edge e in G, but not the MST T, would create one (and only one) cycle if added to T. Every connected graph G has an MST, and furthermore, it can be efficiently computed [64]. If G is not connected, then G can be split into connected components. A *connected component* is a subgraph H of G that is connected, while no strictly bigger subgraph of G is connected. In terms of electrical circuits, the connected components of a network are simply independent circuits and can be analyzed separately. In what follows, we will assume that the graph G representing an electrical circuit is connected.

For an electrical circuit G with an MST T, each edge e in G but not T can be assigned a current variable i_e. The edge e must be given a direction for the current to flow in (say, from $x = \text{start}(e)$ to $y = \text{end}(e)$). Currents in the opposite direction correspond to negative values for i_e. The current in the loop must go in the same direction as in the edge e. Let c_e denote this directed cycle:

$$x \xrightarrow{e} y \xrightarrow{e_1} s_1 \xrightarrow{e_2} s_2 \to \cdots \xrightarrow{e_m} s_m = x,$$

where e_i is in T for $i = 1, 2, \ldots, m$. The current in any directed edge $f \in E(T)$ is then given by

$$i_f = \sum_{e \in E(G) \setminus E(T)} b_{f,e}\, i_e \tag{5.34}$$

with $b_{f,e} = \pm 1$ according to whether the edge f appears in the forward $(+1)$ or opposite (-1) direction in the cycle c_e. If f does not appear in the cycle c_e, then we set $b_{f,e} = 0$.

Each circuit element has an associated transfer function between current velocity, which is also known as the impedance function in terms of Laplace transforms:

$$\mathcal{L}f(s) = \int_0^\infty e^{-st} f(t)\, dt. \tag{5.35}$$

The impedance functions for the different components are given via the following formulas:

$$\begin{array}{lccc} \text{resistor} & v = R\,i, & \mathcal{L}v(s) = R\,\mathcal{L}i(s) & z_R(s) = R, \\ \text{capacitor} & \dfrac{dv}{dt} = C\,i, & s\,\mathcal{L}v(s) - v(0) = C\,\mathcal{L}i(s), & z_C(s) = C\,s^{-1}, \\ \text{inductor} & v = L\,\dfrac{di}{dt}, & \mathcal{L}v(s) = L\,s\,\mathcal{L}i(s) - L\,i(0), & z_L(s) = L\,s. \end{array}$$

We can include voltage sources to obtain the general formulation for an edge e

$$\mathcal{L}v_e(s) = z_e(s)\,\mathcal{L}i_e(s) + \mathcal{L}v_{e,0}(s), \tag{5.36}$$

where $\mathcal{L}v_{e,0}(s)$ is the Laplace transform of the voltage sources plus the additional term due to the initial conditions at time $t = 0$ if appropriate.

For now, let us suppose that our circuit has the property that if we remove all the diodes (and current sources, if any), we are left with a connected network. This is not true for many interesting and useful circuits (such as the bridge rectifier, as shown in Figure 5.6), but it will simplify the analysis. The more complex case will be dealt with later.

Let H be the circuit graph G with all diodes and current sources removed, and suppose that H has an MST T. Note that because H is G with some edges removed, T is also an MST for G. For any edge e of G that is neither a diode nor a current source, $e \in E(H)\backslash E(T)$. This has a unique cycle c_e formed by e and the tree T. From the Kirchhoff voltage law,

$$\sum_{f \in c_e} b_{f,e}\,v_f(t) = 0. \tag{5.37}$$

Since $b_{f,e} = 0$, if $f \notin c_e$, we can write this as

$$\sum_{f \in E(H)} b_{f,e}\,v_f(t) = 0.$$

Note that e is included in the cycle c_e, and that $b_{e,e} = +1$. Taking Laplace transforms gives

$$\sum_{f \in E(H)} b_{f,e}\,\mathcal{L}v_f(s) = 0.$$

Substituting for $\mathcal{L}v_f(s)$ in terms of $\mathcal{L}i_f(s)$ via (5.36) gives

$$\sum_{f \in E(H)} b_{f,e}\,z_f(s)\,\mathcal{L}i_f(s) + \sum_{f \in E(H)} b_{f,e}\,\mathcal{L}v_{e,0}(s) = 0 \qquad \text{for all } e \in E(H)\backslash E(T).$$

Using the representation of i_e in terms of i_g, $g \in E(H)\backslash E(T)$, in (5.34) gives

$$\sum_{f \in E(H)} b_{f,e}\,z_f(s) \sum_{g \in E(H)\backslash E(T)} b_{g,f}\,\mathcal{L}i_g(s) + \sum_{f \in E(H)} b_{f,e}\,\mathcal{L}v_{e,0}(s) = 0$$

for all $e \in E(H)\backslash E(T)$. This can be better understood in matrix-vector terms. Let

$$\begin{aligned} \mathbf{v}_0(t) &= \big[v_{e,0}(t) \mid e \in E(H)\big], \\ \mathbf{i}(t) &= \big[i_e(t) \mid e \in E(H)\backslash E(T)\big], \\ Z_H(s) &= \operatorname{diag}(z_e(s) \mid e \in E(H)), \\ B &= \big[b_{f,e} \mid f \in E(H),\, e \in E(H)\backslash E(T)\big]. \end{aligned}$$

Then

$$B^T Z(s) B \mathcal{L}\mathbf{i}(s) + B^T \mathcal{L}\mathbf{v}_0(s) = 0. \tag{5.38}$$

This system of equations can be solved for $\mathcal{L}\mathbf{i}(s)$: The matrix $Z_H(s)$ is diagonal with positive diagonals for $s > 0$, and so it is positive definite for $s > 0$. The matrix B has linearly independent columns since for each $e \in E(H)\backslash E(T)$ we have $b_{e,e} = +1$, but for any other $g \in E(H)\backslash E(T)$, in column g, $b_{e,g} = 0$ since e is not part of the cycle formed by g and T. Then letting $\mathbf{b}_{\bullet,g}$ be column g of B, if $\sum_{g\in E(H)\backslash E(T)} \alpha_g \mathbf{b}_{\bullet,g} = 0$, taking the component for $e \in E(H)\backslash E(T)$, we get $\sum_{g\in E(H)\backslash E(T)} \alpha_g b_{e,g} = \alpha_e b_{e,e} = \alpha_e = 0$. So B has linearly independent columns: $B\boldsymbol{\alpha} = 0$ implies $\boldsymbol{\alpha} = 0$. Combining positive definiteness of $Z_H(s)$ and linear independence of columns of B gives $B^T Z_H(s) B$ positive definite, and so it is invertible. Thus we can solve the system of equations (5.38) to obtain

$$\mathcal{L}\mathbf{i}(s) = \left(B^T Z_H(s) B\right)^{-1} B^T \mathcal{L}\mathbf{v}_0(s). \tag{5.39}$$

Note that $B^T Z_H(s) B$ is a matrix of rational functions of s, and so $\left(B^T Z_H(s) B\right)^{-1}$ is also a matrix of rational functions of s, and therefore it is the Laplace transform of a matrix of distributions consisting of Dirac-δ functions, their derivatives, and sums of products of polynomials and (possibly complex) exponentials.

5.4.2 Incorporating diodes

We need to include the "external" current sources and diodes which are elements in $E(G)\backslash E(H)$. For given current sources, we do not need to solve any equations to find the current. For the diodes we have the complementarity between the reverse voltage and the forward current. For any edge $g \in E(G)\backslash E(H)$ we have a unique cycle formed by g and T which we call c_g. Recall that H is obtained from G by removing certain edges, but no nodes are removed. For such an edge $g \in E(G)\backslash E(H)$ we define $\widetilde{b}_{f,g} = \pm 1$ for $f \in c_g$ with the sign depending on whether f appears in the forward direction $(+1)$ or in the reverse direction (-1) in c_g, and $\widetilde{b}_{f,g} = 0$ if $f \notin c_g$. These form the matrix $\widetilde{B}$ whose rows are edges in $E(G)$ and columns are edges in $E(G)\backslash E(H)$.

Every edge in H has an associated impedance, but this is not so for the edges in $E(G)\backslash E(H)$ which represent diodes or current sources. There is the possibility that the network may be degenerate in the sense that there is a cycle that does not pass through any impedance element (resistor, capacitor, or inductor). A cycle consisting of a pair of ideal diodes in series as shown in Figure 5.4 is an example. The current passing through such a loop is undefined (although small but nonzero resistances will ensure that this current is zero in any realization of this circuit). We will therefore make the following assumption:

$$\text{any cycle in } G \text{ must pass through an impedance element.} \tag{5.40}$$

If g represents a current source, then $i_g(t)$ is already given to us. If g represents a *diode*, then the sum of voltages around the cycle c_g *minus* g is the *reverse* voltage $-v_g(t)$ on the diode. This is because from Kirchhoff's voltage law the sum of $v_g(t)$ and $\sum_{f\in c_g\backslash\{g\}} \widetilde{b}_{f,g} v_f(t)$ is zero. Because $c_g\backslash\{g\}$ is a path in H, it consists only of R, L, or C

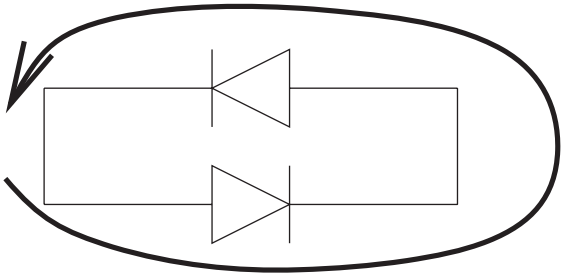


Figure 5.4: Diode loop: the current passing through these ideal diodes is undefined.

elements and voltage sources. So we can compute $\mathcal{L}v_g(s)$ in terms of the currents $\mathcal{L}i_f(s)$, $f \in E(H)$. In particular, wherever g represents a diode we have the equations

$$-\mathcal{L}v_g(s) = \sum_{f \in c_g \setminus \{g\}} \widetilde{b}_{f,g}\, z_f(s)\, \mathcal{L}i_f(s) + \sum_{f \in c_g \setminus \{g\}} \widetilde{b}_{f,g}\, \mathcal{L}v_{f,0}(s),$$

where $v_{f,0}(t)$ is the voltage source on edge $f \in E(H)$. If $g \in E(H)\setminus E(T)$, however, we have the equations

$$0 = \sum_{f \in c_g} b_{f,g}\, z_f(s)\, \mathcal{L}i_f(s) + \sum_{f \in c_g} b_{f,g}\, \mathcal{L}v_{f,0}(s).$$

On the other hand, $i_f(t)$ is determined by the currents in the edges $E(H)\setminus E(T)$ and $E(G)\setminus E(H)$:

$$i_f(t) = \sum_{e \in E(H)\setminus E(T)} b_{f,e}\, i_e(t) + \sum_{k \in E(G)\setminus E(H)} \widetilde{b}_{f,k}\, i_k(t).$$

In matrix-vector terms, if we additionally define

$$\begin{aligned}
\mathbf{v}_0(t) &= \big[v_{e,0}(t) \mid e \in E(H)\big],\\
\mathbf{v}_{ext}(t) &= \big[v_e(t) \mid e \in E(G)\setminus E(H)\big],\\
\mathbf{i}_{ext}(t) &= \big[i_e(t) \mid e \in E(G)\setminus E(H)\big],\\
\widetilde{B} &= \big[\widetilde{b}_{f,e} \mid f \in E(H),\, e \in E(G)\setminus E(H)\big],
\end{aligned}$$

the equations relating Laplace transforms of current and voltage are

$$\begin{bmatrix} 0 \\ \mathcal{L}\mathbf{v}_{ext}(s) \end{bmatrix} = \begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T Z_H(s) \begin{bmatrix} B & \widetilde{B} \end{bmatrix} \begin{bmatrix} \mathcal{L}\mathbf{i}(s) \\ \mathcal{L}\mathbf{i}_{ext}(s) \end{bmatrix} + \begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T \mathcal{L}\mathbf{v}_0(s). \tag{5.41}$$

First note that $Z_H(s)$ is a diagonal matrix with positive diagonal entries for $s > 0$, and so it is symmetric positive definite. It is clear that $\big[B, \widetilde{B}\big]^T Z_H(s) \big[B, \widetilde{B}\big]$ is also symmetric and positive semidefinite. It is positive definite if $\big[B, \widetilde{B}\big]$ has linearly independent columns, or equivalently, that $B\boldsymbol{\alpha} + \widetilde{B}\widetilde{\boldsymbol{\alpha}} = 0$ implies that both $\boldsymbol{\alpha}$ and $\widetilde{\boldsymbol{\alpha}}$ are zero.

If α_e represents the current flowing through cycle c_e for edge $e \in E(H)\backslash E(T)$, and $\widetilde{\alpha}_g$ represents the current flowing through cycle c_g for edge $g \in E(G)\backslash E(H)$, then $\left(B\boldsymbol{\alpha} + \widetilde{B}\widetilde{\boldsymbol{\alpha}}\right)_f$ is the current flowing through edge $f \in E(H)$. Taking $e \in E(H)\backslash E(T)$, we can check that $\left(B\boldsymbol{\alpha} + \widetilde{B}\widetilde{\boldsymbol{\alpha}}\right)_e = \alpha_e$, so $\alpha_e = 0$. These vectors do not have components associated with edges in $E(G)\backslash E(H)$, so we cannot simply take the g component to show that $\widetilde{\alpha}_g = 0$. However, the equation $B\boldsymbol{\alpha} + \widetilde{B}\widetilde{\boldsymbol{\alpha}} = 0$ means that there is no net current in any edge $e \in E(H)$. Therefore, the only net current represented by $\widetilde{\boldsymbol{\alpha}}$ must occur in $E(G)\backslash E(H)$. If this flow is nonzero, there must be a cycle in $E(G)\backslash E(H)$, which is ruled out by (5.40). Thus the current flow in $g \in E(G)\backslash E(H)$ must be $\widetilde{\alpha}_g = 0$, as desired.

Thus $\left[B, \widetilde{B}\right]$ has linearly independent columns, and so $\left[B, \widetilde{B}\right]^T Z_H(s) \left[B, \widetilde{B}\right]$ is positive definite for all $s > 0$.

We can decompose the matrix on the left-hand side of (5.41) in block form:

$$Z(s) = \begin{bmatrix} Z_{11}(s) & Z_{12}(s) \\ Z_{21}(s) & Z_{22}(s) \end{bmatrix} = \begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T Z_H(s) \begin{bmatrix} B & \widetilde{B} \end{bmatrix}.$$

We can then relate $\mathcal{L}\mathbf{v}_{ext}(s)$ to $\mathcal{L}\mathbf{i}_{ext}(s)$ by means of the Schur complement matrix $\widehat{Z}(s) := Z_{22}(s) - Z_{21}(s)\, Z_{11}(s)^{-1}\, Z_{12}(s)$:

$$\mathcal{L}\mathbf{v}_{ext}(s) = \widehat{Z}(s)\, \mathcal{L}\mathbf{i}_{ext}(s) + \left[\widetilde{B}^T - Z_{21}(s)\, Z_{11}(s)^{-1}\, B\right] \mathcal{L}\mathbf{v}_0(s).$$

5.4.3 Bounds on $\widehat{Z}(s)$ and index one

Since the entries of $Z_H(s)$ are rational functions of s, it can be shown that the entries of $\widehat{Z}(s)$ are also rational functions of s. The impulse response represented by $Z_H(s)$ is therefore a linear combination of products of polynomials and (possibly complex) exponentials, and Dirac-δ functions and its derivatives. However, we want to show that $\widehat{Z}(s)$ and $\widehat{Z}(s)^{-1}$ are bounded by $\beta_Z\, s$ for a suitable constant β_Z for large $s > 0$. This is crucial for establishing that these circuit problems are index zero or index one, or a mixture of index one and index zero.

We use the natural ordering on symmetric matrices: $A \preceq B$ if and only if $u^T A u \leq u^T B u$ for all u, and $A \prec B$ if and only if $u^T A u < u^T B u$ for all $u \neq 0$. Note that $0 \preceq A$ if and only if A is positive semidefinite and $0 \prec A$ if and only if A is positive definite. Note also that $A \preceq B$ and $C \preceq D$ imply that $A + C \preceq B + D$; $A \preceq B$ implies that $X^T A X \preceq X^T B X$; $0 \prec A \preceq B$ implies that $0 \prec B^{-1} \preceq A^{-1}$. Also, $\lambda_{min}(A)\, I \preceq A \preceq \lambda_{max}(A)\, I$, where $\lambda_{min}(A)$ and $\lambda_{max}(A)$ are, respectively, the minimum and maximum eigenvalues of A.

For $s \geq 1$, since $Z_H(s)$ is a diagonal matrix with entries of the form either R or $C^{-1}\, s^{-1}$ or $L\, s$, $\zeta_{min}\, s^{-1}\, I \preceq Z_H(s) \preceq \zeta_{max}\, s\, I$ for suitable positive constants ζ_{min} and ζ_{max}. Clearly,

$$\begin{aligned} \begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T Z_H(s) \begin{bmatrix} B & \widetilde{B} \end{bmatrix} &\preceq \begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T \zeta_{max}\, s \begin{bmatrix} B & \widetilde{B} \end{bmatrix} \\ &\preceq \zeta_{max}\, s \begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T \begin{bmatrix} B & \widetilde{B} \end{bmatrix} \\ &\preceq \zeta_{max}\, s\, \sigma_{max}\left(\begin{bmatrix} B & \widetilde{B} \end{bmatrix}\right)^2 I. \end{aligned}$$

On the other hand, using the same techniques,

$$\begin{bmatrix} B & \widetilde{B} \end{bmatrix}^T Z_H(s) \begin{bmatrix} B & \widetilde{B} \end{bmatrix} \succeq \zeta_{min}\, s^{-1}\, \sigma_{min}\left(\begin{bmatrix} B & \widetilde{B} \end{bmatrix}\right)^2 I.$$

Combining these gives positive constants c_{min} and c_{max} where

$$0 \prec c_{min}\, s^{-1}\, I \preceq Z(s) \preceq c_{max}\, s\, I,$$

and so $Z(s)^{-1} \preceq c_{min}^{-1}\, s\, I$. If we write

$$Y(s) = Z(s)^{-1} = \begin{bmatrix} Y_{11}(s) & Y_{12}(s) \\ Y_{21}(s) & Y_{22}(s) \end{bmatrix},$$

then $Y_{22}(s) = \widehat{Z}(s)^{-1}$. Since $Y(s) \preceq c_{min}^{-1}\, s\, I$, it follows that $Y_{22}(s) \preceq c_{min}^{-1}\, s\, I$, and so $\widehat{Z}(s)^{-1} \preceq c_{min}^{-1}\, s\, I$. On the other hand,

$$\begin{aligned} \widehat{Z}(s) &= Z_{22}(s) - Z_{21}(s)\, Z_{11}(s)^{-1}\, Z_{12}(s) \\ &\preceq Z_{22}(s) \preceq c_{max}\, s\, I. \end{aligned}$$

That is,

$$0 \prec c_{min}\, s^{-1}\, I \preceq \widehat{Z}(s) \preceq c_{max}\, s\, I.$$

This means that the diagonal entries $\widehat{z}_{kk}(s)$, which are rational functions of s, satisfy the inequalities

$$0 < c_{min}\, s^{-1} \leq \widehat{z}_{kk}(s) \leq c_{max}\, s.$$

The off-diagonal entries $\widehat{z}_{kl}(s)$ have an upper bound: $|\widehat{z}_{kl}(s)| \leq c_{max}\, s$. Writing $\widehat{z}_{kk}(s)$ as a rational function $p_{kk}(s)/q_{kk}(s)$, this means that the degrees of the polynomials p_{kk} and q_{kk} differ by at most one. For the off-diagonal entries, if $\widehat{z}_{kl}(s) = p_{kl}(s)/q_{kl}(s)$, then the degree of p_{kl} is no more than the degree of q_{kl} plus one. This is the essence of a mixed index-zero and index-one problem. However, if the degree of the numerator is more than the degree of the denominator in any of these entries, the impulse response represented includes derivatives of Dirac-δ functions, which our current theory does not allow.

To simplify the discussions of degrees, define the *relative degree* of a rational function $z(s) = p(s)/q(s)$ to be

$$\operatorname{rdeg} z = \deg p - \deg q. \tag{5.42}$$

The problem, then, is that some entries $\widehat{z}_{kl}(s)$ of $\widehat{Z}(s)$ have $\operatorname{rdeg}\widehat{z}_{kl} > 0$. Since $\widehat{Z}(s)$ is positive definite for $s > 0$, we have $|\widehat{z}_{kl}(s)| \leq \sqrt{\widehat{z}_{kk}(s)\widehat{z}_{ll}(s)}$, so from asymptotics for large s, $\operatorname{rdeg}\widehat{z}_{kl}(s) \leq \frac{1}{2}(\operatorname{rdeg}\widehat{z}_{kk}(s) + \operatorname{rdeg}\widehat{z}_{ll}(s))$. Thus we need only look at diagonal entries to determine whether this is a potential problem. If $\operatorname{rdeg}\widehat{z}_{kk}(s) \leq 0$ for all k, then there is no need to swap currents and voltages.

Assume, for the moment, that for every diode there is a cycle through the diode and H that does not pass through an inductor. Then $Z(s)\,\mathcal{L}\mathbf{i}_{ext}(s)$ represents a convolution $\int_0^\infty M(t-\tau)\mathbf{i}_{ext}(\tau)\,d\tau$ where $\mathcal{L}M(s) = Z(s)$. Note that $M(t) = M_0\,\delta(t) + \widetilde{M}(t)$, where $M_0 = \lim_{s\to\infty} Z(s)$, and $\widetilde{M}(t)$ is a smooth function of t, except at $t = 0$, where there can be a jump discontinuity. Clearly M_0 is a symmetric positive semidefinite matrix. This is the index-zero part of a CCP. Solutions exist and are unique, as can be verified via Theorems 5.6 and 5.7.

If we remove the assumption that every diode is in a cycle that does not pass through an inductor, then $\mathcal{L}^{-1}Z(t)$ may contain derivatives of Dirac-δ functions. To compensate for this, we swap current and voltage variables. In RC diode networks, this is not necessary.

5.4.4 Swapping currents and voltages

The idea here is to swap certain diode currents for the corresponding voltages as the primary variables. Doing this will ensure that the resulting matrix function relating to primary variables is the Laplace transform of a Dirac-δ function plus a well-behaved function of time. The result is a CCP that is a mixed index-zero and index-one problem. However, we will be able to apply the existence and uniqueness theorems of CCPs from Section 4.6 to these diode problems.

We separate the diodes into groups, C diodes, R diodes, and L diodes, as follows. The C diodes are the diodes for which there is a path through H that passes only through capacitors connecting the nodes of the diode; the R diodes are the diodes for which there is no path only through capacitors, but there is a path through H containing no inductors; the L diodes are the remaining diodes where every path through H connecting the diode's nodes passes through an inductor. Let E_{DC}, E_{DR}, and E_{DL} be the set of edges for C diodes, R diodes, and L diodes, respectively. Then, if we separate out the entries of the $\widehat{Z}(s)$ matrix according to this classification, if $e \in E_{DC}$, then $\widehat{z}_{e,e}(s) \sim z^{\infty}_{e,e}/s$ as $s \to \infty$; if $e \in E_{DR}$, then $\widehat{z}_{e,e}(s) \sim z^{\infty}_{e,e}$ as $s \to \infty$; if $e \in E_{DL}$, then $\widehat{z}_{e,e}(s) \sim z^{\infty}_{e,e}\, s$ as $s \to \infty$. We can partition the matrix $\widehat{Z}(s)$ according to this partitioning of $E(G)\backslash E(H) = E_{DC} \cup E_{DR} \cup E_{DL}$:

$$\widehat{Z}(s) = \begin{bmatrix} \widehat{Z}_C(s) & \widehat{Z}_{C,R}(s) & \widehat{Z}_{C,L}(s) \\ \widehat{Z}_{R,C}(s) & \widehat{Z}_R(s) & \widehat{Z}_{R,L}(s) \\ \widehat{Z}_{L,C}(s) & \widehat{Z}_{R,L}(s) & \widehat{Z}_L(s) \end{bmatrix}.$$

Partitioning the diode voltage and current vectors accordingly gives

$$\begin{bmatrix} \mathcal{L}\mathbf{v}_C(s) \\ \mathcal{L}\mathbf{v}_R(s) \\ \mathcal{L}\mathbf{v}_L(s) \end{bmatrix} = \begin{bmatrix} \widehat{Z}_C(s) & \widehat{Z}_{C,R}(s) & \widehat{Z}_{C,L}(s) \\ \widehat{Z}_{R,C}(s) & \widehat{Z}_R(s) & \widehat{Z}_{R,L}(s) \\ \widehat{Z}_{L,C}(s) & \widehat{Z}_{R,L}(s) & \widehat{Z}_L(s) \end{bmatrix} \begin{bmatrix} \mathcal{L}\mathbf{i}_C(s) \\ \mathcal{L}\mathbf{i}_R(s) \\ \mathcal{L}\mathbf{i}_L(s) \end{bmatrix}.$$

From symmetry and positive definiteness of $\widehat{Z}(s)$ it is easy to check that

$$\begin{aligned} \widehat{Z}_{R,C}(s) &= \widehat{Z}_{C,R}(s)^T = \mathcal{O}(s^{-1}), \\ \widehat{Z}_{R,L}(s) &= \widehat{Z}_{L,R}(s)^T = \mathcal{O}(1), \\ \widehat{Z}_{C,L}(s) &= \widehat{Z}_{L,C}(s)^T = \mathcal{O}(1) \end{aligned}$$

as $s \to \infty$. However, the last bound is not sharp. In fact, it can be shown that

$$\widehat{Z}_{C,L}(s) = \widehat{Z}_{L,C}(s)^T = \mathcal{O}(s^{-1}).$$

Consider a jump discontinuity in the current of a C diode. Then this jump in current will flow through the path in H passing only through capacitors connecting the nodes of the diode. Such a jump will not create a jump discontinuity in the voltages at any nodes in the circuit; rather there can be jumps only in the derivatives of the voltages of the nodes. Thus there is no jump in the voltages of any L diode, and hence $\widehat{Z}_{L,C}(s) = \mathcal{O}(s^{-1})$ as $s \to \infty$.

We will now swap the roles of $\mathbf{v}_L$ and $\mathbf{i}_L$: note that (after suppressing dependence on s)

$$\mathcal{L}\mathbf{v}_L = \widehat{Z}_{L,C}\mathcal{L}\mathbf{i}_C + \widehat{Z}_{L,R}\mathcal{L}\mathbf{i}_R + \widehat{Z}_L\mathcal{L}\mathbf{i}_L.$$

Then

$$\mathcal{L}\mathbf{i}_L = \widehat{Z}_L^{-1}\left(\mathcal{L}\mathbf{v}_L - \begin{bmatrix} \widehat{Z}_{L,C} & \widehat{Z}_{L,R} \end{bmatrix}\begin{bmatrix} \mathcal{L}\mathbf{i}_C \\ \mathcal{L}\mathbf{i}_R \end{bmatrix}\right).$$

Now we wish to write $\mathbf{v}_C$ and $\mathbf{v}_R$ in terms of $\mathbf{i}_C$, $\mathbf{i}_R$, and $\mathbf{v}_L$:

$$\begin{aligned}
\begin{bmatrix} \mathcal{L}\mathbf{v}_C \\ \mathcal{L}\mathbf{v}_R \end{bmatrix} &= \begin{bmatrix} \widehat{Z}_C & \widehat{Z}_{C,R} \\ \widehat{Z}_{R,C} & \widehat{Z}_R \end{bmatrix}\begin{bmatrix} \mathcal{L}\mathbf{i}_C \\ \mathcal{L}\mathbf{i}_R \end{bmatrix} + \begin{bmatrix} \widehat{Z}_{C,L} \\ \widehat{Z}_{R,L} \end{bmatrix}\mathcal{L}\mathbf{i}_L \\
&= \left(\begin{bmatrix} \widehat{Z}_C & \widehat{Z}_{C,R} \\ \widehat{Z}_{R,C} & \widehat{Z}_R \end{bmatrix} - \begin{bmatrix} \widehat{Z}_{C,L} \\ \widehat{Z}_{R,L} \end{bmatrix}\widehat{Z}_L^{-1}\begin{bmatrix} \widehat{Z}_{L,C} & \widehat{Z}_{L,R} \end{bmatrix}\right)\begin{bmatrix} \mathcal{L}\mathbf{i}_C \\ \mathcal{L}\mathbf{i}_R \end{bmatrix} \\
&\qquad + \begin{bmatrix} \widehat{Z}_{C,L} \\ \widehat{Z}_{R,L} \end{bmatrix}\widehat{Z}_L^{-1}\mathcal{L}\mathbf{v}_L \\
&= \widehat{Z}_{Schur}\begin{bmatrix} \mathcal{L}\mathbf{i}_C \\ \mathcal{L}\mathbf{i}_R \end{bmatrix} + \begin{bmatrix} \widehat{Z}_{C,L} \\ \widehat{Z}_{R,L} \end{bmatrix}\widehat{Z}_L^{-1}\mathcal{L}\mathbf{v}_L.
\end{aligned}$$

Note that $\widehat{Z}_{Schur}$ is the Schur complement of $\widehat{Z}$ with respect to $\widehat{Z}_L$. Combining these, we get

$$\begin{bmatrix} \mathcal{L}\mathbf{v}_C \\ \mathcal{L}\mathbf{v}_R \\ \mathcal{L}\mathbf{i}_L \end{bmatrix} = \begin{bmatrix} & \widehat{Z}_{Schur} & & \widehat{Z}_{C,L}\widehat{Z}_L^{-1} \\ & & & \widehat{Z}_{R,L}\widehat{Z}_L^{-1} \\ -\widehat{Z}_L^{-1}\widehat{Z}_{L,C} & & -\widehat{Z}_L^{-1}\widehat{Z}_{R,C} & \widehat{Z}_L^{-1} \end{bmatrix}\begin{bmatrix} \mathcal{L}\mathbf{i}_C \\ \mathcal{L}\mathbf{i}_R \\ \mathcal{L}\mathbf{v}_L \end{bmatrix}. \tag{5.43}$$

Since $\widehat{Z}_L(s)^{-1} = \mathcal{O}(s^{-1})$, we have $\widehat{Z}_L(s)^{-1}\widehat{Z}_{L,C}(s) = \mathcal{O}(s^{-1})$ and $\widehat{Z}_L(s)^{-1}\widehat{Z}_{R,C}(s) = \mathcal{O}(s^{-1})$. Also

$$\widehat{Z}_{Schur}(s) \preceq \begin{bmatrix} \widehat{Z}_C(s) & \widehat{Z}_{C,R}(s) \\ \widehat{Z}_{R,C}(s) & \widehat{Z}_R(s) \end{bmatrix} = \mathcal{O}(1).$$

In fact,

$$\lim_{s\to\infty}\widehat{Z}_{Schur}(s) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \widehat{Z}_R^{\infty} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \widehat{Z}_R^{\infty} = \lim_{s\to\infty}\widehat{Z}_R(s).$$

Note that $\widehat{Z}_R^{\infty}$ is the impedance matrix for the R diodes in the circuit where all the capacitors are replaced by short circuits and the inductors are replaced by open circuits. Note that this matrix is symmetric positive definite.

We can turn our dynamic circuit problem into a CCP. The key is that (5.43) is equivalent to

$$\begin{bmatrix} \mathbf{v}_C \\ \mathbf{v}_R \\ \mathbf{i}_L \end{bmatrix} = M * \begin{bmatrix} \mathbf{i}_C \\ \mathbf{i}_R \\ \mathbf{v}_L \end{bmatrix} + \mathbf{q}(t),$$

where $\mathcal{L}M(s) = \widetilde{Z}(s)$. Now $M(t) = M_0\,\delta(t) + m(t)$, where δ is the Dirac-δ function, and $m(t)$ is a smooth function of t. Note that

$$M_0 = \lim_{s\to\infty} \widetilde{Z}(s),$$
$$m(0^+) = \lim_{s\to\infty} s\left(\widetilde{Z}(s) - M_0\right).$$

From our representation of $\widetilde{Z}(s)$, we already have

$$M_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \widehat{Z}_R^\infty & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

To compute $m(0^+)$, we have to check the asymptotics of $\widetilde{Z}(s)$ to order s^{-1}. Let us use the superscript "∞" to indicate limits or asymptotics, such as $\widehat{Z}_C(s) \sim s^{-1}\,\widehat{Z}_C^\infty$, $\widehat{Z}_L(s) \sim s\,\widehat{Z}_L^\infty$, $\widehat{Z}_{R,L}(s) \sim \widehat{Z}_{R,L}^\infty$ as $s \to \infty$. Then

$$m(0^+) = \lim_{s\to\infty} s\left(\widetilde{Z}(s) - M_0\right) = \begin{bmatrix} \widehat{Z}_C^\infty & \widehat{Z}_{C,R}^\infty & 0 \\ \widehat{Z}_{R,C}^\infty & * & +\widehat{Z}_{R,L}^\infty\left(\widehat{Z}_L^\infty\right)^{-1} \\ 0 & -\left(\widehat{Z}_L^\infty\right)^{-1}\widehat{Z}_{R,C}^\infty & \left(\widehat{Z}_L^\infty\right)^{-1} \end{bmatrix}.$$

The conditions for existence of a solution to the CCP as given in Theorem 5.6 are that M_0 is symmetric positive definite, $m(t)$ has bounded variation on finite intervals (which is true), $\mathbf{q}(\cdot)$ belongs to a suitable L^p space, and $m(0^+)$ is, for example, positive definite on range$(M_0)^\perp$. For uniqueness, we also need $m(0^+)$ to be symmetric positive definite on range$(M_0)^\perp$. This is equivalent to requiring that (after dropping the middle block row and column)

$$\begin{bmatrix} \widehat{Z}_C^\infty & 0 \\ 0 & \left(\widehat{Z}_L^\infty\right)^{-1} \end{bmatrix}$$

be positive definite, which is evidently true. The last requirement is that the antisymmetric part of $m(0^+)$ map range(M_0) into range$(M_0)^\perp$ and map range$(M_0)^\perp$ into range(M_0). For the circuit problem, the antisymmetric part of $m(0^+)$ is

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & +\widehat{Z}_{R,L}^\infty\left(\widehat{Z}_L^\infty\right)^{-1} \\ 0 & -\left(\widehat{Z}_L^\infty\right)^{-1}\widehat{Z}_{R,C}^\infty & 0 \end{bmatrix}.$$

Since range(M_0) consists of vectors of the block form

$$\begin{bmatrix} 0 \\ \mathbf{u}_R \\ 0 \end{bmatrix},$$

it can be easily checked that $m(0^+)$ satisfies the conditions for uniqueness as well as existence. Thus we have existence and uniqueness for solutions of the problem of RLC circuits with diodes, provided the graph G with all diodes and current sources removed is connected.

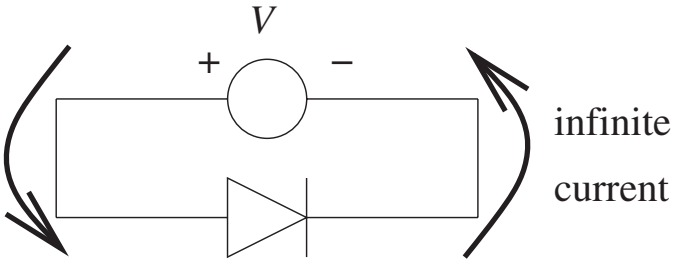


Figure 5.5: Circuit without solution for an ideal diode.

5.4.5 Comparisons with other approaches

This may appear to be a lot of work to show the existence and uniqueness of solutions. However, it is important to remember that this is for ideal diodes. There are other well-known nonlinear models of diodes for which the current passed is a smooth monotone function of the voltage. Why not just use the well-known existence results for differential equations with smooth right-hand sides?

The scale on which large changes in these continuous models is measured is for many applications very small. At room temperature, the forward current increases by a factor of $e \approx 2.718$ with a difference in the voltage of $\approx 26\,\mathrm{mV}$. For small signal analysis, for which voltage variations of μV (10^{-6}V) to mV (10^{-3}V) are typical, the continuous model is appropriate. But for power system applications with voltages ranging from tens to thousands of volts, this is a very small range. In order to use the continuous model as a continuous model, it becomes important to restrict the time steps in the transition region to keep the voltage change during a time step much smaller than $\approx 26\,\mathrm{mV}$. Thus, there is a performance drop for using the continuous model.

On the other hand, the ideal diode model is a limit of singular perturbations, of which the continuous diode model is one. The ideal diode model should give solutions close to the continuous diode model. But there are degenerate situations for which the diode model clearly has no solutions, such as forward biasing a diode with a voltage source (see Figure 5.5). Such situations typically indicate some bad behavior of the continuous model. This might appear in the form of extreme values or extreme sensitivity in solutions of the continuous model. Sensitivity to small perturbations typically appears in ideal diode models as nonuniqueness of solutions. Being able to prove existence and uniqueness of solutions for these (nondegenerate) situations thus tells us some important things about the behavior of solutions of the continuous model, and hopefully practical information about the behavior of real circuits.

Uniqueness is sometimes an undesirable property. For example, if we consider *static* electrical circuits, the techniques of this section that resistor-diode networks with given voltage and current sources indicate that the solution is unique. This means that there are no bistable resistor-diode networks. Thus no memory circuits can be built out of diodes and (linear) resistors. However, transistors can be used to build flip-flops as a basic kind of memory circuit. This requires the property of nonuniqueness, at least for the static case.

Another approach is the use of LCSs (see Section 4.5). These systems offer an almost entirely algebraic approach to understanding these systems. Unfortunately, LCSs restrict

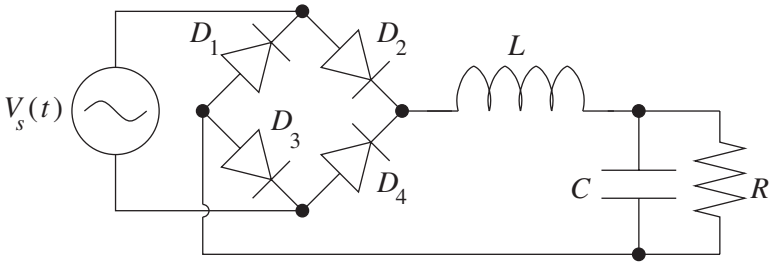


Figure 5.6: Bridge rectifier circuit.

our attention to Bohl distributions, which are linear combinations of Dirac-δ functions and their derivatives, and products of polynomials and (real or complex) exponentials. If we consider voltage or current sources that are not of this form, LCSs cannot help us. Considerable success has been found with the LCS approach for *passive systems* [48]. A passive system is a differential equation with an input and an output,

$$\begin{aligned} \frac{dx}{dt}(t) &= f(x(t), z(t)), \qquad x(t_0) = x_0, \\ w(t) &= h(x(t), z(t)), \end{aligned}$$

together with an "energy storage" function $V(x)$ where for $t_2 \geq t_1$,

$$V(x(t_2)) - V(x(t_1)) \leq \int_{t_1}^{t_2} \langle w(t), z(t) \rangle \, dt.$$

A linear passive system can have a quadratic energy storage function; most interest is when $V(x) = x^T K x$, where K is positive definite. Even so, active devices with external voltage and current sources, such as bipolar junction transistors (BJTs) or field-effect transistors (FETs), need a different approach.

5.4.6 What if H is not a connected subgraph of G?

In the previous sections we assumed that H, the subgraph of the circuit with the resistors, capacitors, and inductors, is a connected subgraph of the entire circuit G. While this assumption helps create the system describing the circuit, it is often not true. An example we have already seen is the bridge rectifier circuit shown in Figure 5.6.

If H is not a connected subgraph of G, then the diode currents are not independent. For example, in the bridge rectifier circuit, if $i_{D,k}$ is the current in the forward direction in diode D_k, then $i_{D,1} + i_{D,3} = i_{D,2} + i_{D,4}$. Let $v_{D,k}$ denote the reverse voltage difference for diode D_k. While

$$0 \leq i_{D,k} \perp v_{D,k} \geq 0 \qquad \text{for each } k,$$

we need to reduce the number of primary current variables. To do this in a systematic way, consider the graph obtained by collapsing the connected components of H down to single nodes. The net current into these nodes must sum to zero, but voltages are not well defined

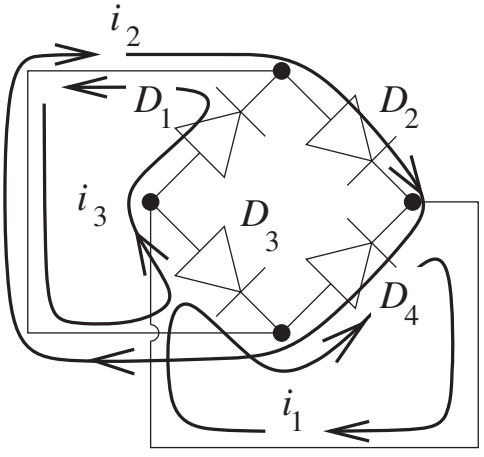


Figure 5.7: Quotient graph G/H for bridge rectifier.

on these collapsed nodes. Call this collapsed or quotient network G/H. We assume that every edge of this quotient network is a diode. We assume that G, the entire circuit, is a connected network, and so G/H is also a connected network. We can, as we did before, find an MST T' for G/H. Every edge e in G/H not in T' defines a unique cycle c'_e in G/H.

Note that if H were connected, then G/H would consist of a single node with a loop for each diode. For the case of the bridge rectifier, H has two connected components: one containing the voltage source, and the other with the R, L, and C components. This circuit has a slight violation of our usual rules in that the current from the voltage source does not pass through a resistor before leaving H; the possibility therefore exists that the voltage source could be short circuited. However, the orientation of the diodes prevents this.

The quotient graph G/H for the bridge rectifier is shown in Figure 5.7. Also shown are the three loops obtained from an MST T' of G/H. Let B' be the matrix $b'_{f,e}$ for $e \in E(G/H)\backslash E(T')$ and $f \in E(G/H)$ where

$$b'_{f,e} = \begin{cases} +1 & \text{if } f \in c'_e \text{ in the forward direction,} \\ -1 & \text{if } f \in c'_e \text{ in the reverse direction,} \\ 0 & \text{if } f \notin c'_e. \end{cases}$$

We use the currents i_e, $e \in E(G/H)\backslash E(T')$ as our primary variables; the current in edge $f \in E(G/H)$ is then

$$i_f = \sum_{e \in E(G/H)\backslash E(T')} b'_{f,e} i_e.$$

If $e \notin E(T')$, this becomes just $i_e = i_e$.

However, each edge $f \in E(G/H)$ represents a diode, and so its current must be non-negative: $i_f \geq 0$. If $\mathbf{i}^* = \left[i_e \mid e \in E(G/H)\backslash E(T')\right]$ is the vector of these primary variables, we have the constraint that $\mathbf{i}_D = B'\mathbf{i}^* \geq 0$, where $\mathbf{i}_D$ is the vector of all diode currents, to

ensure nonnegative current in each diode. In the bridge rectifier example,

$$\begin{aligned} i_2, i_3 &\geq 0, \\ i_1 - i_2 &\geq 0, \\ i_1 - i_3 &\geq 0. \end{aligned}$$

The set of admissible values of $\mathbf{i}^*$ is then

$$K = \left\{ \mathbf{i}^* \mid B'\mathbf{i}^* \geq 0 \right\}.$$

Thus we do not have a standard linear complementarity relationship between $\mathbf{i}^*$ and the associated voltages $\mathbf{v}^*$. However, we do have a generalized complementarity relationship between them, as we will see.

First, K is clearly a closed convex cone. In fact, it is also a polyhedral cone, but this is not the most important thing. For a GCP we need $\mathbf{v}^* \in K^*$, the dual cone to K. By Lemma 2.10,

$$K^* = \left(B'\right)^T \mathbb{R}^n_+.$$

Now the sum of the diode voltage differences around the loop for edge $e \in E(G/H) \backslash E(T')$ is given by

$$\mathbf{v}^* = \left(B'\right)^T \mathbf{v}_D,$$

where $\mathbf{v}_D$ is the vector of reverse diode voltages. Now we must have $\mathbf{v}_D \in \mathbb{R}^n_+$; thus any $\mathbf{v}^* \in K^*$ can be represented by nonnegative diode voltages. Finally,

$$\begin{aligned} \left(\mathbf{i}^*\right)^T \mathbf{v}^* &= \left(\mathbf{i}^*\right)^T \left(B'\right)^T \mathbf{v}_D \\ &= \left(B'\mathbf{i}^*\right)^T \mathbf{v}_D \; = \; \mathbf{i}_D^T \mathbf{v}_D, \end{aligned}$$

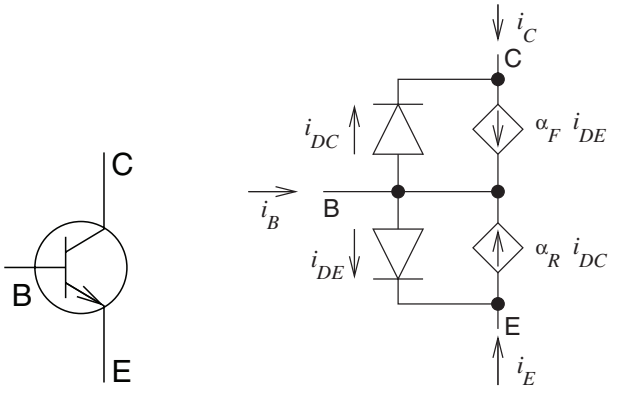
so we have generalized complementarity between $\mathbf{i}^*$ and $\mathbf{v}^*$:

$$K \ni \mathbf{i}^* \perp \mathbf{v}^* \in K^*.$$

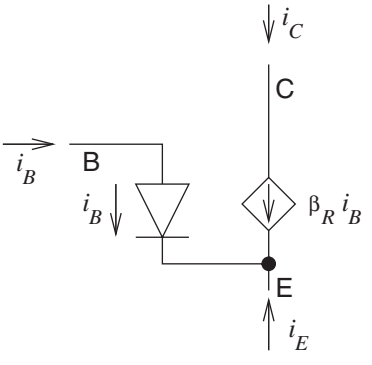
If we write $\mathcal{L}\mathbf{v}^*(s) = Z(s)\left[\mathcal{L}\mathbf{i}^*(s) + \mathcal{L}\mathbf{i}_{ext}(s)\right] + \mathcal{L}\mathbf{v}_{ext}(s)$, we can represent the circuit problem again as a CCP; the kernel function of the CCP is the distribution $M(t)$ where $\mathcal{L}M(s) = Z(s)$. As before, the entries of $Z(s)$ are rational functions and the diagonal entries have relative degree no more than one and no less than minus one. If the relative degrees of the entries of $Z(s)$ are less than or equal to zero, then we can proceed as above and show existence and uniqueness of solutions. If some entries have relative degree one, then we need to find a way to "flip" currents and voltages, as done in Section 5.4.4, to obtain a CCP for which we can show that solutions exist. In the case of RC circuits with diodes, we do not need to perform this "flip," and we immediately have existence and uniqueness of solutions.

5.4.7 Active elements and nonlinear circuits

Active elements such as transistors provide some additional challenges to analysis, as they include new voltage or current sources that depend on currents or voltages elsewhere in the circuit, providing opportunities for feedback that can strongly affect the analysis. As with fixed voltage and current sources, certain configurations are impermissible (such as



(a) NPN transistor (b) Full model



(c) Simplified model

Figure 5.8: Ebers–Moll BJT models.

two current sources in series or two voltage sources in parallel). Current sources in series with a diode, but in the reverse direction, are also impermissible. So we cannot prove results stating that solutions exist for any circuit. Also, the model used for the active device can have an effect on the existence of solutions. In this section we will consider *bipolar junction transistors* as the active devices; these are closely related to diodes, and the models we use are piecewise linear. Figure 5.8 shows Ebers–Moll models for BJTs.

A problem with the simplified Ebers–Moll model (see Figure 5.8(c)), even if we use a threshold model for the diode, is that it does not handle the "saturated" condition of the transistor. This occurs when, in order to make $i_C = \beta i_B$, we need to make the voltage across the current source negative; this means that the current source (inside the transistor) must be an energy source, which is impossible. Rather, the voltage from the emitter (E) to the collector (C) cannot go below the threshold voltage V_T for the diode. In the full Ebers–Moll model (see Figure 5.8(b)), the voltage across the current source between the collector (C) and the base (B) with current $\alpha_F i_{DE}$ cannot go below $-V_T$, and the voltage between the collector and the emitter (E) cannot go below zero.

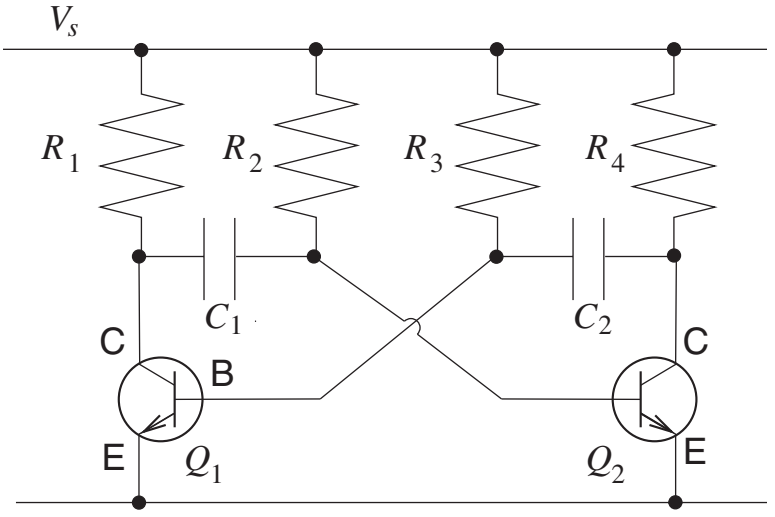


Figure 5.9: Astable flip-flop.

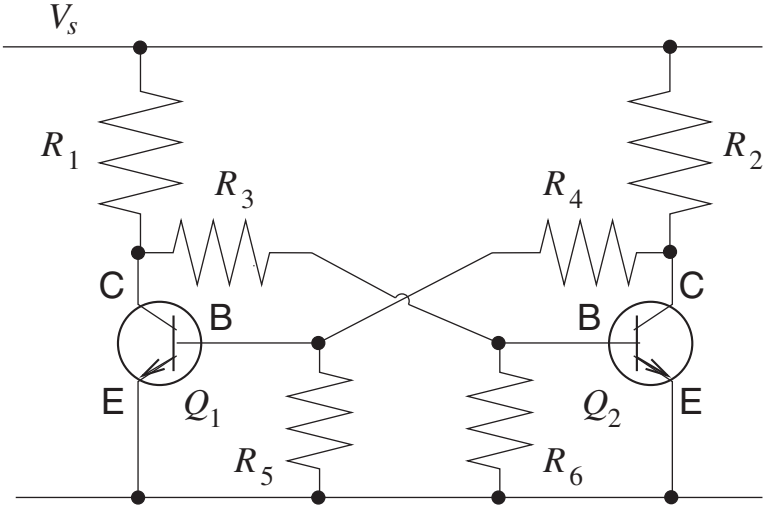


Figure 5.10: Bistable static flip-flop.

For example, to analyze an astable flip-flop as shown in Figure 5.9, we need the full Ebers–Moll model, as we have to deal with saturation of the transistors; that is, when one of the transistors Q_1 or Q_2 is turned "off," the other is turned fully "on."

A crucially important point about the full Ebers–Moll model is that the factors α_R and α_F are both less than one.

An important difference between the case of diodes and transistors is that transistor circuits can be multistable; that is, there can be more than one static solution for a circuit with transistors. An example is the bistable flip-flop as shown in Figure 5.10.

If transistor Q_1 is "on," then there is effectively no resistance between its collector (C) and its emitter (E), and so the voltage at C of Q_1 is essentially zero, so that there is no current flowing through R_3 or R_6. This means that the voltage at the base (B) of Q_2 is below the threshold voltage, and so no current passes through the base of Q_2. Hence no current flows from the collector to the emitter of Q_2, and the voltage at B of Q_1 is $V_s R_5/(R_2 + R_4 + R_5)$. With the current gain of the transistor β sufficiently large and $V_s \gg V_T$, transistor Q_1 is turned "on," completing the feedback loop. Conversely, provided the circuit is symmetric ($R_1 = R_2$, $R_3 = R_4$, and $R_5 = R_6$), the converse situation with Q_2 turned "on" and Q_1 turned "off" is equally possible. Note that the circuit can be "flipped" by supplying an external current to the base of the transistor turned "off." The time needed to do this can be arbitrarily short, since there are no memory elements (capacitors or inductors) in the circuit. In theory, the circuit could "flip" spontaneously. The mathematical formulation does not prevent this. In practice it does not flip, but this can be explained by saying that each transistor has a small but significant pool of electrons or holes at the base of the transistor, and that this pool must be drained of electrons (or holes) before it can switch. Thus the circuit shows hysteretic behavior because of a memory element (effectively a small capacitance) that exists in physical transistors but not in the Ebers–Moll models.

The existence of bistable flip-flop circuits complicates the dynamic analysis in that we must either create a formulation which is explicitly hysteretic or include small memory elements. Including such elements can make the mathematical formulation well defined, but this may be at the cost of requiring numerical methods to use excessively small step sizes.

At the time of this writing, the restrictions that must be placed on idealized BJT circuits in order to obtain existence and uniqueness of solutions is an open question. For example, including a capacitor between the base and emitter of each transistor may be sufficient to obtain existence and uniqueness, as well as having a physical justification.

5.5 Application: Economic networks

The applications we consider in this section are about networks where at each node or vertex of the network there is a constraint, typically related to a resource limit, that makes the system nonsmooth. Dynamic versions of these networks, which are our main concern here, have variables associated with each node and differential equations for these variables. Each node influences neighboring nodes.

These networks are considered "economic" networks in a fairly broad sense. Most of these networks involve decisions being made at the nodes by independent agents, often by performing some kind of short-time optimization. A simple example of this is traffic networks. The static situation is often modeled by the *Wardrop equilibrium*: on a road network, everyone is trying to get to their destinations as fast as possible. However, the time needed to traverse a particular section of road depends on how congested it is. That is, more cars on the road segment makes the traffic go slower. So the decision that each driver needs to make depends on the choices of all the other drivers.

A less "economic" model is that of queuing networks. A queue is a sequence of items which must be "processed" in some way. These items may be people with purchases to make at a supermarket, or employees waiting to find openings in order to advance their careers, or perhaps parts in a factory that must be sorted, assigned, assembled, and finally installed in a finished product. In all cases, there are a number of queues in which items

await some kind of processing. In some cases, the "items" can make their own decisions about which queues to join; in others, the decisions are predetermined. Each queue has its own differential equation relating the number of items in the queue to the rate of processing (or serving) the items in the queue. If the items can choose which queue to join, for example, using a simple rule such as "Join the shortest queue!", then there is additional nonsmoothness in the complete system.

Other sources of economic network problems include market networks where goods are traded between people in neighboring towns or networks where computers "bid" to take part of the load of a large parallel computation. Related problems include differential games, market games (which can describe a small number of competitors trying to maximize profits through price and market share), and Cournot equilibria. There are a large number of these kinds of problems, and they are relevant to a large number of situations involving separate individuals or agents, each trying to maximize their personal reward. (This reward should not necessarily be regarded as material gain; one can imagine humanitarian organizations using different strategies to try to reach and help more, and more isolated, people than other organizations. The point is that there is some kind of competition.)

Whatever kind of situation needs to be described, there are a number of important modeling decisions to be made about dynamic economic models. Care should always be taken with behavioral models. Trying to model intrinsically complex things (such as human beings) by simple and simplistic models should always be taken with a grain of salt. Economists have been surprised by the importance of emotion in traders and businessmen (and women), who are presumably just trying to make money. Also, historical data about behavior often reflects conditions holding at the time the data was collected. Change the conditions, and the behavior changes too, often confounding predictions based on historic data.[7]

Nevertheless, there are often simple rules that seem to be well followed in many situations. "Join the shortest queue!" is perhaps one of them. Taking the quickest route in a road network is probably another. Different people will have slightly different objectives. Some people have more time, and may take a slightly longer route in order to avoid road congestion, or may prefer a more scenic route. If the great majority of people have the same objective (shortest time home, less time in queues), this may not make much difference to the model, and finding appropriate parameters to model this behavior may be considerably more effort than it is worth in terms of modeling the overall system. There is another general behavioral strategy which may be appropriate that can be summarized as "Don't change unless you have to!"[8] Change requires effort, even if it is only mental effort. If a person's objective is to minimize this effort, then they will do what they can to avoid change. This can mean that if the environment (such as road congestion) changes slowly, then decisions do not change until drivers realize that they *have* to change their decisions. This can result in a huge number of drivers suddenly taking new routes, long after they become available. Or perhaps, they recall old routes, long neglected, when they realize that the highway they are accustomed to using takes an hour longer than it used to.

Another issue that is especially important in decision-making situations is the information available and when decisions are made. Consider, for example, the differences between the static traffic equilibrium problem and various versions of the dynamic traffic flow

[7]This is known as the *Lucas critique* regarding economic behavior, especially in regard to macroeconomic policy.

[8]The subject of *viability theory* is essentially devoted to this idea.

problem. In the static problem, the drivers are assumed to have developed a sense for how congested various roads are through trial and error, television reports, and news from their friends. This historical data that each driver has painstakingly built up over the years is necessarily crude and limited, but is reasonably good for modeling the daily commute between home and work. Information and computer technology can change this situation radically. Each day, each driver can plan a separate route to and from work. Further, this can change *during the drive*. Congestion information can be transmitted to vehicles, which can then process the data along with a computerized road network, to arrive at a currently optimal route. Decisions can be taken only *at intersections*, which are nodes in the road network. If the decision of which route to take is made before the journey starts, then there may well be the uncomfortable situation of thousands of drivers converging on a previously empty road since it had no congestion *at the time the decisions were made*. For a workable system, then, we would expect decisions to be made fairly often, based on global and local data.

5.5.1 Traffic networks

As noted in the previous section, the starting point for traffic flow models is usually the *Wardrop equilibrium*. This is a continuum model of traffic flow (fractional cars are allowed), which is also static (flows are assumed to exist for all time), but allows for congestion (time needed to traverse a road segment is a function of the flow rate on that segment). Since, in equilibrium, cars cannot accumulate either on road segments or at intersections, conservation of flow must hold. That is, the flow rate is the same at each point in a road segment, and the net inflow to an intersection or node is zero. It is also assumed that the decisions of a single driver do not significantly change the congestion on each segment.

Wardrop's formulation [267] of the principle for which he is best known was "*The journey time on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route*." To turn this into a more mathematical statement, we need to set up a framework for traffic networks. These we will represent as *directed graphs* or *networks*. That is, the traffic network will be represented by a graph $G = (V, E)$ where V is the set of vertices or nodes and E is the set of edges. Nodes represent intersections while edges represent road links. The edges are directed, in that for each edge $e \in E$ there are nodes start(e) and end(e) representing the starting and ending nodes of e. Two-way roads are regarded as a pair of edges e, g where start(e) = end(g) and start(g) = end(e). The traffic flow along edge e is denoted by f_e for the static problem and $f_e(t)$ for the dynamic problem.

We use directed networks even for a road network consisting only of two-way roads. The alternative of using the net flow through a pair of edges e and g with start(e) = end(g) and start(g) = end(e) given by $f_e - f_g$ does not properly represent the congestion that can occur. For example, the net flow can be zero with no cars traveling in either direction or if there is an equal and large traffic flow in each direction (which would be highly congested).

At each node $x \in V$, there must be conservation of flow, except that certain nodes will act as sources and others will act as sinks for the flow. Thus, the net flow into an edge plus the source for the node must be zero. If we introduce the node-edge incidence matrix

$$w_{x,e} = \begin{cases} +1, & x = \text{end}(e), \\ -1, & x = \text{start}(e), \\ 0 & \text{otherwise}, \end{cases} \tag{5.44}$$

then the net flow into a node x is given by

$$0 = h_x + \sum_{e \in E} w_{x,e} f_e, \tag{5.45}$$

where h_x is the net source of flow for node x. We might, for example, consider a node representing a suburb to be a source node for the morning commute ($h_x > 0$), while it would be sink for the evening commute ($h_x < 0$) when workers are returning home. The sum of the net source flows over all nodes must be zero in order to have conservation of cars:

$$0 = \sum_{x \in V} h_x.$$

Often, traffic models have a single source and a single sink. If there are multiple sinks (that is, destinations), then we need to be careful to identify the flows associated with each destination: $f_{e,d}$ is the flow on edge e with destination d. Let $D \subset V$ be the set of destination nodes. Then

$$f_e = \sum_{d \in D} f_{e,d}. \tag{5.46}$$

Furthermore, we need to ensure that the source flows are labeled with the destination node: $h_{x,d}$ is the amount of flow originating at node x with destination d. Conservation of flow must hold for each destination node:

$$0 = h_{x,d} + \sum_{e \in E} w_{x,e} f_{e,d} \qquad \text{for all } d \in D. \tag{5.47}$$

The time needed to travel edge e depends on the total flow along edge e: $\widehat{\tau}_e = \varphi_e(f_e)$. Here $\varphi_e \colon \mathbb{R} \to \mathbb{R}$ is a nondecreasing function with $\varphi_e(0) > 0$.

Let $\tau_{x,d}$ be the optimal travel time from node x to destination node d. Clearly, $\tau_{d,d} = 0$. Suppose that e is an edge with $x = \text{start}(e)$ and $y = \text{end}(e)$. Then, since $\tau_{x,d}$ is the optimal travel time from x to d,

$$\tau_{x,d} \le \tau_{y,d} + \varphi_e(f_e). \tag{5.48}$$

But drivers with destination d would use this edge e only if we have equality. If there is a strict inequality, then edge e will give a travel time to d greater than the optimal, and therefore drivers going to d would choose another route. Thus we obtain the complementarity conditions

$$0 \le f_{e,d} \perp \tau_{x,d} - \tau_{y,d} + \varphi_e(f_e) \ge 0 \tag{5.49}$$

for all nodes $x, y \in V$, destinations $d \in D$, and edges e where $x = \text{start}(e)$ and $y = \text{end}(e)$. This can be rewritten in terms of the node-edge incidence matrix as

$$0 \le f_{e,d} \perp \sum_{x \in V} w_{x,e} \tau_{x,d} + \varphi_e(f_e) \ge 0 \qquad \text{for all } e \in E,\ d \in D. \tag{5.50}$$

If, for simplicity, we consider a single destination ($D = \{d^*\}$), we can drop the subscript d and put the system into matrix-vector form using the following vectors:

$$\begin{aligned}
\mathbf{f} &= \left[f_{e,d^*} \mid e \in E \right], \\
\boldsymbol{\tau} &= \left[\tau_{x,d^*} \mid x \in V \right], \\
\mathbf{h} &= \left[h_{x,d^*} \mid x \in V \right], \\
\boldsymbol{\varphi}(\mathbf{f}) &= \left[\varphi_e(f_{e,d^*}) \mid e \in E \right], \\
W &= \left[w_{x,e} \mid x \in V, e \in E \right].
\end{aligned}$$

Then

$$0 = \mathbf{h} + W\mathbf{f}, \tag{5.51}$$

$$0 \le \mathbf{f} \quad \perp \quad W^T\boldsymbol{\tau} + \boldsymbol{\varphi}(\mathbf{f}) \ge 0. \tag{5.52}$$

This can be rewritten in terms of VIs as

$$0 = \mathbf{h} + W\mathbf{f}, \tag{5.53}$$

$$\mathbf{f} \ge 0 \;\&\; 0 \le \left\langle \widetilde{\mathbf{f}} - \mathbf{f},\, W^T\boldsymbol{\tau} + \boldsymbol{\varphi}(\mathbf{f}) \right\rangle \qquad \text{for all } \widetilde{\mathbf{f}} \ge 0. \tag{5.54}$$

This is an example of a VI with Lagrange multipliers as described in Section 2.4.4. The vector of Lagrange multipliers is $\boldsymbol{\tau}$, which is associated with the constraint $0 = \mathbf{h} + W\mathbf{f}$. Following the approach of Section 2.4.4, let $K = \{\mathbf{g} \mid \mathbf{h} + W\mathbf{g} = 0, \mathbf{g} \ge 0\}$. Clearly K is closed and convex, and $\mathbf{f} \in K$. We can write $K = L \cap M$, where $M = \{\mathbf{g} \mid \mathbf{h} + W\mathbf{g} = 0\}$ and $L = \mathbb{R}^n_+$. Both L and M are polyhedral sets, so we do not need to worry about constraint qualifications. The equivalent VI version of (5.53)–(5.54) is then

$$\mathbf{f} \in K, \tag{5.55}$$

$$0 \le \left\langle \widetilde{\mathbf{f}} - \mathbf{f},\, \boldsymbol{\varphi}(\mathbf{f}) \right\rangle \qquad \text{for all } \widetilde{\mathbf{f}} \in K. \tag{5.56}$$

Existence and uniqueness of solutions of the VI follow from the fact that $\nabla\boldsymbol{\varphi}(\mathbf{f})$ is a diagonal matrix with positive diagonal entries bounded away from zero, and so $\boldsymbol{\varphi}$ is a strongly monotone function.

5.5.2 Dynamic traffic models

The basic Wardrop model is very powerful, but it describes a static situation. A dynamic version of this is the *Boston traffic equilibrium* model of Friesz et al. [107]. The essential idea is that individual drivers choose the route to take at a given intersection based on the current congestion values. The flow into a road does not necessarily have to equal the flow out of a road, as the road can absorb and release a number a vehicles; however, as the number of cars on a road increases, so does the congestion, and so the speed of the traffic on the road will decrease. Because of this we need to introduce more variables. In particular, we need to distinguish between the flow rate into an edge and the flow rate out of an edge, as these can be different, unlike the static situation.

The model developed here is not the full *Boston traffic equilibrium* model of Friesz et al. [107], as the model developed here does not impose the usual "first-in-first-out" queue

discipline of single lane roads. However, the "first-in-first-out" discipline can be approximated by subdividing a road without intersections into a sequence of shorter segments. The limit as the number of segments goes to infinity is a partial differential equation that is commonly used for modeling traffic on road segments (see, for example, [118]). This partial differential equation does preserve the "first-in-first-out" discipline.

The basic variables are $n_{e,d}(t)$, the number of vehicles on edge $e \in E$ with destination $d \in V$, $f^+_{e,d}(t)$ the rate of inflow of vehicles to edge e with destination d, and $f^-_{e,d}(t)$ the rate of outflow of vehicles from edge e with destination d. The basic law of conservation is that

$$\frac{dn_{e,d}}{dt}(t) = f^+_{e,d}(t) - f^-_{e,d}(t). \tag{5.57}$$

The time to traverse edge e for entering traffic depends only on the total number of vehicles on the edge: $\widehat{\tau}_e = \varphi_e(n_e(t))$, where $n_e(t) = \sum_{d\in D} n_{e,d}(t)$. As before, D is the set of destination nodes. We assume that $\varphi_e(n_e)$ is an increasing function of n_e. Let $\tau_{x,d}(t)$ be the anticipated minimal time (based on current congestion) to travel from node $x \in V$ to the destination $d \in D$. The simplest version of the Wardrop equilibrium is that drivers choose the exit at an intersection (or node) so as to give this anticipated minimum time. That is, the inflow $f^+_{e,d}$ to an edge e of vehicles with destination d can be positive only if $\tau_{y,d} = \widehat{\tau}_e + \tau_{x,d}$ for $y = \text{end}(e)$ and $x = \text{start}(e)$. Otherwise, $f^+_{e,d} = 0$ and $\tau_{y,d} > \widehat{\tau}_e + \tau_{x,d}$. This can be represented as

$$0 \le \widehat{\tau}_e - \sum_{x\in V} w_{x,e}\tau_{x,d} \perp f^+_{e,d} \ge 0 \qquad \text{for all } e \in E, d \in D. \tag{5.58}$$

In addition, $\tau_{d,d} = 0$; that is, the time to reach destination d from node d is zero. These complementarity conditions represent the drivers' decisions.

The outflows $f^-_{e,d}$, on the other hand, depend on the local traffic conditions. For an edge e we will assume that there is a natural speed v_e that depends on the vehicle density $\rho_e := n_e/\ell_e$ where ℓ_e is the length of edge e. The natural flow rate on edge e would then be the product of the density and the speed, which is $\rho_e\, v_e(\rho_e) = (n_e/\ell_e)\, v_e(n_e/\ell_e)$, which we can write as $n_e\, \psi_e(n_e)$. Note that $v_e(\rho_e)$ is a positive, decreasing function of ρ_e. We do not allow $v_e(\rho_e) = 0$ for any value of ρ_e: this would mean that no vehicles could leave the edge. The time to traverse edge e can be determined in terms of the length and natural velocity on the edge: $\widehat{\tau}_e = \varphi_e(n_e) = \ell_e/v_e(n_e/\ell_e)$.

The total outflow $f^-_e = \sum_d f^-_{e,d}$ can then be written as

$$f^-_e = n_e\, \psi_e(n_e).$$

The outflow $f^-_{e,d}$ with destination d is simply f^-_e multiplied by the fraction of vehicles on e with destination d:

$$\begin{aligned} f^-_{e,d} &= \frac{n_{e.d}}{n_e}\, n_e\, \psi_e(n_e) \\ &= n_{e,d}\, \psi_e(n_e). \end{aligned}$$

In addition to these conditions, we need conservation of flows to hold. That is, cars are neither "created" nor "destroyed," but they must be accounted for. Some nodes can act as sources or sinks (parking lots, or residential suburbs, for example). Assuming that

drivers leave their source nodes at predetermined times, the sources have predetermined flow rates: $h_{x,d}(t)$ is the rate at which cars with destination d leave node x at time t. The sinks do not have predetermined flow rates: vehicles arrive when they can, and road conditions will affect the time taken for the journey. But the sink for the flows for cars with destination d is precisely the node d. So we set $f^+_{e,d}(t) = 0$ whenever $d = \text{start}(e)$; that is, no vehicle with destination d leaves node d.

If a node $x \neq d$, then the outflow from x (with destination d) must be equal to the inflow to x (with destination d) plus the source of vehicles (with destination d). Thus

$$\begin{aligned} h_{x,d}(t) &= \sum_{e:\text{start}(e)=x} f^+_{e,d}(t) - \sum_{e:\text{end}(e)=x} f^-_{e,d}(t) \\ &= \sum_{e\in E} \Big[w^+_{x,e} f^+_{e,d}(t) - w^-_{x,e} f^-_{e,d}(t) \Big], \end{aligned} \tag{5.59}$$

where

$$w^+_{x,e} = \begin{cases} 1 & \text{if } x = \text{start}(e), \\ 0 & \text{otherwise,} \end{cases}$$

$$w^-_{x,e} = \begin{cases} 1 & \text{if } x = \text{end}(e), \\ 0 & \text{otherwise.} \end{cases}$$

Note that $w_{x,e} = w^+_{x,e} - w^-_{x,e}$ for all $x \in V$ and $e \in E$.

Several points should be noted about this model:

1. The nodes of the traffic network are, in the short time limit, decoupled because vehicles can be "stored" on the edges connecting the nodes.
2. The problem formulation is asymmetric in that the rules governing $f^+_{e,d}$ are very different from the rules governing $f^-_{e,d}$. This is also very different from the static problem where these must be the same quantity.
3. Generically, for a fixed node $x \in V$, we expect the values $\widehat{\tau}_e + \tau_{y,d}$ with $\text{start}(e) = x$ and $\text{end}(e) = y$ to be different for different neighboring y. If this is true for all $x \in V$, then $f^+_{e,d} = 0$ for all edges e with $\text{start}(e) = x$ and $\text{end}(e) = y$ where y is not the optimal choice from node x, and for the edge e^* that is optimal, $f^+_{e^*,d}$ is determined uniquely by the flow constraints. Thus the model is a piecewise smooth differential equation or inclusion.

Various modifications to this model can be made. One might be to require that "saturated" nodes cannot accept more vehicles. This requirement can be modeled using VIs. However, the conservation conditions (5.59) would need to be modified to allow vehicles at source nodes to stay at the source node if all outgoing edges are saturated. Various other modifications to the models can be made, and DVIs provide a convenient way of doing so.

5.5.3 Existence

To show existence of solutions, we can represent this system as a differential inclusion. First, it should be noted that the $\tau_{x,d}$ are Lipschitz functions of $\mathbf{n} = \big[n_{e,d} \mid e \in E, d \in D\big]$

given recursively by

$$\tau_{x,d} = \min_{e:\text{start}(e)=x} \widehat{\tau}_e(n_e) + \tau_{\text{end}(e),d}, \tag{5.60}$$

$$\tau_{d,d} = 0. \tag{5.61}$$

If we order the nodes $x \in V$ by the number of edges from the destination d, then we can define $\tau_{x,d}$ in terms of $\tau_{y,d}$ where y ranges over previous nodes in the ordering. The conditions on the $f^+_{e,d}$ values are

$$f^+_{e,d} \geq 0, \tag{5.62}$$

$$f^+_{e,d} = 0 \qquad \text{if } \tau_{\text{start}(e),d} < \tau_{\text{end}(e),d} + \widehat{\tau}_e(n_e), \tag{5.63}$$

$$h_{x,d}(t) = \sum_{e \in E} \left[w^+_{x,e} f^+_{e,d}(t) - w^-_{x,e} f^-_{e,d}(t) \right] \qquad \text{for all } x \neq d. \tag{5.64}$$

Thus the set of permissible values of $\mathbf{f}^+ = [f^+_{e,d} \mid e \in E, d \in V]$ forms a closed convex set. Noting that $f^-_{e,d}(t) = n_{e,d}(t)\,\psi_e(n_e(t))$ is a function of $\mathbf{n}(t)$, we can write

$$\frac{d\mathbf{n}}{dt}(t) \in \Phi(t, \mathbf{n}(t)), \qquad \text{where}$$

$$\Phi(t,\mathbf{n}) = \left\{ \left[f^+_{e,d} - n_{e,d}\psi_e(n_e) \right]_{e \in E, d \in D} \mid \mathbf{f}^+ \text{ satisfies (5.62)–(5.64)} \right\}.$$

By continuity of $\tau_{x,d}$ in $\mathbf{n}$, we can show that the graph of $\Phi(t,\cdot)$ is closed. The values $\Phi(t,\mathbf{n})$ are closed convex sets. And finally, for bounded $h_{x,d}(\cdot)$, $\Phi(t,\mathbf{n})$ is contained in a common closed bounded set. Thus, by Theorem 4.3, there are solutions for this differential inclusion for any initial conditions $\mathbf{n}(t_0) = \mathbf{n}_0$.

Existence of solutions can also be shown by means of the DVI theory in this chapter. Again, we can treat $\tau_{x,d}$ as a function of $\mathbf{n}$ via (5.60)–(5.61). However, we note that for given $x \neq d \in V$, the conservation condition (5.59) means that

$$\sum_{e:\text{start}(e)=x} f^+_{e,d} = h_{x,d}(t) + \sum_{e:\text{end}(e)=x} n_{e,d}\,\psi_e(n_e).$$

We can therefore write

$$f^+_{e,d} = \left[h_{x,d}(t) + \sum_{e:\text{end}(e)=x} n_{e,d}\,\psi_e(n_e) \right] \beta_{e,d}$$

with $\beta_{e,d} \geq 0$ and $\sum_{e:\text{start}(e)=x} \beta_{e,d} = 1$. Note that we need $h_{x,d}(t) \geq 0$ for $x \neq d$ to do this. Writing $\boldsymbol{\beta}_{x,d} = \left[\beta_{e,d} \mid \text{start}(e) = x \right] \in \mathbb{R}^{m_{x,d}}$, then $\boldsymbol{\beta}_{x,d}$ belongs to the unit simplex $\Sigma_{x,d}$ in $m_{x,d}$ dimensions, where $m_{x,d}$ is the number of edges e with $\text{start}(e) = x$, provided $x \neq d$. We can then modify the complementarity conditions (5.58) to be a VI in $\boldsymbol{\beta}_{x,d}$:

$$\begin{aligned} &\boldsymbol{\beta}_{x,d} \in \Sigma_{x,d}, \\ &\quad 0 \leq \left\langle \widetilde{\boldsymbol{\beta}}_{x,d} - \boldsymbol{\beta}_{x,d}, \left[\widehat{\tau}_e(n_e) + \tau_{\text{end}(e),d}(\mathbf{n}) \mid \text{start}(e) = x \right] \right\rangle \\ &\qquad\qquad \text{for all } \widetilde{\boldsymbol{\beta}}_{x,d} \in \Sigma_{x,d}. \end{aligned}$$

This shows that the DVI has index one. However, the function

$$\mathbf{G}_{x,d}(\mathbf{n}) = \left[\widehat{\tau}_e(n_e) + \tau_{\mathrm{end}(e),d}(\mathbf{n}) \mid \mathrm{start}(e) = x \right] \tag{5.65}$$

is only Lipschitz, not differentiable everywhere. We can, however, take smooth approximations for the purpose of showing existence of solutions. Using the $\beta_{e,d}$ variables, the differential equation (5.57) should be modified to read as

$$\frac{dn_{e,d}}{dt} = \left[h_{\mathrm{start}(e),d}(t) + \sum_{e':\mathrm{end}(e')=x} n_{e',d}\, \psi_{e'}(n_{e'}) \right] \beta_{e,d} - n_{e,d}\, \psi_e(n_e).$$

Thus the $B(t,\mathbf{n})$ matrix is diagonal with diagonal entries

$$B_{e,d}(t,\mathbf{n}) = h_{\mathrm{start}(e),d}(t) + \sum_{e':\mathrm{end}(e')=\mathrm{start}(e)} n_{e',d}\, \psi_{e'}(n_{e'}),$$

which are positive. The set K over which the DVI is based is

$$K = \prod_{x \in V, d \in D: x \neq d} \Sigma_{x,d},$$

a Cartesian product of simplexes. This set is bounded, and so we can apply Theorem 5.1 to show existence of solutions for smooth approximations to $\mathbf{G}_{x,d}(\mathbf{n})$. Taking limits of the resulting approximate solutions (weakly for $\boldsymbol{\beta}_{x,d}(\cdot)$, uniformly for $\mathbf{n}(\cdot)$), we obtain a solution of the DVI for the dynamic traffic problem here.

5.5.4 Uniqueness

At the time of this writing, there is no proof of uniqueness. However, we can obtain some insights into possible instabilities. Suppose that there are two solutions $\mathbf{n}^{(1)}(\cdot)$ and $\mathbf{n}^{(2)}(\cdot)$ with the same initial values $\mathbf{n}^{(1)}(t_0) = \mathbf{n}^{(2)}(t_0) = \mathbf{n}_0$. Suppose also that there is only one destination: $D = \{d\}$. Let

$$t^* = \sup \left\{ t \mid t \geq t_0 \text{ and } \mathbf{n}^{(1)}(t) = \mathbf{n}^{(2)}(t) \right\},$$

and $\mathbf{n}^* = \mathbf{n}^{(1)}(t^*) = \mathbf{n}^{(2)}(t^*)$ by continuity. Showing uniqueness is equivalent to showing $t^* = +\infty$, and so we just need to show uniqueness for t close to t^*. Clearly we need to consider the matrix $\nabla\mathbf{G}(\mathbf{n})\, B(t,\mathbf{n})$. Again, we note that $\mathbf{G}(\mathbf{n})$ is not smooth in $\mathbf{n}$ in general, which complicates the analysis, but the main problem here is lack of symmetry in $\nabla\mathbf{G}(\mathbf{n})$. The matrix $B(t,\mathbf{n})$ is diagonal with diagonal entries, which we can assume for now to be positive. The structure of $\nabla\mathbf{G}(\mathbf{n})$ is block upper triangular with blocks consistent with the Cartesian product structure of K. To see this, note that we can order the nodes V according to $\tau_{x,d}(\mathbf{n}^*)$, with ties broken arbitrarily. Since $\widehat{\tau}_e(n_e^*) > 0$ for all edges e, $\tau_{x,d}(\mathbf{n})$ can depend only on $\tau_{y,d}(\mathbf{n})$ for $\mathbf{n} \approx \mathbf{n}^*$ if $\tau_{x,d}(\mathbf{n}^*) > \tau_{y,d}(\mathbf{n}^*)$. For any node $x \neq d$, we can write $\mathbf{G}_{x,d}(\mathbf{n}) = \left[G_{e,d}(\mathbf{n}) \mid e : \mathrm{start}(e) = x \right]$ with

$$G_{e,d}(\mathbf{n}) = \widehat{\tau}_e(n_e) + \tau_{\mathrm{end}(e),d}(\mathbf{n}).$$

Thus, for $\mathbf{n} \approx \mathbf{n}^*$, $\mathbf{G}_{x,d}(\mathbf{n})$ depends only on $\mathbf{n}_{x,d}$ and $\mathbf{n}_{y,d}$ with $\tau_{x,d}(\mathbf{n}^*) > \tau_{y,d}(\mathbf{n}^*)$. This dependence is Lipschitz, and $\mathbf{G}_{x,d}(\mathbf{n})$ depends smoothly on $\mathbf{n}_{x,d}$.

If we are focusing on a node $x \neq d$, and the edges e with $\text{start}(e) = x$, with $\mathbf{n}_{x,d} = \left[n_{e,d} \mid \text{start}(e) = x \right]$ we see that $\nabla_{\mathbf{n}_{x,d}} \mathbf{G}_{x,d}(\mathbf{n})$ is diagonal with entries $\widehat{\tau}_e'(n_e)$ for e where $\text{start}(e) = x$. Thus $\nabla_{\mathbf{n}_{x,d}} \mathbf{G}_{x,d}(\mathbf{n})\, B_{x,d}(\mathbf{n})$ is diagonal with positive diagonal entries. If we could consider $\mathbf{n}_{x,d}$ *in isolation*, then we could show uniqueness of solutions.

The problem is that there is a feedback loop in the dynamics from $\mathbf{n}_{y,d}$ with $\tau_{y,d}(\mathbf{n}^*) > \tau_{x,d}(\mathbf{n}^*)$ as well as the forward dependence of $\mathbf{n}_{x,d}$ on $\mathbf{n}_{y,d}$ with $\tau_{y,d}(\mathbf{n}^*) < \tau_{x,d}(\mathbf{n}^*)$. In particular, if $\text{start}(e) = x$, $B_{e,d}(\mathbf{n})$ depends on $n_{e',d}$, where $\text{end}(e') = x$. This dependence, in spite of the fact that it is Lipschitz, has the potential to destroy uniqueness.

We could attempt to generalize Theorem 5.3 to systems where ∇G has block upper triangular structure with symmetric positive definite blocks consistent with a Cartesian product structure of K: $K = \prod_{i=1}^m K_i$. Consider a system

$$\frac{dx_i}{dt} = f_i(\mathbf{x}) + B_i(\mathbf{x}) z_i(t), \qquad x_i(t_0) = x_{i,0}, \tag{5.66}$$

$$z_i(t) \in K_i \quad \& \quad 0 \leq \langle \widetilde{z}_i - z_i(t),\, G_i(\mathbf{x}) \rangle \qquad \text{for all } \widetilde{z}_i \in K_i, \tag{5.67}$$

$i = 1, 2, \ldots, m$, where $\nabla_{x_i} G_i(\mathbf{x})\, B_i(\mathbf{x})$ is symmetric positive definite and $G_i(\mathbf{x})$ depends only on x_j with $j \leq i$. We make the usual assumptions that all functions involved are bounded and Lipschitz. Then there is a symmetric positive definite matrix $Q_i(\mathbf{x})$ that is locally Lipschitz where $\nabla_{x_i} G_i(\mathbf{x}) = B_i(\mathbf{x})^T Q_i(\mathbf{x})$. In what follows, we use the notation $f(s) = \mathcal{O}(g(s))$ as $s \to 0$ to mean that there are constants C and $s_0 > 0$ such that $\| f(s) \| \leq C\, g(s)$ for all s with $|s| \leq s_0$. In this case, the "hidden constants," C and s_0, depend on the bounds and Lipschitz constants of the functions defining the problem, but not on other quantities. We also use the notation $\|u\|_C = \sqrt{\langle u, Cu \rangle}$ for C a symmetric positive definite matrix; if $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are two solutions of the system (5.66)–(5.67) with $\mathbf{u} = \mathbf{x}^{(1)} - \mathbf{x}^{(2)}$, $\mathbf{v} = \mathbf{z}^{(1)} - \mathbf{z}^{(2)}$, we have

$$\begin{aligned}
&\frac{d}{dt} \left(\frac{1}{2} \|u_i\|^2_{Q_i(\mathbf{x}^{(1)})} \right) \\
&\quad \leq \left\langle \frac{du_i}{dt},\, Q_i(\mathbf{x}^{(1)}) u_i \right\rangle + \mathcal{O}\left(\|u_i\|^2 \right) \\
&\quad = \left\langle f_i(\mathbf{x}^{(1)}) - f_i(\mathbf{x}^{(2)}) + B_i(\mathbf{x}^{(1)}) z_i^{(1)} - B_i(\mathbf{x}^{(2)}) z_i^{(2)},\, Q_i(\mathbf{x}^{(1)}) u_i \right\rangle \\
&\qquad\qquad + \mathcal{O}(\|u_i\|\, \|\mathbf{u}\|) \\
&\quad = \left\langle B(\mathbf{x}^{(1)}) z_i^{(1)} - B(\mathbf{x}^{(2)}) z_i^{(2)},\, Q_i(\mathbf{x}^{(1)}) u_i \right\rangle + \mathcal{O}(\|u_i\|\, \|\mathbf{u}\|) \\
&\quad = \left\langle B_i(\mathbf{x}^{(1)}) v_i,\, Q_i(\mathbf{x}^{(1)}) u_i \right\rangle + \mathcal{O}(\|u_i\|\, \|\mathbf{u}\|) \\
&\quad = \left\langle v_i,\, B_i(\mathbf{x}^{(1)})^T Q_i(\mathbf{x}^{(1)}) u_i \right\rangle + \mathcal{O}(\|u_i\|\, \|\mathbf{u}\|).
\end{aligned}$$

On the other hand,

$$0 \geq \left\langle z_i^{(1)} - z_i^{(2)},\, G_i(\mathbf{x}^{(1)}) - G_i(\mathbf{x}^{(2)}) \right\rangle$$

from the VI (5.67). We need to replace $G_i(\mathbf{x}^{(1)}) - G_i(\mathbf{x}^{(2)})$ by terms using $\nabla_{x_i} G_i(\mathbf{x}^{(1)}) = B_i(\mathbf{x}^{(1)})^T Q_i(\mathbf{x}^{(1)})$ and perturbations due to the dependence of $G_i(\mathbf{x})$ on x_j, $j < i$. This

gives

$$G_i(\mathbf{x}^{(1)}) - G_i(\mathbf{x}^{(2)}) = \nabla_{x_i} G_i(\mathbf{x}^{(1)})\, u_i + \mathcal{O}\left(\|u_i\|^2\right) + \sum_{j:j<i} \mathcal{O}\left(\|u_j\|\right).$$

Thus, using boundedness of $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$,

$$\left\langle v_i,\, B_i(\mathbf{x}^{(1)})^T\, Q_i(\mathbf{x}^{(1)}) u_i \right\rangle \leq \mathcal{O}\left(\|u_i\|^2\right) + \sum_{j:j<i} \mathcal{O}\left(\|u_j\|\right),$$

and so

$$\begin{aligned}\frac{d}{dt}\left(\frac{1}{2}\|u_i\|^2_{Q_i(\mathbf{x}^{(1)})}\right) &\leq \mathcal{O}\left(\|u_i\|\,\|\mathbf{u}\|\right) + \sum_{j:j<i} \mathcal{O}\left(\|u_j\|\right) \\ &= \mathcal{O}\left(\|u_i\|_{Q_i(\mathbf{x}^{(1)})}\|\mathbf{u}\|_{Q(\mathbf{x}^{(1)})}\right) + \sum_{j:j<i} \mathcal{O}\left(\|u_j\|_{Q_j(\mathbf{x}^{(1)})}\right).\end{aligned}$$

If we set $\eta_i(t) = \|u_i(t)\|^2_{Q_i(\mathbf{x}^{(1)}(t))}$, then we have the differential inequalities for suitable constant $C > 0$,

$$\frac{d\eta_i}{dt} \leq C\left[\sum_{j=1}^{m} \eta_j + \sum_{j=1}^{i-1} \eta_j^{1/2}\right], \qquad i = 1, 2, \ldots, m,$$

using $\eta_i^{1/2}\eta_j^{1/2} \leq \frac{1}{2}\left(\eta_i + \eta_j\right)$. If $t^* = \sup\left\{\, t \mid \mathbf{x}^{(1)}(t) = \mathbf{x}^{(2)}(t)\,\right\}$, then $\eta_i(t^*) = 0$ for all i. We want to show that $\eta_i(t) = 0$ for all i for at least a small time interval $\left[t^*, t^* + \epsilon\right]$, $\epsilon > 0$. However, for $m = 2$, this is not true. Scaling the time variable by C we can remove this constant, giving the system

$$\begin{aligned}\frac{d\eta_1}{dt} &\leq \eta_1 + \eta_2, \\ \frac{d\eta_2}{dt} &\leq \eta_1^{1/2} + \eta_2\end{aligned}$$

with $\eta_i(t^*) = \eta_2(t^*) = 0$. But the system $d\eta_1/dt = \eta_2$, $d\eta_2/dt = \eta_1^{1/2}$ has the solution $\eta_1(t) = t^4/144$, $\eta_2(t) = t^3/48$ which satisfies the above differential inequalities for $m = 2$. Thus we cannot conclude from this that solutions are unique, but there are strong limits on how the nonuniqueness can arise.

Chapter 6

Index Two: Impact Problems

Mechanical impact problems are a rich source of finite-dimensional and infinite-dimensional DVIs. Unlike resource-constrained problems, these are all at least formally index two since Newton's laws of motion give second order differential equations.

We distinguish between rigid-body dynamics with impact, which give finite-dimensional problems, and elastic-body dynamics with impact, which give infinite-dimensional problems. For elastic-body dynamics there can be contact over the domain of the body, or over all or part of the boundary of the body. Also, a body can be elastic or viscoelastic. For the infinite-dimensional problems, the regularity of the solution both in time and space can be crucial for the existence of solutions and their behavior.

If the (normal) contact force is known, then determining the Coulomb friction forces and the resulting motions can be represented as an index-one problem of a variational kind. However, with both the normal and the Coulomb friction forces to be determined, the problems can no longer be represented as optimization problems. Impact problems with Coulomb friction remain the most challenging problems involving mechanics with constraints.

6.1 Rigid-body dynamics

Rigid-body and particle models of mechanics have been in existence since Newton's *Principia Mathematica*. Impact problems for particles was a topic considered by Newton (see Stronge [253, p. 28]). One of the features of rigid-body impacts is that the contact forces must include impulses, that is, Dirac δ-functions. Including such irregular "functions" complicates the theory for these problems. First, we work in a space of measures, and we look for weak solutions. Second, we need to interpret the differential inclusions in a new way.

Rigid-body models use only a finite number of parameters to describe the state of the system: three for a particle in $\mathbb{R}^3$, three for a rigid body in $\mathbb{R}^2$, and six for a rigid body in $\mathbb{R}^3$. For rigid bodies in three dimensions, the issue of how to represent the orientation of a body is a common problem; the main representations in common use are Euler angles, unit quaternions, 3×3 orthogonal matrices, and Rodrigues parameters. Each method has some advantages and disadvantages: Euler angles have singularities as a coordinate system; unit quaternions use four numbers to represent an orientation, and each orientation is

represented by two different unit quaternions; and Rodrigues parameters (while using three parameters) require occasional transformations to avoid singularities.

Whatever means is used to represent the state of a rigid body, we can represent the state in a generalized coordinate vector $q(t)$ which can contain angles as well as positions of centers of mass, for example. We will also assume that there is a generalized velocity vector, which can contain angular as well as ordinary velocities. Typically $v(t) = dq/dt(t)$, but with different representations of orientations, the relation $dq/dt(t) = G(q(t))\,v(t)$ allows more flexibility. For example, the orientation of a body can be represented by quaternions, while the velocity may use the ordinary angular velocity vector. Then the dimensions of $q(t)$ and $v(t)$ are different: $q(t) \in \mathbb{R}^4$, while $v(t) \in \mathbb{R}^3$ and $G(q(t))$ is not even a square matrix.

However, we represent a system of rigid bodies; when there is impact, the forces can be impulses and the velocities can be discontinuous. This means that Newton's second law, that mass times acceleration is the applied force, must be understood in a generalized or distributional sense, as neither the acceleration nor the applied force is a conventional function of time. To understand such systems, we need to turn to ideas such as MDIs as described in Section 4.4.4. This approach of using measures and MDIs can be found in books by Monteiro Marques [174], Brogliato [42], and Glocker [111].

6.1.1 Lagrangian formulation of mechanics

Rigid-body dynamics without contact is often described in terms of Lagrangian or Hamiltonian mechanics. Hamiltonian mechanics is often preferred by people in theoretical mechanics because of the special properties of the resulting differential equations. Here we will use Lagrangian mechanics, which are a little easier to work with for external, frictional, and dissipative forces.

Lagrangian mechanics without constraints start with a Lagrangian function

$$L(q, v) = T(q, v) - V(q), \tag{6.1}$$

where $T(q, v)$ is the *kinetic energy*, and $V(q)$ is the *potential energy*, associated with configuration (generalized coordinate vector) q and generalized velocity $v = dq/dt$. Note that q can contain angular and orientation as well as translational components, so v can contain angular velocities as well as ordinary translational velocities. Usually the kinetic energy is a quadratic homogeneous function of the velocity

$$T(q, v) = \frac{1}{2} v^T M(q)\, v, \tag{6.2}$$

where $M(q)$ is the *mass matrix*. For systems of particles, $M(q)$ is a constant diagonal matrix with the masses of the particles on the diagonal. For rigid bodies, using a suitable method for representing orientations such as Euler angles or quaternions, together with the coordinates of the center of mass, the mass matrix is partly diagonal (with the masses of the rigid bodies on the diagonal) and partly block diagonal 3×3 or 4×4 matrices (for the moment of inertia matrices).

The kinetic energy function can sometimes be quadratic but not homogeneous in v:

$$T(q, v) = \frac{1}{2} v^T M(q)\, v + b(q)^T v + c(q). \tag{6.3}$$

For example, when rotating or other moving reference frames are used so that $q \equiv$ constant does not mean that the body is stationary this leads to nonzero $b(q)$ or $c(q)$. A common example of this are *Coriolis forces* that arise due to fact that coordinate systems fixed in the Earth are, in fact, rotating. These Coriolis forces are not true forces, but rather are pseudoforces that arise because of rotating coordinate systems. For example, if we use Earth-based coordinate systems with $\dot{\mathbf{q}}(t)$ the position of a particle of mass m (in an Earth-based coordinate system) and the Earth's angular velocity is $\boldsymbol{\Omega}$, the velocity of the particle is $\dot{\mathbf{q}} + \boldsymbol{\Omega} \times \mathbf{q} = \mathbf{v} + \boldsymbol{\Omega} \times \mathbf{q}$, and so its kinetic energy is

$$\frac{1}{2} m \left\| \mathbf{v} + \boldsymbol{\Omega} \times \mathbf{q} \right\|_2^2 = \frac{1}{2} m \mathbf{v}^T \mathbf{v} + m \left(\boldsymbol{\Omega} \times \mathbf{q} \right)^T \mathbf{v} + \frac{1}{2} m \left\| \boldsymbol{\Omega} \times \mathbf{q} \right\|_2^2.$$

Nevertheless, even in these systems, the kinetic energy function will be taken to have the form (6.3). The potential energy function $V(q)$ can come from gravitational, electrical, magnetic, or other forces. As such, there is no general form for $V(q)$.

For the remainder of the chapter we will assume that $T(q, v)$ is quadratic *homogeneous* in v.

The fundamental equations for Lagrangian mechanics come from the so-called *principle of least action*. This name is actually a misnomer, and it should be called the *principle of stationary action*. The action is the functional

$$S[q] := \int_a^b L\left(q(t), \frac{dq}{dt}(t) \right) dt, \tag{6.4}$$

where $a < b$ are arbitrary times.

The principle of stationary action is that the "gradient" of $S[q]$ with respect to q (with $q(a)$ and $q(b)$ fixed) is zero. That is,

$$\frac{d}{d\alpha} S[q + \alpha\eta]\big|_{\alpha=0} = 0 \tag{6.5}$$

for all sufficiently smooth functions $\eta \colon [a,b] \to \mathbb{R}^n$ with $\eta(a) = \eta(b) = 0$. This variational condition is equivalent to the *Euler–Lagrange equations*:

$$0 = \frac{d}{dt} \nabla_v L(q, v) - \nabla_q L(q, v), \tag{6.6}$$

$$v = \frac{dq}{dt}. \tag{6.7}$$

Assuming that the kinetic energy is quadratic homogeneous (6.2), we obtain the differential equations

$$M(q) \frac{dv}{dt} = k(q, v) - \nabla V(q), \tag{6.8}$$

$$\frac{dq}{dt} = v, \tag{6.9}$$

where

$$k_i(q, v) = \frac{1}{2} \sum_{j,k} \left(\frac{\partial m_{jk}}{\partial q_i}(q) - \frac{\partial m_{ik}}{\partial q_j}(q) - \frac{\partial m_{ji}}{\partial q_k}(q) \right),$$

with $m_{ij}(q)$ the (i, j) entry of the mass matrix $M(q)$.

6.1.2 Frictionless problems

Frictionless impact problems for rigid bodies can be represented in terms of inequality constraints on the generalized coordinates:

$$\varphi_i(q) \geq 0, \qquad i = 1, 2, \ldots, m. \tag{6.10}$$

In order to enforce these constraints, we need to introduce some Lagrange multipliers λ. Physically, the Lagrange multipliers represent generalized forces that ensure that the constraints are not violated. Since the constraints $\varphi_i(q) \geq 0$ must be enforced at all times, there must be a new Lagrange multiplier for each time t; that is, λ is a function of t: $\lambda(t)$. Incorporating this into the Lagrangian function gives

$$L(q, v, \lambda) = T(q, v) - V(q) - \lambda^T \varphi(q). \tag{6.11}$$

Naively applying the Karush–Kuhn–Tucker conditions to the action leads to the system

$$M(q)\frac{dv}{dt} = k(q, v) - \nabla V(q) + \nabla\varphi(q)^T\lambda, \tag{6.12}$$

$$\frac{dq}{dt} = v, \tag{6.13}$$

$$0 \leq \lambda \perp \varphi(q) \geq 0. \tag{6.14}$$

This is a DCP with index two:

$$\frac{d}{dt}\varphi(q) = \nabla\varphi(q)\, v,$$

$$\frac{d^2}{dt^2}\varphi(q) = \frac{d}{dt}\left[\nabla\varphi(q)\right] v + \nabla\varphi(q)\, M(q)^{-1}\left[k(q, v) - \nabla V(q) + \nabla\varphi(q)^T\lambda\right],$$

so that λ can be determined from q and v and $\left(d^2/dt^2\right)\varphi(q)$.

A simple example

The example in Section 1.1 of a ball of mass m and radius r colliding with a table-top can be easily treated with this approach. The only generalized coordinate is the height of the ball y above the table-top. The Lagrangian is

$$L(y, v) = \frac{1}{2}mv^2 + mgy.$$

The constraint is

$$\varphi(y) := y - r \geq 0.$$

This gives the system

$$\frac{dv}{dt} = -mg + \lambda,$$

$$\frac{dy}{dt} = v,$$

$$0 \leq \lambda \perp y - r \geq 0.$$

The Lagrange multiplier λ can be easily identified as the normal contact force $N(t)$ from (1.1).

As noted in Section 3.2.4, these problems do not have unique solutions unless we impose some additional conditions. Usually we assume that there is a given *coefficient of restitution* $0 \le e \le 1$ where $dy/(t^+) = -e\,dy/dt(t^-)$ for any t where $y(t) - r = 0$. This is extended to general mechanical impact problems by requiring that

$$n_i(q(t))^T v(t^+) = -e\, n_i(q(t))^T v(t^-) \qquad \text{whenever } \varphi_i(q(t)) = 0.$$

There is more on the issue of modeling partially elastic impacts in Section 6.1.4.

6.1.3 Coulomb friction

The standard Coulomb law for frictional contact [71] can be summarized as follows:

- the friction force is in the opposite direction to the direction of the slip velocity;
- the magnitude of the friction force never exceeds μ (the coefficient of friction) times the normal contact force; and
- if there is nonzero slip, then the magnitude of the friction force is exactly μ times the normal contact force.

A remarkable property of this law is that it implies that the frictional force does not depend on the apparent area of contact; this characteristic was first announced by Amontons [9] in 1699, and it was considered very anti-intuitive at the time [37, p. 14].

Coulomb's law for frictional contact is a semiempirical law, and there are many variations on it; for more details see Section 1.2. The theoretical foundations for this law are weak, and dry friction is a physically complex phenomenon. Engineering practice has led a number of researchers to develop models of dry friction that modify Coulomb's laws. However, in this book, we will stay with slight modifications of Coulomb's basic laws, but allowing for anisotropic friction.

Often these laws are written in a straightforward way: if N_i is the normal contact force for the ith contact point and F_i the corresponding friction force,

$$F_i = -\mu_i N_i \frac{v_{rel}}{\|v_{rel}\|_2} \qquad \text{if } v_{rel} \neq 0, \tag{6.15}$$

$$\|F_i\|_2 \le \mu_i N_i \qquad \text{if } v_{rel} = 0. \tag{6.16}$$

This is often the simplest way to formulate Coulomb friction, but there are other ways.

Even if we stay with Coulomb's basic laws, we can reformulate them in a way that makes the complementarity and variational aspects more visible. One approach is to use the *maximum dissipation principle* of Erdmann [91]. One begins with the set of possible friction forces $\mathcal{F}_i(q)$ for a given contact i in a given configuration q for unit normal contact force at this contact. This set $\mathcal{F}_i(q)$ should depend continuously on q (provided contact i is maintained) and be a closed, bounded, and convex set. For isotropic Coulomb friction, $\mathcal{F}_i(q)$ is a disk centered on the origin with diameter μ_i, the coefficient of friction for contact i; for a particle, the disk lies in the plane orthogonal to the normal direction vector at q. If we use generalized coordinates, then orthogonality can be lost, and the shape is no longer a disk. However, $\mathcal{F}_i(q)$ will remain a closed, bounded, and convex set.

If contact i is broken, then we can take $\mathcal{F}_i(q) = \{0\}$. In this way, $\mathcal{F}_i$ is an upper semicontinuous set-valued map with closed convex and bounded values. In some situations, it can be convenient to allow $\mathcal{F}_i(q)$ to be something other than a disk. For example, for ice skating, the friction force on a skate clearly depends on the angle between the slip velocity and the direction of the blade of the skate. Another use is to include a frictional torque, such as arises if the steering wheel of a car is turned while the car is stationary. In this case, the frictional torque is due to the fact that contact occurs at more than just a single point.

The friction force for the contact i must then satisfy

$$F_i = \arg \max_{F \in N_i \mathcal{F}_i(q)} -\langle v, F \rangle, \tag{6.17}$$

where v is the (slip) velocity of the system at contact i, and N_i is the normal contact force for contact i. This maximal dissipation principle can be expressed as a VI:

$$F_i \in N_i \mathcal{F}_i(q) \quad \& \quad 0 \leq \left\langle \widetilde{F}_i - F_i, v \right\rangle \qquad \text{for all } \widetilde{F}_i \in N_i \mathcal{F}_i(q). \tag{6.18}$$

Another way is to use VIs of the second kind. This is based on the *support function* of $\mathcal{F}_i(q)$: The support function of a closed convex set C is

$$\sigma_C(p) = \sup_{x \in C} \langle x, p \rangle.$$

Note that F_i minimizes $F \mapsto v \cdot F$ over $F \in N_i \mathcal{F}_i(q)$, and so F minimizes the function $F \mapsto v \cdot F + I_{N_i \mathcal{F}_i(q)}(F)$ where I_C is the *indicator function* where $I_C(x) = 0$ if $x \in C$ and $I_C(x) = +\infty$ otherwise. Thus, $0 \in v + \partial I_{N_i \mathcal{F}_i(q)}(F_i)$ or $-v \in \partial I_{N_i \mathcal{F}_i(q)}(F_i)$. Using Fenchel duality, this is equivalent to

$$F_i \in \partial I^*_{N_i \mathcal{F}_i(q)}(-v).$$

The dual of the indicator function I_C is the support function σ_C. So we can formulate the condition for F_i as

$$F_i \in \partial \sigma_{N_i \mathcal{F}_i(q)}(-v).$$

Noting that $\sigma_{\alpha C}(p) = \alpha \sigma_C(p)$ for any $\alpha \geq 0$, we get

$$F_i \in N_i \, \partial \sigma_{\mathcal{F}_i(q)}(-v).$$

From the definition of subdifferential, this can be written as

$$N_i \, \sigma_{\mathcal{F}_i(q)}(-w) \geq N_i \, \sigma_{\mathcal{F}_i(q)}(-v) + \langle F_i, v - w \rangle \tag{6.19}$$

for all w. If $\mathcal{F}_i(q)$ is, for example, the disk of radius μ_i (the coefficient of friction for contact i) in the plane generated by orthonormal vectors d_1 and d_2, then $\sigma_{\mathcal{F}_i(q)}(p) = \mu_i \left\| [d_1, d_2]^T \, p \right\|_2$. Note that the formulation (6.19) is a VI of the second kind. This formulation has become particularly common in the literature for elastic bodies with Coulomb friction.

Note that given N_i, F_i is the solution of a monotone VI. However, in general rigid-body dynamics, we do not know the normal contact forces N_i a priori. The separate problems of determining the normal contact forces N_i, and given the N_i to compute the friction

forces F_i, can both be represented as monotone VIs. The combined problem of finding both the normal contact forces N_i and the friction forces F_i, however, cannot. This makes rigid-body dynamics with Coulomb friction difficult, even from a theoretical point of view.

An important concept in rigid-body dynamics with Coulomb friction is the *friction cone* for a given contact i, which is the set

$$\widetilde{\mathcal{F}}_i(q) := \{ N_i\, n_i(q) + F_i \mid F_i \in N_i\, \mathcal{F}_i(q), N_i \geq 0 \}.$$

This is a closed convex cone in the space of forces. It is the cone generated by the set $n_i(q) + \mathcal{F}_i(q)$. The friction cone for the entire systems with contacts $i = 1, 2, \ldots, N$ is the Cartesian product $\widetilde{\mathcal{F}}(q) = \prod_{i=1}^{N} \widetilde{\mathcal{F}}_i(q)$. This represents all contact forces acting on the system.

An alternative approach to formulating the maximum dissipation principle can be found in [198]. This uses a representation for the set $\widetilde{\mathcal{F}}_i(q)$ in terms of level sets:

$$\widetilde{\mathcal{F}}_i(q) = \left\{ N_i\, n_i(q) + F_i \mid \phi_{ij}(q, N_i, F_i) \leq 0,\ j = 1, 2, \ldots, m_i \right\},$$

where each ϕ_{ij} has a number of important properties apart from smoothness conditions:

- $\phi_{ij}(q, N_i, F_i)$ is convex in F_i;
- $\phi_{ij}(q, N_i, 0) \leq 0$ for all $N_i \geq 0$ with equality if and only if $N_i = 0$;
- if $\phi_{ij}(q, 0, F_i) \leq 0$ for all j, then $F_i = 0$;
- $\phi_{ij}(q, N_i, 0)$ is positively homogeneous in N_i with order $\gamma_i \geq 1$ (that is, $\phi_{ij}(q, N_i, 0) = N_i^{\gamma_i}\, \phi_{ij}(q, 1, 0)$).

For example, the standard friction cone can be represented by $\phi_{i1}(q, N_i, F_i) = \|F_i\|_2^2 - (\mu_i N_i)^2 \leq 0$ with $F_i \perp n_i(q)$, which satisfies the above conditions with $\gamma_i = 2$. If we have such a representation of the friction cone, then the maximum dissipation principle can be represented by a specially structured nonlinear CP. In [198] it is shown that solutions exist for *static* or *incremental* frictional contact problems represented in this way. Such formulations can be used for time-stepping methods in dynamic problems, for example.

6.1.4 Modeling of partially elastic restitution

The modeling of partly elastic impacts cannot be reduced to modeling with DVIs, differential inclusions, or related techniques. As noted in Section 3.2.4, since this is an index-two DVI, we cannot expect uniqueness just from specifying the DVI. Instead, we need to impose an additional constraint to handle coefficients of restitution. Note that coefficients of restitution come into the formulation only when there are impulsive forces. Often it is better to give a complementarity or VI formulation of the impact law than to simply write, for example, $\langle n, v(t^+) \rangle = -e \langle n, v(t^-) \rangle$ for *Newton's impact law* with coefficient of restitution e. This is particularly true when there are multiple simultaneous impacts and it becomes unclear if the condition can even be satisfied. Take, for example, a ball colliding with a frictional wall after rolling, as shown in Figure 6.1.

Because the wall is rigid, there will be an impulse to the left at contact 1. Because of friction and because the ball is rolling, there will be an *upward* frictional impulse, also

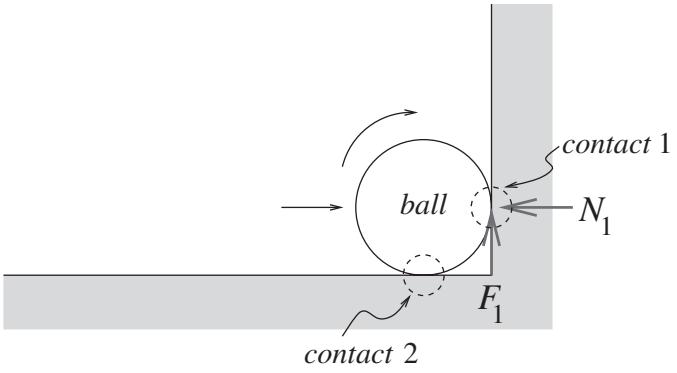


Figure 6.1: Ball rolling into a corner: an example of failure of a naive model of inelastic impact.

at contact 1. Because of this, the ball will have an upward velocity after impact, and there will be no impulsive forces at contact 2. However, $\langle n_2, v(t^-)\rangle = 0$ (n_2 being the normal inward direction vector), so naive application of Newton's impact law would imply that $\langle n_2, v(t^+)\rangle = 0$ as well. Instead, we should use the formulation

$$0 \le N_i^* \perp \langle n_i(q(t)), v(t^+) + e\, v(t^-)\rangle \ge 0 \quad \text{for all } i \text{ where } \varphi_i(q(t)) = 0, \tag{6.20}$$

where N_i^* is the impulse at the ith contact. This way, we can still have $\langle n_2, v(t^+)\rangle > 0$ in Figure 6.1, as long as there is no impulse at contact 2 ($N_2^* = 0$).

In *Poisson's impact law*, the impact is divided into two parts: the compression phase and the expansion phase. The compression phase is essentially purely inelastic. At the end of the compression phase, the normal component of the velocity is zero. The total impulse of the normal contact force during expansion is taken to be the coefficient of restitution times the total impulse during compression. If we write the total impulse during the compression for contact i as $N_i^{(c)}$ and the total impulse for expansion as $N_i^{(x)}$, then following [12] we can use the formulation

$$\begin{aligned} 0 \le N_i^{(c)} &\perp \left\langle n_i(q(t)), v(t^{(c)})\right\rangle \ge 0, \\ 0 \le N_i^{(x)} - e\, N_i^{(c)} &\perp \langle n_i(q(t)), v(t^+)\rangle \ge 0 \\ &\text{for all } i \text{ where } \varphi_i(q(t)) = 0, \end{aligned} \tag{6.21}$$

where $v(t^{(c)})$ is the velocity after the compression phase ($N_i^{(c)}$) and $v(t^+)$ is the velocity after both compression and expansion phases ($N_i^{(c)} + N_i^{(x)}$). Normally, $N_i^{(x)} = e\, N_i^{(c)}$. However, there can be situations in which the second complementarity condition is needed to prevent interpenetration.

Both Newton's and Poisson's impact laws can violate conservation of energy. In the case of Newton's law of impact, the problem is when the direction of slip reverses during

impact due to friction, as described by Stronge [252, 253]. Stronge's model of the motion of a rigid body in contact is a little different from being perfectly rigid. Rather, Stronge essentially solves a singular perturbation problem: the obstacle is not taken to be rigid, but to have a spring with very large stiffness which becomes active when there is contact, and we look at the limit as the stiffness of this spring goes to infinity. From the way it is solved in [252], we write the velocity as a function of the impulse-so-far $0 \le P \le N_1^*$ for a single contact and treat the configuration $q(t)$ as constant during impact. Applying the equations of motion then gives a differential equation for the velocities in terms of dv/dP.

To handle this, Stronge introduced an energy-based impact law [253, p. 69]. In this formulation, the impact is separated into two phases: the compression phase and the expansion (or decompression) phase. At the end of the compression phase the relative normal velocity at the contact point is zero, which determines the work done by the normal contact force during the compression phase ($-W_c$). The total impulse at contact i is then $N_i^{(c)}$. The additional impulse $N_i^{(x)} \ge 0$ is then determined so that the work done by the normal contact force during this phase W_x is a given fraction e_*^2 of W_c.

The trouble with this approach, as with all other attempts to create a law of restitution for rigid-body models, is that it requires the imposition of a physical law a priori, which did not previously exist. Ideally, the model of restitution is a result of the model, not an input to it. The essential problem is the lack of uniqueness inherent to index-two DVIs.

In fact, there are a number of reasons why any rigid-body restitution law would be inadequate in giving physically realistic results in general situations. Evidence of the difficulty in setting up such a mechanism can be seen in the experimental results of Stoianovici and Hurmuzlu [250]. In their experimental setup, slender steel bars were dropped onto a flat hard steel anvil. High-speed video cameras captured the motion of the bars, and wires were connected to both the bar and the anvil in order to determine when there is true contact between bar and anvil. When the bars were dropped while oriented vertically, the ratio of the postimpact normal velocity and the preimpact normal velocity was close to one, indicating essentially perfectly elastic impacts according to the Newton impact law. However, as the angle of the bars from the vertical was increased, the observed Newton coefficient of restitution dropped from near one to less than half and, for more slender rods, to around 0.1. For the more slender rods, this drop occurred over a smaller change in angle and was more dramatic. After this drop, the observed Newton coefficient of restitution started increasing and oscillated erratically about 0.6 to 0.7 for angles far from vertical. This behavior was found not only in the experimental results, but in numerical simulations as well, using a finite-dimensional approximation of the elastic behavior of the rod. Experimental and computational results are shown in Figure 6.2. Qualitatively similar results that compare well were computed by Paoli and Schatzman [203].

Clearly, from these results, the Newton impact law is far from being physically correct, even for slender steel bars. However, if we look at the remarkable drop in the observed Newton coefficient of restitution for the most slender bars from around 0.9 for being dropped vertically to 0.1 for an angle of about 16° from vertical, it should become clear that both the Poisson model of impact and the Stronge energy-based model of impact are also incorrect. In fact, if there is no friction, the velocity depends affinely on the impulse parameter P since the configuration changes negligibly during an impact. Thus the Poisson coefficient and the Stronge coefficient are identical with the Newton coefficient for frictionless impacts.

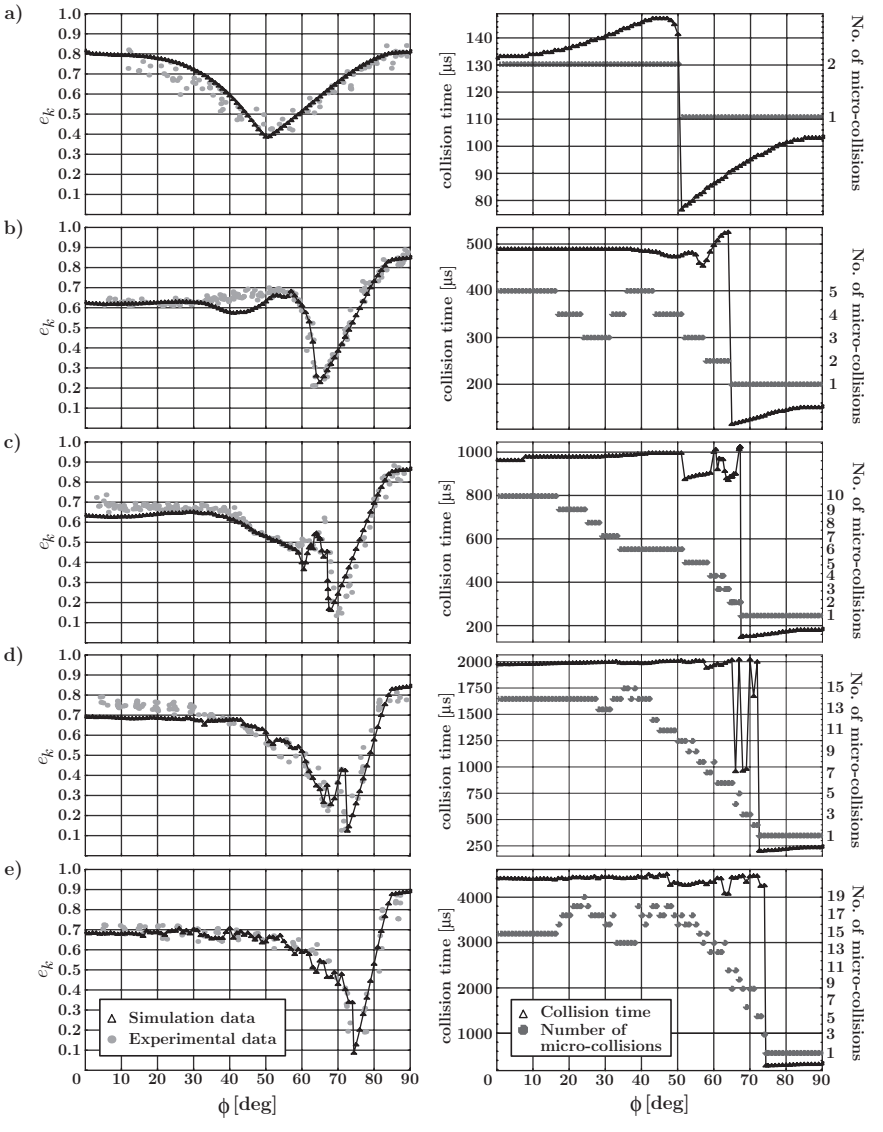


Figure 6.2: Results from [250] for steel bars impacting a steel anvil at different angles from horizontal. All bars have width 12.7 mm with lengths (a) 100 mm, (b) 200 mm, (c) 300 mm, (d) 400 mm, (e) 600 mm. Reprinted with permission.

Also, the simplistic division of an impact into a compression phase and an expansion phase is often far from reality. The number of "microcollisions" (periods of electrical contact between bar and anvil) reported by Stoianovici and Hurmuzlu is commonly much more than one (up to 19 for the most slender bar). Also, the impact time goes from a minimum near 100 μs to a maximum of over 4 ms, again for the most slender bar; the ratio between largest to smallest impact times is about 40. Both the number of "microimpacts"

and the contact times indicate complex dynamics within an apparently simple impact. The erratic behavior of the apparent coefficient of restitution for angles far from vertical also indicates complex dynamics that cannot be represented by simple algebraic relations.

If none of the available models of restitution is applicable, what is to be done? There is one situation in which all models essentially agree, which is the case of zero coefficient of restitution: perfectly inelastic impacts. Beyond this case, the answer is to incorporate elastic vibrations into the models of impact. Here there are difficulties, both computational and theoretical, but these will be discussed in the sections on elastic and viscoelastic impacts.

6.1.5 Technical issues

Even though for rigid bodies we are essentially dealing with *ordinary* differential equations, our contact conditions for the normal contact forces N_i are index two. On the other hand, the conditions for the Coulomb friction forces F_i are index one. Index-two problems, as we have seen in Section 3.2.4, have two technical obstacles we will have to deal with. The first is that the solutions can contain Dirac-δ functions. This occurs very naturally in mechanical contact problems for rigid bodies, as their velocity is unaffected until the bodies make contact, when the velocities of one or both bodies must change instantaneously. This clearly means that the instantaneous acceleration must contain Dirac-δ functions. This makes our contact equations impulsive. Unlike many other approaches to dealing with impulsive systems, our contact problems do not have impulses at a priori known times.

The natural way of handling these impulsive systems mathematically is in terms of measures and spaces of measures. This leads to a different issue, which is how to interpret these kinds of equations with measures when the right-hand side of our differential "equations" are sets. This is a natural way of handling the Coulomb friction part of the problem: we use *differential inclusions* (see Section 4.1). For handling the combination of differential inclusions and impulses, the theory of *MDIs* was set up, as described in Section 4.4.2.

The interaction of the normal and Coulomb friction forces can be very important. In fact, a famous paradox due to Painlevé [194], which was claimed to show the *nonexistence* of solutions to rigid-body dynamics with contact and Coulomb friction forces, is due to this interaction. The resolution which is hinted at in, for example, Delassus [76, 77] involves impulses without collisions. This idea finds fuller expression in the works of Moreau [179, 181], Monteiro Marques [174], and Stewart [237, 238]. This paradox has been a stumbling block for many people working on rigid-body dynamics with friction. The resolution of the paradox, however, is something many people have experienced writing on a blackboard: when the chalk goes in the "wrong" direction, it can jump and jitter, leaving a trail of chalk dots on the board. These dots are evidence of (approximate) Dirac-δ functions. Indeed, anyone wanting to understand how parts of mechanical systems can suddenly jam instead of smoothly sliding should understand the interaction of normal and frictional contact forces.

6.1.6 Painlevé's paradox

Painlevé's paradox is normally described as appearing when a rod is sliding in the "wrong" direction across a frictional surface. The situation is essentially that shown in Figure 6.3.

Let us proceed in a naive way, as Painlevé did, assuming that all the forces are bounded and the solutions are smooth. Then the equations of motion for the rod can be

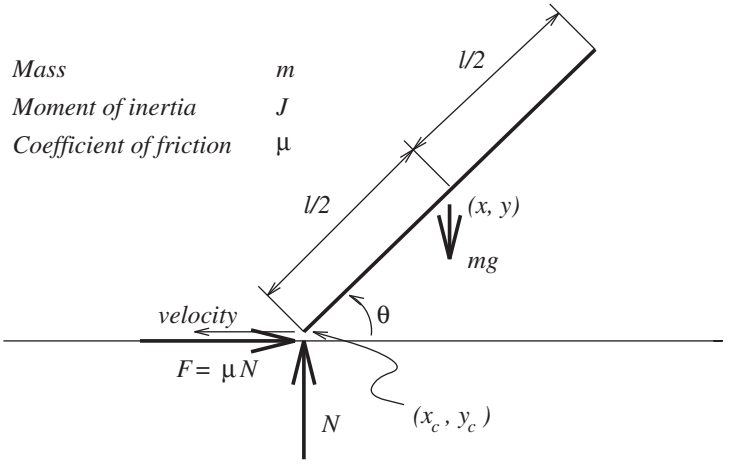


Figure 6.3: Painlevé's paradox.

written as

$$
\begin{aligned}
m\frac{d^2x}{dt^2} &= F,\\
m\frac{d^2y}{dt^2} &= N - mg,\\
J\frac{d^2\theta}{dt^2} &= \frac{\ell}{2}F\sin\theta - \frac{\ell}{2}N\cos\theta.
\end{aligned}
$$

Note that J is the moment of inertia for the rod. For the rod sliding to the left (the "wrong" direction), Coulomb's laws say that $F = \mu N$, where μ is the coefficient of friction. Substituting this into our equations of motion gives

$$
\begin{aligned}
m\frac{d^2x}{dt^2} &= \mu N,\\
m\frac{d^2y}{dt^2} &= N - mg,\\
J\frac{d^2\theta}{dt^2} &= \frac{\ell}{2}\left[\sin\theta\,\mu - \cos\theta\right]N.
\end{aligned}
$$

For the contact conditions we need to follow the position of the contacting point (x_c, y_c) which is at

$$
\begin{aligned}
x_c &= x - \frac{\ell}{2}\cos\theta,\\
y_c &= y - \frac{\ell}{2}\sin\theta.
\end{aligned}
$$

The second derivative gives

$$\begin{aligned}\frac{d^2y_c}{dt^2} &= \frac{d^2y}{dt^2} - \frac{\ell}{2}\cos\theta\,\frac{d^2\theta}{dt^2} + \frac{\ell}{2}\sin\theta\left(\frac{d\theta}{dt}\right)^2 \\ &= \frac{N}{m} - g - \frac{\ell}{2}\cos\theta\,\frac{\ell}{2J}\left[\sin\theta\,\mu - \cos\theta\right]N + \frac{\ell}{2}\sin\theta\left(\frac{d\theta}{dt}\right)^2 \\ &= \frac{1}{m}\left[1 - \frac{m\ell^2}{4J}\cos\theta\,(\sin\theta\,\mu - \cos\theta)\right]N + \frac{\ell}{2}\sin\theta\left(\frac{d\theta}{dt}\right)^2 - g.\end{aligned}$$

If we choose $0 < \theta < \pi/2$, μ sufficiently large, and $J/(m\ell^2)$ sufficiently small, then we can make

$$1 - \frac{m\ell^2}{4J}\cos\theta\,(\sin\theta\,\mu - \cos\theta) < 0.$$

Assuming that $y_c = 0$, $dy_c/dt = 0$, and $d\theta/dt = 0$ at some time, then no matter how large N is, we will always get $dy_c/dt < 0$: penetration is inevitable! This contradicts the basic assumptions of rigid-body dynamics, and so there cannot be any solution to this problem.

Or so Painlevé thought.

6.1.7 Resolution of Painlevé's paradox

The problem with this approach is that it implicitly excludes the possibility of impulsive forces: all functions must be sufficiently differentiable. Clearly there must be impulsive forces in a collision in rigid-body dynamics, but it is less clear that there can be impulsive forces in other situations. However, this is one of those situations.

How can impulsive forces lead to a solution? Even if we allow N to contain Dirac-δ functions, we still get the inequality going the wrong way for d^2y_c/dt^2. The flaw in the argument is the assumption that "$F = \mu N$." This is true as long as $dx_c/dt < 0$, but if we have $dx_c/dt = 0$, then we require only that "$|F| \le \mu N$." We can then allow other, smaller values for F/N, and so we obtain a solution.

To see that we really do get a solution, we can set up the problem as a 4×4 linear CP for the impulsive forces. We can follow the approach of using complementarity-based time-stepping methods proposed for this problem, such as can be found in [238, 247, 248], but with the step size $h > 0$ set to zero. Alternatively, we can derive such conditions directly. We will assume that $N(t) = N^*\,\delta(t - t^*) + N_1(t)$ and $F(t) = F^*\,\delta(t - t^*) + F_1(t)$, where t^* is the time of the impulsive forces, and N_1 and F_1 are smooth functions or, at worst, measures but with no impulsive component at $t = t^*$.

We can obtain a solution by assuming that the maximum dissipation principle applies to the *postimpact* velocity:

$$\begin{aligned}&F^* = \arg\min x_c'(t^{*+})F \\ &\text{over all } F \in [-\mu N^*, +\mu N^*].\end{aligned}$$

This can be written as a CP if we write $F^* = F_+^* - F_-^*$: $0 \le F_\pm^*$, and we include an additional

variable λ. Our complementarity conditions can be written as

$$\begin{aligned} 0 \le F_+^* \perp \lambda + x_c'(t^{*+}) \ge 0, \\ 0 \le F_-^* \perp \lambda - x_c'(t^{*+}) \ge 0, \\ 0 \le \lambda \perp \mu N^* - F_+^* - F_-^* \ge 0. \end{aligned}$$

On the other hand, the normal contact force can be described in terms of a CP in which we now need to include a coefficient of restitution e: for $y_c(t^*) - r = 0$ we have

$$0 \le e\, y_c'(t^{*-}) + y_c'(t^{*+}) \perp N^* \ge 0.$$

To complete the system we have to add the effects of the impulsive forces on $x_c'(t^{*+})$ and $y_c'(t^{*+})$:

$$\begin{aligned} x'(t^{*+}) &= x'(t^{*-}) + \frac{1}{m} F^*, \\ y'(t^{*+}) &= y'(t^{*-}) + \frac{1}{m} N^*, \\ \theta'(t^{*+}) &= \theta'(t^{*-}) + \frac{\ell}{2J} \big[F^* \sin\theta - N^* \cos\theta \big], \end{aligned}$$

so, using $x_c' = x' + (\ell/2)\sin\theta\,\theta'$, $y_c' = y' - (\ell/2)\cos\theta\,\theta'$, we have

$$\begin{aligned} x_c'(t^{*+}) &= x_c'(t^{*-}) + \frac{1}{m} F^* + \frac{\ell^2}{4J} \sin\theta \big[F^* \sin\theta - N^* \cos\theta \big], \\ y_c'(t^{*+}) &= y_c'(t^{*-}) + \frac{1}{m} N^* - \frac{\ell^2}{4J} \cos\theta \big[F^* \sin\theta - N^* \cos\theta \big]. \end{aligned}$$

The LCP generated is

$$\begin{aligned} 0 \le & \begin{bmatrix} N^* \\ F_+^* \\ F_-^* \\ \lambda \end{bmatrix} \\ & \perp \begin{bmatrix} +a_{11}(\theta) & +a_{12}(\theta) & -a_{12}(\theta) & 0 \\ +a_{12}(\theta) & +a_{22}(\theta) & -a_{22}(\theta) & 1 \\ -a_{12}(\theta) & -a_{22}(\theta) & +a_{22}(\theta) & 1 \\ \mu & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} N^* \\ F_+^* \\ F_-^* \\ \lambda \end{bmatrix} + \begin{bmatrix} (1+e) y_c'(t^{*-}) \\ +x_c'(t^{*-}) \\ -x_c'(t^{*-}) \\ 0 \end{bmatrix} \ge 0, \end{aligned}$$

where

$$\begin{aligned} a_{11}(\theta) &= \frac{1}{m} + \frac{\ell^2}{4J} \cos^2\theta, \\ a_{12}(\theta) &= -\frac{\ell^2}{4J} \sin\theta\,\cos\theta, \\ a_{22}(\theta) &= \frac{1}{m} + \frac{\ell^2}{4J} \sin^2\theta. \end{aligned}$$

The question naturally arises: Do solutions exist for this CP? The answer is given, in fact, by Lemke's algorithm. To see this, note that the matrix

$$\begin{bmatrix} a_{11}(\theta) & a_{12}(\theta) & -a_{12}(\theta) \\ a_{12}(\theta) & a_{22}(\theta) & -a_{22}(\theta) \\ -a_{12}(\theta) & -a_{22}(\theta) & a_{22}(\theta) \end{bmatrix}$$

is symmetric positive semidefinite, and its null space is one dimensional, generated by the vector $[0, 1, 1]^T$. First, the matrix

$$M = \begin{bmatrix} +a_{11}(\theta) & +a_{12}(\theta) & -a_{12}(\theta) & 0 \\ +a_{12}(\theta) & +a_{22}(\theta) & -a_{22}(\theta) & 1 \\ -a_{12}(\theta) & -a_{22}(\theta) & +a_{22}(\theta) & 1 \\ \mu & -1 & -1 & 0 \end{bmatrix}$$

of the CP is copositive: if $z \geq 0$, then $z^T Mz \geq 0$. The inequality holds because the upper left principal 3×3 matrix is positive semidefinite, and, apart from the $\mu > 0$ entry, the remainder of the matrix is antisymmetric. If $z \geq 0$ and $z^T Mz = 0$, then

$$z = \alpha \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \alpha, \beta \geq 0, \qquad \text{while} \quad b = \begin{bmatrix} (1+e)y_c'(t^{*-}) \\ +x_c'(t^{*-}) \\ -x_c'(t^{*-}) \\ 0 \end{bmatrix}.$$

From this, $\langle z, b \rangle = 0$, and by Theorem 6.1 in Section 6.1.9 there is a solution to the LCP. This solution provides a resolution of the paradox of Painlevé.

6.1.8 Approaches to the general problem of existence

There are two main approaches commonly used to establish existence of solutions to rigid-body dynamics with Coulomb friction:

- use a penalty approximation for the normal contact force and take the rigid limit (where stiffness goes to $+\infty$);
- use a time-stepping method which respects the no-interpenetration condition (or a linearization of it), and take the limit as the time step goes to zero.

Both involve a limiting process for which considerable analysis is needed. There is also some difference in how we set up the approximations that can have an important effect on the limiting solution; in particular, the coefficient of restitution of the solutions obtained by different processes can be quite different. It is more natural to set up penalty methods that conserve energy, and hence give coefficient of restitution $e = 1$ in the limit, while for implicit time-stepping methods it is much more natural to obtain a coefficient of restitution $e = 0$ in the limit. There are, however, ways of incorporating different coefficients of restitution into either method. See, for example, [200, 201, 202, 204] for penalty methods and [12, 13, 14] for time-stepping methods that incorporate coefficients of restitution between zero and one. Both approaches are amenable to numerical treatment, but the penalty approach becomes a two-stage method: first approximate the differential equations and then

solve the differential equations using some (usually standard) time-stepping method. However, as the penalty parameter approaches its limit, the time step used must be decreased accordingly to prevent numerical instability. This "tuning" of the time step, if not done carefully, can have disastrous effects on a simulation. On the other hand, using the complementarity framework for performing a time step usually leads to more difficult problems, especially when Coulomb friction is included. However, recent work seems to have made some progress on both of these problems [206, 256], and it is hard to see at this time which approach will become the dominant one, or if either will dominate the other.

Whatever approach is used, there are several steps we can take to obtain existence of solutions. The first is to obtain energy bounds, from which we obtain momentum bounds. This, in turn, can be used to bound integrals of the normal contact forces. Then we can obtain weak* convergence of the normal contact forces as measures, pointwise convergence of the velocities, and uniform convergence of the trajectories. The task is then to show that the limits indeed satisfy all the conditions for a solution.

6.1.9 Proving existence with Coulomb friction

We will set up an MDI formulation of the problem with Coulomb friction with inelastic impacts. The treatment of this section follows [238], which provides a complete proof of the existence of solutions to rigid-body dynamics that includes Coulomb friction (and Painlevé's problem), at least for one contact. Partially elastic impacts without friction are treated in [163, 201, 202, 205], while inelastic impacts with friction for particles is treated in [174]. For full details of the proof, see [238].

The objective

The objective of the proof is to show the existence of solutions to the system (understood in the DVI and MDI senses):

$$M(q)\frac{dv}{dt} = k(q,v) - \nabla V(q) + n(q)\,N(t) + D(q)\,\beta(t), \tag{6.22}$$

$$\frac{dq}{dt} = v, \tag{6.23}$$

$$0 \le \varphi(q(t)) \perp N(t) \ge 0, \tag{6.24}$$

$$\beta(t) \in N(t)\,K \;\;\&\;\; 0 \le v(t)^T D(q)\left[\widetilde{\beta} - \beta(t)\right] \quad \text{for all } \widetilde{\beta} \in N(t)\,K. \tag{6.25}$$

We suppose that $\varphi\colon \mathbb{R}^n \to \mathbb{R}$ is a *scalar* function. Recall that

$$k_i(q,v) = -\frac{1}{2}\sum_{j,k}\left[\frac{\partial m_{ij}}{\partial q_k} + \frac{\partial m_{ik}}{\partial q_j} - \frac{\partial m_{jk}}{\partial q_i}\right] v_j\, v_k.$$

Here, $M(q)$ is the mass matrix, $V(q)$ is the potential energy function, $n(q) = \nabla\varphi(q)$ is the *inward* normal vector for the feasible set $\{q \mid \varphi(q) \ge 0\}$, $D(q)$ is the matrix of direction vectors of the friction set (for unit normal contact force), and K is the set generating the set of friction forces for unit normal contact force: $\{D(q)\beta \mid \beta \in K\}$. For the proof, K is taken to be the unit simplex $\left\{\beta \mid \beta \ge 0,\ e^T\beta = 1\right\}$ where e is the vector of ones of the appropriate size. The function $k(q,v)$ is the generalized pseudoforce due to a nonlinear

representation of the configuration of the system. The normal contact force is $n(q(t))\,N(t)$, and the friction force is $D(q(t))\,\beta(t)$. Both $\beta(t)$ and $N(t)$ can be measures, $v(t)$ is a function of bounded variation, $M(q)$ is symmetric positive definite for all q, and all functions are at least Lipschitz and are typically smooth.

For the impact law, we assume that the impact law is perfectly inelastic: if $\varphi(q(t)) = 0$,

$$0 \le n(q(t))^T v(t^+) \perp N(t) \ge 0. \tag{6.26}$$

Since $v(\cdot)$ can be discontinuous, we interpret the Coulomb friction law (6.25) as applying to the postimpact velocity. We have already seen that something like this is needed to resolve Painlevé's paradox. We thus require that

$$\beta(t) \in N(t)\,K \quad \& \quad 0 \le v(t^+)^T D(q(t)) \left[\widetilde{\beta} - \beta(t)\right] \text{ for all } \widetilde{\beta} \in N(t)\,K. \tag{6.27}$$

In all that follows, $h > 0$ is the step size for a discrete (or numerical) approximation.

Time stepping

Time stepping is basic to the method of proof. We use approximations $q^\ell \approx q(t_\ell)$, $v^\ell \approx v(t_\ell)$ with $t_\ell = t_0 + \ell h$. At each time step we solve a CP, which we list here for one contact, which represents the above system with K the unit simplex:

$$M(q^{\ell+1})(v^{\ell+1} - v^\ell) = n(q^\ell)N^{\ell+1} + D(q^\ell)\beta^{\ell+1} + h\left[k(q^\ell, v^\ell) - \nabla V(q^\ell)\right], \tag{6.28}$$

$$q^{\ell+1} - q^\ell = h v^{\ell+1}, \tag{6.29}$$

$$0 \le \beta^{\ell+1} \perp \lambda^{\ell+1} e + D(q^\ell)^T v^{\ell+1} \ge 0, \tag{6.30}$$

$$0 \le N^{\ell+1} \perp n(q^\ell)^T v^{\ell+1} \ge 0, \tag{6.31}$$

$$0 \le \lambda^{\ell+1} \perp \mu N^{\ell+1} - e^T \beta^{\ell+1} \ge 0, \tag{6.32}$$

provided $\varphi(q^l) + h\,n(q^l)^T v^l \le 0$. If $\varphi(q^l) + h\,n(q^l)^T v^l > 0$, then we assume that there is no contact, and so $N^{\ell+1} = 0$ and $\beta^{\ell+1} = 0$. The variable $\lambda^{\ell+1}$ is a Lagrange multiplier associated with the maximal dissipation principle for Coulomb friction.

That solutions exist for this problem is based on an LCP

$$0 \le \begin{bmatrix} N \\ \beta \\ \lambda \end{bmatrix} \perp \begin{bmatrix} n^T M^{-1} n & n^T M^{-1} D & 0 \\ D^T M^{-1} n & D^T M^{-1} D & e \\ \mu & -e^T & 0 \end{bmatrix} \begin{bmatrix} N \\ \beta \\ \lambda \end{bmatrix} + \begin{bmatrix} n^T b \\ D^T b \\ 0 \end{bmatrix} \ge 0 \tag{6.33}$$

with $b = h\,M^{-1}(k(q,v) - \nabla V(q))$. Solutions exist for this problem, and if we fix M, n and D can be computed by Lemke's algorithm. The proof of this is instructive, as the matrix above is copositive but not copositive plus. Note that cols(D) is the set of columns of D.

Theorem 6.1. *If $n \notin \text{span cols}(D)$, M is symmetric positive definite, and $\mu > 0$, then solutions exist for* (6.33).

This proof requires results from Section 2.2.1, which the reader may refer to in order to understand the proof.

Proof. From the reversibility lemma (Lemma 2.9) applied to Lemke's algorithm, Lemke's algorithm for the LCP

$$0 \le z \perp \widetilde{M} z + q \ge 0$$

can fail only if there is an unbounded ray

$$(z, w, s) = (z_0, w_0, s_0) + \alpha\,(z_\infty, w_\infty, s_\infty), \qquad \alpha \ge 0,$$

for the system

$$0 \le z \perp w = \widetilde{M} z + s\,d + q \ge 0$$

with $s_0 > 0$, $s_\infty = 0$, and $z_\infty \neq 0$. Here d is a vector with only positive entries used for starting Lemke's algorithm. The matrix

$$\widetilde{M} = \begin{bmatrix} n^T M^{-1} n & n^T M^{-1} D & 0 \\ D^T M^{-1} n & D^T M^{-1} D & e \\ \mu & -e^T & 0 \end{bmatrix}$$

is copositive ($z \ge 0$ implies $z^T \widetilde{M} z \ge 0$) since the upper left 2×2 block $[n, D]^T M^{-1} [n, D]$ is positive semidefinite, $\mu > 0$, and the remainder of the matrix is antisymmetric. However, if $z \ge 0$ and $z^T \widetilde{M} z = 0$, then with $z^T = \left[N, \beta^T, \lambda\right]$, we have

$$0 = (Nn + D\beta)^T M^{-1} (Nn + D\beta) + \mu N \lambda,$$

giving $Nn + D\beta = 0$ and $N\lambda = 0$. Since $n \notin \text{span cols}(D)$, we have $N = 0$ and $D\beta = 0$. However, it is still possible to have $\lambda > 0$, so $\widetilde{M}$ is not strictly copositive.

From the properties of the unbounded ray,

$$\begin{aligned} 0 = z_\infty^T w_\infty &= z_\infty^T \widetilde{M} z_\infty, \\ 0 = z_\infty^T w_0 &= z_\infty^T \left(\widetilde{M} z_0 + s_0 d + q\right), \\ 0 = z_0^T w_\infty &= z_0^T \widetilde{M} z_\infty. \end{aligned}$$

If we write $z_\infty^T = \left[N_\infty, \beta_\infty^T, \lambda_\infty\right]$, then from the previous calculations, $N_\infty = 0$, $D\beta_\infty = 0$. But

$$\begin{aligned} \left(\widetilde{M} + \widetilde{M}^T\right) z_\infty &= \begin{bmatrix} n^T M^{-1} (nN_\infty + D\beta_\infty) + \mu\lambda_\infty \\ D^T M^{-1} (nN_\infty + D\beta_\infty) \\ \mu N_\infty \end{bmatrix} \\ &= \begin{bmatrix} \mu\lambda_\infty \\ 0 \\ 0 \end{bmatrix} \ge 0. \end{aligned}$$

Thus $\widetilde{M}^T z_\infty \geq -\widetilde{M} z_\infty$, and since $z_0 \geq 0$, $z_\infty^T \widetilde{M} z_0 = z_0^T \widetilde{M}^T z_\infty \geq -z_0^T \widetilde{M} z_\infty = 0$, we have

$$\begin{aligned} 0 &= z_\infty^T \left(\widetilde{M} z_0 + s_0 d + q\right) \\ &\geq s_0 z_\infty^T d + z_\infty^T q. \end{aligned}$$

But $z_\infty^T q = [N_\infty, \beta_\infty^T, \lambda_\infty]\left[n^T b, \left(D^T b\right)^T, 0\right]^T = (N_\infty n + D\beta_\infty)^T b = 0$; thus $0 \geq s_0 z_\infty^T d$. As $s_0 > 0$, we have $z_\infty^T d = 0$. But d is a vector with strictly positive entries and $z_\infty \geq 0$, so $z_\infty = 0$. Thus $w_\infty = 0$. Combined with $s_\infty = 0$, we see that $(z_\infty, w_\infty, s_\infty) = 0$, and so we do not really have an unbounded ray. Thus Lemke's method does not fail, but rather succeeds in finding a solution of the complementarity problem (6.33), and so a solution exists. □

It is possible to use the results for this LCP to solve the system (6.28)–(6.32) with $M = M(q^{\ell+1})$, as there is some nonlinear feedback from the solution of the LCP and $M(q^{\ell+1})$ in (6.28). Details can be found in [238].

Bounds on the discrete-time solutions

The first and most important bound is the energy bound. If we write $M(q^{\ell+1}) = M^{\ell+1}$, and $k^\ell = k(q^\ell, v^\ell) - \nabla V(q^\ell)$, then a discrete energy bound is proved first:

$$\frac{1}{2}(v^{\ell+1})^T M^{\ell+1} v^{\ell+1} + (k^\ell)^T q^{\ell+1} \leq \frac{1}{2}(v^\ell)^T M^{\ell+1} v^\ell + (k^\ell)^T q^\ell.$$

To show this, start with

$$\begin{aligned} (v^{\ell+1})^T M^{\ell+1}(v^{\ell+1} - v^\ell) = \frac{1}{2}\Big[(v^{\ell+1})^T M^{\ell+1} v^{\ell+1} - v^{\ell\,T} M^{\ell+1} v^\ell\Big] \\ + \frac{1}{2}(v^{\ell+1} - v^\ell)^T M^{\ell+1}(v^{\ell+1} - v^\ell) \end{aligned}$$

and then substitute the right-hand side of (6.28) for $M^{\ell+1}(v^{\ell+1} - v^\ell)$. With this, bounds on the kinetic energy $KE^\ell = \frac{1}{2} v^{\ell\,T} M^\ell v^\ell$ can be found of the form

$$KE^{\ell+1} \leq KE^\ell + h\left[a + b\left(KE^\ell\right)^{1/2} + c\left(KE^\ell\right)^{3/2}\right]$$

for positive constants a, b, and c which depend only on the problem data. A discrete nonlinear Gronwall lemma like Lemma 5.2 can then be applied to show *short-time* bounds, independent of $h > 0$, on the velocities $\left\|v^\ell\right\| \leq B_v$ for $0 \leq \ell h \leq T^*$ for some $T^* > 0$. We can define numerical trajectories: $q_h(\cdot)$ is the piecewise linear interpolant of $q_h(t_\ell) = q^\ell$, and $v_h(\cdot)$ is the piecewise constant interpolant $v_h(t) = v^{\ell+1}$ for $t_\ell < t \leq t_{\ell+1}$. Then these functions are uniformly bounded as $h \downarrow 0$ and $q_h(\cdot)$ are uniformly Lipschitz on $\left[0, T^*\right]$.

The next step is to show that the variation of $v_h(\cdot)$ is uniformly bounded on $\left[0, T^*\right]$. This is equivalent to showing that $\int_0^{T^*} N_h(t)\,dt$ is uniformly bounded, where $N_h(\cdot)$ is the piecewise constant interpolant $N_h(t) = N^{\ell+1}/h$ for $t_\ell < t \leq t_{\ell+1}$. To prove this we need an additional condition: the cone

$$\widetilde{\mathcal{F}}(q) = \{n(q)\,N + D(q)\,\beta \mid \beta \in N\,K\}$$

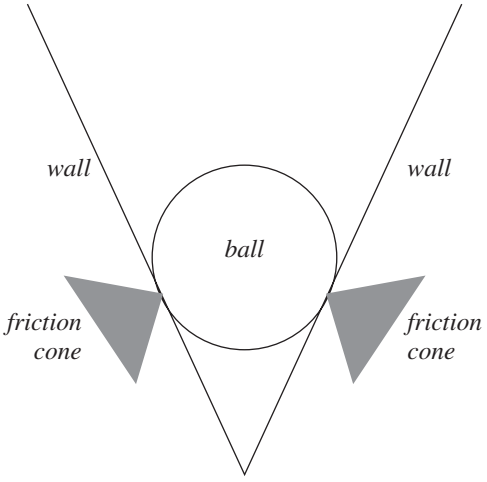


Figure 6.4: Example of jamming. The friction coefficient is large enough that a horizontal contact force from the left can be matched by an opposing force on the right without causing the ball to move or spin. Reprinted with permission.

must be a *pointed* cone. That is, we require that $\widetilde{\mathcal{F}}(q)$ not contain a vector subspace other than $\{0\}$. Without this it is possible to have *jamming* where there is no bound on the contact forces. An example of jamming is shown in Figure 6.4.

Since $\widetilde{\mathcal{F}}(q)$ is a pointed cone in a finite-dimensional space, there is a vector ζ where for all $w \in \widetilde{\mathcal{F}}(q)$, $\zeta^T w \geq \|w\|$. Now

$$\begin{aligned}
\zeta^T \left(v^{\ell+1} - v^0\right) &= \sum_{j=0}^{\ell} \zeta^T \left[h\,k^j + n(q^j)\,N^j + D(q^j)\,\beta^j\right] \\
&\geq \zeta^T h \sum_{j=0}^{\ell} k^j + \sum_{j=0}^{\ell} \left\| n(q^j)\,N^j + D(q^j)\,\beta^j \right\| \\
&\geq \|\zeta\|\,(t_{\ell+1} - t_0) \max_{0\leq j\leq \ell} \left\|k^j\right\| + \eta \sum_{j=0}^{\ell} N^j,
\end{aligned}$$

where $\eta > 0$ is the distance between $n(q)$ and $\{\, D(q)\beta \mid \beta \in K \,\}$. Since $v^{\ell+1}$ is bounded from the kinetic energy bounds, and k^j are uniformly bounded for $0 \leq j\,h \leq T^*$, we see that $\sum_{j=0}^{\ell} N^j$ are also uniformly bounded for $\ell\,h \leq T^*$. Thus $\int_0^{T^*} N_h(t)\,dt$ are uniformly bounded, and $\sum_{j=0}^{\ell} \left\|v^{j+1} - v^j\right\| = \bigvee_0^{T^*} v_h(\cdot)$ is uniformly bounded as $h \downarrow 0$. Also, if we take $\beta_h(t) = \beta^\ell$ for $t_\ell < t \leq t_{\ell+1}$, then we also have $\int_0^{T^*} \beta_h(t)\,dt$ uniformly bounded as $\left\|\beta^\ell\right\| \leq N^\ell \max_{w\in K} \|w\|$.

These results hold only on a "sufficiently short" time interval $\left[0, T^*\right]$. To extend this further, we need stronger bounds to prevent "blowup" of the velocities in finite time. This involves using energy bounds refined using uniform bounded variation of $v_h(\cdot)$. The

bounds obtained have the form

$$v^{\ell\,T} M^\ell v^\ell + V(q^\ell) \le v^{0\,T} M^0 v^0 + V(q^0) + \mathcal{O}(h)\left[1 + \sum_{j=0}^{\ell-1} \left\| v^{j+1} - v^j \right\|\right].$$

Mild conditions on $V(q)$ then ensure that the numerical (and continuous) problems have solutions whose bounds do not go to infinity in finite time. Bootstrapping these results allows us to obtain uniform bounds on both $v_h(\cdot)$ and its variation over any finite interval $[0,T]$.

Obtaining limits

Thus on the interval $\left[0,T^*\right]$ we can use the Arzela–Ascoli theorem, Alaoglu's theorem, and Helly's selection theorem to prove the existence of a subsequence of $h \downarrow 0$ in which

$$\begin{aligned}
q_h(\cdot) &\to q(\cdot) && \text{uniformly,}\\
N_h(\cdot) &\rightharpoonup^* N(\cdot) && \text{weak* as measures,}\\
\beta_h(\cdot) &\rightharpoonup^* \beta(\cdot) && \text{weak* as measures,}\\
v_h(\cdot) &\to v(\cdot) && \text{pointwise almost everywhere,}
\end{aligned}$$

with $q(\cdot)$ continuous, $N(\cdot)$ and $\beta(\cdot)$ measures, and $v(\cdot)$ having bounded variation.

From the theory of MDIs we have

$$\begin{aligned}
M(q)\frac{dv}{dt} &\in k(q,v) - \nabla V(q) + \widetilde{\mathcal{F}}(q(t)),\\
\frac{dq}{dt} &= v,
\end{aligned}$$

with the first inclusion understood in the sense of MDIs, by Theorem 4.9.

It should also be noted that $\beta(t) \in N(t)\,K$ in the sense that the Radon–Nikodym derivative $d\beta/dN(t) \in K$ for N-almost all t.

Inelastic impacts

To show that we have inelastic impacts in the single-contact case, at least in the one-contact case, we first show that

$$(n^{\ell+1})^T v^{\ell+1} \le \max\left(0, (n^\ell)^T v^\ell\right) + K\,h$$

for some number K independent of h. So over a number of steps we have

$$n(q_h(t+\epsilon))^T v_h(t+\epsilon) \le \max\left(0, n(q_h(t-\epsilon))^T v_h(t-\epsilon)\right) + K\,(2\epsilon + h).$$

Taking limits as $h \downarrow 0$ gives

$$n(q(t+\epsilon))^T v(t+\epsilon) \le \max\left(0, n(q(t-\epsilon))^T v(t-\epsilon)\right) + 2\,K\,\epsilon.$$

Finally, taking $\epsilon \downarrow 0$,

$$n(q(t))^T v(t^+) \leq \max\left(0, n(q(t))^T v(t^-)\right).$$

Then, if $\varphi(q(t)) = 0$, we must have $n(q(t))^T v(t^+) \geq 0$ and $n(q(t))^T v(t^-) \leq 0$. Thus

$$n(q(t))^T v(t^+) = 0,$$

as desired.

Coulomb friction in the limit

Showing that the Coulomb friction law holds in the limit requires a number of steps. The essential part of this is to accurately estimate the changes in energy over an arbitrarily small time interval. The preparation for this involves showing that

$$\begin{aligned} n(q^{\ell+1})^T v^{\ell+1} - \mu \left\| D(q^{\ell+1})^T v^{\ell+1} \right\|_\infty \\ \geq n(q^\ell)^T v^\ell - \mu \left\| D(q^\ell)^T v^\ell \right\|_\infty + \mathcal{O}(h). \end{aligned}$$

This can be done using the complementarity formulation of the time stepping. Taking limits of the difference over many steps, we then have

$$\begin{aligned} n(q(t+\epsilon))^T v(t+\epsilon) - \mu \left\| D(q(t+\epsilon))^T v(t+\epsilon) \right\| \\ \geq n(q(t))^T v(t) - \mu \left\| D(q(t))^T v(t) \right\| + \mathcal{O}(\epsilon). \end{aligned}$$

We then apply the following lemma.

Lemma 6.2. *Suppose that $\mu_n \rightharpoonup^* \mu$ weak* as measures, $\mu_n \geq 0$, and $\theta_n \to \theta$ pointwise. Suppose also that θ_n are uniformly bounded, and for all $\epsilon > 0$ sufficiently small $\theta_n(t+\epsilon) \geq \theta_n(t) - K\epsilon$ for all t (K independent of t, n, and $\epsilon > 0$). Then, if $\theta_n \mu_n \rightharpoonup^* \nu$ as measures, $\nu \geq \mu\, \theta^+$.*

With this, if $N_h \rightharpoonup^* N$ and $N_h \left[n(q_h)^T v_h - \left\| D(q_h)^T v_h \right\|_\infty \right] \rightharpoonup^* \nu$ in a suitable subsequence, then $\nu \geq N \left[n(q)^T v^+ - \left\| D(q) v^+ \right\|_\infty \right]$. This can then be applied to obtain an energy balance.

The energy function is

$$\begin{aligned} E(t) &= E(q(t), v(t)), \qquad \text{where} \\ E(q, v) &= \frac{1}{2} v^T M(q) v + V(q). \end{aligned}$$

The function $t \mapsto E(t)$ is a function of bounded variation because $q(\cdot)$ is Lipschitz and $v(\cdot)$ has bounded variation. The main problem in computing the differential measure dE is $v^T M(q) v$ because it involves a product of functions of bounded variations. There is, however, the product rule discovered by Moreau [177, 180] (Lemma 4.10): If u and v have

bounded variation and $\psi(t) = \langle u(t), v(t)\rangle$, then $d\psi = \langle du, v^+\rangle + \langle u^-, dv\rangle = \langle du, v^-\rangle + \langle u^+, dv\rangle$. With this rule,

$$\begin{aligned} d\left(v^T M(q)v\right) &= dv^T M(q)\,v^- + \left(v^+\right)^T d\left(M(q)\,v\right) \\ &= dv^T M(q)\,v^- + \left(v^+\right)^T \left(d(M(q))\,v + M(q)\,dv\right) \\ &= \left(v^- + v^+\right)^T M(q)\,dv + v^T\, d(M(q))\,v. \end{aligned}$$

The measure differential $d(M(q))$ can be written out in terms of its components:

$$\begin{aligned} d(m_{ij}(q)) &= \sum_k \frac{\partial m_{ij}}{\partial q_k}(q(t))\,\frac{dq_k}{dt}(t)\,dt \\ &= \sum_k \frac{\partial m_{ij}}{\partial q_k}(q(t))\,v_k(t)\,dt, \end{aligned}$$

where "dt" is the Lebesgue measure. So

$$\begin{aligned} v^T\, d(M(q))\,v &= \sum_{i,j,k} \frac{\partial m_{ij}}{\partial q_k}(q(t))\,v_i(t)\,v_j(t)\,v_k(t)\,dt \\ &= -2\,v^T k(q,v)\,dt. \end{aligned}$$

Thus

$$\begin{aligned} dE &= \frac{1}{2}\left(v^- + v^+\right)^T M(q)\,dv + \frac{1}{2} v^T\, d(M(q))\,v + v^T \nabla V(q)\,dt \\ &= \left(v^+\right)^T M(q)dv + \left[v^T \nabla V(q) - 2v^T k(q,v)\right] dt - \frac{1}{2}\left(v^+ - v^-\right)^T M(q)\,dv \\ &= \left(v^+\right)^T \left[n(q)\,N + D(q)\beta + k(q,v)dt - \nabla V(q)\,dt\right] \\ &\qquad + \left[v^T \nabla V(q) - v^T k(q,v)\right] dt - \frac{1}{2}\left(v^+ - v^-\right)^T M(q)\,dv \\ &= \left(v^+\right)^T \left[n(q)\,N + D(q)\beta\right] - \frac{1}{2}\left(v^+ - v^-\right)^T M(q)\,dv. \end{aligned}$$

Since $n(q(t))^T v^+(t) = 0$ whenever $\varphi(q(t)) = 0$, it follows that $\left(v^+\right)^T n(q)\,N = 0$.

If $E_h(t) = E(q_h(t), v_h(t))$, then from the discrete formulation,

$$\begin{aligned} &E_h(t_2') - E_h(t_1') \\ &\quad = -\frac{1}{2} \sum_{\ell\, h \in (t_1', t_2')} (v^{\ell+1} - v^\ell)^T M(q^{\ell+1})(v^{\ell+1} - v^\ell) \\ &\qquad + \sum_{\ell\, h \in (t_1', t_2')} N^{\ell+1} [n(q^{\ell+1})^T v^{\ell+1} - \mu \| D(q^{\ell+1})^T v^{\ell+1} \|_\infty] + \mathcal{O}(h) \\ &\quad = -\int_{t_1'}^{t_2'} N_h\, \mu \| D(q_h)^T v_h \|_\infty\, dt - \frac{1}{2} \int_{t_1'}^{t_2'} \left(v_h^+ - v_h^-\right)^T M(q_h)\,dv_h + \mathcal{O}(h). \end{aligned}$$

The functional $u \mapsto \int_{[t_1',t_2']} \left(u^+ - u^-\right)^T M(q)\,du$ is a nonnegative quadratic (and therefore convex) functional. The difficulty we have to face is that although $v_h \rightharpoonup^* v$, we cannot conclude that

$$\int_{[t_1',t_2']} \left(v^+ - v^-\right)^T M(q)\,dv \le \liminf_{h\downarrow 0} \int_{[t_1',t_2']} \left(v_h^+ - v_h^-\right)^T M(q)\,dv_h,$$

as we might expect from Mazur's lemma. The reason is that we have only weak* convergence, rather than weak convergence, and Mazur's lemma applies to weakly convergent sequences.

Let us suppose that (restricting to a further subsequence if necessary)

$$\begin{aligned} \left(v_h^+ - v_h^-\right)^T M(q_h)\,dv_h &\rightharpoonup^* \sigma, \\ N_h\left[n(q_h)^T v_h - \mu \| D(q_h)^T v_h \|_\infty\right] &\rightharpoonup^* \nu. \end{aligned}$$

We first note that $\sigma \ge 0$. We already know that $\nu \le N\left[n(q)^T v^+ - \mu \| D(q)^T v^+ \|_\infty\right]$. Thus

$$\begin{aligned} dE_h &\rightharpoonup^* dE \qquad \text{(since } E_h \to E \text{ pointwise)} \\ &= -\frac{1}{2}\sigma + \nu. \end{aligned}$$

Thus $dE \le N\left[n(q)^T v^+ - \mu \left\| D(q)^T v^+ \right\|_\infty\right] = -\mu N \left\| D(q)^T v^+ \right\|_\infty$. Recall that

$$\begin{aligned} dE &= \left(v^+\right)^T \left[n(q)\,N + D(q)\beta\right] - \frac{1}{2}\left(v^+ - v^-\right)^T M(q)\,dv \\ &= \left(v^+\right)^T D(q)\beta - \frac{1}{2}\left(v^+ - v^-\right)^T M(q)\,dv \\ &\le -\mu N \left\| D(q)^T v^+ \right\|_\infty. \end{aligned}$$

At any t where v is continuous, $\left(v^+ - v^-\right)^T M(q)\,dv = 0$, and since $e^T\beta \le \mu N$ (interpreted in the sense of measures if necessary), we have

$$\left(v^+\right)^T D(q)\beta = -\mu N \left\| D(q)^T v^+ \right\|_\infty,$$

which is the maximum dissipation principle for Coulomb friction. Note that we do not need v to be *absolutely continuous* in order to obtain this result; continuity of v is sufficient.

Another condition under which we can prove the maximum dissipation principle is under *Erdmann's condition* [91]:

$$0 < n(q)^T M(q)^{-1}\left[N\,n(q) + D(q)\,\beta\right] \tag{6.34}$$

whenever $e^T\beta \le \mu N$, $N \ne 0$. Geometrically, this means that accelerations due to the contact forces must be in the admissible region. Erdmann noted that under this condition, Painlevé's paradox cannot occur. For particles (as discussed by Monteiro Marques [174]), $M(q) = m\,I$ and the columns of $D(q)$ are orthogonal to $n(q)$ so that (6.34) holds.

To prove that the maximal dissipation principle holds in this case, it is important that if there is a jump in the velocity, then this jump occurs in one time step of the discretization rather than being spread out over several time steps. To explain in more detail, let

$$0 < \gamma(q) := \min \left\{ n(q)^T M(q)^{-1} \left[n(q) + D(q)\beta \right] \mid e^T \beta \le \mu \right\}$$

and let $\gamma^* > 0$ be a lower bound of $\gamma(q)$ for q in a given bounded region containing the numerical and limiting trajectories. Then

$$n(q(t))^T \left[v(t+\epsilon) - v(t^+) \right] \ge \frac{\gamma^*}{2} \int_{(t,t+\epsilon)} N + \mathcal{O}(\epsilon). \tag{6.35}$$

But if $q(t)$ is on the boundary of the admissible region, then $n(q(t))^T v(t^+) = 0$; indeed, $n(q(t))^T v(t+\epsilon) = \mathcal{O}(\epsilon)$, so $\int_{(t,t+\epsilon)} N = \mathcal{O}(\epsilon)$ and N is bounded over sufficiently small intervals $(t, t+\epsilon)$. Thus v is Lipschitz over such intervals, and so we can conclude that the maximum dissipation principle holds over such intervals.

The problem remaining is to treat the case where $q(t)$ is on the boundary of the admissible region, but $n(q(t))^T v(t^-) < 0$. In the time-stepping method, if ${n^\ell}^T v^\ell < 0$ but $N^\ell > 0$, then ${n^\ell}^T v^{\ell+1} = 0$. The bound (6.35) can be applied to the time-stepping method to show that $N^{\ell+k}$ is bounded for $0 \le k\,h < \epsilon$, and so $\left\| v^{\ell+k} - v^\ell \right\| = \mathcal{O}(k\,h) = \mathcal{O}(\epsilon)$. Taking the time step $h \downarrow 0$, the limiting velocity jump occurs essentially in one time step. Since the maximum dissipation principle is built into the time-stepping method, the maximum dissipation principle holds in the limit at the velocity jump.

The final task is to show that even if Erdmann's condition fails but the friction force is one dimensional (which is the case in the Painlevé paradox), the maximum dissipation principle still holds. Again, the only point of difficulty is at velocity jumps. If there is a velocity jump at time t and $D(q(t))^T v(t^+) = 0$, then any atom $\beta(\{t\})$ satisfying the constraint $e^T \beta(\{t\}) \le \mu\, N(\{t\})$ satisfies the maximum dissipation principle. So we consider the case $D(q(t))^T v(t^+) \ne 0$. In the one-dimensional friction case, $D(q) = \left[+d_1(q), -d_1(q) \right]$. The main task is then to show that in a sufficiently small interval $(t-\epsilon, t+\epsilon)$ the slip velocity $d_1(q^\ell)^T v^\ell$ does not change sign. If this were to happen for arbitrarily small $h > 0$, then in the limit we must have

$$0 = n(q(t))^T M(q(t))^{-1} \left[n(q(t)) \pm \mu d_1(q(t)) \right],$$

which would imply $n(q(t))^T M(q(t))^{-1} d_1(q(t)) = 0$, and Erdmann's condition would hold; a contradiction. Thus solutions exist even for the Painlevé paradox.

Thus Painlevé's paradox does not lead to a contradiction in rigid-body dynamics. Rather it can result in impulsive or unbounded forces without a collision.

6.1.10 Limits of rigid-body models

Clearly all real materials have some elasticity; no material is perfectly rigid, just as no material is perfectly elastic or viscoelastic, or even a continuum. But the problem with rigid-body dynamics with impact is deeper than this. The fundamental problem is a lack of uniqueness of solutions of index two, which is why the coefficient of restitution is introduced (see Section 6.1.4).

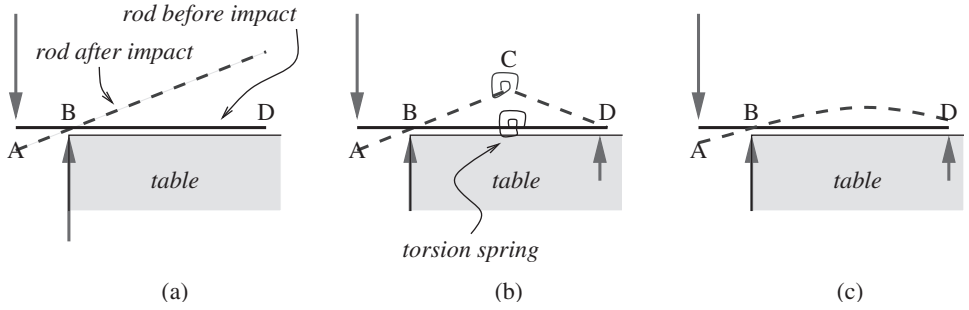


Figure 6.5: Chatterjee's example showing that rigid-body dynamics is not the limit as stiffness goes to infinity; (a) rigid-body model; (b) torsion spring model; (c) fully elastic model.

It is natural to try to justify a model, or to truly understand it, by trying to derive it as a limit of a more sophisticated but difficult to analyze model. The natural path here is to incorporate elasticity or some approximation to it, and to see if we can recover the rigid-body model by taking the stiffness of the elastic elements to infinity. This could at least be used to determine an intellectually justifiable basis for models of restitution. However, the limit as stiffness goes to infinity of more sophisticated models with elastic elements is *not* rigid-body dynamics under any model of restitution. This can be seen in the example of Chatterjee [50]. Although Chatterjee claims to be critiquing the use of complementarity conditions for modeling restitution, the example in fact undermines any kind of algebraic restitution law.

The example in [50] is essentially a rod that is lying on a table with one end extended beyond the edge of the table. An impulse is applied to the end of the rod not supported by the table. This is illustrated in Figure 6.5.

To approximate the elasticity of the rod, consider a torsion spring located at C. However, the hinge at C means that the section of the rod CD rotates clockwise. If BC were less than CD, then the point D would rotate into the table without an impulse at D, no matter how stiff the torsion spring at C is. Furthermore, the strength of the impulse at D is independent of the stiffness of the torsion spring. Taking the stiffness of the torsion spring to infinity still gives an impulse at D, contradicting the behavior expected from a rigid-body model. Although only the torsion spring model is analyzed in [50], physical experiments indicate that the same behavior is seen in real rods which are fully elastic.

If rigid-body models do not accurately reflect physical reality, we should consider the next step in sophistication for our models: elastic and viscoelastic bodies. Because these result in partial differential equations, we must work in infinite-dimensional spaces and deal with some of their additional technical difficulties.

6.2 Elastic bodies in impact

Elastic bodies are governed by partial differential equations rather than ordinary differential equations; typically the main variable in these equations is the displacement field $\mathbf{u}(t,\mathbf{x}) \in \mathbb{R}^d$ where the point $\mathbf{x}$ in the undeformed body $\Omega \subset \mathbb{R}^d$ is deformed to $\psi(t,\mathbf{x}) = \mathbf{x} + \mathbf{u}(t,\mathbf{x})$. From this displacement field we construct a *strain tensor* $\varepsilon(t,\mathbf{x})$ which measures the local

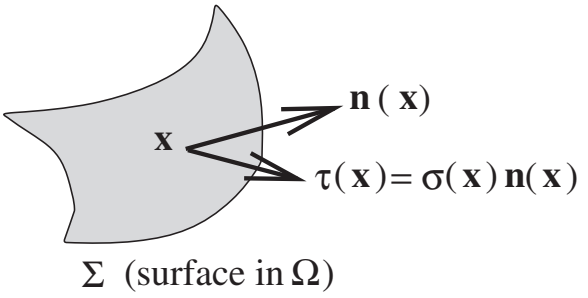


Figure 6.6: Relationship between stress and traction: traction $\tau(\mathbf{x})$ is the force per unit area acting on surface Σ at $\mathbf{x} \in \Sigma$.

deformation, which in turn is used to determine the local *stress tensor* $\sigma(t,\mathbf{x})$. The stress tensor indicates the forces per unit area acting on small pieces of surface inside the material, as illustrated in Figure 6.6. Note that the force acting on a given surface $\Sigma \subset \overline{\Omega}$ of dimension $d-1$ is given by the integral

$$\int_\Sigma \tau(\mathbf{x})\, dS(\mathbf{x}) = \int_\Sigma \sigma(\mathbf{x})\mathbf{n}(\mathbf{x})\, dS(\mathbf{x}),$$

where $\tau(\mathbf{x})$ is the traction at $\mathbf{x}$ and $\sigma(\mathbf{x})$ is the rank-two stress tensor at $\mathbf{x}$. We can consider the stress tensor to be a $d \times d$ matrix, and $\sigma(\mathbf{x})\mathbf{n}(\mathbf{x})$ to be ordinary matrix-vector multiplication. Note that this force is acting on the body in the $-\mathbf{n}(\mathbf{x})$ direction of the surface Σ.

Details of how to formulate contact conditions can be found in [121, 229], for example.

There are several important properties of both the strain and stress tensors: they are both rank-two tensors: $\varepsilon(t,\mathbf{x}) = \left[\varepsilon_{ij}(t,\mathbf{x})\right]_{i,j=1}^{d}$ and $\sigma(t,\mathbf{x}) = \left[\sigma_{ij}(t,\mathbf{x})\right]_{i,j=1}^{d}$, and they are both symmetric as rank-two tensors: $\varepsilon_{ij}(t,\mathbf{x}) = \varepsilon_{ji}(t,\mathbf{x})$ and $\sigma_{ij}(t,\mathbf{x}) = \sigma_{ji}(t,\mathbf{x})$. The relationship between stress and strain tensors at a given point in a material is called the *constitutive relation* and defines the nature of the material. Usually there is a simple functional relationship, but sometimes other relationships are used to model memory effects, viscous behavior, and plastic behavior. These issues are part of *continuum mechanics*; more on this area can be found in textbooks such as [51, 112, 233, 261].

In this book we will focus on linearized elasticity and viscoelasticity where the following relationships are assumed to hold:

$$\varepsilon_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right), \tag{6.36}$$

$$\sigma_{ij} = \sum_{k,l=1}^{d} A_{ijkl}\,\varepsilon_{ij} \qquad \text{for pure elasticity,} \tag{6.37}$$

$$\sigma_{ij} = \sum_{k,l=1}^{d} A_{ijkl}\,\varepsilon_{ij} + \sum_{k,l=1}^{d} B_{ijkl}\,\frac{\partial \varepsilon_{ij}}{\partial t} \tag{6.38}$$

for Kelvin–Voigt viscoelasticity.

Note that (6.36) is actually the infinitesimal strain tensor. For large displacement problems, there is the possibility of geometric nonlinearities, and this must be replaced with a nonlinear function of $\nabla \mathbf{u}$, such as the Cauchy strain tensor [261, p. 42]. The rank-four tensors A_{ijkl} and B_{ijkl} are assumed to have a number of important properties: there is $\eta_A > 0$, where

$$A_{ijkl} = A_{jikl} = A_{ijlk} = A_{klij},$$
$$\sum_{i,j,k,l=1}^{d} A_{ijkl}\xi_{ij}\xi_{kl} \geq \eta_A \sum_{i,j=1}^{d} \xi_{ij}^2$$

for all symmetric rank-two tensors. This assumes a linear relationship between the stress and strain tensors. In general, there can be material nonlinearities, and these relationships need to be replaced by a nonlinear function $\sigma = \sigma(\varepsilon)$.

An important special case is where the elastic (or viscoelastic) materials are assumed to be *isotropic*. This means that the constitutive equations are invariant under rotations. That is, if Q is an orthogonal matrix of determinant one, then under the transformation $\xi'_{kl} = \sum_{i,j} q_{ki} q_{lj} \xi_{ij}$ and $\eta'_{kl} = \sum_{i,j} q_{ki} q_{lj} \eta_{ij}$ we have

$$\sum_{i,j,k,l} A_{ijkl}\eta'_{ij}\xi'_{kl} = \sum_{i,j,k,l} A_{ijkl}\eta_{ij}\xi_{kl}.$$

This greatly reduces the number of free parameters down to two:

$$A_{ijkl} = \lambda \delta_{ij}\delta_{kl} + \mu\left(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\right),$$

where δ_{ij} is the Kronecker δ function:

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

The parameters λ and μ are known as Lamé parameters. The formula for the stress tensor can be simplified to

$$\sigma_{ij} = \lambda \sum_k \varepsilon_{kk}\,\delta_{ij} + 2\mu\,\varepsilon_{ij}$$

in the linear isotropic elastic case.

The equations of motion for elasticity for an elastic or viscoelastic body can be written in the form

$$\rho(\mathbf{x})\frac{\partial^2 \mathbf{u}}{\partial t^2} = \operatorname{div}\sigma(t,\mathbf{x}) + \mathbf{f}(t,\mathbf{x}) \qquad \text{inside } \Omega, \tag{6.39}$$

where $\rho(\mathbf{x})$ is the mass density, and $\mathbf{f}(t,\mathbf{x})$ is the density of the other forces acting throughout the body (with respect to volume), such as gravity or electromagnetism. Note that $\operatorname{div}\sigma$ is a vector field given by the formula

$$(\operatorname{div}\sigma)_i = \sum_{j=1}^{d} \frac{\partial \sigma_{ij}}{\partial x_j}. \tag{6.40}$$

In the case of isotropic linear elasticity this can be reduced to

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = (\lambda + \mu) \nabla (\nabla \cdot \mathbf{u}) + \mu \nabla^2 \mathbf{u} + \mathbf{f}(t, \mathbf{x}). \tag{6.41}$$

As with other partial differential equations, we also need boundary conditions to determine the behavior of the body. Usually these are one of two types:

$$\begin{aligned} \mathbf{u}(t, \mathbf{x}) &= \mathbf{g}(t, \mathbf{x}), && \text{given deformation, or} \\ \sigma(t, \mathbf{x}) \mathbf{n}(x) &= \tau(t, \mathbf{x}), && \text{given traction} \end{aligned}$$

boundary conditions. For the traction boundary conditions, $\tau(t, \mathbf{x})$ is the density (with respect to surface area) of the forces acting on the boundary of the body; $\mathbf{n}(\mathbf{x})$ is the outward unit normal vector on the boundary of the body Ω.

6.2.1 Formulating the contact conditions

We will denote the region of the boundary with given displacement by $\Gamma_D \subseteq \partial\Omega$, and the region with given traction by $\Gamma_N \subseteq \partial\Omega$. We cannot deal with both conditions applied at the same point, so $\Gamma_D \cap \Gamma_N = \emptyset$. What remains is the region of potential contact

$$\Gamma_C = \partial\Omega \backslash (\Gamma_D \cup \Gamma_N).$$

For linearized elastic bodies in contact with a rigid obstacle, we have the following linearized contact conditions:

$$\sigma(t, \mathbf{x}) \mathbf{n}(\mathbf{x}) = -N(t, \mathbf{x}) \mathbf{n}(\mathbf{x}) + \mathbf{F}(t, \mathbf{x}),$$

where $N(t, \mathbf{x})$ is the normal contact force at $\mathbf{x}$ at time t, which must be inward to the body, and $\mathbf{F}(t, \mathbf{x}) \perp \mathbf{n}(\mathbf{x})$ is the frictional force at $\mathbf{x}$ at time t. The fact that $N(t, \mathbf{x})$ must be inward to the body is the reason for the negative sign, as $\mathbf{n}(\mathbf{x})$ is the *outward* normal direction vector at $\mathbf{x} \in \partial\Omega$.

All that remains is to give the relationships between the displacement field $\mathbf{u}(t, \mathbf{x})$ on the boundary and the contact forces. The usual Signorini conditions (see Section 2.6) for the normal contact force $N(t, \mathbf{x})$ are

$$\begin{aligned} 0 \leq N(t, \mathbf{x}) \perp \mathbf{n}(\mathbf{x}) \cdot \mathbf{u}(t, \mathbf{x}) - \varphi(\mathbf{x}) \geq 0 \\ \text{for all } \mathbf{x} \in \Gamma_C, \quad t \geq 0. \end{aligned} \tag{6.42}$$

Here $\varphi(\mathbf{x})$ is the *gap function* which measures the distance between the undeformed object at $\mathbf{x}$ and the rigid obstacle. This is illustrated in Figure 6.7.

The standard Coulomb friction law can be written in several different forms, as is the case for rigid-body dynamics. One of the most direct formulations is

$$\sigma(t, \mathbf{x}) \mathbf{n}(\mathbf{x}) = -N(t, \mathbf{x}) \mathbf{n}(\mathbf{x}) + \mathbf{F}(t, \mathbf{x}) \qquad \text{on } \Gamma_C, \tag{6.43}$$

$$\mathbf{F}(t, \mathbf{x}) \perp \mathbf{n}(\mathbf{x}) \qquad \text{on } \Gamma_C, \tag{6.44}$$

$$\frac{\partial \mathbf{u}}{\partial t}(t, \mathbf{x}) \cdot \mathbf{F}(t, \mathbf{x}) = -\mu N(t, \mathbf{x}) \left\| \frac{\partial \mathbf{u}}{\partial t}(t, \mathbf{x}) \right\| \qquad \text{on } \Gamma_C, \tag{6.45}$$

$$\|\mathbf{F}(t, \mathbf{x})\| \leq +\mu N(t, \mathbf{x}). \tag{6.46}$$

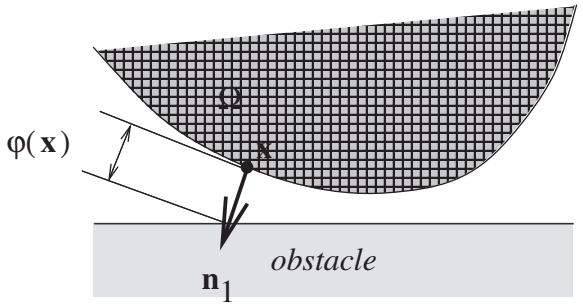


Figure 6.7: Illustration of an elastic body in contact with a rigid obstacle.

Alternatively, we can use a VI formulation:

$$\sigma(t,\mathbf{x})\mathbf{n}(\mathbf{x}) = -N(t,\mathbf{x})\mathbf{n}(\mathbf{x}) + \mathbf{F}(t,\mathbf{x}) \qquad \text{on } \Gamma_C, \tag{6.47}$$

$$\mathbf{F}(t,\mathbf{x}) \perp \mathbf{n}(\mathbf{x}) \qquad \text{on } \Gamma_C, \tag{6.48}$$

$$\mu\, N(t,\mathbf{x})\, \|\mathbf{w}(t,\mathbf{x})\| \geq \mu\, N(t,\mathbf{x}) \left\| \frac{\partial \mathbf{u}}{\partial t}(t,\mathbf{x}) \right\| + \mathbf{F}(t,\mathbf{x}) \cdot \left(\frac{\partial \mathbf{u}}{\partial t}(t,\mathbf{x}) - \mathbf{w}(t,\mathbf{x}) \right) \qquad \text{on } \Gamma_C \tag{6.49}$$

for any sufficiently smooth field $\mathbf{w}(t,\mathbf{x})$.

Note that most of these conditions are given in terms of integrals over Γ_C in order to put them in the weakest form possible. This is commonly done by means of a VI of the second kind (see Section 2.3.1):

$$\int_{\Gamma_C} \mathbf{F} \cdot \mathbf{w}\, dS \leq \int_{\Gamma_C} \mu\, N \left(\left| \frac{\partial \mathbf{u}_T}{\partial t} + \mathbf{w}_T \right| - \left| \frac{\partial \mathbf{u}_T}{\partial t} \right| \right) dS \tag{6.50}$$

for all $\mathbf{w}$, where $\mathbf{z}_T = \mathbf{z} - (\mathbf{z}\cdot\mathbf{n})\mathbf{n}$ is the tangential component of $\mathbf{z}$, $|\mathbf{v}|$ is the Euclidean or 2-norm of the vector $\mathbf{v} \in \mathbb{R}^d$, and $\mathbf{n}$ is the outward unit normal vector to Γ_C. The reformulation (6.91) replaces (6.48)–(6.49). Integration over time with smooth $\mathbf{w}\colon [0,T] \to H^1(\Omega)$ makes for an even weaker formulation.

6.2.2 Formulating contact between two bodies

This formulation can be extended between two elastic or viscoelastic bodies that undergo small deformations. Then we take $\Omega = \Omega_1 \cup \Omega_2$, where Ω_1 and Ω_2 are disjoint regions in $\mathbb{R}^d$ (see Figure 6.8). To formulate the contact conditions for N, for a given point $\mathbf{x}_1 \in \Gamma_{C,1}$, the potential contact region of $\partial\Omega_1$, we can use the nearest point $\mathbf{x}_2 = \pi(\mathbf{x}_1) \in \partial\Omega_2$. Here π is simply the nearest point projection onto $\partial\Omega_2$. The potential contact region $\Gamma_{C,2} \subseteq \partial\Omega_2$ must be consistent with $\Gamma_{C,1} \subseteq \partial\Omega_1$, so we assume that $\Gamma_{C,2} = \pi(\Gamma_{C,1})$.

Let $\mathbf{u}_1$ be the displacement field on Ω_1 and $\mathbf{u}_2$ the displacement field on Ω_2. Furthermore, let $N_1(t,\mathbf{x}_1)\mathbf{n}_1(\mathbf{x}_1) + \mathbf{F}_1(t,\mathbf{x}_1)$ the contact force at $\mathbf{x}_1$ and $N_2(t,\mathbf{x}_2)\mathbf{n}_2(\mathbf{x}_2) + \mathbf{F}_2(t,\mathbf{x}_2)$ the contact force at $\mathbf{x}_2$. Since $\mathbf{x}_1 \approx \mathbf{x}_2$, we have nearly opposite normal direction vectors

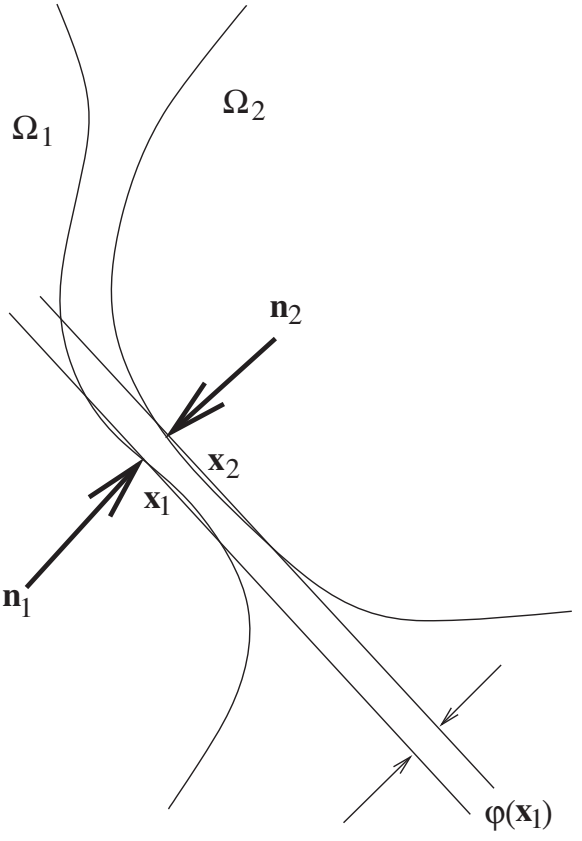


Figure 6.8: Contact between two elastic bodies.

$\mathbf{n}_1(\mathbf{x}_1) \approx -\mathbf{n}_2(\mathbf{x}_2)$. Then we can use a gap function $\varphi(\mathbf{x}_1) = \mathbf{n}_1(\mathbf{x}_1) \cdot (\mathbf{x}_2 - \mathbf{x}_1)$ on the potential contact $\Gamma_{C,1}$ region of $\partial\Omega_1$. From Newton's third law, that every action has an equal and opposite reaction, the contact force at $\mathbf{x}_1 \in \partial\Omega_1$ must be the negative of the contact force at $\mathbf{x}_2 \in \partial\Omega_2$. We can represent this (approximately) by having the frictionless components satisfy

$$N_1(t,\mathbf{x}_1) = N_2(t,\mathbf{x}_2), \qquad \mathbf{x}_1 \in \Gamma_{C,1},\ \mathbf{x}_2 = \pi(\mathbf{x}_1) \in \Gamma_{C,2}, \tag{6.51}$$

and the frictional components satisfy

$$\mathbf{F}_1(t,\mathbf{x}_1) = -\mathbf{F}_2(t,\mathbf{x}_2), \qquad \mathbf{x}_1 \in \Gamma_{C,1},\ \mathbf{x}_2 = \pi(\mathbf{x}_1) \in \Gamma_{C,2}. \tag{6.52}$$

The conditions for frictionless contact can then be represented by

$$\begin{aligned} 0 \le N_1(t,\mathbf{x}_1) &= N_2(t,\mathbf{x}_2) \\ &\perp\ \mathbf{n}_1(\mathbf{x}_1)^T\,(\mathbf{u}_2(t,\mathbf{x}_2) - \mathbf{u}_1(t,\mathbf{x}_2)) - \varphi(\mathbf{x}_1) \ge 0 \\ &\qquad \text{for all } \mathbf{x}_1 \in \Gamma_{C,1},\ \mathbf{x}_2 = \pi(\mathbf{x}_1). \end{aligned} \tag{6.53}$$

For frictional contact we use (6.53) for the normal contact forces, and for the friction forces

we modify (6.50) as follows:

$$\int_{\Gamma_{C,1}} \mathbf{F}_1(t,\mathbf{x}_1)^T \mathbf{w}(\mathbf{x}_1)\, dS(\mathbf{x}_1) \leq \int_{\Gamma_{C,1}} \mu N_1 \left(\left| \frac{\partial \mathbf{u}_{1T}}{\partial t} - \frac{\partial \mathbf{u}_{2T}}{\partial t} + \mathbf{w}_T \right| - \left| \frac{\partial \mathbf{u}_{1T}}{\partial t} - \frac{\partial \mathbf{u}_{2T}}{\partial t} \right| \right) dS(\mathbf{x}_1) \qquad (6.54)$$

for all smooth $\mathbf{w}(\mathbf{x}_1)$. In this expression, $\mathbf{u}_2$ is evaluated at $\mathbf{u}_2(t,\mathbf{x}_2) = \mathbf{u}_2(t,\pi(\mathbf{x}_1))$. As noted above, $\mathbf{F}_2(t,\mathbf{x}_2) = -\mathbf{F}_1(t,\mathbf{x}_1)$, where $\mathbf{x}_1 = \pi(\mathbf{x}_1)$.

6.2.3 Technical issues

Since dealing with elastic or viscoelastic bodies means using partial differential equations (and their generalizations) rather than ordinary differential equations (and their generalizations), we have to deal with infinite-dimensional spaces and unbounded operators. These have their own difficulties and leave an imprint on the theory used to deal with the associated DVIs. Usually these difficulties appear in the form of compactness or (pseudo)monotonicity conditions.

There is one respect in which things are actually easier than the finite-dimensional (rigid body) theory. Since stronger impulse responses for the operators involved result in less singular solutions, the solutions for elastic and viscoelastic impact problems tend to have less singular solutions for the contact forces than those that exist for rigid-body problems. The contact forces in rigid-body dynamics typically include Dirac-δ functions, while in elastic-body dynamics the forces are typically integrable or L^2 in time. However, even with this, finding a suitable space in which the normal contact forces must reside is difficult. While the contact forces must be in the space of measures on $[0,T] \times \Gamma_C$, this is too large a space for many purposes, including showing that solutions exist.

Some of these issues can be resolved if Kelvin–Voigt viscoelasticity is used. With this model of viscoelasticity, the equations without the contact conditions essentially become parabolic. This has a great many advantages from the point of view of proving existence results; many of these stem from the fact that the solution map for the differential operator $u(0) \mapsto u(t)$ is a compact operator. There are a number of other features of these viscoelastic equations which give tighter bounds on crucial quantities. The improved regularity of the solutions helps to show existence results. The downside of using viscoelasticity is that the contact forces generally become more singular. To avoid dealing with this, the problems are reformulated as VIs in which the normal contact forces do not appear.

The inclusion of Coulomb friction generally makes the difficulties much worse. If the frictional and normal contact forces can be decoupled, then we can obtain the results that we would expect. Once the frictionless problem is solved, we can use the solution to find the friction forces via a monotone PVI. This is essentially the same situation as arises in the so-called *Tresca friction*, where the normal force is given a priori and contact is assumed.

However, with rare exceptions, problems in elasticity couple together the normal and frictional (tangential) forces which result in crucial kinds of instabilities. There is, indeed, a *nonexistence* result which has been shown for the special case of a two-dimensional hyperelastic[9] neo-Hookean (that is, the Cauchy stress tensor is proportional to the Cauchy

[9]Hyperelastic materials, such as rubber, do not change volume. For small deformation problems, this amounts to requiring that $\nabla \cdot \mathbf{u} = 0$.

strain tensor) elastic body with Coulomb friction, *provided that the coefficient of friction exceeds a certain threshold value* [212].

To circumvent some of the difficulties associated with these issues, various modifications of the Coulomb friction, and even the Signorini conditions, have been used in order to obtain existence of solutions. One of the simplest modifications is to replace the Signorini conditions with a penalty or a stiff spring approximation [165, 193]. Another is to write the Signorini conditions in terms of the normal velocity. Combined with the use of Kelvin–Voigt viscoelasticity, this reduces the frictionless problem to essentially a PVI. Even so, existence has been shown only for the velocity-based "Signorini" conditions, provided the coefficient of friction does not exceed a certain threshold value [88, 89]. An alternative is to use a nonlocal friction law, where the normal contact force is replaced by a suitable smoothed local average [56, 154]. None of these modifications has a compelling physical basis, and it may be that solutions, even for Kelvin–Voigt viscoelasticity, do not exist beyond a certain threshold value of the coefficient of friction. If so, then this would be an instance where mathematical *nonexistence* results have important physical implications: the observed macroscopic value of Coulomb friction coefficients may well be a result of the limits due to frictional instabilities, rather than to microscopic behavior.

6.2.4 Routh's rod

This is the simplest elastic-body impact problem, where a one-dimensional body (moving in a line) impacts a rigid obstacle at one end. This is illustrated in Figure 1.2. This problem was first solved by Routh in [216, pp. 442–444] in 1860 for the case where the rod is initially moving with a constant, uniform velocity until impact. The more general problem of showing that the rod impacting a rigid obstacle with initially finite energy has a solution which is unique (and can be approximated numerically) is much more recent. Since the rod impacts the obstacle only at an end ($x = 0$) of the domain ($\Omega = (0, \ell)$), this is called a *thin obstacle* problem since contact can occur only on a set that has Lebesgue measure zero of the domain Ω. In fact, it is sometimes called a *boundary thin obstacle* problem since contact can occur only on the boundary $\partial\Omega$, rather than in the interior of the domain of the problem.

The equations of motion in a one-dimensional linear elastic medium reduce to the wave equation:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < \ell, \quad t > 0, \tag{6.55}$$

where $u = u(t,x)$ is the displacement field. The free end has no traction, so $\partial u/\partial x(t,\ell) = 0$ for all t. At the left end, which can contact the obstacle, we have Signorini-type conditions:

$$-\frac{\partial u}{\partial x}(t,0) = N(t), \tag{6.56}$$

$$0 \le N(t) \perp u(t,0) \ge 0 \tag{6.57}$$

for all t. The approach we take is to represent this problem as a CCP. To do this, we

consider the impulse response function:

$$\begin{aligned} \frac{\partial^2 w}{\partial t^2} &= c^2 \frac{\partial^2 w}{\partial x^2}, \qquad 0 < x < \ell, \quad t \geq 0, \\ -c^2 \frac{\partial w}{\partial x}(t,0) &= \delta(t), \qquad \frac{\partial w}{\partial x}(t,\ell) \;=\; 0, \qquad t \geq 0, \\ w(0,x) &= 0, \qquad \frac{\partial w}{\partial t}(0,x) \;=\; 0, \qquad 0 < x < \ell. \end{aligned}$$

This can be done fairly easily using d'Alembert solutions:

$$w(t,x) = c^{-1}\left[\sum_{k=0}^{\infty} H(ct - x - 2k\ell) + \sum_{k=0}^{\infty} H(ct + x - (2k+2)\ell)\right],$$

where $H(s) = 1$ if $s > 0$ and $H(s) = 0$ if $s < 0$ is the Heaviside function. Also, there is the solution due to the initial conditions:

$$\begin{aligned} \frac{\partial^2 \widehat{u}}{\partial t^2} &= c^2 \frac{\partial^2 \widehat{u}}{\partial x^2}, \\ -\frac{\partial \widehat{u}}{\partial x}(t,0) &= 0, \qquad \frac{\partial \widehat{u}}{\partial x}(t,\ell) \;=\; 0, \\ \widehat{u}(0,x) &= u_0(x), \qquad \frac{\partial \widehat{u}}{\partial t}(0,x) \;=\; v_0(x), \end{aligned}$$

where u_0 is the initial displacement and v_0 is the initial velocity. Using the standard theory for linear differential equations,

$$u(t,x) = \widehat{u}(t,x) + \int_0^t N(\tau)\, w(t-\tau, x)\, d\tau.$$

Substituting this for (6.57) gives

$$0 \leq N(t) \perp \widehat{u}(t,0) + \int_0^t w(t-\tau,0)\, N(\tau)\, d\tau \geq 0 \tag{6.58}$$

for $t \geq 0$. We have reduced the problem to a CCP. The kernel of the CCP is

$$w(t,0) = c^{-1}\left[H(ct) + \sum_{k=1}^{\infty} 2\, H(ct - 2k\ell)\right]. \tag{6.59}$$

In any finite interval, $t \mapsto w(t,0)$ has bounded variation and $w(0^+,0) = c^{-1} > 0$. So, applying Theorem 5.6, we see that solutions exist. In the interval $(0, 2\ell/c)$, $w(\cdot,0)$ is constant, and so we can apply Theorem 5.7 to show that solutions are unique.

It should be noted that although this problem is *formally* an index-two problem, the CCP we obtain is an index-one CCP, and this is the secret of the success of the CCP approach. The core of the approach is the *Neumann to Dirichlet operator* for the wave equation. The Neumann to Dirichlet map takes $\partial u/\partial n|_{\partial\Omega} \mapsto u|_{\partial\Omega}$ for solutions to the given homogeneous partial differential equation.

Of particular interest here is the matter of conservation of energy. It is well known that the wave equation by itself conserves energy:

$$E[u,\partial u/\partial t] = \frac{1}{2}\int_0^\ell \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial t}\right)^2\right] dx, \tag{6.60}$$

at least if we have only fixed displacement or fixed traction boundary conditions. However, it is not clear if energy should be conserved. For rigid-body dynamics we must have a coefficient of restitution. But in this model there is no coefficient of restitution, and yet we have uniqueness of solutions. Do we have conservation of energy? The answer is yes, and showing this involves a differentiation lemma for CPs.

Let us start by looking at the rate of change of energy:

$$\begin{aligned}
\frac{d}{dt}E[u,\partial u/\partial t] &= \frac{d}{dt}\frac{1}{2}\int_0^\ell \left[\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial t}\right)^2\right] dx \\
&= \int_0^\ell \left[\frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial t\,\partial x} + \frac{\partial u}{\partial t}\frac{\partial^2 u}{\partial t^2}\right] dx \\
&= \int_0^\ell \left[\frac{\partial u}{\partial x}\frac{\partial^2 u}{\partial t\,\partial x} + \frac{\partial u}{\partial t}\frac{\partial^2 u}{\partial x^2}\right] dx \\
&= \int_0^\ell \left[\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\frac{\partial u}{\partial t}\right) - \frac{\partial^2 u}{\partial x^2}\frac{\partial u}{\partial t} + \frac{\partial u}{\partial t}\frac{\partial^2 u}{\partial x^2}\right] dx \\
&= \frac{\partial u}{\partial x}\frac{\partial u}{\partial t}\bigg|_{x=0}^{x=\ell} = N(t)\frac{\partial u}{\partial t}(0,t).
\end{aligned}$$

Physically, this says that the rate of change in the energy of the rod is simply the rate at which the normal contact force does work on the rod. This, in turn, is zero. The reason is that we have the CP

$$0 \le N(t) \perp u(t,0) \ge 0. \tag{6.61}$$

Further, we have regularity results. Since $t \mapsto \widehat{u}(t,0)$ is in $H^1(0,T)$ (that is, $\partial\widehat{u}/\partial t(\cdot,0) \in L^2(0,T)$), from CCP theory we have $N \in L^2(0,T)$ by Theorem 5.6. In turn, $\partial u/\partial t(\cdot,0)$ is also in $L^2(0,T)$. Thus $N\,\partial u/\partial t|_{x=0}$ is in $L^1(0,T)$ and the energy $E[u]$ is an absolutely continuous function of t. But we can apply the differentiation lemma (Lemma 3.2) to show that (6.61) implies

$$0 = N(t)\frac{\partial u}{\partial t}(0,t) \qquad \text{for almost all } t.$$

Thus the normal contact force does no work on the rod, and hence energy is conserved: $E[u,\partial u/\partial t] = \text{constant}$.

6.2.5 Vibrating string

A vibrating string making contact with a rigid obstacle is another example of contact with elastic bodies. As before, we have a one-dimensional elastic body and the wave equation

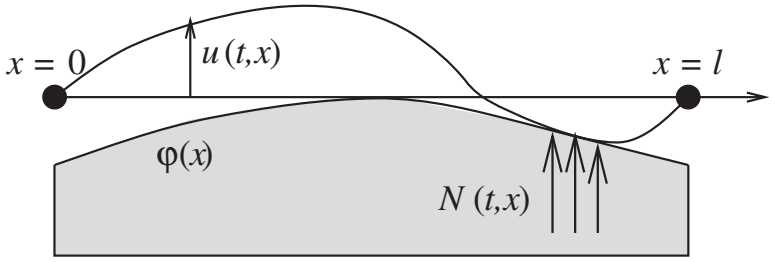


Figure 6.9: Vibrating string example.

holds within the body, at least where there is no contact. Unlike the previous example, we allow contact over most of the domain of the partial differential equation. This means that we cannot use finite-dimensional CCPs to describe the dynamics of the system with contact. This problem has been treated in a number of papers, such as [2, 8, 52, 221, 222]. For our analysis, we have a gap function $\varphi(x)$ which represents an obstacle $u(t,x) \geq \varphi(x)$, as illustrated in Figure 6.9.

The equations of motion then become

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2} + f(t,x) + N(t,x), \tag{6.62}$$

$$0 = u(t,0) \;=\; u(t,\ell), \tag{6.63}$$

$$0 \leq N(t,x) \perp u(t,x) - \varphi(x) \geq 0, \tag{6.64}$$

$$u(0,x) = u_0(x), \qquad \frac{\partial u}{\partial t}(0,x) \;=\; v_0(x) \tag{6.65}$$

for $t \geq 0$ and $0 < x < \ell$. This is an example of a *thick obstacle* problem: contact can occur over a part of the domain that has positive Lebesgue measure.

For consistency we have to assume that $\varphi(0), \varphi(\ell) \leq 0$. In fact, we will need slightly stronger conditions to prove existence:

$$\varphi(0), \varphi(\ell) < 0. \tag{6.66}$$

To begin the process of showing the *existence* of a solution, we will choose to start with the penalty approach; that is, we will obtain approximate solutions $u_\epsilon \approx u$ and $N_\epsilon \approx N$. We approximate the constraint $u - \varphi \geq 0$ by a system of stiff springs:

$$N_\epsilon(t,x) = \frac{1}{\epsilon}(u_\epsilon(t,x) - \varphi(x))_-, \tag{6.67}$$

where $s_- = \max(-s, 0)$ is the negative part of s. Since this is an approximation to the constraint $u - \varphi \geq 0$, the solutions obtained must be approximate as well:

$$\frac{\partial^2 u_\epsilon}{\partial t^2} = \frac{\partial^2 u_\epsilon}{\partial x^2} + f(t,x) + N_\epsilon(t,x), \qquad 0 < x < \ell, \tag{6.68}$$

$$0 = u_\epsilon(t,0) \;=\; u_\epsilon(t,\ell). \tag{6.69}$$

The penalty term $(u_\epsilon(t,x)-\varphi(x))_-/\epsilon$ has its own contribution to the energy of the string:

$$E_\epsilon[u_\epsilon,\partial u_\epsilon/\partial t]=\int_0^\ell\left[\frac{1}{2}\left(\frac{\partial u_\epsilon}{\partial t}\right)^2+\frac{1}{2}\left(\frac{\partial u_\epsilon}{\partial x}\right)^2+\frac{1}{2\epsilon}\left[(u_\epsilon(t,x)-\varphi(x))_-\right]^2\right]dx. \tag{6.70}$$

As the penalty parameter goes to zero, note that the penalty term of the energy penalizes violations of the constraint $u-\varphi\geq 0$ more and more strongly. However, this can cause difficulties in trying to prove conservation of energy in the limit.

The first result that we need to establish is that solutions exist (in appropriate spaces) for the penalty approximation (6.67)–(6.69). This is typically not hard to show from a more abstract point of view. Let $\mathcal{A}=-\partial^2/\partial x^2$ be the partial differential operator acting on the Sobolev space $X:=H_0^1(0,\ell)=\left\{\,w\in H^1(0,\ell)\mid w(0)=w(\ell)=0\,\right\}$. Let $H=L^2(0,\ell)$ be the pivot space in a Gelfand triple $X=H_0^1(0,\ell)\subset H=L^2(0,\ell)=H'\subset X'=H_0^1(0,\ell)'$. Then (6.68)–(6.69) can be written in the form

$$\frac{\partial^2 u_\epsilon}{\partial t^2}=-\mathcal{A}u_\epsilon+f(t)+\frac{1}{\epsilon}\psi\left[u_\epsilon\right] \tag{6.71}$$

with $\psi:X\to H$ the Lipschitz operator given by $\psi[w](x)=(w(x)-\varphi(x))_-$. The function $f:[0,T]\to H$ is given by $f(t)(x)=f(t,x)$ from the data to the problem. We assume that $f\in L^2(0,T;H)$. With initial conditions $u(0,x)=u_0(x)$ and $\partial u/\partial t(0,x)=v_0(x)$, we can use an abstract version of the "variation or parameters" method for ordinary differential equations:

$$u_\epsilon(t)=\cos\left(\mathcal{A}^{1/2}t\right)u_0+\mathcal{A}^{-1/2}\sin\left(\mathcal{A}^{1/2}t\right)v_0+\int_0^t\mathcal{A}^{-1/2}\sin\left(\mathcal{A}^{1/2}(t-\tau)\right)\left[f(\tau)+\frac{1}{\epsilon}\psi\left[u_\epsilon(\tau)\right]\right]d\tau.$$

Note that $\mathcal{A}^{1/2}$ is a well-defined self-adjoint elliptic operator since $\mathcal{A}$ is a self-adjoint elliptic operator. Note also that since $f:[0,T]\to H$ and $\psi:X\to H$ are Lipschitz, then $\mathcal{A}^{-1/2}f:[0,T]\to X$ is in $L^2(0,T;X)$ and $\mathcal{A}^{-1/2}\psi:X\to X$ is Lipschitz. Also, $\sin(\mathcal{A}^{1/2}t)$ and $\cos(\mathcal{A}^{1/2}t)$ are bounded linear operators $X\to X$ and $H\to H$, in fact with norm less than or equal to 1 for all t. We can then use the Picard iteration to show the existence of solutions: consider the map $C(0,T;X)\to C(0,T;X)$ given by $u\mapsto w$ where

$$w(t)=\cos\left(\mathcal{A}^{1/2}t\right)u_0+\mathcal{A}^{-1/2}\sin\left(\mathcal{A}^{1/2}t\right)v_0+\int_0^t\mathcal{A}^{-1/2}\sin\left(\mathcal{A}^{1/2}(t-\tau)\right)\left[f(\tau)+\frac{1}{\epsilon}\psi\left[u(\tau)\right]\right]d\tau.$$

A Lipschitz constant for this map is $(1/\epsilon)\,L_\psi\,\|\mathcal{A}^{-1/2}\|_{\mathcal{L}(H,X)}\,T$, where L_ψ is the Lipschitz constant for $\psi:X\to H$. This can be determined from the Lipschitz constant for the imbedding of $X=H_0^1(0,\ell)$ into $H=L^2(0,\ell)$. The map $H=L^2(0,\ell)\to H=L^2(0,\ell)$ given by $u\mapsto(u-\varphi)_-$ has Lipschitz constant one since the maps $\mathbb{R}\to\mathbb{R}$ given by $s\mapsto(s-\varphi(x))_-$

have Lipschitz constant one. Whatever the precise values of these constants (which can depend on which of many equivalent norms we choose for $H_0^1(0,\ell)$), what is important is that for any $\epsilon > 0$ we can choose $0 < T < \epsilon/(L_\psi \|\mathcal{A}^{-1/2}\|_{\mathcal{L}(H,X)})$. This makes the map $u \mapsto w$ a contraction, so by the contraction mapping theorem, there is a unique fixed point which solves the differential equation (6.71).

We should now turn to the question of energy bounds for the penalty problem. First we do some calculations:

$$\begin{aligned}
\frac{d}{dt}E_\epsilon[u_\epsilon, \partial u_\epsilon/\partial t] &= \frac{d}{dt}\int_0^\ell \left[\frac{1}{2}\left(\frac{\partial u_\epsilon}{\partial t}\right)^2 + \frac{1}{2}\left(\frac{\partial u_\epsilon}{\partial x}\right)^2 + \frac{1}{2\epsilon}\left[(u_\epsilon(t,x)-\varphi(x))_-\right]^2\right] dx \\
&= \int_0^\ell \left[\frac{\partial u_\epsilon}{\partial t}\frac{\partial^2 u_\epsilon}{\partial t^2} + \frac{\partial u_\epsilon}{\partial x}\frac{\partial^2 u_\epsilon}{\partial t\,\partial x} - \frac{1}{\epsilon}(u_\epsilon(t,x)-\varphi(x))_- \frac{\partial u_\epsilon}{\partial t}\right] dx \\
&= \int_0^\ell \left[\frac{\partial u_\epsilon}{\partial t}\left(\frac{\partial^2 u_\epsilon}{\partial x^2} + \frac{1}{\epsilon}(u_\epsilon(t,x)-\varphi(x))_-\right)\right. \\
&\qquad \left. + \frac{\partial u_\epsilon}{\partial x}\frac{\partial^2 u_\epsilon}{\partial t\,\partial x} - \frac{1}{\epsilon}(u_\epsilon(t,x)-\varphi(x))_- \frac{\partial u_\epsilon}{\partial t}\right] dx \\
&= \int_0^\ell \frac{\partial}{\partial x}\left(\frac{\partial u_\epsilon}{\partial t}\frac{\partial u_\epsilon}{\partial x}\right) dx \;=\; \left.\frac{\partial u_\epsilon}{\partial t}\frac{\partial u_\epsilon}{\partial x}\right|_{x=0}^{x=\ell} \;=\; 0
\end{aligned}$$

since $u_\epsilon(t,0) = u_\epsilon(t,\ell) = 0$ for all t. Thus the energy with the penalty term is conserved by solutions to the penalty equations. Since, by assumption, the initial conditions $u_0(x) \geq \varphi(x)$ for all x, we see that $E_\epsilon[u_\epsilon(0,\cdot), \partial u_\epsilon/\partial t(0,\cdot)] = E[u_\epsilon(0,\cdot), \partial u_\epsilon/\partial t(0,\cdot)]$. Thus

$$\begin{aligned}
E\left[u_\epsilon(t,\cdot), \partial u_\epsilon/\partial t(t,\cdot)\right] &\leq E_\epsilon\left[u_\epsilon(t,\cdot), \partial u_\epsilon/\partial t(t,\cdot)\right] \\
&= E_\epsilon\left[u_\epsilon(0,\cdot), \partial u_\epsilon/\partial t(0,\cdot)\right] \\
&= E\left[u_\epsilon(0,\cdot), \partial u_\epsilon/\partial t(0,\cdot)\right] = E\left[u_0, v_0\right],
\end{aligned}$$

which is bounded independently of $\epsilon > 0$.

From the energy bounds we have immediate bounds, independent of $\epsilon > 0$, on

1. the kinetic energy, which is $\frac{1}{2}\|\partial u_\epsilon/\partial t\|^2_{L^2(0,\ell)}$,
2. the elastic energy, which is $\frac{1}{2}\|\partial u_\epsilon/\partial x\|^2_{L^2(0,\ell)}$, from which we obtain bounds on $\|u_\epsilon\|_{H^1(0,\ell)}$, and
3. the penalty energy, which is $\left\|(u_\epsilon-\varphi)_-\right\|^2_{L^2(0,\ell)}/(2\epsilon)$, so that $\left\|(u_\epsilon-\varphi)_-\right\|_{L^2(0,\ell)} = \mathcal{O}(\epsilon^{1/2})$.

Let $E^* = E[u_0, v_0]$ be this bound on the energy. The kinetic energy bounds will be used to obtain momentum bounds (essentially bounds on the integral of the velocity), which in turn will give integral bounds on the normal contact forces. In particular, we will take $w(x) = x(\ell - x)$, which is strictly positive on $(0,\ell)$. Then, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
\int_0^\ell w(x)\frac{\partial u_\epsilon}{\partial t}(t,x)\,dx &\leq \left[\int_0^\ell w(x)^2\,dx\right]^{1/2}\left[\int_0^\ell \left(\frac{\partial u_\epsilon}{\partial t}(t,x)\right)^2 dx\right]^{1/2} \\
&\leq \text{constant}\left(E^*\right)^{1/2} \qquad \text{for all } t.
\end{aligned}$$

On the other hand, if we write $N_\epsilon(t,x) = (u_\epsilon(t,x) - \varphi(x))_- / \epsilon$,

$$\begin{aligned}\frac{d}{dt}\int_0^\ell w(x)\frac{\partial u_\epsilon}{\partial t}(t,x)\,dx &= \int_0^\ell w(x)\frac{\partial^2 u_\epsilon}{\partial t}(t,x)\,dx \\ &= \int_0^\ell w(x)\left[\frac{\partial^2 u_\epsilon}{\partial x^2}(t,x) + N_\epsilon(t,x)\right]dx \\ &= w(x)\frac{\partial u_\epsilon}{\partial x}(t,x)\bigg|_{x=0}^{x=\ell} - \int_0^\ell w'(x)\frac{\partial u_\epsilon}{\partial x}(t,x)\,dx \\ &\qquad + \int_0^\ell w(x)\,N_\epsilon(t,x)\,dx.\end{aligned}$$

The first term in the final expression is zero since $w(0) = w(\ell) = 0$. The second term is bounded by $\left[\int_0^\ell w'(x)^2\,dx\right]^{1/2}(E^*)^{1/2}$. Integrating the above expression over the time interval $[0,T]$, we get

$$\begin{aligned}\int_0^\ell w(x)\,\frac{\partial u_\epsilon}{\partial t}(t,x)\bigg|_{t=0}^{t=T} dx &= -\int_0^T\int_0^\ell w'(x)\frac{\partial u_\epsilon}{\partial x}(t,x)\,dx \\ &\qquad + \int_0^T\int_0^\ell w(x)\,N_\epsilon(t,x)\,dx\,dt.\end{aligned}$$

Using the bounds we have already obtained,

$$\begin{aligned}&\left|\int_0^T\int_0^\ell w(x)\,N_\epsilon(t,x)\,dx\,dt\right| \\ &\qquad \le \left(2\left[\int_0^\ell w(x)^2\,dx\right]^{1/2} + T\left[\int_0^\ell w'(x)^2\,dx\right]^{1/2}\right)\left(E^*\right)^{1/2}.\end{aligned}$$

Since $w \ge 0$ and $N_\epsilon \ge 0$, we have integral bounds on N_ϵ. We would rather not have $w(x)$ in the integral bounds. If we can show that $N_\epsilon(t,x) = 0$ for $x < \eta$ or $x > L - \eta$ for some $\eta > 0$ (independent of ϵ), then we can bound

$$\int_0^T\int_0^\ell N_\epsilon(t,x)\,dx\,dt \le \frac{1}{w(\eta)}\int_0^T\int_0^\ell w(x)\,N_\epsilon(t,x)\,dx\,dt.$$

In fact, this is true from the compatibility condition (6.66) and using the Cauchy–Schwarz inequality: for $x < y$,

$$\begin{aligned}|u_\epsilon(t,x) - u_\epsilon(t,y)| &\le \int_x^y \left|\frac{\partial u_\epsilon}{\partial x}(t,z)\right| dz \\ &\le \left[\int_x^y 1^2\,dz\right]^{1/2}\left[\int_x^y \left|\frac{\partial u_\epsilon}{\partial x}(t,z)\right|^2 dz\right]^{1/2} \\ &\le |y-x|^{1/2}\left\|\frac{\partial u_\epsilon}{\partial x}(t,\cdot)\right\|_{L^2(0,\ell)} \le |y-x|^{1/2}\left(E^*\right)^{1/2}.\end{aligned}$$

Thus u_ϵ is uniformly Hölder continuous with exponent $1/2$. In particular, since $u_\epsilon(t,0)=0$ for all t, $|u_\epsilon(t,x)| \le x^{1/2}(E^*)^{1/2}$. For $x>0$ sufficiently small we must have $u_\epsilon(t,x) > \varphi(x)$ for all t. Thus, there is an $\eta>0$ such that $N_\epsilon(t,x)=0$ for $0 \le x \le \eta$. A similar argument gives the corresponding result that there is an $\eta>0$ such that $N_\epsilon(t,x)=0$ for $\ell-\eta \le x \le \ell$.

Thus we have a bound on $\int_0^T \int_0^\ell N_\epsilon(t,x)\,dx\,dt$ independent of $\epsilon>0$. Treating N_ϵ as a measure on $[0,T]\times[0,\ell]$, we can apply Alaoglu's theorem to show that there is a weakly* convergent subsequence $N_\epsilon \rightharpoonup^* \widehat{N}$ in the space of measures $\mathcal{M}([0,T]\times[0,\ell])$. We restrict our attention to this subsequence, which we also denote by N_ϵ. At the same time, the u_ϵ are uniformly bounded in $L^\infty(0,T;X)$. Thus there is a weakly* convergent subsequence $u_\epsilon \rightharpoonup^* \widehat{u}$ in $L^\infty(0,T;X)$. We now wish to show that $u=\widehat{u}$ and $N=\widehat{N}$ solve (6.62)–(6.65).

We need to show that $N \ge 0$; that is, N is a nonnegative measure. Since each $N_\epsilon \ge 0$ and we have weak* convergence $N_\epsilon \rightharpoonup^* \widehat{N}$ in $\mathcal{M}([0,T]\times[0,\ell])$, for any nonnegative continuous function $\psi:\ [0,T]\times[0,\ell]\to\mathbb{R}$ we have

$$0 \le \int_0^T \int_0^\ell \psi(t,x)\,N_\epsilon(t,x)\,dx\,dt \to \int_{[0,T]\times[0,\ell]} \psi\,\widehat{N},$$

and so $\int_{[0,T]\times[0,\ell]} \psi\,\widehat{N} \ge 0$ for all continuous $\psi \ge 0$. Treating $\mathcal{M}([0,T]\times[0,\ell])$ as the dual space of $C([0,T]\times[0,\ell])$, we see that N is a nonnegative measure.

On the other hand, $\|(u_\epsilon-\varphi)_-\|_{L^\infty(0,T;L^2(0,\ell))} \le \text{constant}\,\epsilon^{1/2}$. Now $u_\epsilon \rightharpoonup^* \widehat{u}$ in $L^\infty(0,T;X) = L^\infty(0,T;H_0^1(0,\ell))$ means that for any $\psi \in L^1(0,T;H^{-1}(0,\ell))$,

$$\int_0^T \int_0^\ell \psi(t,x)\,u_\epsilon(t,x)\,dx\,dt \to \int_0^T \int_0^\ell \psi(t,x)\,\widehat{u}(t,x)\,dx\,dt.$$

To get around the fact that L^∞ spaces are not reflexive, note that $L^\infty(0,T;X) \subset L^2(0,T;X)$, and weak* convergence in $L^\infty(0,T;X)$ implies weak* convergence in $L^2(0,T;X)$. Since $L^2(0,T;X)$ *is* reflexive (the dual is $L^2(0,T;X')$), weak* and weak convergence are the same in $L^2(0,T;X)$. Now the function

$$u \mapsto \int_0^T \int_0^\ell \left[(u-\varphi)_-\right]^2 dx\,dt$$

is a *convex* and continuous function on $H_0^1(0,\ell)$. By Mazur's lemma,

$$\begin{aligned} 0 \le \int_0^T \int_0^\ell \left[(\widehat{u}-\varphi)_-\right]^2 dx\,dt &\le \liminf_{\epsilon\downarrow 0} \int_0^T \int_0^\ell \left[(u_\epsilon-\varphi)_-\right]^2 dx\,dt \\ &\qquad \text{in the subsequence} \\ &\le \liminf_{\epsilon\downarrow 0} T\,\text{constant}\,\epsilon^{1/2} \;=\; 0. \end{aligned}$$

Thus $(\widehat{u}-\varphi)_- = 0$ for almost all (t,x); or equivalently, $u(t,x) \ge \varphi(x)$ for almost all (t,x).

We need to show that $\widehat{u}$ and $\widehat{N}$ satisfy the differential equation (6.62). First note that since $\widehat{N} \in \mathcal{M}([0,T]\times[0,\ell]) = C([0,T]\times[0,\ell])' = C(0,T;C[0,\ell])'$, and since $C[a,b]$ (the space of continuous functions $[a,b]\to\mathbb{R}$) contains $H^1(a,b)$, then we get $C(0,T\ C[0,\ell]) \supset C(0,T;H_0^1(0,\ell))$. Thus $\mathcal{M}([0,T]\times[0,\ell]) \subset C(0,T;H_0^1(0,\ell))'$. So, if $\psi:\ [0,T]\to X = H_0^1(0,\ell)$ is a smooth (and therefore continuous) function, then we get

$\int_0^T \int_0^\ell \psi\, u_\epsilon \, dx\, dt \to \int_0^T \int_0^\ell \psi\, \widehat{u}\, dx\, dt$ and $\int_0^T \int_0^\ell \psi\, N_\epsilon \, dx\, dt \to \int_0^T \int_0^\ell \psi\, \widehat{N}\, dx\, dt$ as $\epsilon \downarrow 0$ in the subsequence. Then we use integration by parts to show that (6.62) is satisfied in the limit. Suppose that $\psi\colon [0,T] \to X = H_0^1(0,\ell)$ is indeed a function in $C^m(0,T;X)$ for sufficiently large m, which can be determined later. For now, let us also suppose that $\psi(T,x) = \partial\psi/\partial t(T,x) = 0$. Then

$$\begin{aligned}
0 &= \int_0^T \int_0^\ell \psi \left[\frac{\partial^2 u_\epsilon}{\partial t^2} + \mathcal{A} u_\epsilon + N_\epsilon \right] dx\, dt \\
&= \int_0^\ell \psi \, \frac{\partial u_\epsilon}{\partial t}\bigg|_{t=0}^{t=T} dx - \int_0^\ell \frac{\partial \psi}{\partial t} u_\epsilon \bigg|_{t=0}^{t=T} dx \\
&\qquad + \int_0^T \int_0^\ell \frac{\partial^2 \psi}{\partial t^2} u_\epsilon \, dx\, dt + \int_0^T \int_0^\ell \psi(t,x)\, \mathcal{A} u_\epsilon(t,x)\, dx\, dt \\
&\qquad + \int_0^T \int_0^\ell \psi(t,x)\, N_\epsilon(t,x)\, dx\, dt.
\end{aligned}$$

Each integral over $[0,T]\times[0,\ell]$ converges to the appropriate limit. The integrals over $[0,\ell]$ with $t=0$ and $t=T$ reduce to

$$\begin{aligned}
\int_0^\ell \psi \, \frac{\partial u_\epsilon}{\partial t}\bigg|_{t=0}^{t=T} dx &= \int_0^\ell \left[\psi(T,x) \frac{\partial u_\epsilon}{\partial t}(T,x) - \psi(0,x) \frac{\partial u_\epsilon}{\partial t}(0,x) \right] dx \\
&= -\int_0^\ell \psi(0,x)\, v_0(x)\, dx, \\
\int_0^\ell \frac{\partial \psi}{\partial t} u_\epsilon \bigg|_{t=0}^{t=T} dx &= \int_0^\ell \left[\frac{\partial \psi}{\partial t}(T,x)\, u_\epsilon(T,x) - \frac{\partial \psi}{\partial t}(0,x)\, u_\epsilon(0,x) \right] dx \\
&= -\int_0^\ell \frac{\partial \psi}{\partial t}(x,0)\, u_0(x)\, dx.
\end{aligned}$$

Thus taking limits in the subsequence gives

$$\begin{aligned}
0 &= \int_0^\ell \left(\frac{\partial \psi}{\partial t}\bigg|_{t=0} u_0 - \psi|_{t=0}\, v_0 \right) dx \\
&\qquad + \int_0^T \int_0^\ell \frac{\partial^2 \psi}{\partial t^2} \widehat{u}\, dx\, dt + \int_0^T \int_0^\ell \psi(t,x)\, \mathcal{A}\widehat{u}(t,x)\, dx\, dt \\
&\qquad + \int_0^T \int_0^\ell \psi(t,x)\, \widehat{N}(t,x)\, dx\, dt.
\end{aligned}$$

Undoing the integration by parts we did before now gives

$$0 = \int_0^T \int_0^\ell \psi \left[\frac{\partial^2 \widehat{u}}{\partial t^2} + \mathcal{A}\widehat{u} + \widehat{N} \right] dx\, dt,$$

so that $\widehat{u}$ and $\widehat{N}$ do indeed satisfy the differential equation (6.62).

The only remaining task to show that $\widehat{u}$ and $\widehat{N}$ solve our problem is to show that $\widehat{N} \perp \widehat{u} - \varphi$.

Now, for every $\epsilon > 0$, $\int_0^T \int_0^\ell N_\epsilon \, (u_\epsilon - \varphi) \, dx \, dt \le 0$, as $N_\epsilon(t,x) = 0$ whenever $u_\epsilon(t,x) > \varphi(x)$. We want to take limits (within some subsequence) to get $\int_0^T \int_0^\ell N \, (u - \varphi) \, dx \, dt \le 0$ for the limits N and u. Then, since $u - \varphi \ge 0$ and $N \ge 0$, we would have to conclude that $\int_0^T \int_0^\ell N \, (u - \varphi) \, dx \, dt = 0$, and we would have a solution of the problem. The trouble with this is that we have weak* convergence of $N_\epsilon \rightharpoonup^* \widehat{N}$ and $u_\epsilon \rightharpoonup^* \widehat{u}$, and this is not enough to conclude that the limit of the integral is the integral of the limit.

To complete the result, we need an additional *compactness* result or property to use. To do this we need to introduce some other spaces in which we can get the right kind of convergence. Let us review the spaces (and kinds of convergence) we have:

$$\begin{aligned} N_\epsilon &\rightharpoonup^* \widehat{N} \qquad \text{weak}^* \text{ in } \mathcal{M}([0,T] \times [0,\ell]) \subset \mathcal{M}(0,T; H_0^1(0,\ell)'), \\ u_\epsilon &\rightharpoonup^* \widehat{u} \qquad \text{weak}^* \text{ in } L^\infty(0,T; H_0^1(0,\ell)). \end{aligned}$$

Currently these spaces do not even have a duality pairing: L^∞ is the dual of L^1, not the larger space $\mathcal{M}$ of measures. However, we can do better than $\mathcal{M}([0,T] \times [0,\ell]) \subset \mathcal{M}(0,T; H^{-1}(0,\ell))$; $\mathcal{M}[0,\ell]$, the space of measures on $[0,\ell]$, is also in $H^{-1/2-\eta}(0,\ell)$ for any $\eta > 0$ since every function in $H^{1/2+\eta}(0,\ell)$ is continuous (Theorem A.8). Thus we can imbed

$$\mathcal{M}([0,T] \times [0,\ell]) = \mathcal{M}(0,T; \mathcal{M}[0,\ell]) \subset \mathcal{M}\left(0,T; H^{-1/2-\eta}(0,\ell)\right).$$

Even noting the compact imbedding $H_0^1(0,\ell) \subset H^{1/2+\eta}(0,\ell)$ we have the problem that $\mathcal{M}[a,b]$ is the dual of $C[a,b]$, not $L^\infty(a,b)$, and we have only weak* convergence in L^∞.

We need some compactness results for the spaces $L^p(a,b;X)$ or $C(a,b;X)$ where X is a Banach or Hilbert space compactly imbedded in another. There is an extra ingredient that we have not used yet. The velocity $v = \partial u/\partial t$ is bounded from the energy bounds: $\partial u_\epsilon/\partial t$ is uniformly bounded in $L^\infty\left(0,T; L^2(0,\ell)\right) = L^\infty(0,T; H)$, independently of $\epsilon > 0$. Here we use the compactness theorem of Seidman (Theorem A.6).

Since we know that $\partial u_\epsilon/\partial t$ is uniformly bounded in $L^\infty\left(0,T; L^2(0,\ell)\right)$, and $X = H_0^1(0,\ell) \subset H^{3/4}(0,\ell) = Z \subset L^2(0,\ell) = Y$ with $X \subset Z$ compact and $Z \subset Y$ continuous, we can apply this theorem to show that u_ϵ belongs to a compact subset of $C(0,T; H^{3/4}(0,\ell))$. Possibly by restricting to a further subsequence, we can show that $u_\epsilon \to \widehat{u}$ *strongly* in $C(0,T; H^{3/4}(0,\ell))$. Taking $\eta = 1/4$, we have $N_\epsilon \rightharpoonup \widehat{N}$ weakly* in $\mathcal{M}([0,T]; H^{-3/4}(0,\ell))$. Now we have duality pairing between $\widehat{N}$ and $\widehat{u}$ as well as strong convergence of at least one of the functions. Thus we can take limits and get $\langle \widehat{N}, \widehat{u} - \varphi \rangle = 0$, and so $\widehat{u}$ and $\widehat{N}$ are a solution.

Note that our result is a pure *existence* result. No one has been able to prove uniqueness except in special circumstances (see, for example, Schatzman [221]). In general, we still need a coefficient of restitution, which has not been included in the formulation, and conservation does not hold in general for solutions of (6.62)–(6.65). For example, consider $u_0(x) = \sin(\pi x/\ell)$ and $v_0(x) = 0$ for all x, while $\varphi(x) = 0$ for $\eta \le x \le \ell - \eta$ but $\varphi(x) < 0$ if either $x < \eta$ or $x > \ell - \eta$ with $0 < \eta < \ell/2$. Then the solution $u(t,x) = \cos(\pi t/\ell)$ for $0 \le t \le \ell/(2\pi)$. But at $t^* = \ell/(2\pi)$ we have $u(t^*,x) = 0$ for all x. At contact, we could set $\partial u/\partial t(t^{*+},x) = 0$ for $\eta \le x \le \ell - \eta$, for example. Since the wave equation has finite speed of propagation, all the energy in $\eta < x < \ell - \eta$ would be lost. Or we could set

$\partial u/\partial t(t^{*+},x) = -e\,\partial u/\partial t(t^{*-},x)$ with a coefficient of restitution e for $\eta < x < \ell - x$ and obtain different solutions for different $0 \le e \le 1$.

Our method of construction should lead to conservative solutions since solutions of our penalty method conserve the penalty energy

$$E_\epsilon\left[u_\epsilon, \frac{\partial u_\epsilon}{\partial t}\right] = \int_0^\ell \left[\frac{1}{2}\left(\frac{\partial u_\epsilon}{\partial t}\right)^2 + \frac{1}{2}\left(\frac{\partial u_\epsilon}{\partial x}\right)^2 + \frac{1}{2\epsilon}\left[(u_\epsilon(t,x)-\varphi(x))_-\right]^2\right] dx.$$

We have used the energy bound to show that the energy in the penalty part of the energy

$$\int_0^\ell \frac{1}{2\epsilon}\left[(u_\epsilon(t,x)-\varphi(x))_-\right]^2 dx$$

is bounded in order to show that $\|(u-\varphi)_-\|_{L^\infty(0,T;L^2(0,\ell))} = \mathcal{O}(\epsilon^{1/2})$. However, we do not know if this energy goes to zero, or goes to zero uniformly, as $\epsilon \downarrow 0$. For the example of the previous paragraph with $u(t,x) = \cos(\pi t/\ell)\sin(\pi x/\ell)$ for $0 \le t \le t^*$, we could see that in the penalty approximation, we would get the same solution for $0 \le t \le t^*$, but once contact is made, there would be considerable energy transferred to this penalty part. Since we have only weak* convergence in $L^\infty\left(0,T;H_0^1(0,\ell)\right)$, we cannot conclude that the nonpenalty energy $E[u_\epsilon, \partial u_\epsilon/\partial t]$ converges to $E[\widehat{u}, \partial\widehat{u}/\partial t]$. By using Mazur's lemma, all we can really tell is that $E[\widehat{u}(t,\cdot), \partial\widehat{u}/\partial t] \le E^*$ for all $t \ge 0$, which does not imply that the solution is necessarily even *dissipative*.

Schatzman's paper [221] is remarkable in that it shows that for *concave* obstacles it is possible to find solutions which do indeed conserve energy. The coefficient of restitution is explicitly set to one: $\partial u/\partial t(t^+,x) = -\partial u/\partial t(t^-,x)$ for any (t,x) where $u(t,x) = \varphi(t,x)$. The method of analysis is via the method of characteristics, and the technical difficulties of applying this approach to general obstacles (where there can be infinitely many separate reflections in finite time) meant that the obstacle was restricted to being concave. Whether there are solutions which conserve energy for arbitrary smooth obstacles is an open question.

Another issue that the reader may wonder about is the insistence for the strict inequalities $\varphi(0), \varphi(\ell) < 0$. While this may appear to be simply for mathematical convenience, there is a physical reason for this restriction. If we allow, for example, $\varphi(0) = 0$, there can be reflections at $x = 0$ from contact occurring arbitrarily close to $x = 0$. This can result in a feedback loop with reflections causing contacts causing reflections, and so on, in arbitrarily short times. This effect could be strong enough to destroy solutions, although that is not clear at present.

While this example is very simple and can be treated by some simpler or more specific methods (such as by Schatzman [221]), the principles used here are much more general and can be used in a wider class of problems.

6.2.6 Abstract treatment of a class of elastic bodies

There is a wider class of problems than both the Routh rod and vibrating string problems that can be treated in this fashion. It does not include all, or even the most important, elastic-body impact problems. However, it is an important class of problems and provides important insights. The development of this section follows [3].

We start with two Gelfand triples, one for the displacements and another for the forces. This can be later reduced to a single Gelfand triple, but using both can make the way the method works clearer. These Gelfand triples are

$$X \subset H = H' \subset X',$$
$$W \subset Z = Z' \subset W'$$

with the displacement field $u(t) \in X$ and the normal contact forces $N(t) \in W'$. The imbeddings in the Gelfand triples are assumed to be compact.

The set of admissible displacements is given by $u(t) - \varphi \in K$ with K a closed convex cone in X. Connecting them is a continuous linear operator $\beta\colon X \to W$ with $\beta(X)$ dense in W. The equations of motion are

$$\rho \frac{\partial^2 u}{\partial t^2} = -\mathcal{A}u(t) + f(t) + \beta^* N(t), \tag{6.72}$$
$$K^* \ni N(t) \perp \beta u(t) - \varphi \in K \tag{6.73}$$

with $\mathcal{A}\colon X \to X'$ a linear elliptic or semielliptic operator, $f\colon [0,T] \to H$ in $L^2(0,T;H)$, and K^* the dual cone to the cone K.

Example 6.1 (Routh's rod). Take $X = H^1(0,\ell)$ and $H = L^2(0,\ell)$ for the displacement field, and take $W = Z = \mathbb{R}$ for the contact force. The connecting map is the trace operator $\beta\colon H^1(0,\ell) \to \mathbb{R}$ given by $w \mapsto w(0)$ because we have contact only at the left endpoint of the interval $[0,\ell]$. The cone $K = \mathbb{R}_+$ to indicate that $u(t,0) \geq 0$. The operator $\mathcal{A} = -\partial^2/\partial x^2$. Note that the adjoint operator $\beta^*\colon W' = \mathbb{R} \to X' = H^{-1}(0,\ell)$ is simply multiplication by the Dirac-δ function $\beta^* N(t) = N(t)\delta(x)$, which is a force concentrated at the left endpoint of the rod. ∎

Example 6.2 (Vibrating string). Take $X = H_0^1(0,\ell)$ and $H = L^2(0,\ell)$ for the displacement field, but this time, take $W = H^1(\eta, \ell - \eta)$ for some $\eta > 0$. (See the compatibility constraint (6.66) and the discussion of it in the previous section for more explanation.) Take $Z = L^2(\eta, \ell - \eta)$. The connecting map $\beta\colon X \to W$ is the restriction to the subinterval: $w \mapsto w|_{(\eta,\ell-\eta)}$. The cone K is the cone of nonnegative functions in $H_0^1(0,\ell)$. The operator $\mathcal{A}$ is again $-\partial^2/\partial x^2$. The adjoint operator $\beta^*\colon W' = H^{-1}(\eta,\ell-\eta) \to X' = H_0^1(0,\ell)'$ is simply the extension of a distribution by zero:

$$\int_0^\ell \beta^* N(t,x)\,\psi(x)\,dx = \int_\eta^{\ell-\eta} N(t,x)\,\psi(x)\,dx. \quad ∎$$

Example 6.3 (Euler–Bernoulli beam). The equations of motion for an Euler–Bernoulli beam without contact are fourth order in space:

$$\rho A \frac{\partial^2 u}{\partial t^2} = -E I \frac{\partial^4 u}{\partial x^4} + f(t,x), \qquad 0 < x < \ell.$$

Here E is Young's modulus, A is the cross-sectional area of the beam, I is the second moment of area of the cross section of the beam, and ρ is the density. Since we have a fourth order equation, we need a higher order Sobolev space for the displacements. We

include in this example clamped boundary conditions at $x = 0$, where the displacement field $u(t,0) = \partial u/\partial x(t,0) = 0$, and free boundary conditions at $x = \ell$:

$$X = H^2_{cf}(0,\ell) = \left\{ w \in H^2(0,\ell) \mid w(0) = \partial w/\partial x(0) = 0 \right\}.$$

Again, $H = L^2(0,\ell)$. For contact at the free end, we take $W = Z = \mathbb{R}$ with $\beta\colon X \to W$ to be the trace operator given by $w \mapsto w(\ell)$. The cone $K = \mathbb{R}_+$ for the condition $u(t,\ell) - \varphi \geq 0$. If we have contact over the length of the beam, then we take $W = H^2(\eta,\ell)$, $Z = L^2(\eta,\ell)$ with $\beta\colon X \to W$ given by the restriction $w \mapsto w|_{(\eta,\ell)}$. In this case K is the set of nonnegative functions in $H^2(\eta,\ell)$. ∎

Example 6.4 (General linear frictionless elastic bodies). Here we take $X = H^1\left(\Omega;\mathbb{R}^d\right)$, where d is the dimension of the body (usually two or three) and $H = L^2(\Omega)$. The spaces for the contact forces are on the boundary $\Gamma_c \subseteq \partial\Omega$, so we take $W = H^{1/2}(\Gamma_c)$ and $Z = L^2(\Gamma_c)$ with the connecting map $\beta\colon X = H^1\left(\Omega;\mathbb{R}^d\right) \to W = H^{1/2}(\Gamma_c)$ given by $\beta\mathbf{w}(\mathbf{x}) = \mathbf{n}(\mathbf{x}) \cdot \mathbf{w}(\mathbf{x})$ for $\mathbf{x} \in \Gamma_c$. Note that, apart from the dot product with the unit outward normal vector, this is essentially the *trace* map $H^1(\Omega) \to H^{1/2}(\Gamma_c)$ onto a part of the boundary of Ω. To represent the condition that $\mathbf{n}(\mathbf{x}) \cdot \mathbf{u}(t,\mathbf{x}) - \varphi(\mathbf{x}) \geq 0$ on Γ_c, we take K to be the cone of nonnegative functions in $H^{1/2}(\Gamma_c)$. The operator $\mathcal{A}$ is the negative of the usual elasticity operator. For isotropic elasticity this is $\mathcal{A}\mathbf{u} = -(\lambda+\mu)\nabla(\nabla\cdot\mathbf{u}) - \mu\nabla^2\mathbf{u}$, where λ and μ are the usual Lamé parameters. Note that the adjoint operator $\beta^*\colon W' = H^{-1/2}(\Gamma_c) \to X' = H^{-1}\left(\Omega;\mathbb{R}^d\right)$ is multiplication by the surface measure ν_{Γ_c} times the outward normal vector $\mathbf{n}(\mathbf{x})$. Thus $\beta^* N(t,\mathbf{x})$ is a surface measure with values in $\mathbb{R}^d$. ∎

6.2.7 Proving existence

We can prove existence of solutions for problems in this framework (6.72)–(6.73) under some important conditions. These conditions are inspired by the one-dimensional examples above. Unfortunately, they do not extend to general frictionless elastic bodies.

The critical condition is that the dual cone K^* is strongly pointed; that is, K is a solid cone, with nonempty interior. This holds in the one-dimensional examples essentially because $H^1(0,\ell)$ and $H^2(0,\ell)$ are contained in the space of continuous functions, so that the norm is stronger than the supremum norm $\sup_{0\leq x\leq\ell}|f(x)|$. This means that cones of nonnegative functions can have nonempty interior. However, in $H^1(\Omega)$ for Ω a domain in $\mathbb{R}^d$ for $d \geq 2$, this is not the case. Similarly, $H^{1/2}(\partial\Omega)$ is contained in $C(\partial\Omega)$ if $d = 1$ but not for $d \geq 2$. At the time of this writing, there are as yet no existence proofs for general linearly elastic bodies contacting rigid obstacles. See the next section for more information.

Penalty approximation

We can create a penalty approximation to the rigid obstacle problem. For a closed convex set C, we can use the Yosida approximation for the normal cone function $u \mapsto \lambda^{-1}(u - \Pi_C(u))$ where Π_C is the nearest point projection in the pivot space H of the Gelfand triple. For the vibrating string problem, this gives exactly the penalty approximation used in Section 6.2.5. Since $u(t) - \varphi \in K$ for all t in our problem, we take $C = \varphi + K$. Then we have the differential equation

$$\frac{\partial^2 u_\epsilon}{\partial t^2} = -\mathcal{A}u_\epsilon(t) + f(t) - \frac{1}{\epsilon}\beta^*\left(\beta u_\epsilon - \Pi_{\varphi+K}(\beta u_\epsilon)\right). \tag{6.74}$$

To show the existence of solutions to the penalty differential equation, we use the same method as for the vibrating string problem: from the variation of parameters formula,

$$u_\epsilon(t) = \cos\left(\mathcal{A}^{1/2}t\right)u_0 + \mathcal{A}^{-1/2}\sin\left(\mathcal{A}^{1/2}t\right)v_0 + \int_0^t \mathcal{A}^{-1/2}\sin\left(\mathcal{A}^{1/2}(t-\tau)\right)\left[f(\tau) + \frac{1}{\epsilon}\psi\left[u_\epsilon(\tau)\right]\right]d\tau,$$

where $\psi[w] = -\beta^*\left(\beta w - \Pi_{\varphi+K}(\beta w)\right)$. Because $\beta^*\beta$ is a bounded operator $X \to X'$ but not $X \to X$ or even $X \to H$, we need a further step to show existence of solutions for the penalized problem. We use the Galerkin method. Let X_m be a finite-dimensional subspace of X for $m = 1, 2, 3, \ldots$ with $X_m \subset X_{m+1}$ for all m and $\overline{\bigcup_{m=1}^{\infty} X_m} = X$. We will use the particular choice $X_m = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_m\}$ where $(\mathcal{A}+\mathcal{A}^*)\phi_i = \lambda_i\phi_i$, where λ_i is the ith smallest eigenvalue of $\mathcal{A}+\mathcal{A}^*$. We look for solutions $u_{m,\epsilon}\colon [0,T] \to X_m$ which satisfy

$$\frac{d^2}{dt^2}\langle u_{\epsilon,m}, v\rangle = \left\langle -\mathcal{A}u_{\epsilon,m} + f(t) - \frac{1}{\epsilon}\beta^*\left(\beta u_{\epsilon,m} - \Pi_{\varphi+K}\left(\beta u_{\epsilon,m}\right)\right), v\right\rangle \tag{6.75}$$

for all $v \in X_m$. Note that we are using duality pairings, which are equivalent to inner products in H. Let $\Pi_m = \Pi_{X_m}$, the orthogonal (or nearest point) projection onto X_m using the inner product in H. Then (6.75) is equivalent to

$$\frac{d^2u_{\epsilon,m}}{dt^2} = \Pi_m\left(-\mathcal{A}u_{\epsilon,m} + f(t) - \frac{1}{\epsilon}\beta^*\left(\beta u_{\epsilon,m} - \Pi_{\varphi+K}\left(\beta u_{\epsilon,m}\right)\right)\right).$$

This is now a finite-dimensional ordinary differential equation with a Lipschitz right-hand side, and so it has solutions for initial conditions $u_{\epsilon,m}(0) = \Pi_m u_0$ and $du_{\epsilon,m}/dt(0) = \Pi_m v_0$.

Energy bounds

The energy functional for the penalty term has to be included as it was for the vibrating string problem:

$$E_\epsilon\left[u_{\epsilon,m}, \frac{\partial u_{\epsilon,m}}{\partial t}\right] = \frac{1}{2}\left\langle\frac{\partial u_{\epsilon,m}}{\partial t}, \frac{\partial u_{\epsilon,m}}{\partial t}\right\rangle + \frac{1}{2}\langle u_{\epsilon,m}, \mathcal{A}u_{\epsilon,m}\rangle + \frac{1}{2\epsilon}d\left(\beta u_{\epsilon,m}, \varphi+K\right)^2.$$

Note that $d(a, B) = \min_{b\in B}\|a-b\|_H$ is the distance from the point a to the (closed convex) set B; the distance is measured in the space H. The rate of change of the energy is then

$$\frac{d}{dt}E_\epsilon\left[u_{\epsilon,m}, \frac{\partial u_{\epsilon,m}}{\partial t}\right] = \left\langle\frac{\partial u_{\epsilon,m}}{\partial t}, \frac{\partial^2 u_{\epsilon,m}}{\partial t^2}\right\rangle + \left\langle\frac{\partial u_{\epsilon,m}}{\partial t}, \mathcal{A}u_{\epsilon,m}\right\rangle + \frac{1}{\epsilon}\left\langle\beta u_{\epsilon,m} - \Pi_C(\beta u_\epsilon), \frac{\partial}{\partial t}\left(\beta u_{\epsilon,m} - \Pi_C(\beta u_{\epsilon,m})\right)\right\rangle,$$

where $C = \varphi + K$. Now note that for any $w \in H$, $\langle w - \Pi_C(w), \Pi_C(w) - z\rangle \geq 0$ for all $z \in C$. So take $w = x(t)$ and $z = \Pi_C(x(t \pm h))$ with $h > 0$. Then,

$$\langle x(t) - \Pi_C(x(t)), \Pi_C(x(t)) - \Pi_C(x(t \pm h))\rangle \geq 0.$$

If t is a point of differentiability of $\Pi_C(x(t))$, then dividing by h and taking $h \downarrow 0$ give

$$\left\langle x(t) - \Pi_C(x(t)), \pm\frac{d}{dt}\Pi_C(x(t)) \right\rangle \geq 0.$$

That is, $\langle x(t) - \Pi_C(x(t)), (d/dt)\Pi_C(x(t))\rangle = 0$. In particular, since $\partial u_{\epsilon,m}/\partial t(t) \in X_m$ for all t,

$$\begin{aligned}
\frac{d}{dt}E_\epsilon\left[u_{\epsilon,m}, \frac{\partial u_{\epsilon,m}}{\partial t}\right] &= \left\langle \frac{\partial u_{\epsilon,m}}{\partial t}, \frac{\partial^2 u_{\epsilon,m}}{\partial t^2} \right\rangle + \left\langle \frac{\partial u_{\epsilon,m}}{\partial t}, \mathcal{A}u_{\epsilon,m} \right\rangle \\
&\qquad + \frac{1}{\epsilon}\left\langle \beta u_{\epsilon,m} - \beta\Pi_C(\beta u_{\epsilon,m}), \beta\frac{\partial u_{\epsilon,m}}{\partial t} \right\rangle \\
&= \left\langle \frac{\partial u_{\epsilon,m}}{\partial t}, -\mathcal{A}u_{\epsilon,m} - \frac{1}{\epsilon}\beta^*\left(\beta u_{\epsilon,m} - \Pi_C(\beta u_{\epsilon,m})\right) \right\rangle \\
&\qquad + \left\langle \frac{\partial u_{\epsilon,m}}{\partial t}, \mathcal{A}u_{\epsilon,m} \right\rangle + \frac{1}{\epsilon}\left\langle \beta u_{\epsilon,m} - \Pi_C(\beta u_{\epsilon,m}), \beta\frac{\partial u_{\epsilon,m}}{\partial t} \right\rangle \\
&= 0.
\end{aligned}$$

That is, $E_\epsilon\left[u_{\epsilon,m}, \partial u_{\epsilon,m}/\partial t\right]$ is constant, and so it is equal to the initial energy

$$\begin{aligned}
E_m^* &= \frac{1}{2}\left\langle \Pi_{X_m} v_0, \Pi_{X_m} v_0 \right\rangle + \frac{1}{2}\left\langle \Pi_{X_m} u_0, \mathcal{A}\Pi_{X_m} u_0 \right\rangle \\
&\leq \frac{1}{2}\langle v_0, v_0\rangle + \frac{1}{2}\langle u_0, \mathcal{A}u_0\rangle = E^*
\end{aligned}$$

for $X_m = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_m\}$. We do not really need $X_m = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_m\}$, just that the projections $\Pi_m \colon H \to X_m$ have uniformly bounded norms as operators $X \to X_m \subset X$.

Once we have bounds $E_\epsilon\left[u_{\epsilon,m}, \partial u_{\epsilon,m}/\partial t\right] \leq E^*$ we automatically have bounds for $E\left[u_{\epsilon,m}, \partial u_{\epsilon,m}/\partial t\right] = \frac{1}{2}\left\|\partial u_{\epsilon,m}/\partial t\right\|_H^2 + \frac{1}{2}\left\langle u_{\epsilon,m}, \mathcal{A}u_{\epsilon,m}\right\rangle$ that are independent of $\epsilon > 0$ and m. These are the energy bounds that we need.

Momentum and contact impulse bounds

Here we need to use strong pointedness of K^*. With this assumption there is a $w \in K$ where

$$\langle w, \psi\rangle \geq \|\psi\|_{X'} \qquad \text{for all } \psi \in K^*.$$

Actually we need something a little stronger:

$$\beta^*\left(K^*\right) \cap X_{-\theta} \quad \text{strongly pointed in } X_{-\theta} \text{ with } 0 < \theta < 1, \tag{6.76}$$

where $X_{-\theta} = X_\theta'$ is the dual interpolation space as defined in Section 3.3.2. Thus we want a $w \in K$ where

$$\langle w, \psi\rangle \geq \|\psi\|_{X_{-\theta}} \qquad \text{for all } \psi \in \beta^*\left(K^*\right) \cap X_{-\theta}. \tag{6.77}$$

Note that a consequence of this is that $\beta^*(K^*) \subset X_{-\theta}$.

We can, without loss of generality, take $w \in X_m$ for sufficiently large m. Since $\Pi_m w \to w$ in X_θ as $m \to \infty$, for sufficiently large m, $\|w - \Pi_m w\|_{X_\theta} < 1$, and we can then scale $\Pi_m w$ by $1/\left(1 - \|w - \Pi_m w\|_{X_\theta}\right)$ to obtain the strong pointedness condition (6.77).

Now we can take the penalty equations and take duality pairings with w:

$$\begin{aligned}\left\langle w, \frac{\partial^2 u_{\epsilon,m}}{\partial t^2}\right\rangle &= \langle w, -\mathcal{A}u_{\epsilon,m}\rangle + \langle w, \beta^* N_{\epsilon,m}(t)\rangle \\ &= -\langle \mathcal{A}w, u_{\epsilon,m}\rangle + \langle w, \beta^* N_{\epsilon,m}(t)\rangle.\end{aligned}$$

Integrating over $[0,t]$ gives

$$\left\langle w, \frac{\partial u_{\epsilon,m}}{\partial t}(t)\right\rangle - \langle w, v_0\rangle = -\int_0^t \langle \mathcal{A}w, u_{\epsilon,m}(\tau)\rangle\, d\tau + \int_0^t \langle w, \beta^* N_{\epsilon,m}(\tau)\rangle\, d\tau.$$

The left-hand side is bounded from the energy bounds by $2\|w\|_H (E^*)^{1/2}$. Also the integrand $\langle \mathcal{A}w, u_{\epsilon,m}(\tau)\rangle$ is bounded by $\|\mathcal{A}w\|_{X'} \|u_{\epsilon,m}(\tau)\|_X \le C\|\mathcal{A}w\|_{X'}(E^*)^{1/2}$ for some constant C independent of ϵ and τ. This leaves $\int_0^t \langle w, \beta^* N_{\epsilon,m}(\tau)\rangle\, d\tau$ bounded. Using (6.77), we then have a bound

$$\int_0^T \|\beta^* N_{\epsilon,m}(\tau)\|_{X_{-\theta}}\, d\tau \le \text{constant}.$$

Thus, by Alaoglu's theorem, we have weak* convergence of a subsequence in the space of measures with values in $X_{-\theta}$, $\mathcal{M}(0,T;X_{-\theta})$.

Taking limits

Take a subsequence of $\epsilon \downarrow 0$ where $\beta^* N_{\epsilon,m} \rightharpoonup^* \beta^* \widehat{N}$ weakly* in $\mathcal{M}(0,T;X_{-\theta})$. Now, by Seidman's theorem (Theorem A.6), since $u_{\epsilon,m}$ is uniformly bounded in $L^\infty(0,T;X)$ and $\partial u_{\epsilon,m}/\partial t$ is uniformly bounded in $L^\infty(0,T;H)$, we can find a further subsequence $u_{\epsilon,m} \to \widehat{u}$ in $C(0,T;X_\theta)$, as $X = X_1$ is compactly imbedded in X_θ for $0 < \theta < 1$.

Now we show that $\beta\widehat{u}(t) - \varphi \in K$ for all (or almost all) t, and that $\widehat{N}$ is a measure with values in K^*, or equivalently, $\beta^*\widehat{N}$ has values in $\beta^*(K^*)$.

Now we can further restrict our attention to subsequences where $u_{\epsilon,m} \rightharpoonup \widehat{u}$ weakly in $L^p(0,T;X)$ for any $1 < p < \infty$. Since these are reflexive Banach spaces, weak and weak* convergence are the same in them. Consider the functional

$$\Psi[u] = \int_0^T \frac{1}{2} d(\beta u(t), \varphi + K)^2\, dt,$$

where $d(x,C)$ is the distance from x to C using the norm for H. This is a convex continuous functional on $L^2(0,T;X)$, and so by Mazur's lemma it is weakly lower semicontinuous on this space. Also, from the energy bounds, $\Psi[u_{\epsilon,m}] \le \text{constant}\,\epsilon^{1/2}$. Thus taking limits in the weakly convergent subsequence gives $\Psi[\widehat{u}] \le 0$. Since $\Psi[u] \ge 0$ for all u, it follows that $\Psi[\widehat{u}] = 0$, and that $d(\beta\widehat{u}(t), \varphi + K) = 0$ for almost all t; as $\varphi + K$ is closed, this means that $\beta\widehat{u}(t) - \varphi \in K$ for almost all t.

To see that $\widehat{N}$ is a measure with values in K^*, suppose that $z\colon [0,T] \to K$ is continuous. Since $\beta^* N_\epsilon \rightharpoonup^* \beta^* \widehat{N}$ weakly* in $\mathcal{M}(0,T; X_{-\theta}) \subset \mathcal{M}\big(0,T; X'\big)$,

$$0 \le \int_0^T \big\langle \beta^* N_\epsilon(t), z(t) \big\rangle \, dt \to \int_{[0,T]} \big\langle \beta^* \widehat{N}(t), z(t) \big\rangle \, dt.$$

Since this is true for all continuous $z\colon [0,T] \to K$, $\widehat{N}$ is a measure with values in K^*.

Finally, we have to show that $\int_0^T \big\langle \widehat{u}(t) - \varphi, \beta^* \widehat{N}(t) \big\rangle \, dt = 0$. Now we know that $\big\langle u_{\epsilon,m}(t) - \varphi, \beta^* N_\epsilon(t) \big\rangle \le 0$, as the definition of $N_{\epsilon,m}$ implies that $N_{\epsilon,m}(t) \in N_K\big(\beta u_{\epsilon,m}(t) - \varphi\big)$, using the inner product in H. However, we already have $u_{\epsilon,m} \to \widehat{u}$ strongly in $C(0,T; X_\theta)$ and $\beta^* N_{\epsilon,m} \rightharpoonup^* \beta^* \widehat{N}$ in $\mathcal{M}(0,T; X_{-\theta})$. Thus

$$\begin{aligned} 0 \;&\ge\; \int_0^T \big\langle u_{\epsilon,m}(t) - \varphi, \beta^* N_{\epsilon,m}(t) \big\rangle \, dt \\ &\to \int_0^T \big\langle \widehat{u}(t) - \varphi, \beta^* \widehat{N}(t) \big\rangle \, dt \;\ge\; 0, \end{aligned}$$

so we obtain complementarity in the limit.

6.2.8 General elastic bodies

The problem of impact of three-dimensional elastic bodies with rigid obstacles still remains out of reach. However, there are a number of partial results and techniques that are important for handling this and related problems.

An important technique for handling boundary thin obstacle problems is to remove the normal contact forces from consideration. If we have a complementarity formulation

$$\begin{aligned} \rho \frac{\partial^2 u}{\partial t^2} &= -\mathcal{A}u + f(t) + \beta^* N(t), \\ K^* \ni N(t) \perp \beta u(t) - \varphi \in K \qquad &\text{for all } t, \end{aligned}$$

then if K^* is *not* strongly pointed (or equivalently, if K is not a solid cone), we cannot get the bounds on $\int \|N(t)\| \, dt$ needed to complete the proof of existence. In fact, $\int \|N(t)\| \, dt$ may be infinite, so that N is not a measure of bounded variation. Rather than try to work through the theory of measures that have only weakly bounded variation (where $\int |\langle \psi, N(t) \rangle| \, dt \le C$ for all ψ with $\|\psi\| \le 1$), we can reformulate boundary contact problems as VIs. In fact, even in other circumstances, this is a common technique [56, 120, 152, 154].

If we set

$$\widetilde{K} = \{\, w \in X \mid \beta w \in K \,\},$$

then $\widetilde{K}$ is a closed convex cone in X and our CP becomes

$$\widetilde{K}^* \ni \beta^* N(t) \perp u(t) - \widetilde{\varphi} \in \widetilde{K} \qquad \text{for all } t,$$

where $\beta\widetilde{\varphi} = \varphi$. That is, $\widetilde{\varphi}$ is an extension of φ. The equivalent VI is that

$$\begin{aligned} &u(t) \in \widetilde{\varphi} + \widetilde{K} \quad \text{for all } t \qquad \& \\ &0 \le \left\langle w(t) - u(t), \rho \frac{\partial^2 u}{\partial t^2}(t) + \mathcal{A}u(t) - f(t) \right\rangle \quad \text{for all } t \text{ and } w(t) \in \widetilde{K}. \end{aligned}$$

Let $\mathcal{K} = \left\{ z \in W^{1,p}(0,T;X) \mid z(t) \in \widetilde{\varphi} + \widetilde{K} \text{ for all } t \right\}$ for some p which will be fixed later. An integral version of the VI is

$$u \in \mathcal{K} \quad \& \quad \int_0^T \psi(t) \left\langle w(t) - u(t), \rho \frac{\partial^2 u}{\partial t^2}(t) + \mathcal{A}u(t) - f(t) \right\rangle dt \ \geq \ 0$$

for all $w \in \mathcal{K}$ and smooth $\psi \colon [0,T] \to \mathbb{R}_+$. Now assume that $w \colon [0,T] \to \widetilde{K}$ is smooth and $\psi(T) = 0$. We can later shift T if we need to avoid technical issues having to do with the solution at $t = T$. Now we can use integration by parts to reformulate the problem as

$$\begin{aligned} & u \in K \quad \& \\ & 0 \leq \psi(0) \langle u_0 - w(0), \rho v_0 \rangle - \int_0^T \psi(t) \left\langle \frac{\partial w}{\partial t}(t) - \frac{\partial u}{\partial t}(t), \rho \frac{\partial u}{\partial t}(t) \right\rangle dt \\ & \qquad + \int_0^T \psi(t) \langle w(t) - u(t), \mathcal{A}u(t) - f(t) \rangle \, dt \\ & \qquad - \int_0^T \psi'(t) \left\langle w(t) - u(t), \rho \frac{\partial u}{\partial t}(t) \right\rangle dt \qquad \text{for all } w \in \mathcal{K}. \end{aligned} \tag{6.78}$$

With this reformulation, we have removed the normal contact forces, and we have only first order derivatives in time appearing in the integrands.

We could attempt to repeat the process described in the previous section, but since we do not have N, we do not (apparently) need strong pointedness. Then we should be able to prove existence of solutions. We need to check what will happen with a few quadratic integrals with the weak convergence $u_{\epsilon,m} \rightharpoonup \widehat{u}$ in $L^p(0,T;X)$ and $\partial u_{\epsilon,m}/\partial t \rightharpoonup \partial \widehat{u}/\partial t$ in $L^p(0,T;H)$. These integrals are

$$\begin{aligned} & + \int_0^T \psi(t) \left\langle \frac{\partial u}{\partial t}(t), \rho \frac{\partial u}{\partial t}(t) \right\rangle dt, \\ & - \int_0^T \psi(t) \, \langle u(t), \mathcal{A}u(t) \rangle \, dt, \\ & - \int_0^T \psi'(t) \left\langle u(t), \rho \frac{\partial u}{\partial t}(t) \right\rangle dt. \end{aligned}$$

The third is not a problem since by compactness of X in H we can use Seidman's theorem (Theorem A.6) to show that $u_{\epsilon,m} \to \widehat{u}$ strongly in $C(0,T;H)$. This combined with weak convergence $\partial u_{\epsilon,m}/\partial t \rightharpoonup \partial \widehat{u}/\partial t$ in $L^p(0,T;H)$ for any $1 < p < \infty$ gives convergence of this integral. The second integral is also not a problem since the integral is a concave function of u, so by Mazur's lemma,

$$\limsup_{\epsilon \downarrow 0,\, m \to \infty} \ - \int_0^T \psi(t) \left\langle u_{\epsilon,m}(t), \mathcal{A}u_{\epsilon,m}(t) \right\rangle dt \leq - \int_0^T \psi(t) \, \langle \widehat{u}(t), \mathcal{A}\widehat{u}(t) \rangle \, dt,$$

which keeps the inequalities pointing in the same direction, which would allow us to show that the VI holds in the limit. On the other hand, the first integral is convex, so using Mazur's lemma would result in inequalities going in the wrong direction. If we could show

that the velocities $\partial u_{\epsilon,m}/\partial t$ were uniformly bounded in X_θ for any $\theta > 0$, we would be able to obtain convergent subsequences and hence show the existence of solutions in the limit.

Some existence theorems of this type have been shown. However, there are usually some special limitations, or the problem is modified, usually by incorporating Kelvin–Voigt viscoelasticity. Introducing Kelvin–Voigt viscoelasticity makes the viscoelasticity operator essentially a parabolic operator, so that the solutions of the viscoelastic system tend to be much smoother. However, as we will see, it also makes the contact forces more singular in response. This can complicate the analysis.

6.2.9 Wave equation: Existence via compensated compactness

This existence result is due to Kim [146]. The problem here is apparently fairly general. The wave equation applies in the domain, and we have contact on at least part of the boundary. It is however a scalar system. More formally,

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= \nabla^2 u \qquad \text{in } \Omega, \\ N(t,\mathbf{x}) &= -\frac{\partial u}{\partial n}(t,\mathbf{x}), \qquad \mathbf{x} \in \partial\Omega, \\ 0 \le N(t,\mathbf{x}) \perp u(t,\mathbf{x}) - \varphi(\mathbf{x}) &\ge 0, \qquad \mathbf{x} \in \partial\Omega, \end{aligned}$$

with given initial values $u(0,\mathbf{x}) = u_0(\mathbf{x})$, $\partial u/\partial t(0,\mathbf{x}) = v_0(\mathbf{x})$, and with fixed displacement ($u(t,\mathbf{x}) = g(\mathbf{x})$ for $\mathbf{x} \in \Gamma_d$) or given traction ($\partial u/\partial n(t,\mathbf{x}) = \tau(\mathbf{x})$ for $\mathbf{x} \in \Gamma_t$) boundary conditions on the remainder of the boundary.

Kim starts in the same way as is done above: use a sequence of finite-dimensional Galerkin approximations and obtain energy bounds for these approximations. This shows that the Galerkin approximations $u_{\epsilon,m}$ are uniformly bounded in time in the energy space $H^1(\Omega)$. Since Kim uses the VI formulation, he is not concerned about bounds for the normal contact forces N. By Alaoglu's theorem, there is a weak* limit $\widehat{u}$ of a suitable subsequence of $u_{\epsilon,m}$. Taking weak limits as $m \to \infty$ for $\epsilon > 0$ fixed gives a solution u_ϵ of the wave equation satisfying the penalty approximation.

The problem is now to show that the weak* limit $u_\epsilon \rightharpoonup^* \widehat{u}$ satisfies the VI (6.78). The problem, as mentioned above, is that $\int_\Omega (\partial u_\epsilon/\partial t)^2\,dx$ and $\langle u_\epsilon, \mathcal{A}u_\epsilon\rangle = \int_\Omega |\nabla u_\epsilon|^2\,dx$ do not converge to the appropriate limits under weak or weak* convergence. However, Kim invokes the *div-curl lemma* of *compensated compactness*, which implies, from the fact that $\partial^2 u_\epsilon/\partial t^2 - \nabla^2 u_\epsilon = 0$ and the boundedness of $\partial u_\epsilon/\partial t$ and ∇u_ϵ in $L^2(\Omega)$, that $\int_0^T \int_\Omega \psi[(\partial u_\epsilon/\partial t)^2 - |\nabla u_\epsilon|^2]dx\,dt$ *does converge to* $\int_0^T \int_\Omega \psi[(\partial u/\partial t)^2 - |\nabla u|^2]dx\,dt$, at least for ψ smooth with compact support in Ω. This deals with the difficult integrals in the VI (6.78). There are some additional technicalities in dealing with the fact that ψ has to be zero in a neighborhood of the boundary $\partial\Omega$. Kim deals with this by showing that the integrals close to the boundary are small as the thickness of this boundary layer goes to zero. The result is that the limit indeed satisfies (6.78).

If the div-curl lemma could be generalized to the elasticity operators (even for just the isotropic constant coefficient version), then frictionless contact for elastic bodies would be a solved problem. Although considerable effort has been put into this, as yet there is no sign that this can be done. Other ideas are needed to find a solution to this problem.

6.2.10 Wave equation: In a half-space

A different approach was taken by Lebeau and Schatzman [156], which treats the wave equation in a half-space $\Omega = \mathbb{R}_+ \times \mathbb{R}^{n-1}$ with Signorini contact conditions on $\partial\Omega = \{0\} \times \mathbb{R}^{n-1}$. For $\mathbf{x} \in \overline{\Omega}$, put $\mathbf{x} = (x_1, \mathbf{x}')$ with $\mathbf{x}' \in \mathbb{R}^{n-1}$. The basic idea is to use the Neumann to Dirichlet operator for the wave equation, representing the operator in terms of Fourier integrals. The approach below is a loose description of their results; for more details see their paper.

The wave equation can be written as

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x_1^2} + \nabla_{\mathbf{x}'}^2 u.$$

Taking Fourier transforms of this equation with respect to t and $\mathbf{x}'$ gives

$$-\omega^2 \widehat{u}(\omega, x_1, \boldsymbol{\xi}') = \frac{\partial^2 \widehat{u}}{\partial x_1^2}(\omega, x_1, \boldsymbol{\xi}') - |\boldsymbol{\xi}'|^2 \widehat{u}(\omega, x_1, \boldsymbol{\xi}').$$

That is,

$$\frac{\partial^2 \widehat{u}}{\partial x_1^2}(\omega, x_1, \boldsymbol{\xi}') = \left(|\boldsymbol{\xi}'|^2 - \omega^2\right) \widehat{u}(\omega, x_1, \boldsymbol{\xi}').$$

Solving this differential equation in x_1 gives

$$\begin{aligned}\widehat{u}(\omega, x_1, \boldsymbol{\xi}') = \alpha_+(\omega, \boldsymbol{\xi}') \exp\left(+\left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2} x_1\right) \\ + \alpha_-(\omega, \boldsymbol{\xi}') \exp\left(-\left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2} x_1\right).\end{aligned}$$

Using the convention that $\operatorname{Re}\left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2} \geq 0$, which is equivalent to choosing the principal branch of the complex square root function, we must have $\alpha_+(\omega, \boldsymbol{\xi}') = 0$ for $\widehat{u}(\omega, x_1, \boldsymbol{\xi}')$ to be a tempered distribution. Thus we write

$$\widehat{u}(\omega, x_1, \boldsymbol{\xi}') = \alpha(\omega, \boldsymbol{\xi}') \exp\left(-\left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2} x_1\right).$$

Now $-\partial u(t, \mathbf{x})/\partial x_1(t, 0, \mathbf{x}') = N(t, \mathbf{x}')$, so

$$\begin{aligned}\mathcal{F}N(\omega, \boldsymbol{\xi}') &= \alpha(\omega, \boldsymbol{\xi}') \left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2}, \qquad \text{so that} \\ \widehat{u}(\omega, x_1, \boldsymbol{\xi}') &= \frac{\mathcal{F}N(\omega, \boldsymbol{\xi}')}{\left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2}} \exp\left(-\left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{1/2} x_1\right).\end{aligned}$$

In particular,

$$\widehat{u}(\omega, 0, \boldsymbol{\xi}') = \left(|\boldsymbol{\xi}'|^2 - \omega^2\right)^{-1/2} \mathcal{F}N(\omega, \boldsymbol{\xi}').$$

The map $N(\cdot,\cdot) \mapsto u(\cdot,0,\cdot)$ is the *Neumann to Dirichlet operator* for the wave equation on the half-space $\left\{ \left(x_1, \mathbf{x}'\right) \mid x_1 \geq 0 \right\}$. Care must be taken with this Fourier representation of the Neumann to Dirichlet operator, as $\left(\left|\boldsymbol{\xi}'\right|^2 - \omega^2\right)^{-1/2}$ has a branch cut with endpoints at $\omega = \pm|\boldsymbol{\xi}'|$. To make sure that we stay on the principal branch, we can replace ω with the limit $\omega + ia$ as $a \downarrow 0$. Apart from a factor of e^{-at}, the operator represented by the Fourier multiplier $\left(\left|\boldsymbol{\xi}'\right|^2 - \left(\omega + ia\right)^2\right)^{-1/2}$ is independent of $a > 0$, thanks to the Cauchy residue theorem of complex analysis. It is particularly important to choose the correct branch, as one choice gives the causal operator ($u(t,0,\mathbf{x}')$ depends on $N(\tau,\mathbf{y}')$ for $\tau < t$), while the other is anticausal ($u(t,0,\mathbf{x}')$ depends on $N(\tau,\mathbf{y}')$ for $\tau > t$). We choose, of course, the causal operator.

In fact, Lebeau and Schatzman do not deal with the Neumann to Dirichlet operator, but rather with its inverse, the *Dirichlet to Neumann operator*, which is represented by the Fourier multiplier $\left(\left|\boldsymbol{\xi}'\right|^2 - \omega^2\right)^{1/2}$, understood using the principal branch of the square root function applied to $|\boldsymbol{\xi}'|^2 - (\omega + ia)^2$ with $a > 0$. As $\omega \to \pm\infty$, $\left(\left|\boldsymbol{\xi}'\right|^2 - \omega^2\right)^{1/2} \sim i\omega$. The difference is

$$\left(\left|\boldsymbol{\xi}'\right|^2 - \omega^2\right)^{1/2} - i\omega = \frac{\left|\boldsymbol{\xi}'\right|^2}{\left(\left|\boldsymbol{\xi}'\right|^2 - \omega^2\right)^{1/2} + i\omega}.$$

The denominator (using the principal branch) is bounded above and below by multiples of $\max\left(|\omega|, \left|\boldsymbol{\xi}'\right|\right)$. If $\mathcal{A}$ is the Dirichlet to Neumann operator, then we note that $i\omega$ is the Fourier multiplier for the operator $\partial/\partial t$. The operator

$$\mathcal{C} = \mathcal{A} - \frac{\partial}{\partial t}$$

is represented by the Fourier multiplier

$$\left|\boldsymbol{\xi}'\right|^2 / \left[\left(\left|\boldsymbol{\xi}'\right|^2 - \omega^2\right)^{1/2} + i\omega\right].$$

The operator $\mathcal{A}$ can be restricted in time to form $\mathcal{A}_T$ given by

$$\mathcal{A}_T u = \left(\mathcal{A}[E_T u]\right) \chi_{[0,T]}$$

with E_T the operator that extends a function with domain $[0,T]$ to domain $\mathbb{R}$ by zero. The corresponding operator $\mathcal{C}_T$ is a (causal) operator

$$L^2(0,T; H^{1/2}(\mathbb{R}^{n-1})) \to L^2(0,T; H^{-1/2}(\mathbb{R}^{n-1})).$$

While neither $\mathcal{A}_T$ nor $\mathcal{C}_T$ is elliptic on a suitable space, there is a positivity property for $\mathcal{A}_T$:

$$\begin{aligned}\langle w, \mathcal{A}_T w\rangle = &\int_{\{(\omega,\boldsymbol{\xi}') \mid |\boldsymbol{\xi}'| \geq |\omega|\}} \sqrt{\left|\boldsymbol{\xi}'\right|^2 - \omega^2}\, \left|\widehat{w}(\omega,\boldsymbol{\xi}')\right| \, d\omega\, d\boldsymbol{\xi}' \\ &+ \frac{1}{2}\int_{\mathbb{R}^{n-1}} \left|w(T,\mathbf{x}')\right|^2 \, d\mathbf{x}'.\end{aligned}$$

We use a penalty approximation which we can represent via the trace operator $\beta\colon H^1(\mathbb{R}^{n-1}\times\mathbb{R}_+)\to H^{1/2}(\mathbb{R}^{n-1})$ as

$$\frac{\partial^2 u_\epsilon}{\partial t^2} = \nabla^2 u_\epsilon + \beta^* N_\epsilon,$$
$$0 \le \beta u_\epsilon(t) - \varphi \perp \epsilon\, \beta u_\epsilon(t) + N_\epsilon(t) \ge 0.$$

By means of energy bounds for the penalty approximation we can show that $\partial u_\epsilon/\partial t$ and ∇u_ϵ are uniformly bounded in $L^2(0,T;L^2(\mathbb{R}^{n-1}\times\mathbb{R}_+))$, and therefore $w_\epsilon := \beta u_\epsilon$ are uniformly bounded in $L^2(0,T;H^{1/2}(\mathbb{R}^{n-1}))$ and $\partial w_\epsilon/\partial t$ are uniformly bounded in $L^2(0,T;H^{-1/2}(\mathbb{R}^{n-1}))$. Thus there is a subsequence $\epsilon\downarrow 0$ in which $w_\epsilon \rightharpoonup \widehat{w}$ weakly in $L^2(0,T;H^{1/2}(\mathbb{R}^{n-1}))$ and $\partial w_\epsilon/\partial t \rightharpoonup \partial\widehat{w}/\partial t$ weakly in $L^2(0,T;H^{-1/2}(\mathbb{R}^{n-1}))$. Since $w\mapsto\langle w,\mathcal{A}_T w\rangle$ is a nonnegative quadratic function, it is convex, and so by Mazur's lemma,

$$\langle\widehat{w},\mathcal{A}_T\widehat{w}\rangle \le \liminf_{\epsilon\downarrow 0}\langle w_\epsilon,\mathcal{A}_T w_\epsilon\rangle\,.$$

Thus, taking limits in the subsequence,

$$\begin{aligned} 0 &= \liminf_{\epsilon\downarrow 0}\langle w_\epsilon-\varphi,\mathcal{A}_T w_\epsilon\rangle \\ &\ge \langle\widehat{w}-\varphi,\mathcal{A}_T\widehat{w}\rangle \;\ge\; 0, \end{aligned}$$

with the last inequality holding because we can show $\widehat{w}-\varphi\ge 0$ and $\mathcal{A}_T\widehat{w}\ge 0$ as weak limits of nonnegative functions and distributions are nonnegative. Thus we have

$$0\le\widehat{w}-\varphi\perp\mathcal{A}_T\widehat{w}\ge 0,$$

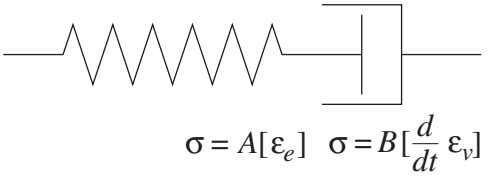
and we have a solution of the contact problem for the wave equation on a half-space.

Not only can we show that there is a solution, but it can also be shown that energy is conserved. This can be done by writing the work done on the half-space by the contact force in terms of $\widehat{w}$ and $\mathcal{A}_T\widehat{w}$ and using a differentiation lemma for CPs.

It is difficult to extend this approach to other problems. Firstly, to apply the method to something other than a half-space or a slab requires much more difficult and delicate Fourier analysis. The technique of "straightening the boundary" can be applied to this problem, but the resulting partial differential equation has varying coefficients and can also have infinitely many reflections in finite time. On the other hand, even if we keep the half-space geometry but try to apply the method to the equations of elasticity, there is the problem of multiple wave speeds, and it is unclear what

$$\int_{\{(\omega,\boldsymbol{\xi}')\,|\,|\boldsymbol{\xi}'|\ge|\omega|\}}\sqrt{|\boldsymbol{\xi}'|^2-\omega^2}\,\left|\widehat{w}(\omega,0,\boldsymbol{\xi}')\right|\,d\omega\,d\boldsymbol{\xi}'$$

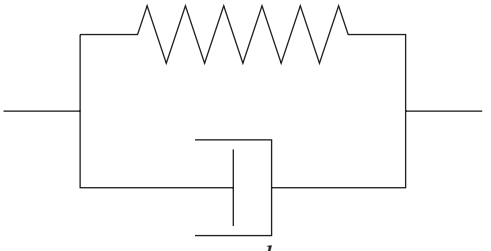
should be replaced with in the positivity result. The positivity result could be weakened if we can show that the negative part is in some way "compact" with respect to the main "positive" part of the operator.



$$\sigma = A[\varepsilon_e] \quad \sigma = B[\frac{d}{dt}\,\varepsilon_v]$$

(a) Maxwell model

$$\sigma_e = A[\varepsilon]$$



$$\sigma_v = B[\frac{d}{dt}\,\varepsilon]$$

(b) Kelvin–Voigt model

Figure 6.10: Maxwell and Kelvin–Voigt models of viscoelasticity.

At the time of this writing, there is no proof of existence of solutions to the equations of elasticity with Signorini contact conditions in more than one dimension. However, if we allow viscoelasticity of a Kelvin–Voigt type, we can go much further, as we will see in the next section.

6.3 Viscoelastic bodies

Practical dynamic models of real elastic materials have to include the effects of viscosity. Viscosity is due to energy losses arising from the rate of "stretching" or "compression" of the material. There are a number of models of viscoelasticity, the simplest of which are the Maxwell and Kelvin–Voigt models. These are often represented in continuum mechanics textbooks with a spring (representing elastic forces) either in series (Maxwell model) or in parallel (Kelvin–Voigt model) with a damper (representing viscosity). This is illustrated in Figure 6.10. These can be represented in terms of the relationship between the stress (σ) and strain (ε) tensors.

If the spring and damper are in series as in the Maxwell model, then the strain or deformation tensor $\varepsilon = \varepsilon_e + \varepsilon_v$, the sum of the strain tensors for the elastic and viscous deformation, but the stress tensor is the same for both the elastic and viscous terms. If the spring and damper are in parallel as in the Kelvin–Voigt model, the stress tensor $\sigma = \sigma_e + \sigma_v$, the sum of the elastic and viscous stresses, but the strain tensors are the same for both the elastic and viscous terms. The operators A and B shown in Figure 6.10 are linear operators mapping tensors to tensors $\mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ satisfying the same symmetry and positive definiteness properties as for the pure elasticity operator as discussed in Section 6.2. Note that

we can represent A and B in terms of components:

$$(A\,\varepsilon)_{ij} = \sum_{k,l} a_{ijkl}\varepsilon_{kl},$$
$$\left(B\,\frac{\partial \varepsilon}{\partial t}\right)_{ij} = \sum_{k,l} b_{ijkl}\frac{\partial \varepsilon_{kl}}{\partial t}.$$

In terms of components, the symmetry and positive definiteness conditions are

$$a_{ijkl} = a_{jikl} = a_{ijlk} = a_{klij},$$
$$b_{ijkl} = b_{jikl} = b_{ijlk} = b_{klij},$$
$$\sum_{i,j,k,l} \varepsilon_{ij} a_{ijkl}\varepsilon_{kl} \geq \eta_A \sum_{i,j} \varepsilon_{ij}\varepsilon_{ij},$$
$$\sum_{i,j,k,l} \frac{\partial \varepsilon_{ij}}{\partial t} b_{ijkl} \frac{\partial \varepsilon_{kl}}{\partial t} \geq \eta_B \sum_{i,j} \frac{\partial \varepsilon_{ij}}{\partial t}\frac{\partial \varepsilon_{ij}}{\partial t}$$

with $\eta_A, \eta_B > 0$.

The Maxwell model can be formulated as follows: $\varepsilon = \varepsilon_e + \varepsilon_v$ with $\sigma = A\,\varepsilon_e = B\,\partial\varepsilon_v/\partial t$ so that $\partial\varepsilon_v/\partial t = B^{-1}A\,\varepsilon_e$. Then the total rate of strain tensor is $\partial\varepsilon/\partial t = \partial\varepsilon_e/\partial t + \partial\varepsilon_v/\partial t = A^{-1}\,\partial\sigma/\partial t + B^{-1}\sigma$. Solving this differential equation for σ gives

$$\sigma(t,x) = e^{-AB^{-1}t}\,\sigma(0,x) + \int_0^T e^{-AB^{-1}(t-\tau)}\,A\,\frac{\partial \varepsilon}{\partial t}(\tau,x)\,d\tau. \tag{6.79}$$

Recall the total strain tensor $\varepsilon = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$ where $\mathbf{u}$ is the (total) displacement field. Then we can use σ in the standard equations for the displacement field $\mathbf{u}(t,\mathbf{x})$:

$$\rho\frac{\partial \mathbf{u}}{\partial t} = \operatorname{div}\sigma + \mathbf{f}(t,\mathbf{x}).$$

Note that the short-time behavior of the Maxwell model is essentially the same as the purely elastic model; over time there is dissipation of energy. The theoretical behavior of Maxwell viscoelasticity in terms of existence of solutions is essentially the same as for pure elasticity, and it suffers the same kind of difficulties.

On the other hand, the Kelvin–Voigt model is easier to formulate:

$$\begin{aligned}\rho\frac{\partial \mathbf{u}}{\partial t} &= \operatorname{div}\sigma + \mathbf{f}(t,\mathbf{x}) \\ &= \operatorname{div}\left(A\,\varepsilon + B\,\frac{\partial \varepsilon}{\partial t}\right) + \mathbf{f}(t,\mathbf{x}).\end{aligned}$$

From the mathematical viewpoint, the advantage of the Kelvin–Voigt model is that it is essentially a parabolic partial differential equation instead of a hyperbolic differential equation. That means that the solution operator for the Kelvin–Voigt model is a compact operator. This makes it easier to prove existence of solutions for frictionless impact problems, but uniqueness is still beyond reach.

6.3.1 Frictionless impact for Kelvin–Voigt viscoelastic bodies

For displacement field $u\colon \Omega \to \mathbb{R}^d$ with Ω a domain in $\mathbb{R}^d$, the linearized strain tensor is $\varepsilon[\mathbf{u}] = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$, which defines the linearized strain *operator*. We can define the differential operators $\mathcal{A}$ and $\mathcal{B}$ from $H^1(\Omega)$ to $H^{-1}(\Omega)$ given by

$$\mathcal{A}\mathbf{u} = -\operatorname{div}(A\,\varepsilon[\mathbf{u}]), \tag{6.80}$$

$$\mathcal{B}\mathbf{v} = -\operatorname{div}(B\,\varepsilon[\mathbf{v}]). \tag{6.81}$$

For boundary conditions we use

$$\mathbf{u}(t,\mathbf{x}) = \mathbf{g}(\mathbf{x}), \qquad \mathbf{x} \in \Gamma_D, \tag{6.82}$$

$$\sigma(t,\mathbf{x})\mathbf{n}(\mathbf{x}) = \tau(t,\mathbf{x}), \qquad \mathbf{x} \in \Gamma_N, \tag{6.83}$$

$$\sigma(t,\mathbf{x})\mathbf{n}(\mathbf{x}) = -N(t,\mathbf{x})\mathbf{n}(\mathbf{x}), \qquad \mathbf{x} \in \Gamma_C, \tag{6.84}$$

$$0 \le N(t,\mathbf{x}) \perp -\mathbf{n}(\mathbf{x})\cdot\mathbf{u}(t,\mathbf{x}) + \varphi(\mathbf{x}) \ge 0, \qquad \mathbf{x} \in \Gamma_C. \tag{6.85}$$

A formal representation of the linearized Kelvin–Voigt viscoelastic impact problem without friction as a VI is

$$\begin{aligned}
&\mathbf{u}(t) \in K,\\
&\quad 0 \le \left\langle \widetilde{\mathbf{u}} - \mathbf{u}(t),\, \rho\frac{\partial^2\mathbf{u}}{\partial t^2}(t) + \mathcal{A}\mathbf{u}(t) + \mathcal{B}\frac{\partial\mathbf{u}}{\partial t}(t) - \mathbf{f}(t)\right\rangle \quad \text{for all } \widetilde{\mathbf{u}} \in K,\\
&\mathbf{u}(0) = \mathbf{u}_0, \qquad \frac{\partial\mathbf{u}}{\partial t}(0) = \mathbf{v}_0,
\end{aligned}$$

where $K = \left\{\mathbf{u} \in H^1(\Omega) \mid \mathbf{u} \text{ satisfies (6.82) and } \beta\mathbf{u} \ge \varphi\right\}$ and $\beta\mathbf{u} = -\mathbf{n}\cdot\mathbf{u}|_{\Gamma_C}$. The function $\mathbf{f}\colon [0,T] \to H^{-1}(\Omega)$ represents the external (that is, nonviscoelastic) forces and the natural or traction boundary conditions.

If $\widetilde{\mathbf{u}}\colon [0,T] \to H^1(\Omega)$ is smooth, then we can use integration by parts to create a weaker formulation of the VI: choose $\psi\colon [0,T] \to \mathbb{R}$, where ψ is smooth, $\psi(t) \ge 0$ for all t, $\psi(t) = 1$ for $t \in [0, T-2\eta]$, and $\psi(t) = 0$ for $t \in [T-\eta, T]$ for some $\eta > 0$. Then

$$\begin{aligned}
&\mathbf{u}(\cdot) \in \mathcal{K}, \qquad \text{and}\\
&\quad 0 \le \int_0^T \psi(t)\left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}(t),\, \rho\frac{\partial^2\mathbf{u}}{\partial t^2}(t) + \mathcal{A}\mathbf{u}(t) + \mathcal{B}\frac{\partial\mathbf{u}}{\partial t}(t) - \mathbf{f}(t)\right\rangle dt\\
&\quad = \int_0^T \psi(t)\left[\left\langle \frac{\partial\widetilde{\mathbf{u}}}{\partial t}(t) - \frac{\partial\mathbf{u}}{\partial t}(t),\, \rho\frac{\partial\mathbf{u}}{\partial t}(t)\right\rangle\right.\\
&\qquad\qquad \left. + \left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}(t),\, \mathcal{A}\mathbf{u}(t) + \mathcal{B}\frac{\partial\mathbf{u}}{\partial t}(t) - \mathbf{f}(t)\right\rangle\right] dt\\
&\qquad + \psi(t)\left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}(t),\, \rho\frac{\partial\mathbf{u}}{\partial t}(t)\right\rangle\Big|_{t=0}^{t=T}\\
&\qquad - \int_0^T \psi'(t)\left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}(t),\, \rho\frac{\partial\mathbf{u}}{\partial t}(t)\right\rangle dt \qquad \text{for all } \widetilde{\mathbf{u}} \in \mathcal{K},
\end{aligned} \tag{6.86}$$

where

$$\mathcal{K} = \left\{\mathbf{w} \in L^2(0,T;H^1(\Omega)) \mid \beta\mathbf{w}(t) + \varphi \ge 0 \text{ and } \mathbf{w}(0) = \mathbf{u}_0\right\}.$$

To create an approximate problem with solution $\mathbf{u}_\epsilon$ of lower order, set

$$N_\epsilon(t,\mathbf{x}) = [-\mathbf{n}(\mathbf{x})\cdot\mathbf{u}_\epsilon(t,\mathbf{x})+\varphi(\mathbf{x})]_-/\epsilon \quad \text{on } \Gamma_C$$

with $\epsilon > 0$. In more abstract terms, $N_\epsilon = -\epsilon^{-1}\,\Psi'\circ(\beta\mathbf{u}_\epsilon+\varphi)$, where $\Psi(s) = \frac{1}{2}\left[s_-\right]^2$ is a convex function. Then we set

$$\rho\frac{\partial^2\mathbf{u}_\epsilon}{\partial t^2} = -\mathcal{A}\mathbf{u}_\epsilon(t) - \mathcal{B}\frac{\partial\mathbf{u}_\epsilon}{\partial t}(t) + \mathbf{f}(t) + \beta^* N_\epsilon(t). \tag{6.87}$$

Theorem 6.3. *Suppose that* $\mathbf{f}\in L^2(0,T;L^2(\Omega))+W^{1,2}(0,T;H^{-1}(\Omega))$ *and the density* ρ *is bounded away from zero. Let* $K=\left\{\mathbf{z}\in H^1(\Omega)\mid \beta\mathbf{z}+\varphi\geq 0\right\}$. *Then solutions exist for* (6.86) *for given initial displacement (*$\mathbf{u}(0)=\mathbf{u}_0\in K\subset H^1(\Omega)$*) and velocity (*$\partial\mathbf{u}/\partial t(0)=\mathbf{v}_0\in L^2(\Omega)$*) for any time interval* $[0,T]$. *Furthermore, the solution*

$$\mathbf{u}\in L^\infty(0,T;H^1(\Omega))\cap W^{1,2}(0,T;H^1(\Omega))\cap W^{2,2}(0,T;H_0^1(\Omega)').$$

Essentially this result was shown by Cocou and Ricaud [58, 59], although the method of proof used there is based on a Ky Fan minimax theorem. Subsequent papers, which extended this result, include Cocou [56] and Kuttler and Shillor [154], both of which incorporate a nonlocal Coulomb friction law. The method of proof used here follows [154].

Proof. We show existence on a sufficiently small interval $[0,T]$ with $T>0$. Throughout the proof we will consider $t\in[0,T]$. The extension to showing existence of a solution on an arbitrary time interval can be accomplished by continuation arguments.

Let E_ϵ be the approximate energy function

$$E_\epsilon[\mathbf{u},\mathbf{v}] = \frac{1}{2}\langle\mathbf{v},\rho\mathbf{v}\rangle + \frac{1}{2}\langle\mathbf{u},\mathcal{A}\mathbf{u}\rangle + \frac{1}{\epsilon}\int_{\Gamma_C}\Psi\circ(\beta\mathbf{u}+\varphi)\,dS.$$

Since $H^1(\Omega)$ is a separable Hilbert space, we choose a basis $\{\phi_1,\phi_2,\phi_3,\dots\}$ for $H^1(\Omega)$ where $\phi_i\in K$ for each i and set $X_m=\operatorname{span}\{\phi_1,\phi_2,\dots,\phi_m\}$. We use $\phi_1=\mathbf{u}_0$. The Galerkin approximation $\mathbf{u}_{m,\epsilon}$ is then given by

$$\left\langle\mathbf{w},\rho\frac{\partial^2\mathbf{u}_{m,\epsilon}}{\partial t^2}\right\rangle = \left\langle\mathbf{w},-\mathcal{A}\mathbf{u}_{m,\epsilon}(t)-\mathcal{B}\frac{\partial\mathbf{u}_{m,\epsilon}}{\partial t}(t)+\mathbf{f}(t)+\beta^*N_{m,\epsilon}(t)\right\rangle \quad \text{for all } \mathbf{w}\in X_m, \tag{6.88}$$

$$N_{m,\epsilon} = -\frac{1}{\epsilon}\,\Psi'\circ\left(\beta\mathbf{u}_{m,\epsilon}+\varphi\right). \tag{6.89}$$

This is a finite-dimensional Lipschitz differential equation, and so it has solutions (which

are unique). Then, with $\mathbf{v}_{m,\epsilon} = \partial \mathbf{u}_{m,\epsilon}/\partial t$, at time t,

$$\begin{aligned}
\frac{d}{dt} E_\epsilon\left[\mathbf{u}_{m,\epsilon}, \mathbf{v}_{m,\epsilon}\right] &= \left\langle \mathbf{v}_{m,\epsilon}, \rho \frac{\partial \mathbf{v}_{m,\epsilon}}{\partial t} \right\rangle + \left\langle \mathbf{v}_{m,\epsilon}, \mathcal{A}\mathbf{u}_{m,\epsilon} \right\rangle \\
&\qquad + \frac{1}{\epsilon} \int_{\Gamma_C} \Psi' \circ \left(\beta \mathbf{u}_{m,\epsilon} + \varphi\right) \cdot \beta \mathbf{v}_{m,\epsilon}\, dS \\
&= \left\langle \mathbf{v}_{m,\epsilon}, -\mathcal{A}\mathbf{u}_{m,\epsilon} - \mathcal{B}\mathbf{v}_{m,\epsilon} + \mathbf{f}(t) + \beta^* N_{m,\epsilon} \right\rangle \\
&\qquad + \left\langle \mathbf{v}_{m,\epsilon}, \mathcal{A}\mathbf{u}_{m,\epsilon} \right\rangle + \frac{1}{\epsilon} \int_{\Gamma_C} \Psi' \circ \left(\beta \mathbf{u}_{m,\epsilon} + \varphi\right) \cdot \beta \mathbf{v}_{m,\epsilon}\, dS \\
&\le \left\langle \mathbf{v}_{m,\epsilon}, \mathbf{f}(t) \right\rangle + \left\langle \beta \mathbf{v}_{m,\epsilon}, N_{m,\epsilon} + \epsilon^{-1} \Psi' \circ \left(\beta \mathbf{u}_{m,\epsilon} + \varphi\right) \right\rangle \\
&= \left\langle \mathbf{v}_{m,\epsilon}, \mathbf{f}(t) \right\rangle.
\end{aligned}$$

Integrating gives the inequality

$$E_\epsilon\left[\mathbf{u}_{m,\epsilon}(t), \mathbf{v}_{m,\epsilon}(t)\right] \le E_\epsilon\left[\mathbf{u}_0, \mathbf{v}_0\right] + \int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathbf{f}(\tau) \right\rangle d\tau.$$

We wish to turn this into a bound on $\mathbf{u}_{m,\epsilon}$ and $\mathbf{v}_{m,\epsilon}$ that is independent of m in suitable spaces. First, note that if we write $\mathbf{f}(t) = \mathbf{f}_1(t) + \mathbf{f}_2(t)$ with $\mathbf{f}_1 \in L^2(0,T; L^2(\Omega))$ and $\mathbf{f}_2 \in W^{1,2}(0,T; H^{-1}(\Omega))$, then

$$\begin{aligned}
&\int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathbf{f}_1(\tau) \right\rangle d\tau \le \left\| \mathbf{v}_{m,\epsilon} \right\|_{L^2(0,t; L^2(\Omega))} \left\| \mathbf{f}_1 \right\|_{L^2(0,T; L^2(\Omega))}, \\
&\int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathbf{f}_2(\tau) \right\rangle d\tau \\
&\qquad = \left\langle \mathbf{u}_{m,\epsilon}(\tau), \mathbf{f}_2(\tau) \right\rangle \Big|_{\tau=0}^{\tau=t} - \int_0^t \left\langle \mathbf{u}_{m,\epsilon}(\tau), \mathbf{f}_2'(\tau) \right\rangle d\tau \\
&\qquad \le \left\| \mathbf{u}_0 \right\|_{H^1(\Omega)} \left\| \mathbf{f}_2(0) \right\|_{H^{-1}(\Omega)} + \left\| \mathbf{u}_{m,\epsilon}(t) \right\|_{H^1(\Omega)} \left\| \mathbf{f}_2(t) \right\|_{H^{-1}(\Omega)} \\
&\qquad\qquad + \left\| \mathbf{u}_{m,\epsilon} \right\|_{L^2(0,T; H^1(\Omega))} \left\| \mathbf{f}_2 \right\|_{W^{1,2}(0,T; H^{-1}(\Omega))}.
\end{aligned}$$

Since $\rho \ge \rho_0 > 0$ over Ω, where ρ_0 is a constant, we have a bound of the form

$$\begin{aligned}
&\frac{\rho_0}{2} \left\langle \mathbf{v}_{m,\epsilon}(t), \mathbf{v}_{m,\epsilon}(t) \right\rangle + \frac{1}{2} \left\langle \mathbf{u}_{m,\epsilon}(t), \mathcal{A}\mathbf{u}_{m,\epsilon}(t) \right\rangle \\
&\qquad + \frac{1}{\epsilon} \int_{\Gamma_C} \Psi \circ \left(\beta \mathbf{u}_{m,\epsilon}(t) + \varphi\right) dS \\
&\quad \le C \left(1 + \left\| \mathbf{v}_{m,\epsilon} \right\|_{L^2(0,t; L^2(\Omega))} + \left\| \mathbf{u}_{m,\epsilon}(t) \right\|_{H^1(\Omega)} + \left\| \mathbf{u}_{m,\epsilon} \right\|_{L^2(0,t; H^1(\Omega))}\right)
\end{aligned}$$

for all m, $\epsilon > 0$ and $t > 0$, and C is a constant that is independent of m, ϵ, and t, provided $0 \le t \le T$. In what follows, C will continue to be a quantity that is independent of m, ϵ, and t, provided $0 \le t \le T$, but its value may be different in different occurrences. Since $\mathcal{A}$ is semielliptic, $\langle \mathbf{z}, \mathcal{A}\mathbf{z} \rangle \ge \eta_A \left\| \mathbf{z} \right\|_{H^1(\Omega)}^2 - \mu_A \left\| \mathbf{z} \right\|_{L^2(\Omega)}^2$ for suitable constants $\eta_A > 0$ and

$\mu_A \geq 0$. Also, $\Psi(s) \geq 0$ for all s. Thus

$$\begin{aligned}
&\frac{\rho_0}{2}\left\|\mathbf{v}_{m,\epsilon}(t)\right\|^2_{L^2(\Omega)} + \frac{\eta_A}{2}\left\|\mathbf{u}_{m,\epsilon}(t)\right\|^2_{H^1(\Omega)} - \frac{\mu_A}{2}\left\|\mathbf{u}_{m,\epsilon}(t)\right\|^2_{L^2(\Omega)} \\
&\quad \leq C\left(1 + \left\|\mathbf{v}_{m,\epsilon}\right\|_{L^2(0,t;L^2(\Omega))} + \left\|\mathbf{u}_{m,\epsilon}(t)\right\|_{H^1(\Omega)} + \left\|\mathbf{u}_{m,\epsilon}\right\|_{L^2(0,t;H^1(\Omega))}\right) \\
&\quad \leq C\left(1 + T^{1/2}\left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,T;L^2(\Omega))} + T^{1/2}\left\|\mathbf{u}_{m,\epsilon}\right\|_{L^\infty(0,T;H^1(\Omega))}\right. \\
&\qquad\qquad \left. + \left\|\mathbf{u}_{m,\epsilon}(t)\right\|_{H^1(\Omega)}\right).
\end{aligned}$$

But $\mathbf{u}_{m,\epsilon}(t) = \mathbf{u}_0 + \int_0^t \mathbf{v}_{m,\epsilon}(\tau)\,d\tau$, so $\left\|\mathbf{u}_{m,\epsilon}(t)\right\|_{L^2(\Omega)} \leq \|\mathbf{u}_0\|_{L^2(\Omega)} + T\left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,t;L^2(\Omega))}$. Then taking the maximum over $t \in [0,T]$ gives

$$\begin{aligned}
&\frac{\rho_0}{2}\left\|\mathbf{v}_{m,\epsilon}\right\|^2_{L^\infty(0,T;L^2(\Omega))} + \frac{\eta_A}{2}\left\|\mathbf{u}_{m,\epsilon}\right\|^2_{L^\infty(0,T;H^1(\Omega))} \\
&\qquad - \frac{\mu_A}{2}\left(\|\mathbf{u}_0\|_{L^2(\Omega)} + T\left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,T;L^2(\Omega))}\right)^2 \\
&\quad = \frac{\rho_0 - \mu_A T}{2}\left\|\mathbf{v}_{m,\epsilon}\right\|^2_{L^\infty(0,T;L^2(\Omega))} \frac{\eta_A}{2}\left\|\mathbf{u}_{m,\epsilon}\right\|^2_{L^\infty(0,T;H^1(\Omega))} \\
&\qquad - \frac{\mu_A}{2}\left(\|\mathbf{u}_0\|^2_{L^2(\Omega)} + 2T\left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,T;L^2(\Omega))}\right) \\
&\quad \leq C\left(1 + T^{1/2}\left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,T;L^2(\Omega))} + T^{1/2}\left\|\mathbf{u}_{m,\epsilon}\right\|_{L^\infty(0,T;H^1(\Omega))}\right. \\
&\qquad\qquad \left. + \left\|\mathbf{u}_{m,\epsilon}\right\|_{L^\infty(0,T;H^1(\Omega))}\right).
\end{aligned}$$

Choosing $T > 0$ sufficiently small so that $\rho_0 - \mu_A T \geq \rho_0/2$ we can remove the "μ_A" term from the above inequality to obtain

$$\begin{aligned}
&\frac{\rho_0}{4}\left\|\mathbf{v}_{m,\epsilon}\right\|^2_{L^\infty(0,T;L^2(\Omega))} + \frac{\eta_A}{2}\left\|\mathbf{u}_{m,\epsilon}\right\|^2_{L^\infty(0,T;H^1(\Omega))} \\
&\qquad \leq C\left(1 + \left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,T;L^2(\Omega))} + \left\|\mathbf{u}_{m,\epsilon}\right\|_{L^\infty(0,T;H^1(\Omega))}\right).
\end{aligned}$$

Thus both $\left\|\mathbf{v}_{m,\epsilon}\right\|_{L^\infty(0,T;L^2(\Omega))}$ and $\left\|\mathbf{u}_{m,\epsilon}\right\|_{L^\infty(0,T;H^1(\Omega))}$ are bounded independently of m and $\epsilon > 0$.

Since each term of $E_\epsilon[\mathbf{u}, \mathbf{v}]$ is nonnegative, this means that

$$\frac{1}{\epsilon}\int_{\Gamma_C} \Psi \circ \left(\beta \mathbf{u}_{m,\epsilon}(t) + \varphi\right) dS \leq E_0 + \int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathbf{f}(\tau)\right\rangle d\tau;$$

note that the right-hand side is bounded independently of m, ϵ, and $t \in [0,T]$. There is a constant C where

$$\int_{\Gamma_C} \Psi \circ \left(\beta \mathbf{u}_{m,\epsilon}(t) + \varphi\right) dS \leq C\epsilon.$$

There is an additional bound that we can extract using the loss of energy through viscosity. Note that

$$\frac{d}{dt}E_\epsilon\left[\mathbf{u}_{m,\epsilon}, \mathbf{v}_{m,\epsilon}\right] = \left\langle \mathbf{v}_{m,\epsilon}(t), \mathbf{f}(t)\right\rangle - \left\langle \mathbf{v}_{m,\epsilon}(t), \mathcal{B}\mathbf{v}_{m,\epsilon}(t)\right\rangle.$$

Thus

$$E_\epsilon\left[\mathbf{u}_{m,\epsilon}(t), \mathbf{v}_{m,\epsilon}(t)\right] = E_0 + \int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathbf{f}(\tau)\right\rangle d\tau - \int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathcal{B}\mathbf{v}_{m,\epsilon}(\tau)\right\rangle d\tau,$$

and so

$$\int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathcal{B}\mathbf{v}_{m,\epsilon}(\tau)\right\rangle d\tau \le E_0 + \int_0^t \left\langle \mathbf{v}_{m,\epsilon}(\tau), \mathbf{f}(\tau)\right\rangle d\tau$$

since $E_\epsilon[\mathbf{u}, \mathbf{v}] \ge 0$ for all $\mathbf{u}$ and $\mathbf{v}$. Using $\langle \mathbf{z}, \mathcal{B}\mathbf{z}\rangle \ge \eta_B \|\mathbf{z}\|^2_{H^1(\Omega)} - \mu_B \|\mathbf{z}\|^2_{L^2(\Omega)}$ for constants $\eta_B > 0$ and $\mu_B \ge 0$ and the bound on $\|\mathbf{v}_{m,\epsilon}\|_{L^\infty(0,T;L^2(\Omega))}$, we can obtain a bound

$$\|\mathbf{v}_{m,\epsilon}\|_{L^2(0,T;H^1(\Omega))} \le C$$

independent of m and $\epsilon > 0$.

There is an additional bound that we will need on $\partial \mathbf{v}_{m,\epsilon}/\partial t$ in $L^2(0,T;H_0^1(\Omega)')$. To do this we use duality: suppose $\boldsymbol{\phi}\colon [0,T]\times\Omega \to \mathbb{R}^d$ is a smooth function that is zero in a neighborhood of the boundary $\partial\Omega$. Then

$$\begin{aligned}
&\int_0^T \int_\Omega \boldsymbol{\phi}(t,\mathbf{x}) \cdot \rho(\mathbf{x}) \frac{\partial \mathbf{v}_{m,\epsilon}}{\partial t}(t,\mathbf{x}) d\mathbf{x}\, dt \\
&\quad = \int_0^T \int_\Omega \boldsymbol{\phi}(t,\mathbf{x}) \cdot \left[-\mathcal{A}\mathbf{u}_{m,\epsilon}(t,\mathbf{x}) - \mathcal{B}\mathbf{v}_{m,\epsilon}(t,\mathbf{x}) + \mathbf{f}(t,\mathbf{x})\right] d\mathbf{x}\, dt \\
&\quad = \int_0^T \int_\Omega \boldsymbol{\phi}(t,\mathbf{x}) \cdot \left[-\operatorname{div}\left(A\varepsilon\left[\mathbf{u}_{m,\epsilon}\right]\right) - \operatorname{div}\left(B\varepsilon\left[\mathbf{v}_{m,\epsilon}\right]\right) + \mathbf{f}\right] d\mathbf{x}\, dt \\
&\quad = \int_0^T \int_\Omega \left(\varepsilon[\boldsymbol{\phi}] : A\varepsilon\left[\mathbf{u}_{m,\epsilon}\right] + \varepsilon[\boldsymbol{\phi}] : B\varepsilon\left[\mathbf{v}_{m,\epsilon}\right] + \boldsymbol{\phi}\cdot\mathbf{f}\right) d\mathbf{x}\, dt \\
&\quad \le C\|\boldsymbol{\phi}\|_{L^2(0,T;H_0^1(\Omega))} \left(\|\mathbf{u}_{m,\epsilon}\|_{L^2(0,T;H^1(\Omega))} + \|\mathbf{v}_{m,\epsilon}\|_{L^2(0,T;H^1(\Omega))} \right. \\
&\qquad \left. + \|\mathbf{f}\|_{L^2(0,T;H^{-1}(\Omega))}\right) \\
&\quad \le C\|\boldsymbol{\phi}\|_{L^2(0,T;H_0^1(\Omega))}.
\end{aligned}$$

Taking the supremum over all $\boldsymbol{\phi}$ smooth and zero in a neighborhood of $\partial\Omega$ shows that $\rho\,\partial\mathbf{v}_{m,\epsilon}/\partial t$ is bounded independently of m, $\epsilon > 0$ in $L^2(0,T;H_0^1(\Omega))'$, so $\rho\,\partial\mathbf{v}_{m,\epsilon}/\partial t$ is bounded in $L^2(0,T;H_0^1(\Omega)')$ independently of m, and $\epsilon > 0$.

With these bounds, by Alaoglu's theorem, there are weakly* convergent subsequences (also denoted by $\mathbf{v}_{m,\epsilon}$ and $\mathbf{u}_{m,\epsilon}$, respectively)

$$\begin{aligned}
\mathbf{v}_{m,\epsilon} &\rightharpoonup^* \widehat{\mathbf{v}} \quad \text{in } L^\infty(0,T;L^2(\Omega)), \\
\frac{\partial \mathbf{v}_{m,\epsilon}}{\partial t} &\rightharpoonup^* \widehat{\mathbf{z}} \quad \text{in } L^2(0,T;H_0^1(\Omega)'), \\
\mathbf{u}_{m,\epsilon} &\rightharpoonup^* \widehat{\mathbf{u}} \quad \text{in } L^\infty(0,T;H^1(\Omega)) \qquad \text{as } m\to\infty \text{ and } \epsilon \downarrow 0
\end{aligned}$$

in the subsequence. Clearly this implies weak convergence:

$$\begin{aligned}
\mathbf{v}_{m,\epsilon} &\rightharpoonup \widehat{\mathbf{v}} \quad \text{in } L^2(0,T;L^2(\Omega)) \text{ and } L^2(0,T;H^1(\Omega)),\\
\frac{\partial \mathbf{v}_{m,\epsilon}}{\partial t} &\rightharpoonup \widehat{\mathbf{z}} \quad \text{in } L^2(0,T;H_0^1(\Omega)'),\\
\mathbf{u}_{m,\epsilon} &\rightharpoonup \widehat{\mathbf{u}} \quad \text{in } L^2(0,T;H^1(\Omega)).
\end{aligned}$$

Note that for $s<t$ we have $\widehat{\mathbf{v}}(t)-\widehat{\mathbf{v}}(s)=\int_s^t \widehat{\mathbf{z}}(\tau)\,d\tau$ taking limits in $H_0^1(\Omega)'$. Thus we can justify $\widehat{\mathbf{w}}=\partial\widehat{\mathbf{v}}/\partial t$. By Seidman's theorem and Simon's theorem (Theorems A.6 and A.7) we can have, perhaps by taking further subsequences, $\mathbf{v}_{m,\epsilon}\to\widehat{\mathbf{v}}$ strongly in $C(0,T;H^{-\delta}(\Omega))$ for any $\delta>0$, $\mathbf{v}_{m,\epsilon}\to\widehat{\mathbf{v}}$ strongly in $L^2(0,T;L^2(\Omega))$, $\mathbf{u}_{m,\epsilon}\to\widehat{\mathbf{u}}$ strongly in $C(0,T;H^{1-\delta}(\Omega))$ for any $\delta>0$.

We can then apply Mazur's lemma to $(\mathbf{u},\mathbf{v})\mapsto\int_0^T E_\epsilon[\mathbf{u}(t),\mathbf{v}(t)]\,\theta(t)\,dt$, where θ is a continuous nonnegative function. Since $E[\mathbf{u},\mathbf{v}]$ is a convex function of $\mathbf{u}$ and $\mathbf{v}$, so is this functional on $L^2(0,T;L^2(\Omega))\times L^2(0,T;H^1(\Omega))$. Hence

$$\begin{aligned}
\int_0^T E[\widehat{\mathbf{u}}(t),\widehat{\mathbf{v}}(t)]\,\theta(t)\,dt &\le \liminf_{m\to\infty}\int_0^T E\left[\mathbf{u}_{m,\epsilon}(t),\mathbf{v}_{m,\epsilon}(t)\right]\theta(t)\,dt\\
&\le \liminf_{m\to\infty}\int_0^T\left[E_0+\int_0^t\left\langle\mathbf{v}_{m,\epsilon}(\tau),\mathbf{f}(\tau)\right\rangle d\tau\right]\theta(t)\,dt\\
&=\int_0^T\left[E_0+\int_0^t\langle\widehat{\mathbf{v}}(\tau),\mathbf{f}(\tau)\rangle\,d\tau\right]\theta(t)\,dt,
\end{aligned}$$

where $E_0=E_\epsilon[\mathbf{u}_0,\mathbf{v}_0]=E[\mathbf{u}_0,\mathbf{v}_0]$ for $\beta\mathbf{u}_0+\varphi\ge 0$. Since this is true for all continuous nonnegative θ, for almost all t,

$$E[\widehat{\mathbf{u}}(t),\widehat{\mathbf{v}}(t)]\le E_0+\int_0^t\langle\widehat{\mathbf{v}}(\tau),\mathbf{f}(\tau)\rangle\,d\tau.$$

We now wish to show that the limits $(\widehat{\mathbf{u}},\widehat{\mathbf{v}})$ satisfy the VI (6.86). First we assume that $\widetilde{\mathbf{u}}\colon[0,T]\to X_m\cap K$ is in $W^{1,2}(0,T;H^1(\Omega))$ and that $\widetilde{\mathbf{u}}(0)=\mathbf{u}_0$. Now from the Galerkin approximation (6.88)–(6.89),

$$0=\left\langle\widetilde{\mathbf{u}}(t)-\mathbf{u}_{m,\epsilon}(t),\,\rho\frac{\partial^2\mathbf{u}_{m,\epsilon}}{\partial t^2}(t)+\mathcal{A}\mathbf{u}_{m,\epsilon}(t)+\mathcal{B}\frac{\partial\mathbf{u}_{m,\epsilon}}{\partial t}-\mathbf{f}(t)-\beta^*N_{m,\epsilon}(t)\right\rangle.$$

The term $\left\langle\widetilde{\mathbf{u}}(t)-\mathbf{u}_{m,\epsilon}(t),\,\beta^*N_{m,\epsilon}(t)\right\rangle\ge 0$ since

$$\begin{aligned}
&\left\langle\widetilde{\mathbf{u}}(t)-\mathbf{u}_{m,\epsilon}(t),\,\beta^*N_{m,\epsilon}(t)\right\rangle\\
&\quad=\left\langle\beta\widetilde{\mathbf{u}}(t)-\beta\mathbf{u}_{m,\epsilon}(t),\,N_{m,\epsilon}(t)\right\rangle\\
&\quad=\left\langle(\beta\widetilde{\mathbf{u}}(t)+\varphi)-\left(\beta\mathbf{u}_{m,\epsilon}(t)+\varphi\right),\,N_{m,\epsilon}(t)\right\rangle.
\end{aligned}$$

Since $N_{m,\epsilon}(t)\ge 0$ and $\beta\widetilde{\mathbf{u}}(t)+\varphi\ge 0$ on Γ_C, $\left\langle\beta\widetilde{\mathbf{u}}(t)+\varphi,\,N_{m,\epsilon}(t)\right\rangle\ge 0$. On the other hand,

$$\begin{aligned}
-\left\langle\beta\mathbf{u}_{m,\epsilon}(t)+\varphi,\,N_{m,\epsilon}(t)\right\rangle&=\epsilon^{-1}\left\langle\beta\mathbf{u}_{m,\epsilon}+\varphi,\,\Psi'\circ\left(\beta\mathbf{u}_{m,\epsilon}+\varphi\right)\right\rangle\\
&=\epsilon^{-1}\int_{\Gamma_C}\left(\beta\mathbf{u}_{m,\epsilon}+\varphi\right)\cdot\Psi'\circ\left(\beta\mathbf{u}_{m,\epsilon}+\varphi\right)dS.
\end{aligned}$$

But $s\,\Psi'(s) = -s\,s_- \geq 0$ for $s < 0$ and zero for $s \geq 0$. So the integral over Γ_C must be nonnegative. Combining the two inequalities gives $\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}_{m,\epsilon}(t),\, \beta^* N_{m,\epsilon}(t) \rangle \geq 0$.

Thus

$$0 \leq \left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}_{m,\epsilon}(t),\, \rho \frac{\partial^2 \mathbf{u}_{m,\epsilon}}{\partial t^2}(t) + \mathcal{A}\mathbf{u}_{m,\epsilon}(t) + \mathcal{B}\frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t} - \mathbf{f}(t) \right\rangle.$$

Multiplying by $\psi(t) \geq 0$ and integrating over $[0,T]$ give

$$\begin{aligned}
0 &\leq \int_0^T \psi(t) \left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}_{m,\epsilon}(t),\, \rho \frac{\partial^2 \mathbf{u}_{m,\epsilon}}{\partial t^2}(t) + \mathcal{A}\mathbf{u}_{m,\epsilon}(t) + \mathcal{B}\frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t} - \mathbf{f}(t) \right\rangle dt \\
&= \int_0^T \psi(t) \left\langle \frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t}(t) - \frac{\partial \widetilde{\mathbf{u}}}{\partial t}(t),\, \rho \frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t}(t) \right\rangle dt \\
&\qquad + \psi(t) \left\langle \mathbf{u}_{m,\epsilon}(t) - \widetilde{\mathbf{u}}(t),\, \rho \frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t}(t) \right\rangle \Bigg|_{t=0}^{t=T} \\
&\qquad - \int_0^T \psi'(t) \left\langle \mathbf{u}_{m,\epsilon}(t) - \widetilde{\mathbf{u}}(t),\, \rho \frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t}(t) \right\rangle dt \\
&\qquad + \int_0^T \psi(t) \left\langle \widetilde{\mathbf{u}}(t) - \mathbf{u}_{m,\epsilon}(t),\, \mathcal{A}\mathbf{u}_{m,\epsilon}(t) + \mathcal{B}\frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t} - \mathbf{f}(t) \right\rangle dt. \qquad (6.90)
\end{aligned}$$

The term

$$\psi(t) \left\langle \mathbf{u}_{m,\epsilon}(t) - \widetilde{\mathbf{u}}(t),\, \rho \frac{\partial \mathbf{u}_{m,\epsilon}}{\partial t}(t) \right\rangle \Bigg|_{t=0}^{t=T} = 0,$$

provided $\widetilde{\mathbf{u}}(0) = \mathbf{u}_0$ since $\psi(T) = 0$.

From

$$\begin{aligned}
&\partial \mathbf{u}_{m,\epsilon}/\partial t = \mathbf{v}_{m,\epsilon} \rightharpoonup \widehat{\mathbf{v}} \quad \text{weakly in } L^2(0,T; H^1(\Omega)), \\
&\mathbf{v}_{m,\epsilon} \to \widehat{\mathbf{v}} \quad \text{strongly in } L^2(0,T; H^{1-\delta}(\Omega)), \\
&\mathbf{u}_{m,\epsilon} \rightharpoonup \widehat{\mathbf{u}} \quad \text{weakly in } L^2(0,T; H^1(\Omega)),
\end{aligned}$$

we have, with $\widetilde{\mathbf{v}} = \partial \widetilde{\mathbf{u}}/\partial t$,

$$\begin{aligned}
\int_0^T \psi(t) \left\langle \mathbf{v}_{m,\epsilon}(t) - \widetilde{\mathbf{v}}(t),\, \rho \mathbf{v}_{m,\epsilon}(t) \right\rangle dt &\to \int_0^T \psi\, \langle \widehat{\mathbf{v}} - \widetilde{\mathbf{v}},\, \rho \widehat{\mathbf{v}} \rangle\, dt, \\
\int_0^T \psi'(t) \left\langle \mathbf{u}_{m,\epsilon}(t) - \widetilde{\mathbf{u}}(t),\, \rho \mathbf{v}_{m,\epsilon}(t) \right\rangle dt &\to \int_0^T \psi'\, \langle \widehat{\mathbf{u}} - \widetilde{\mathbf{u}},\, \rho \widehat{\mathbf{v}} \rangle\, dt, \\
\int_0^T \psi(t) \left\langle \widetilde{\mathbf{u}},\, \mathcal{A}\mathbf{u}_{m,\epsilon}(t) + \mathcal{B}\mathbf{v}_{m,\epsilon} \right\rangle dt &\to \int_0^T \psi\, \langle \widetilde{\mathbf{u}},\, \mathcal{A}\widehat{\mathbf{u}} + \mathcal{B}\widehat{\mathbf{v}} \rangle\, dt, \\
\int_0^T \psi(t) \left\langle \mathbf{u}_{m,\epsilon}(t),\, \mathbf{f}(t) \right\rangle dt &\to \int_0^T \psi\, \langle \widehat{\mathbf{u}},\, \mathbf{f} \rangle\, dt
\end{aligned}$$

as $m \to \infty$ and $\epsilon \downarrow 0$.

The remaining terms require Mazur's lemma: for $\int_0^T \psi \left\langle \mathbf{u}_{m,\epsilon},\, \mathcal{A}\mathbf{u}_{m,\epsilon} \right\rangle dt$,

$$\int_0^T \psi\, \langle \widehat{\mathbf{u}},\, \mathcal{A}\widehat{\mathbf{u}} \rangle\, dt \leq \liminf_{m \to \infty,\, \epsilon \downarrow 0} \int_0^T \psi \left\langle \mathbf{u}_{m,\epsilon},\, \mathcal{A}\mathbf{u}_{m,\epsilon} \right\rangle dt.$$

For $\int_0^T \psi \left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{v}_{m,\epsilon} \right\rangle dt$ we use the fact that $\mathcal{B}$ is self-adjoint:

$$\begin{aligned}\int_0^T \psi \left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{v}_{m,\epsilon} \right\rangle dt &= \int_0^T \psi \frac{d}{dt}\frac{1}{2}\left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{u}_{m,\epsilon} \right\rangle dt \\ &= \psi(t)\frac{1}{2}\left\langle \mathbf{u}_{m,\epsilon}(t), \mathcal{B}\mathbf{u}_{m,\epsilon}(t) \right\rangle \Big|_{t=0}^{t=T} \\ &\qquad - \frac{1}{2}\int_0^T \psi' \left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{u}_{m,\epsilon} \right\rangle dt \\ &= -\frac{1}{2}\langle \mathbf{u}_0, \mathcal{B}\mathbf{u}_0 \rangle - \frac{1}{2}\int_0^T \psi' \left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{u}_{m,\epsilon} \right\rangle dt.\end{aligned}$$

Now, since ψ is nonincreasing, $\psi' \le 0$, and so

$$\int_0^T -\psi' \langle \widehat{\mathbf{u}}, \mathcal{B}\widehat{\mathbf{u}} \rangle \, dt \le \liminf_{m\to\infty,\, \epsilon\downarrow 0} \int_0^T -\psi' \left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{u}_{m,\epsilon} \right\rangle dt.$$

Reversing the integration by parts for $\int_0^T -\psi' \langle \widehat{\mathbf{u}}, \mathcal{B}\widehat{\mathbf{u}} \rangle \, dt$, we get

$$\int_0^T \psi \langle \widehat{\mathbf{u}}, \mathcal{B}\widehat{\mathbf{v}} \rangle \, dt \le \liminf_{m\to\infty,\, \epsilon\downarrow 0} \int_0^T \psi \left\langle \mathbf{u}_{m,\epsilon}, \mathcal{B}\mathbf{v}_{m,\epsilon} \right\rangle dt.$$

Combining the above inequalities and taking the liminf of (6.90), we get

$$0 \le \int_0^T \psi(t) \left\langle \widetilde{\mathbf{u}}(t) - \widehat{\mathbf{u}}(t), \rho \frac{\partial^2 \widehat{\mathbf{u}}}{\partial t^2}(t) + \mathcal{A}\widehat{\mathbf{u}}(t) + \mathcal{B}\frac{\partial \widehat{\mathbf{u}}}{\partial t} - \mathbf{f}(t) \right\rangle dt$$

for all $\widetilde{\mathbf{u}}\colon [0,T] \to X_m \cap K$ that are in $W^{1,2}(0,T; H^1(\Omega))$ and $\widetilde{\mathbf{u}}(0) = \mathbf{u}_0$, and $\psi\colon [0,T] \to [0,1]$ smooth and nonincreasing with $\psi(t) = 1$ for $t \in [0, T-2\eta]$ and $\psi(t) = 0$ for $t \in [T-\eta, T]$. Since

$$\rho \frac{\partial^2 \widehat{\mathbf{u}}}{\partial t^2} + \mathcal{A}\widehat{\mathbf{u}} + \mathcal{B}\frac{\partial \widehat{\mathbf{u}}}{\partial t} - \mathbf{f} \in L^2(0,T; H_0^1(\Omega)')$$

for any $\widetilde{\mathbf{u}}\colon [0,T] \to K$ that is in $C(0,T; H^1(\Omega))$, we have approximations $\widetilde{\mathbf{u}}_m\colon [0,T] \to X_m \cap K$ in $W^{1,2}(0,T; H^1(\Omega))$ where $\widetilde{\mathbf{u}}_m \to \widetilde{\mathbf{u}}$ as $m \to \infty$. Hence

$$0 \le \int_0^T \psi(t) \left\langle \widetilde{\mathbf{u}}(t) - \widehat{\mathbf{u}}(t), \rho \frac{\partial^2 \widehat{\mathbf{u}}}{\partial t^2}(t) + \mathcal{A}\widehat{\mathbf{u}}(t) + \mathcal{B}\frac{\partial \widehat{\mathbf{u}}}{\partial t} - \mathbf{f}(t) \right\rangle dt$$

holds for any $\widetilde{\mathbf{u}}\colon [0,T] \to K$ that is in $L^2(0,T; H^1(\Omega))$ since $C(0,T; H^1(\Omega))$ is dense in $L^2(0,T; H^1(\Omega))$. Furthermore, we can take limits over ψ converging pointwise to $\chi_{[0,T)}$ to establish that $\widehat{\mathbf{u}}$ is a solution of the VI

$$\begin{aligned}&\widehat{\mathbf{u}}(t) \in K \qquad \& \\ &\quad 0 \le \left\langle \widetilde{\mathbf{u}} - \widehat{\mathbf{u}}(t), \rho \frac{\partial^2 \widehat{\mathbf{u}}}{\partial t^2}(t) + \mathcal{A}\widehat{\mathbf{u}}(t) + \mathcal{B}\frac{\partial \widehat{\mathbf{u}}}{\partial t} - \mathbf{f}(t) \right\rangle \qquad \text{for all } \widetilde{\mathbf{u}} \in K\end{aligned}$$

for almost all t. □

While this establishes existence for a general class of impact problems for viscoelastic bodies, uniqueness is an open problem as of the time of this writing. The methods can be easily adapted for other impact problems which have a similar structure, and the operators $\mathcal{A}$ and $\mathcal{B}$ need not be second order, such as for a viscoelastic version of the Euler–Bernoulli beam. However, in this approach, it is important that the contact forces are applied on the boundary of Ω.

The reader interested in viscoelastic bodies in frictionless impact should also consider the papers [207, 208] by Petrov and Schatzman. Both of these papers consider viscoelastic wave equations of Kelvin–Voigt type:

$$\frac{\partial^2 u}{\partial t^2} = \nabla^2 u + \alpha \nabla^2 \frac{\partial u}{\partial t} + f(t, \mathbf{x}) \qquad \text{in } \Omega$$

with Signorini conditions on the boundary. In [207], the problem is a viscoelastic rod, and so Ω is one dimensional, with contact at one end. The approach in the analysis is to treat the problem as a CCP to be solved for the normal contact force, which is a scalar function of time. In [208], the problem is in a half-space $\Omega = (0, \infty) \times \mathbb{R}^{d-1}$. The analysis in both of these papers is based on Fourier transforms, and they give some deep results. In particular, it should be noted that in the former paper [207], not only is existence of solutions proved, but it is shown that all solutions satisfy an energy balance. That is, it is shown that the change in the energy $\frac{1}{2}\int_0^\ell \left[\left(\partial u/\partial t\right)^2 + \left(\partial u/\partial x\right)^2\right] dx$ is the work done by the external forces $f(t,x)$ minus the losses due to the viscosity. More details on this approach will be given in the next chapter. In [208], the authors are not quite able to obtain this result for the viscous wave equation in a half-space, but are still able to obtain strong regularity results for the trace of u on the boundary $\partial\Omega = \{0\} \times \mathbb{R}^{d-1}$.

6.3.2 Coulomb friction

At the time of this writing, existence results for viscoelastic impact with friction have been proven only for modified friction laws [56, 154] or under non-Signorini contact conditions [87, 88, 89]. Alternatively, static and quasi-static contact problems have been shown to have solutions [10, 57, 60, 85, 86, 148, 183, 230]. No results have been shown for the dynamic contact problem with purely *elastic* bodies under Signorini contact conditions and the standard local Coulomb friction laws.

In fact, there is an important *nonexistence* result due to Renardy [212] for a linearized model of a slab of a two-dimensional hyperelastic[10] material sliding over a frictional surface if the friction coefficient exceeds a certain threshold. Careful study of Renardy's results shows that there is a frictional instability resulting from a feedback loop where the friction forces cause displacements in the normal direction, resulting in changes to the normal contact forces, which in turn change the friction forces. As the spatial scale goes to zero, the time scale for the instability also goes to zero, so that in the high (spatial) frequency limit, the exponential rate of the instability goes to infinity. Application of Agmon's condition [270, p. 280 ff.] (or alternatively, via the Lopatinsky–Shapiro conditions—see [270, p. 148 ff.] or [128]) shows that there are no solutions to such partial differential equation except for extremely specific initial and boundary conditions.

[10] A material is *hyperelastic* if it does not change volume, or equivalently in the linearized case, $\operatorname{div} \mathbf{u} = 0$.

As noted in the section on rigid-body dynamics with impact, the Signorini conditions and the Coulomb friction laws are each monotone *separately*, but combined they can be highly nonmonotone.

We will use the time-integrated version of (6.50) as our formulation of Coulomb friction:

$$\int_0^T \int_{\Gamma_C} \mathbf{F} \cdot \mathbf{w}\, dS \leq \int_0^T \int_{\Gamma_C} \mu N \left(\left| \frac{\partial \mathbf{u}_T}{\partial t} + \mathbf{w}_T \right| - \left| \frac{\partial \mathbf{u}_T}{\partial t} \right| \right) dS \tag{6.91}$$

for all $\mathbf{w}$, where $N = -\mathbf{n} \cdot \sigma[\mathbf{u}] \cdot \mathbf{n}$ is the inward normal contact force on Γ_C, and

$$\mathbf{F} = \sigma\left[\mathbf{u}, \partial \mathbf{u}/\partial t\right] \cdot \mathbf{n} + N\,\mathbf{n}$$

is the friction force on Γ_C. The regularity results from the existence results for frictionless Kelvin–Voigt viscoelasticity with impact (Theorem 6.3) are not sufficient to show that $N = -\mathbf{n} \cdot \sigma\left[\mathbf{u}, \partial \mathbf{u}/\partial t\right] \cdot \mathbf{n}$ even makes sense: $\mathbf{u} \in L^\infty(0,T; H^1(\Omega))$ means that the strain tensor $\varepsilon[\mathbf{u}] \in L^\infty(0,T; L^2(\Omega))$, and so $\sigma\left[\mathbf{u}, \partial \mathbf{u}/\partial t\right] \in L^\infty(0,T; L^2(\Omega))$, and there is no "trace" of an $L^2(\Omega)$ function on a part of the boundary such as $\Gamma_C \subseteq \partial\Omega$.

The approach of Cocou [56] and Kuttler and Shillor [154] is to use a smoothing operator to create a nonlocal version of the Coulomb friction law. This approach is to use (6.91) with

$$N(t) = \left| \mathbf{n} \cdot \mathcal{R}\sigma\left[\mathbf{u}(t), \partial \mathbf{u}/\partial t(t)\right] \cdot \mathbf{n} \right| \qquad \text{or}$$
$$N(t) = \max\left(0, -\mathbf{n} \cdot \mathcal{R}\sigma\left[\mathbf{u}(t), \partial \mathbf{u}/\partial t(t)\right] \cdot \mathbf{n}\right),$$

where $\mathcal{R}\colon L^2(\Omega) \to H^1(\Omega)$ is a compact nonlocal smoothing operator. One way of creating a suitable $\mathcal{R}$ is to use an extension operator $E\colon L^2(\Omega) \to L^2(\mathbb{R}^d)$ where $E\phi|\Omega = \phi$, followed by convolution with a suitable smooth nonnegative function ψ with compact support: $\mathcal{R}\phi = \psi * E\phi$. (The extension operator can be extension by zero: $E\phi(\mathbf{x}) = \phi(\mathbf{x})$ if $\mathbf{x} \in \Omega$ and $E\phi(\mathbf{x}) = 0$ if $\mathbf{x} \notin \Omega$.)

As with the frictionless case, we can use a Galerkin approach combined with a penalty approximation to obtain a set of approximate solutions $\mathbf{u}_{m,\epsilon}$. The resulting finite-dimensional DVI has solutions: it is a differential inclusion of Filippov type. The problem then is to show that some subsequence converges in a suitable sense, and that the limit indeed solves the problem, at least in terms of a suitably weak variational formulation.

Most of the argument follows that of the frictionless case. The standard energy bounds give weak* convergence $\mathbf{u}_{m,\epsilon} \rightharpoonup^* \widehat{\mathbf{u}}$ in $C^{1/2}(0,T; H^1(\Omega))$ (the space of Hölder continuous functions $[0,T] \to H^1(\Omega)$ with exponent $1/2$) and $\mathbf{v}_{m,\epsilon} \rightharpoonup^* \widehat{\mathbf{v}}$ in $L^2(0,T; H^1(\Omega)) \cap L^\infty(0,T; L^2(\Omega))$, where $\mathbf{v}_{m,\epsilon} = \partial \mathbf{u}_{m,\epsilon}/\partial t$ and $\widehat{\mathbf{v}} = \partial \widehat{\mathbf{u}}/\partial t$. For the friction forces, we use strong convergence of $N_{m,\epsilon} = \left| \mathbf{n} \cdot \mathcal{R}\sigma\left[\mathbf{u}_{m,\epsilon}, \partial \mathbf{u}_{m,\epsilon}/\partial t\right] \cdot \mathbf{n} \right| \to \widehat{N}$ in $L^2(0,T; H^{1/2}(\Gamma_C))$ coming from compactness of $\mathcal{R}$.

Chapter 7

Fractional Index Problems

> Perhaps, it ... prompted l'Hospital to ask [about $d^n y/dx^n$] "*What if n be 1/2?*". Leibnitz, in 1695, replied "*It will lead to a paradox*" but added ... "*From this apparent paradox, one day useful consequences will be drawn.*"
>
> *Virginia Kiryakova*

Fractional index differential variational inequalities (DVIs) are not mixed-index DVIs. In fact, of the problems considered in this book, only convolution complementarity problems (CCPs) can be considered to have a noninteger index. However, they can arise naturally, as in the example of Petrov and Schatzman [207] for a viscoelastic rod striking a rigid surface at one end. Using the notation $(\cdot)_t = \partial(\cdot)/\partial t$ and $(\cdot)_x = \partial(\cdot)/\partial x$, the equations of motion for this situation are

$$u_{tt} = u_{xx} + \beta u_{txx}, \qquad t > 0, \quad 0 < x < L, \tag{7.1}$$

$$0 = u_x(t,L) + \beta u_{tx}(t,L), \qquad t > 0, \tag{7.2}$$

$$N(t) = u_x(t,0) + \beta u_{tx}(t,0), \qquad t > 0, \tag{7.3}$$

$$0 \le N(t) \perp u(t,0) \ge 0, \qquad t > 0. \tag{7.4}$$

The orthogonality condition "$N(t) \perp u(t,0)$" means that $N(t)u(t,0) = 0$ for almost all t; together with the nonnegativity conditions, this is equivalent to

$$0 = \int_0^T N(t)u(t,0)\,dt.$$

From the fundamental solution for this partial differential equation we can construct a CCP for $N(t)$:

$$u(t,0) = \int_0^t k(t-\tau)N(\tau)\,d\tau + q(t),$$
$$0 \le N(t) \perp u(t,0) \ge 0, \qquad t > 0.$$

The problem is that for this problem, $k(t) \sim k_0 t^{1/2}$ for $t > 0$ small, so the theory of Section 4.6.2 is not applicable.

Another example comes from the heat equation with a source at the origin controlled by a thermostat set at temperature U_0, also at the origin [249]:

$$u_t = u_{xx} + z(t)\,\delta(x), \tag{7.5}$$

$$0 \le z(t) \perp u(t,0) - U_0 \ge 0. \tag{7.6}$$

As usual, $\delta(\cdot)$ is the Dirac-δ function. The unknown $z(\cdot)$ is the rate at which the source produces heat. This also leads to a CCP:

$$u(t,0) = \int_0^t k(t-\tau)z(\tau)\,d\tau + q(t),$$
$$0 \le z(t) \perp u(t,0) - U_0 \ge 0.$$

The function $q(t)$ is the value that $u(t,0)$ would have if $z(\cdot) \equiv 0$.

This time we have $k(t) \sim k_0\, t^{-1/2}$ for $t > 0$ small.

The theory of fractional differentiation and fractional integration [145, 147] can be used to identify the index of these problems.

7.1 Fractional differentiation and integration

Fractional differentiation and integration are operators that can be represented in terms of convolutions with particular distributions. In particular, indefinite integration is convolution with the constant one. The main tool to define these is the Laplace transform (see Section C.2). Using the property that Laplace transforms of convolutions are products of the Laplace transforms ($\mathcal{L}\big[f * g\big] = \mathcal{L}f \cdot \mathcal{L}g$), we can investigate fractional integration and differentiation in terms of Laplace transforms. In particular, for the constant function one, $\mathcal{L}1(s) = s^{-1}$. Convolution with the Dirac-δ function does not change anything: ($f * \delta = f$) and $\mathcal{L}\delta(s) = 1$. On the other hand, $\delta' * f = f'$ and $\mathcal{L}\big[\delta'\big](s) = s$. So if the convolution $\delta^{(\alpha)} * f$ represents the αth derivative of f, then $\mathcal{L}\delta^{(\alpha)}(s) = s^\alpha$. For α negative, the result is a fractional indefinite integral.

Put $\alpha = -\beta < 0$. Then $\delta^{(-\beta)}$ is an ordinary function $[0,\infty) \to \mathbb{R}$ given by

$$\delta^{(-\beta)}(t) = \Gamma(\beta)^{-1}\, t^{\beta-1}, \qquad t > 0,$$

where $\Gamma(\beta)$ is Euler's Γ-function:

$$\Gamma(\beta) = \int_0^\infty e^{-t}\, t^{\beta-1}\, dt.$$

Also note that $\delta^{(-\beta)}(t) = 0$ for $t < 0$. The Fourier transform of $\delta^{(-\beta)}$ is interpreted in the sense of distributions, and

$$\mathcal{F}\Big[\delta^{(-\beta)}\Big](\omega) = (i\,\omega)^{-\beta}$$

with $i\omega$ understood as having argument $\pm\pi/2$ as a complex number. Thus, for $\omega > 0$, we understand $(i\,\omega)^{-\beta}$ to be $e^{-i\beta\pi/2}\omega^{-\beta}$, and for $\omega < 0$ to be $e^{+i\beta\pi/2}\,|\omega|^{-\beta}$.

For $\alpha > 0$, the distribution $\delta^{(\alpha)}$ can be understood as the inverse Fourier transform of $\psi_\alpha(\omega) = (i\,\omega)^\alpha$ using an appropriate branch of the function (that is, for $\omega > 0$, $(i\,\omega)^\alpha =$

$e^{+i\alpha\pi/2}\,\omega^\alpha$, and for $\omega < 0$, $(i\,\omega)^\alpha = e^{-i\alpha\pi/2}\,|\omega|^\alpha$). Alternatively, we can split $\alpha = m - \gamma$ with m an integer and $0 \le \gamma < 1$ and use

$$\begin{aligned}\delta^{(\alpha)} * f(t) &= \delta^{(m)} * \delta^{(-\gamma)} * f(t)\\ &= \frac{d^m}{dt^m}\int_0^t \frac{(t-\tau)^{\gamma-1}}{\Gamma(\gamma)} f(\tau)\,d\tau.\end{aligned}$$

In this way we can calculate αth derivatives or integrals of a function f for any real value of α.

Recall that for a CCP,

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^*,$$

where K^* is the dual cone to K (see (B.8)), we say this is an index-zero CCP if $m(t) = m_0\,\delta(t) + m_1(t)$ with m_0 nonsingular and m_1 is a measure with no atom (or Dirac-δ function) at $t = 0$; it is index one if $m(t)$ is a function of bounded variation (locally) and $m(0^+)$ is nonsingular. This can be generalized to say that the CCP has index α if $\delta^{(\alpha)} * m(t) = m_0\,\delta(t) + m_1(t)$, where m_0 is nonsingular and m_1 is a measure with no atom at $t = 0$. Simply put, if $m(t) \sim m_0\,t^{\alpha-1}$ as $t \downarrow 0$ with m_0 nonsingular, then the CCP has index α. Thus the problem of Petrov and Schatzman (7.1)–(7.4) has index 3/2, while (7.5)–(7.6) has index 1/2.

7.2 Existence and uniqueness

There are two cases of fractional indexes that must be considered separately: index between zero and one, and index between one and two. We already have existence results for index-zero and index-one CCPs. We can construct existence proofs for fractional index problems by approximating a fractional index CCP with an index-zero or index-one CCP as appropriate. The crucial tool is the Fourier transform (see Section C.4). In particular, we note that

$$\mathcal{F}\delta^{(\alpha)}(\omega) = (i\,\omega)^\alpha\,.$$

If α is an integer, this is well defined, but for fractional α we have the problem that there may be branch cuts in the complex plane:

$$\arg i\,\omega = \begin{cases} +\pi/2 + 2m\pi & \text{if } \omega > 0,\\ -\pi/2 + 2m\pi & \text{if } \omega < 0.\end{cases}$$

If α is irrational, then $(i\,\omega)^\alpha = e^{i(\pm 1/2 + 2m)\alpha\pi}\,|\omega|^\alpha$ can be made arbitrarily close to *any* complex number with magnitude $|\omega|^\alpha$ by a suitable choice of m. We first need to show that we can set $m = 0$; this is the *principal branch* of $\arg i\,\omega$.

If we consider $\beta < 0$, then the integral $\int_{-\infty}^{+\infty} e^{-i\omega t}\,\delta^{(\beta)}(t)\,dt$ is not convergent, as $\delta^{(\beta)}$ is not an integrable function on the real line. But, for $\beta < 0$, the function $t \mapsto e^{-\varepsilon t}\,\delta^{(\beta)}(t)$ is integrable. In fact, for $\beta < 0$,

$$\int_{-\infty}^{+\infty} e^{-i\omega t}\,e^{-\varepsilon t}\,\delta^{(\beta)}(t)\,dt = \frac{1}{\Gamma(|\beta|)}\int_0^\infty e^{-(i\omega+\varepsilon)t}\,t^{-1+|\beta|}\,dt.$$

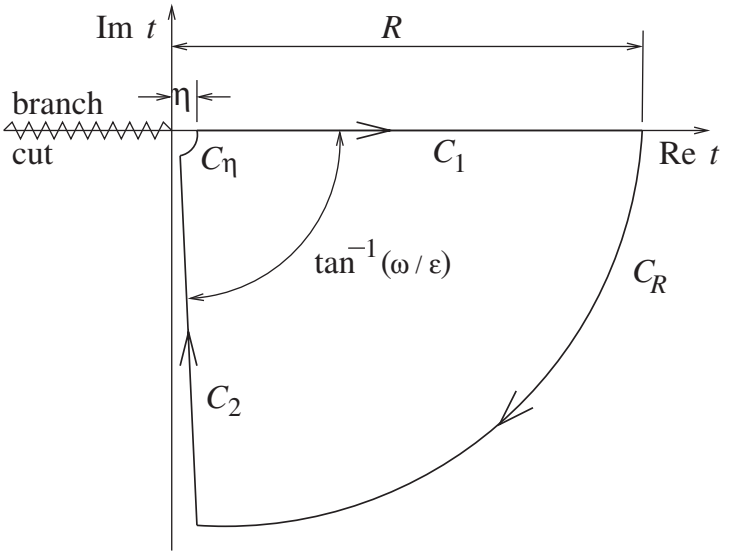


Figure 7.1: Contour integral for change of variables.

This integral can be exactly computed via a change of (complex) variables which can be justified (for $\omega > 0$) via the contour integral shown in Figure 7.1 with $0 < \eta, R$. From Cauchy's residue theorem, the integral around the contour is zero. Taking $\eta \downarrow 0$ and $R \to \infty$, we get $\int_{C_R} e^{-(i\omega+\varepsilon)t}\, t^{-1+|\beta|}\, dt$ and $\int_{C_\eta} e^{-(i\omega+\varepsilon)t}\, t^{-1+|\beta|}\, dt$ both go to zero. Thus the differences between the integrals over C_1 and C_2 go to zero. On C_2 set $t = (-i+\varepsilon/\omega)s$. Then, for $\omega > 0$,

$$\begin{aligned}
&\int_0^\infty e^{-(i\omega+\varepsilon)t}\, t^{-1+|\beta|}\, dt \\
&\quad = \int_0^\infty e^{-(i\omega+\varepsilon)(-i+\varepsilon/\omega)s}\, (-i+\varepsilon/\omega)^{-1+|\beta|}\, s^{-1+|\beta|}\, (-i+\varepsilon/\omega)\, ds \\
&\quad = (-i+\varepsilon/\omega)^{|\beta|} \int_0^\infty e^{-\left(\omega+\varepsilon^2/\omega\right)s}\, s^{-1+|\beta|}\, ds \\
&\quad = (-i+\varepsilon/\omega)^{|\beta|} \left(\omega+\varepsilon^2/\omega\right)^{-|\beta|} \Gamma(|\beta|).
\end{aligned}$$

Taking $\varepsilon \downarrow 0$ gives

$$\mathcal{F}\delta^{(\beta)}(\omega) = i^{-|\beta|}\, \omega^{-|\beta|} \;=\; (i\omega)^\beta \qquad \text{for } \beta < 0.$$

Note that thanks to the branch cut, we remain in the principal branch of $z \mapsto z^\beta$, and that $(i\omega)^\beta$ should be understood in this sense.

This may appear to be a diversion from our main concern, but we will need the sign of the Fourier transform to have positive real part, and this requires the correct branch of $z \mapsto z^\beta$.

Consider first the CCP

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^*, \qquad \text{where}$$
$$m(t) = m_0 t^{\alpha-1}, \qquad 0 < \alpha < 1,$$

and m_0 is a symmetric positive definite matrix. Now $\mathcal{F}(m * z) = (2\pi)^{-1} \mathcal{F}m \cdot \mathcal{F}z$. Applying Plancherel's theorem (3.34) to $z(t) \perp (m * z)(t) + q(t)$, we have

$$\begin{aligned} 0 &= \int_0^\infty \langle z(t), (m * z)(t) + q(t) \rangle \, dt \\ &= \frac{1}{2\pi} \mathrm{Re} \int_{-\infty}^{+\infty} \langle \mathcal{F}z(\omega), \mathcal{F}m(\omega)\mathcal{F}z(\omega) + \mathcal{F}q(\omega) \rangle \, d\omega. \end{aligned}$$

But here $\mathcal{F}m(\omega) = m_0 \Gamma(\alpha) (i\omega)^{-\alpha}$ using the principal branch of $z \mapsto z^{-\alpha}$. In this branch, for $\omega > 0$, $(i\omega)^{-\alpha} = e^{-i\pi\alpha/2} \omega^{-\alpha}$, whose real part is $\cos(\pi\alpha/2)\, \omega^{-\alpha}$. Similarly, for $\omega < 0$, the real part is $\cos(\pi\alpha/2)\, |\omega|^{-\alpha}$. For $0 < \alpha < 1$, this is always a positive quantity. For symmetric positive definite m_0, there is an $\eta_0 > 0$ where $\langle w, m_0 w \rangle \geq \eta_0 \|w\|^2$ for all w. Thus, for $0 < \alpha < 1$,

$$\begin{aligned} 0 &\geq \frac{1}{2\pi} \eta_0 \Gamma(\alpha) \cos(\pi\alpha/2) \int_{-\infty}^{+\infty} \|\mathcal{F}z(\omega)\|^2 \, |\omega|^{-\alpha} \, d\omega \\ &\qquad - \frac{1}{2\pi} \int_{-\infty}^{+\infty} \|\mathcal{F}z(\omega)\| \left(1 + \omega^2\right)^{-\alpha/2} \|\mathcal{F}q(\omega)\| \left(1 + \omega^2\right)^{+\alpha/2} d\omega \\ &\geq \eta_0 \cos(\pi\alpha/2)\, \Gamma(\alpha) \, \|z\|^2_{H^{-\alpha/2}} - \|z\|_{H^{-\alpha/2}} \, \|q\|_{H^{\alpha/2}}, \end{aligned}$$

which gives a uniform bound on the solution z in $H^{-\alpha/2}(\mathbb{R}; X)$ in terms of the norm of q in $H^{+\alpha/2}(\mathbb{R}; X')$. If the index α exceeds one, then $\cos(\pi\alpha/2)$ can be negative, and this does not give any bound on z. Note that we need *symmetry* of m_0 since for $w = u + iv$, $\omega > 0$, and $\lambda = (i\omega)^{-\alpha} = \rho + i\sigma$,

$$\begin{aligned} &\mathrm{Re}\, \langle w, \lambda m_0 w \rangle \\ &\quad = \mathrm{Re}\, \overline{w}^T m_0 \lambda w \\ &\quad = \mathrm{Re}\, \frac{1}{2} \overline{w}^T \left[m_0 \lambda + \overline{m_0 \lambda}^T \right] w \\ &\quad = \mathrm{Re}\, \frac{1}{2} \overline{(u + iv)}^T \left[m_0 (\rho + i\sigma) + m_0^T (\rho - i\sigma) \right] (u + iv) \\ &\quad = \frac{\rho}{2} \overline{w}^T \left(m_0 + m_0^T \right) w + \frac{\sigma}{2} v^T \left(m_0 - m_0^T \right) u. \end{aligned}$$

If $0 < \alpha < 1$ and $m_0 = m_0^T$, we have

$$\mathrm{Re} \left\langle w, (i\omega)^{-\alpha} m_0 w \right\rangle \geq \cos(\alpha\pi/2)\, |\omega|^{-\alpha} \, \|w\|^2 > 0 \tag{7.7}$$

for $w \neq 0$. But if $m_0 \neq m_0^T$, then we have to restrict α further. In fact, we can obtain a bound

$$\mathrm{Re} \left\langle w, (i\omega)^{-\alpha} m_0 w \right\rangle \geq \eta \, \|w\|^2 \tag{7.8}$$

with $\eta > 0$, provided

$$0 < \alpha < \frac{2}{\pi} \tan^{-1} \left(\frac{\lambda_{min} \left(m_0 + m_0^T \right)}{\left\| m_0 - m_0^T \right\|} \right) \tag{7.9}$$

for $m_0 \neq m_0^T$, where $\lambda_{min}(B)$ is the minimum eigenvalue for a symmetric matrix B. Note that for existence of solutions to index-zero problems we need only m_0 to be positive definite, while uniqueness for index one requires that m_0 also be symmetric.

To turn these observations into a reasonably general existence and uniqueness theorem, we need to do a few things:

- create an approximate index-zero problem with kernel $m_\epsilon(t) = \epsilon\, \delta(t)\, I + m(t)$;
- remove the restriction that $m(t) = m_0\, t^{\alpha-1}$ so that we require only that $\mathcal{F}m(\omega) \sim m_0\, (i\omega)^{-\alpha}$ for large $|\omega|$;
- restrict our attention to a finite interval $[0, T]$ for compactness in the appropriate Sobolev space;
- restrict $X = \mathbb{R}^n$.

Theorem 7.1. *Suppose that $m(t) = m_0\, t^{\alpha-1} + m_1(t)$, where $0 < \alpha < 1$ and $m_1 \colon [0,T] \to \mathbb{R}^{n\times n}$ have the following properties:*

- *$\omega \mapsto \omega^\alpha \mathcal{F}\left(m_1 \chi_{[0,T]}\right)(\omega)$ converges to zero uniformly as $T \downarrow 0$;*
- *$\left\| \mathcal{F}\left(m_1 \chi_{[0,T]}\right)(\omega) \right\| \leq C\, |\omega|^{-\beta}$ for some $\beta > \alpha$ for all $T > 0$;*
- *m_0 is positive definite satisfying (7.9) (taking $\alpha < 1$ if m_0 is symmetric); and*
- *$q \in H^{\alpha/2}(0, T; \mathbb{R}^n)$;*

then there exists one and only one solution $z \in H^{-\alpha/2}(0, T; \mathbb{R}^n)$ to the CCP

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^* \qquad \text{for all } t. \tag{7.10}$$

A proof of a slightly weaker version of this result can be found in [249].

Proof. We prove this result for sufficiently small $T > 0$. Once we have this result on a sufficiently small interval, we can extend the solution since for $z_1 \in H^{-\alpha/2}(0, T; \mathbb{R}^n)$, we have $m * z_1 \in H^{+\alpha/2}(0, T; \mathbb{R}^n)$. Let $\chi_E(t) = 1$ if $t \in E$ and $\chi_E(t) = 0$ if $t \notin E$. Let $\psi(t) = m_0\, t^{\alpha-1}$. Then, using (7.8), $\langle w, \mathcal{F}\psi(\omega) w \rangle \geq \eta \|w\|^2$ for all w for some $\eta > 0$.

Consider the index-zero approximate problem

$$K \ni z_\epsilon(t) \perp (\epsilon\, I\, \delta + m) * z_\epsilon(t) + q(t) \in K^*.$$

Solutions exist and are unique for this problem, provided $\epsilon > 0$ by the results in Section 5.1.1. We wish to show that the z_ϵ have a convergent subsequence as $\epsilon \downarrow 0$. Let $\gamma_T = \max_\omega \left\| \omega^\alpha \mathcal{F}\left(m_1 \chi_{[0,T]}\right)(\omega) \right\|$; $\gamma_T \to 0$ as $T \downarrow 0$. Note that

$$
\begin{aligned}
0 &= \int_0^T \langle z_\epsilon(t), (\epsilon\, I\, \delta + m) * z_\epsilon(t) + q(t) \rangle \, dt \\
&= \int_{-\infty}^{+\infty} \left\langle z_\epsilon(t)\chi_{[0,T]}(t), \left(\epsilon\, I\, \delta + \psi_0 + m_1 \chi_{[0,T]}\right) * \left(z_\epsilon \chi_{[0,T]}\right)(t) + q(t) \right\rangle dt \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Big\langle \mathcal{F}\left(z_\epsilon \chi_{[0,T]}\right)(\omega), \\
&\qquad \mathcal{F}\left(\epsilon\, I\, \delta + \psi_0 + m_1 \chi_{[0,T]}\right)(\omega)\, \mathcal{F}\left(z_\epsilon \chi_{[0,T]}\right)(\omega) + \mathcal{F}q(\omega) \Big\rangle \, d\omega \\
&= \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Bigg\langle \mathcal{F}\left(z_\epsilon \chi_{[0,T]}\right)(\omega), \\
&\qquad \left(\epsilon\, I + \frac{m_0}{\Gamma(\alpha)} \omega^{-\alpha} + \mathcal{F}\left(m_1 \chi_{[0,T]}\right)(\omega)\right) \mathcal{F}\left(z_\epsilon \chi_{[0,T]}\right)(\omega) + \mathcal{F}q(\omega) \Bigg\rangle \, d\omega \\
&\geq \frac{1}{2\pi} \left[\epsilon \left\| \mathcal{F}\left(z_\epsilon \chi_{[0,T]}\right) \right\|_{L^2}^2 + (\eta - \gamma_T) \int_{-\infty}^{+\infty} \omega^{-\alpha} \left\| \mathcal{F}\left(z_\epsilon \chi_{[0,T]}\right)(\omega) \right\|^2 d\omega \right. \\
&\qquad \left. - \left\| z_\epsilon \chi_{[0,T]} \right\|_{H^{-\alpha/2}} \|q\|_{H^{\alpha/2}} \right] \\
&\geq \frac{1}{2\pi} \left[(\eta - \gamma_T) \left\| z_\epsilon \chi_{[0,T]} \right\|_{H^{-\alpha/2}}^2 - \left\| z_\epsilon \chi_{[0,T]} \right\|_{H^{-\alpha/2}} \|q\|_{H^{\alpha/2}} \right].
\end{aligned}
$$

Thus, for sufficiently small $T > 0$, $\gamma_T < \eta$, and so

$$
\left\| z_\epsilon \chi_{[0,T]} \right\|_{H^{-\alpha/2}} \leq \frac{1}{\eta - \gamma_T} \|q\|_{H^{\alpha/2}}.
$$

Since $H^{-\alpha/2}(0,T;\mathbb{R}^n)$ is a Hilbert space, there is a weakly convergent subsequence (which we also denote by z_ϵ) by Alaoglu's theorem. Let $z \in H^{-\alpha/2}(0,T;\mathbb{R}^n)$ be the limit of such a subsequence. Then by Mazur's lemma we can show that $z(t) \in K$ for almost all t via weak convergence. On the other hand, $m * z_\epsilon + q$ converges weakly to $m * z + q$, so for any smooth $\xi : [0,T] \to K$ we have

$$
\begin{aligned}
0 &\leq \int_0^T \langle \epsilon z_\epsilon(t) + (m * z_\epsilon)(t) + q(t), \xi(t) \rangle \, dt \\
&\to \int_0^T \langle (m * z)(t) + q(t), \xi(t) \rangle \, dt,
\end{aligned}
$$

taking limits $\epsilon \downarrow 0$ in the subsequence. Thus $(m * z)(t) + q(t) \in K^*$ for all $t > 0$. Note that by positivity, the quadratic map $u \mapsto \int_0^T \langle u, m * u \rangle \, dt$ is a convex function. Then, again by

Mazur's lemma,

$$\begin{aligned}
&\int_0^T \langle z(t), (m*z)(t)+q(t)\rangle\, dt \\
&\qquad \le \liminf_{\epsilon\downarrow 0} \int_0^T \langle z_\epsilon(t), (m*z_\epsilon)(t)+q(t)\rangle\, dt \\
&\qquad \le \liminf_{\epsilon\downarrow 0} \int_0^T \langle z_\epsilon(t), \epsilon z_\epsilon(t)+(m*z_\epsilon)(t)+q(t)\rangle\, dt \\
&\qquad = 0,
\end{aligned}$$

so $\int_0^T \langle z(t), (m*z)(t)+q(t)\rangle\, dt \le 0$. It cannot be negative since $z(t) \in K$ and $(m*z)(t)+q(t) \in K^*$ for almost all t. Thus we have existence of a solution on an interval $[0,T]$ for $T>0$ sufficiently small. This can be extended to any interval by means of shifting $t=T$ to $t=0$ and incorporating the solution on $[0,T]$ into q.

To show uniqueness, suppose

$$\begin{aligned}
K \ni z_1(t) \perp (m*z_1)(t)+q_1(t) \in K^*, \\
K \ni z_2(t) \perp (m*z_2)(t)+q_2(t) \in K^*.
\end{aligned}$$

Then, if $\zeta = z_1 - z_2$ and $\theta = q_1 - q_2$, we have

$$\begin{aligned}
0 &\ge \int_0^T \langle \zeta(t), (m*\zeta)(t)+\theta(t)\rangle\, dt \\
&\ge (\eta-\gamma_T)\left\|\zeta\,\chi_{[0,T]}\right\|^2_{H^{-\alpha/2}} - \left\|\zeta\,\chi_{[0,T]}\right\|_{H^{-\alpha/2}} \|\theta\|_{H^{\alpha/2}},
\end{aligned}$$

and so $\left\|\zeta\,\chi_{[0,T]}\right\|_{H^{-\alpha/2}} \le \|\theta\|_{H^{\alpha/2}}/(\eta-\gamma_T)$. In particular, if $q_1=q_2$, then $\theta=0$, and so $\zeta = z_1 - z_2 = 0$, establishing uniqueness. □

Clearly the uniqueness result can be extended to show that the map $q \mapsto z$ is a Lipschitz continuous map $H^{+\alpha/2}(0,T;\mathbb{R}^n) \to H^{-\alpha/2}(0,T;\mathbb{R}^n)$. This is the best regularity result that we can expect as convolution with m maps $H^{-\alpha/2}(0,T;\mathbb{R}^n)$ to $H^{+\alpha/2}(0,T;\mathbb{R}^n)$.

7.3 Further regularity results

Differentiability lemmas can be used to get stronger regularity results with stronger assumptions on q. In particular, for q smoother and $q(0) \in K^*$ we can get stronger regularity on the solution z. However, unlike most linear problems, making q arbitrarily smooth and satisfying compatibility conditions cannot make the solution z arbitrarily smooth. There is a natural limit to how smooth z can be, as we saw for DVIs in Section 3.2.

Theorem 7.2. *If $q \in H^{1+\alpha/2}(0,T;\mathbb{R}^n)$ and $q(0) \in K^*$ and the CCP* (7.10) *satisfies the conditions of Theorem* 7.1*, the solution z is in $H^{1-\alpha/2}(0,T;\mathbb{R}^n)$.*

Proof. For the approximate index-zero CCP

$$K \ni z_\epsilon(t) \perp \epsilon z_\epsilon(t) + (m*z_\epsilon)(t) + q(t) \in K^*, \tag{7.11}$$

there are solutions that are absolutely continuous since q is absolutely continuous. We use the differentiation property, Lemma 3.2: $K \ni a(t) \perp b(t) \in K^*$ for all t implies that $\langle a'(t), b'(t)\rangle \leq 0$, provided a and b are absolutely continuous. Applying this to (7.11) gives

$$0 \geq \left\langle z_\epsilon'(t), \epsilon z_\epsilon'(t) + \left(m * z_\epsilon'\right)(t) + q'(t)\right\rangle.$$

Integrating over $[0, T]$ and applying the bounds obtained in Theorem 7.1 for $T > 0$ sufficiently small,

$$0 \geq \epsilon \left\|z_\epsilon'\right\|_{L^2}^2 + (\eta - \gamma_T)\left\|z_\epsilon'\right\|_{H^{-\alpha/2}}^2 - \left\|z_\epsilon'\right\|_{H^{-\alpha/2}} \left\|q'\right\|_{H^{\alpha/2}}$$

so that

$$\left\|z_\epsilon'\right\|_{H^{-\alpha/2}} \leq \frac{1}{\eta - \gamma_T}\left\|q'\right\|_{H^{\alpha/2}}.$$

To obtain the bound on z_ϵ itself, we use the fact that $q(0) \in K$ to get $z_\epsilon(0) = 0$. Thus z_ϵ is uniformly bounded in $H^{1-\alpha/2}(0, T; \mathbb{R}^n)$. Thus by Alaoglu's theorem there is a weakly convergent subsequence; let $z \in H^{1-\alpha/2}(0, T; \mathbb{R}^n)$ be the weak limit in such a subsequence. Using the techniques of Theorem 7.1 we can show that this weak limit is in fact a solution of the CCP (which is unique by Theorem 7.1). Thus solutions lie in $H^{1-\alpha/2}(0, T; \mathbb{R}^n)$, as desired. □

This proof is one of the few occasions in which a differentiation lemma involving two derivatives is useful.

7.4 Index between one and two

If the index α is between one and two, then $\cos(\alpha\pi/2) < 0$ and the arguments above do not work. This is unfortunate since, as noted in the introduction to this chapter, index $\alpha = 3/2$ naturally arises in studying impact of a viscoelastic rod. It is, however, possible to obtain existence results, though it is unclear at the time of this writing whether uniqueness holds for these problems or not. The problem then is a CCP of the form

$$K \ni z(t) \perp (m * z)(t) + q(t) \in K^*$$

with $m(t) \sim m_0 t^{\alpha-1}$ with $1 < \alpha < 2$ for t small and positive, and m_0 a positive definite symmetric matrix.

To show existence we also use an index reduction strategy, but instead of reducing the index to zero as was done for $0 < \alpha < 1$, we reduce it to one. Let

$$m_\epsilon(t) = \epsilon\, I\, H(t) + m(t),$$

where $H(t) = 1$ for $t > 0$ and $H(t) = 0$ for $t < 0$ is the *Heaviside function*. Provided $m(\cdot)$ and $q(\cdot)$ are sufficiently smooth with $q(0) \in K^*$, solutions to the index-one CCP

$$K \ni z_\epsilon(t) \perp (m_\epsilon * z_\epsilon)(t) + q(t) \in K^*$$

exist and are unique by Theorem 5.6. Using a one-derivative differentiation lemma (Lemma 3.2),

$$0 = \left\langle z_\epsilon(t), \epsilon\, z_\epsilon(t) + \left(m' * z_\epsilon\right)(t) + q'(t)\right\rangle.$$

Integrating over a sufficiently small interval $[0, T]$ with $T > 0$, we note that

$$\mathcal{F}\left[m'\right](\omega) = i\omega \mathcal{F}m(\omega) \sim \Gamma(\alpha) m_0 \, (i\omega)^{1-\alpha}$$

in the principal branch. Now the real part of $i^{1-\alpha}$ is $\cos((\alpha - 1)\pi/2)$, which is positive for $1 < \alpha < 2$. Then we can apply the methods of Theorem 7.1 to show existence via boundedness of z_ϵ in $H^{-\alpha/2}(0, T; \mathbb{R}^n)$, provided $q \in H^{1+\alpha/2}(0, T; \mathbb{R}^n)$, but not uniqueness. The usual approach to showing uniqueness is to suppose that z_1 and z_2 are two solutions of the CCP and then set $\zeta = z_1 - z_2$; from linearity of the convolution and the fact that K and K^* are dual cones, we have $\int_{[0,T]} \langle \zeta, m * \zeta \rangle \leq 0$. If the convolution operator $\zeta \mapsto m * \zeta$ is elliptic or positive definite, we can conclude that $\zeta \equiv 0$, so $z_1 \equiv z_2$. However, in this case the leading part of the Fourier transform of m is $(i\omega)^{-\alpha} \, m_0$ and the real part of $(i\omega)^{-\alpha} = \cos(\alpha\pi/2) \, |\omega|^{-\alpha}$ is negative, so this convolution operator is definitely *not* elliptic or positive definite.

Further details can be found in [243].

Chapter 8

Numerical Methods

8.1 Choices

Numerical methods for dynamic problems with inequality constraints take several forms. The main families of methods are

- penalty methods, or the related index reduction methods,
- active set methods which track which inequality constraints are "active" (that is, where the inequalities are equalities), and
- time-stepping methods, in which for each time step, a CP or VI is solved.

Penalty methods aim to turn a nonsmooth or discontinuous differential equation into one that is smooth, and so we can use standard methods for differential equations. The true trajectory is often made up of smooth pieces joined by "kinks" or "jumps," so active set methods aim to find the smooth pieces and the points at which a kink or jump occurs; once the kink or jump is reached, a new smooth differential equation is set up for the next piece. Time-stepping methods have the largest computational effort per time step, but they can be very effective when the active inequalities change frequently.

Penalty methods are perhaps the most common methods used in practice, although the other techniques (particularly time-stepping methods) are gaining popularity. Penalty methods work by replacing the nonsmoothness of the original problem with a smooth, or at least smoother, approximation. Then standard smooth ordinary differential equation solvers can be applied to the smooth approximation. This naturally depends on both the time step and the accuracy of the smooth approximation. Typically, the smooth approximation is a *stiff* differential equation, which often means that small time steps are needed for accurate solutions.

Active set methods can give the greatest accuracy, since the only errors are typically those due to the smooth differential equation solver used for each piece. However, the main difficulty in using them arises if the active set (the set of inequalities that happen to be equalities) changes often. At each change of the active set, some special calculation must be performed to identify the new active set. There can be problems if there is some degeneracy in the problem so that determining the new active set can depend on the data of

the problem in a sensitive way. Problems in which the active set changes infinitely often in a finite time interval pose a particular challenge to these methods.

Time-stepping methods can be the most computationally expensive methods, but they are usually not vulnerable to problems with degeneracies or rapid changes in the active set. The main difficulty is that a CP or VI must be solved with each step. Fortunately, advances in techniques for solving CPs and VIs have brought these methods to the forefront. In particular, the development of nonsmooth Newton methods along with reformulations of CPs and VIs as (nonsmooth) systems of equations means that for many time steps only a single linear system must be solved. This brings the computational cost close to or less than that for penalty or active set methods.

To explain these methods and give the reader a way of comparing these approaches, we will focus mainly on the problems of mechanical impact and Coulomb friction. Mechanical impact problems are essentially index two, while Coulomb friction problems are index one. These two problems give an overview and a means of comparing numerical methods.

8.1.1 Methods for smooth differential equations

Numerical methods for smooth differential equations

$$\frac{dx}{dt} = f(t, x), \qquad x(t_0) = x_0$$

have been around at least since Euler and his method. These methods were improved by the work of Heun [126], Runge [218], and Kutta [150], which led to modern Runge–Kutta methods [18, 46]. Another set of methods that have been widely used are multistep methods, of which there are three main families: Adams–Bashforth methods [26], Adams–Moulton methods [184], and the backward difference (BDF) methods [69].

Higher order methods all assume higher orders of smoothness in the differential equation and its solution. For many of the problems that we consider here, the *solution* is often smooth (at least on certain intervals) even though the differential equation or inclusion definitely is *not* smooth. For such problems, often it is appropriate to use methods for smooth but stiff differential equations. Methods for such problems include many implicit Runge–Kutta methods and BDF multistep methods and can be found in [18, 46, 119]. The books [46, 119] in particular provide a great deal of information about their behavior both theoretically and practically, and about advanced implementations including, for example, adaptive control of step sizes.

An extreme example of a stiff differential equation

$$\begin{aligned} \frac{dx}{dt} &= f(t, x(t), y(t)), \qquad x(t_0) = x_0, \\ \epsilon \frac{dy}{dt} &= g(t, x(t), y(t)), \qquad y(t_0) = y_0 \end{aligned}$$

is the case where ϵ goes to zero, giving the ordinary differential algebraic equations (DAEs)

$$\begin{aligned} \frac{dx}{dt} &= f(t, x(t), y(t)), \qquad x(t_0) = x_0, \\ 0 &= g(t, x(t), y(t)), \qquad y(t_0) = y_0. \end{aligned}$$

As with DVIs, the concept of index is very important for understanding DAEs and their numerical methods. The index used for DAEs is the smallest integer $k \geq 0$ such that the equations $\left(d^j/dt^j\right) g(t, x(t), y(t)) = 0$ for $j = 0, 1, \ldots, k$ (substituting $dx/dt = f(t,x,y)$) are sufficient to obtain a *differential equation* for $y(t)$. As a result, the index for a DAE is one more than the index for the corresponding DVI.

Methods for such problems were first developed by Gear and Petzold [110] based on BDF multistep methods. More recently Runge–Kutta methods have been applied to these problems. For a thorough discussion of methods for DAEs, see [15, 38, 46, 119], although more recent Runge–Kutta methods for higher index problems can be found in [131, 132, 133, 134].

8.2 Penalty and index reduction methods

Consider the mechanical impact illustrated in Figure 1.1:

$$\begin{aligned} m\frac{d^2y}{dt^2} &= N - mg, \\ 0 \leq N(t) \perp y(t) - r &\geq 0. \end{aligned}$$

This is an index-two problem, and it can have solutions with $dy/dt(t)$ discontinuous and $N(t)$ impulsive.

The simplest approach to setting up a penalty approximation is to set $N_\epsilon = (y-r)_-/\epsilon$, where $\epsilon > 0$ is a small constant, where $s_- = \max(0, -s)$. This can be represented as the solution of the CP

$$0 \leq N_\epsilon(t) \perp \epsilon N_\epsilon(t) + y_\epsilon(t) - r \geq 0.$$

With this change, we have changed the index of the problem from two for the original problem to zero for the penalty approximation. The penalty approximation is then

$$m\frac{d^2y_\epsilon}{dt^2} = \frac{1}{\epsilon}(y_\epsilon - r)_- - mg.$$

This is a Lipschitz differential equation, and so it has solutions which are unique, given the initial conditions $y_\epsilon(t_0)$ and $dy_\epsilon/dt(t_0)$. These solutions can be found by standard numerical methods, such as Euler's method or a Runge–Kutta method. However, lack of smoothness means that the rate of convergence of these methods is kept down at first order. This can be improved by replacing $(y_\epsilon - r)_-$ with a smooth function $\varphi(r - y_\epsilon)$ where φ is nondecreasing, $\varphi(s) = 0$ for $s \leq 0$, $\varphi(s) > 0$ for $s > 0$, and $\varphi(s) \to \infty$ as $s \to \infty$. Then the corresponding smooth penalty approximation is

$$m\frac{d^2y_\epsilon}{dt^2} = \frac{1}{\epsilon}\varphi(r - y_\epsilon) - mg.$$

For Coulomb friction with known contact forces, consider the "brick on a ramp" problem illustrated by Figure 1.4, given by the differential inclusion

$$m\frac{d^2x}{dt^2} \in -\mu mg\,(\cos\theta)\,\mathrm{Sgn}\left(\frac{dx}{dt}\right) + mg\sin\theta,$$

where $\mathrm{Sgn}(v) = \{+1\}$ if $v > 0$, $\{-1\}$ if $v < 0$, and $[-1, +1]$ if $v = 0$. A Lipschitz penalty approximation can be obtained from the DVI formulation

$$\begin{aligned} m\frac{d^2x}{dt^2} &= F + mg\sin\theta, \\ \frac{dx}{dt}\cdot\left(\widetilde{F} - F\right) \geq 0 \qquad &\text{for all } \widetilde{F} \leq\in [-\mu N, +\mu N], \\ F &\in [-\mu N, +\mu N], \\ N &= mg\cos\theta. \end{aligned}$$

To do this we again reduce the index from one to zero by adding ϵF to the VI term:

$$\begin{aligned} m\frac{d^2x_\epsilon}{dt^2} &= F_\epsilon + mg\sin\theta, \\ \left(\epsilon F_\epsilon + \frac{dx_\epsilon}{dt}\right)\cdot\left(\widetilde{F} - F_\epsilon\right) \geq 0 \qquad &\text{for all } \widetilde{F} \in [-\mu N, +\mu N], \\ F_\epsilon &\in [-\mu N, +\mu N], \\ N &= mg\cos\theta. \end{aligned}$$

This is an index-zero DVI, and we can write $F_\epsilon = -\mu N\,\mathrm{sat}(\epsilon^{-1}\,dx_\epsilon/dt)$, where

$$\mathrm{sat}(s) = \begin{cases} +1, & +1 \leq s, \\ s, & -1 \leq s \leq +1, \\ -1, & s \leq -1. \end{cases}$$

A smooth approximation can be constructed using a smooth increasing function $\varphi(s)$ where $\varphi(s) \to +1$ as $s \to +\infty$ and $\varphi(s) \to -1$ as $s \to -\infty$. Then with a penalty parameter $\epsilon > 0$ we have the smoothed approximation

$$m\frac{d^2x_\epsilon}{dt^2} = -\mu mg\,(\cos\theta)\,\varphi\left(\frac{1}{\epsilon}\,\frac{dx_\epsilon}{dt}\right) + mg\sin\theta.$$

We can apply numerical methods for smooth ordinary differential equations to this problem. But we have to be careful that the step size we use goes to zero at the right rate to match the size of ϵ as $\epsilon \downarrow 0$. In particular, we should have a step size $h = h(\epsilon) = o(\epsilon)$; that is, $h(\epsilon)/\epsilon \to 0$ as $\epsilon \downarrow 0$.

8.3 Piecewise smooth methods

The basic idea of these methods is to decompose the solution into segments, each of which is smooth. These pieces can then be joined as the solution is computed. Since the solution is assumed to be smooth on each segment, we can use standard efficient numerical methods for smooth problems on each segment.

8.3.1 Index-zero problems

To be more specific, consider the DVI

$$\frac{dx}{dt} = f(x(t), z(t)), \qquad x(t_0) = x_0, \tag{8.1}$$

$$z(t) \in K \quad \& \quad 0 \le \langle \widetilde{z} - z(t), F(x(t), z(t)) \rangle \qquad \text{for all } \widetilde{z} \in K. \tag{8.2}$$

We will assume that the convex set K has a representation in terms of smooth convex functions:

$$K = \{ w \mid \phi_i(w) \le 0, i = 1, 2, \ldots, m \}. \tag{8.3}$$

We will assume that the Slater constraint qualification (B.22) holds: for some z^*,

$$\phi_i(z^*) < 0 \qquad \text{for all } i.$$

Then the equivalent condition to the VI (8.2) is

$$0 \in F(x(t), z(t)) + N_K(z(t)).$$

Using the representation

$$N_K(z) = \operatorname{co} \{ \nabla \phi_i(z) \mid \phi_i(z) = 0 \},$$

we can write the equivalent formulation of the VI as

$$0 = F(x(t), z(t)) + \sum_{i=1}^{m} \lambda_i(t) \nabla \phi_i(z(t)), \tag{8.4}$$

$$0 \le \lambda_i(t) \perp -\phi_i(z(t)) \ge 0. \tag{8.5}$$

For a given $z \in K$ we have the active set

$$\mathcal{I}(z) = \{ i \mid \phi_i(z) = 0 \}.$$

The crucial assumption is that there are times $t_0 < t_1 < t_2 < \cdots$ where $\mathcal{I}(z(t)) = \mathcal{I}_k$ for all $t \in (t_k, t_{k+1})$. Since $z(t) \in K$ for all t, if $j \notin \mathcal{I}_k$, then $\lambda_j(t) = 0$ for all $t \in (t_k, t_{k+1})$. On the interval (t_k, t_{k+1}) we have the DAEs

$$\frac{dx}{dt} = f(x(t), z(t)), \qquad x(t_0) = x_0, \tag{8.6}$$

$$0 = F(x(t), z(t)) + \sum_{i \in \mathcal{I}_k} \lambda_i(t) \nabla \phi_i(z(t)), \tag{8.7}$$

$$0 = \phi_i(z(t)), \qquad i \in \mathcal{I}_k, \tag{8.8}$$

$$0 = \lambda_i(t), \qquad i \notin \mathcal{I}_k. \tag{8.9}$$

If $x(t) \in \mathbb{R}^n$ and $z(t) \in \mathbb{R}^m$, then the number of equations and number of unknowns are both $n + m + |\mathcal{I}_k|$. Let $w_{\mathcal{J}} = \left[w_j \mid j \in \mathcal{J} \right]$, where $\mathcal{J}$ is a finite set of indices and w is a vector of the appropriate size.

For an index-zero DVI we expect $\nabla_z F(x,z)$ to be positive definite for all x and z. The Jacobian matrix of the *algebraic* part of the DAEs with respect to $\left[z^T, \lambda_{\mathcal{I}_k}^T\right]^T$ is

$$\begin{bmatrix} \nabla_z F + \sum_{i\in\mathcal{I}_k} \lambda_i^T \operatorname{Hess}\phi_i & \nabla\phi_{\mathcal{I}_k}^T \\ \nabla\phi_{\mathcal{I}_k} & 0 \end{bmatrix} \tag{8.10}$$

evaluated at $(x(t), z(t), \lambda(t))$. Since $\nabla_z F(x,z)$ is assumed to be positive definite, $\lambda_i(t) \geq 0$ for all i, and ϕ_i is convex for all i, we can see that $\nabla_z F(x,z) + \lambda_{\mathcal{I}_k}^T \operatorname{Hess}\phi_{\mathcal{I}_k}(z)$ should also be positive definite. Provided also that $\nabla\phi_{\mathcal{I}_k}(z(t))$ has full rank, the matrix (8.10) is nonsingular, so that the system (8.6)–(8.9) is a solvable system of DAEs. The index of this system as a system of DAEs is one [15, 38].

8.3.2 Index-one problems

For index-one DVIs, we will assume that $F(x,z) = G(x)$. Then (8.6)–(8.9) is still a system of DAEs, but now with a higher index. Typically, we differentiate (8.7)–(8.9) with respect to time to obtain equations for dz/dt and $d\lambda/dt$.

To illustrate this idea, consider the piecewise smooth but discontinuous differential equations

$$\frac{dx}{dt} = f_i(x(t)), \qquad \text{when } h_i(x(t)) < h_j(x(t)),\ j \neq i.$$

The h_i functions are called indicator functions, as they indicate which right-hand side to use for the differential equation. We assume that the functions f_i and h_i are smooth. This can be represented in terms of DVIs as

$$\frac{dx}{dt} = \sum_{i=1}^{m} \theta_i(t)\, f_i(x(t)), \qquad \theta(t) \in \Sigma_m, \tag{8.11}$$

$$0 \leq \left\langle \widetilde{\theta} - \theta(t), h(x(t)) \right\rangle \qquad \text{for all } \widetilde{\theta} \in \Sigma_m, \tag{8.12}$$

where

$$\Sigma_m = \left\{ \theta \in \mathbb{R}^m \mid \theta_i \geq 0 \text{ for all } i,\ \sum_{i=1}^{m} \theta_i = 1 \right\}$$

is the standard unit simplex in $\mathbb{R}^m$. The normal cone to Σ_m is given by

$$N_{\Sigma_m}(\theta) = -\operatorname{cone}\{e_i \mid \theta_i = 0\} + \mathbb{R}e,$$

where e is the vector of ones of the appropriate size. The equivalent condition for solving the VI that $0 \in h(x(t)) + N_{\Sigma_m}(\theta)$ can be parametrized as finding $\lambda_i(t)$ and $\mu(t)$ such that

$$\begin{aligned} 0 &= h_i(x(t)) - \lambda_i(t) + \mu(t) && \text{for all } i, \\ 0 \leq \theta_i(t) &\perp \lambda_i(t) \geq 0 && \text{for all } i. \end{aligned}$$

If we let $\mathcal{I}(t) = \left\{ i \mid h_i(x(t)) = \min_j h_j(x(t)) \right\}$ be the active set at time t, then $\lambda_i(t) = 0$ for all $i \notin \mathcal{I}(t)$. Also, if $i \in \mathcal{I}(t)$, $h_i(x(t)) + \mu(t) = 0$. If $\mathcal{I}(t) = \mathcal{I}_k$ for all $t \in (t_k, t_{k+1})$ and all functions involved are smooth, then differentiating this equation gives

$$0 = \nabla h_i(x(t)) \frac{dx}{dt}(t) + \mu'(t) \qquad \text{for all } i \in \mathcal{I}_k,\ t \in (t_k, t_{k+1}).$$

Substituting $dx/dt = \sum_i \theta_i(t) f_i(x(t))$ gives the system of equations

$$\begin{aligned} 0 &= \sum_{j \in \mathcal{I}_k} \nabla h_i(x(t)) f_j(x(t)) \theta_j(t) + \mu'(t) \qquad \text{for all } i \in \mathcal{I}_k, \\ 1 &= \sum_{j \in \mathcal{I}_k} \theta_j(t). \end{aligned}$$

A way of solving this system of $(|\mathcal{I}_k| + 1) \times (|\mathcal{I}_k| + 1)$ linear equations is to solve a slightly smaller $|\mathcal{I}_k| \times |\mathcal{I}_k|$ system $M_{\mathcal{I}_k}(x(t))\widehat{\theta}(t) = e$ where $M(x) = \nabla h(x) f(x) + \alpha e e^T$ with $f(x) = \left[f_1(x), f_2(x), \ldots, f_m(x) \right]$ and α chosen to make $M(x)$ nonsingular. Then set $\theta_i(t) = \widehat{\theta}_i(t) / \sum_{j \in \mathcal{I}_k} \widehat{\theta}_j(t)$ for $i \in \mathcal{I}_k$ and $\theta_i(t) = 0$ for $i \in \mathcal{I}_k$. This can be substituted into (8.11) to give a smooth differential equation for $x(t)$. This approach of differentiating the constraints is common in treating DAEs, but it suffers from the problem of *drift*. That is, the solution (while $t \in (t_k, t_{k+1})$) should satisfy $h_i(x(t)) = h_j(x(t))$ for all $i, j \in \mathcal{I}_k$. But due to the limitations of numerical solution methods, inevitably this equality becomes false. Worse, the differences $h_i(x(t)) - h_j(x(t))$ can grow exponentially until they become large, and the numerical solution loses all validity. There are a number of ways of dealing with this, as described in [15, 38], for example.

The next step is to identify if there is a change in the active set in the current step and then to accurately locate the switching time within the current step.

8.3.3 Switching for index-zero problems

The task is now to identify the new active set when it changes. From the theory of index-zero DVIs, $z(t)$ is Lipschitz continuous in t. From (8.4) we note that provided the vectors $\nabla \phi_i(z(t))$ for $i \in \mathcal{I}_k^*$ are linearly independent, we have local Lipschitz continuity of the $\lambda_i(t)$ as well. Thus, if $\phi_i(z(t)) < 0$, we have $\phi_i(z(t')) < 0$ for all t' sufficiently close to t; if $\lambda_i(t) > 0$, we have $\lambda_i(t') > 0$ for all t' sufficiently close to t. The only way there can be a change in the active set is if there is an i where $\lambda_i(t) = 0 = \phi_i(z(t))$. Let $\mathcal{I}_k^0 = \{ i \mid \lambda_i(t_k) = 0 = \phi_i(z(t_k)) \}$. Thus, at a switching time t_k, $\mathcal{I}_k^0 \neq \emptyset$. The next switching time t_{k+1} must be a zero of $t \mapsto \min(\min_{i \in \mathcal{I}_k} \lambda_i(t), \min_{i \notin \mathcal{I}_k} -\phi_i(z(t)))$.

There are many methods that can be used for locating the switching time. Bracketing methods such as bisection [17, 44], Brent's method [39], and Dekker's method [75] are most suitable for this task; the further point b from the final bounding interval $[a, b]$ can be used for the new starting time.

The last part of the method that needs to be implemented is to identify the new active set after arriving at a switching time. Suppose that t_k is a switching time, and let $\mathcal{I}_k^*$ be the

active set at $t = t_k$. For the index-zero case, let

$$\begin{aligned}\mathcal{I}_k^* &= \{i \mid \phi_i(z(t_k)) = 0\},\\ \mathcal{I}_k^0 &= \left\{i \in \mathcal{I}_k^* \mid \lambda_i(t_k) = 0\right\},\\ \mathcal{I}_k^+ &= \left\{i \in \mathcal{I}_k^* \mid \lambda_i(t_k) > 0\right\}.\end{aligned}$$

From continuity arguments, $\mathcal{I}_k \subseteq \mathcal{I}_k^*$. To determine $\mathcal{I}_k$ we need to look at the *direction* in which the solution is moving. We first consider the index-zero case. For a function $f\colon (a,b) \to \mathbb{R}^m$ let $f'_+(t) = \lim_{h\downarrow 0}(f(t+h) - f(t))/h$ be the forward directional derivative.

Recall that if $\phi_i(z(t)) < 0$, then $\lambda_i(t) = 0$. So, if $i \notin \mathcal{I}_k^*$, $\lambda_i(t) = 0$ for all t sufficiently close to t_k. If $i \notin \mathcal{I}_k$, then $\phi_i(z(t)) < 0$ for all $t > t_k$ sufficiently close to t_k, and so $\lambda'_{i+}(t_k) = 0$.

Since $\phi_i(z(t)) \le 0$ for all i and $\phi_i(z(t_k)) = 0$ for $i \in \mathcal{I}_k^*$, $(\phi_i \circ z)'_+(t_k) \ge 0$ for all $i \in \mathcal{I}_k^*$. Differentiating equation (8.7),

$$\begin{aligned}0 &= F(x(t), z(t)) + \sum_{i\in\mathcal{I}_k^*} \lambda_i(t)\,\nabla\phi_i(z(t))^T \qquad \text{gives}\\ 0 &= \nabla_x F(x(t), z(t))\,x'_+(t) + \nabla_z F(x(t), z(t))\,z'_+(t)\\ &\qquad + \sum_{i\in\mathcal{I}_k^*} \left(\lambda'_{i+}(t)\,\nabla\phi_i(z(t))^T + \lambda_i(t)\,\mathrm{Hess}\,\phi_i(z(t))\,z'_+(t)\right).\end{aligned}$$

Now $x'_+(t_k) = x'(t_k) = f(x(t_k), z(t_k))$. If we write $x(t_k) = x_k$, $z(t_k) = z_k$, $z'_+(t_k) = z'_k$, $\lambda'_{i+}(t_k) = \lambda'_{i,k}$, then

$$\begin{aligned}0 &= \nabla_x F(x_k, z_k)\,f(x_k, z_k) + \nabla_z F(x_k, z_k)\,z'_k\\ &\qquad + \sum_{i\in\mathcal{I}_k^*} \left(\lambda'_{i,k}\,\nabla\phi_i(z_k) + \lambda_{i,k}\,\mathrm{Hess}\,\phi_i(z_k)\,z'_k\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}0 \le \lambda'_{i,k} \perp -\nabla\phi_i(z_k)\,z'_k \ge 0 \qquad &\text{for all } i \in \mathcal{I}_k^0,\\ 0 = -\nabla\phi_i(z_k)\,z'_k = 0 \qquad &\text{for all } i \in \mathcal{I}_k^+.\end{aligned}$$

To tie these equations together, let

$$\begin{aligned}\mu_k^0 &= \left[\lambda'_{i,k} \mid i \in \mathcal{I}_k^0\right],\\ \mu_k^+ &= \left[\lambda'_{i,k} \mid i \in \mathcal{I}_k^+\right],\\ \nabla\phi_k^0 &= \left[\nabla\phi_i(z_k) \mid i \in \mathcal{I}_k^0\right],\\ \nabla\phi_k^+ &= \left[\nabla\phi_i(z_k) \mid i \in \mathcal{I}_k^+\right].\end{aligned}$$

Note that the matrices $\nabla\phi_k^0$ and $\nabla\phi_k^+$ are formed by stacking the row vectors $\nabla\phi_i(z_k)$ vertically. For simplicity of notation, let $\nabla_x F(x_k, z_k) = \nabla_x F_k$, $\nabla_z F(x_k, z_k) = \nabla_z F_k$, $f_k =$

$f(x_k, z_k)$, and $A_k = \nabla_z F_k + \sum_i \lambda_{i,k} \operatorname{Hess} \phi_i(z_k)$. In matrix-vector form, the conditions become

$$\begin{bmatrix} A_k & \left(\nabla \phi_k^+\right)^T \\ -\nabla \phi_k^+ & 0 \end{bmatrix} \begin{bmatrix} z_k' \\ \mu_k^+ \end{bmatrix} = \begin{bmatrix} -\nabla_x F_k \, f_k - \left(\nabla \phi_k^0\right)^T \mu_k^0 \\ 0 \end{bmatrix}, \tag{8.13}$$

$$0 \le \mu_k^0 \perp -\nabla \phi_k^0 z_k' \ge 0. \tag{8.14}$$

Solving the first linear system gives $z_k' = A_S \left(-\nabla_x F_k \, f_k - \left(\nabla \phi_k^0\right)^T \mu_k^0 \right)$, where

$$A_S = A_k^{-1} - A_k^{-1} \left(\nabla \phi_k^+\right)^T \left[\nabla \phi_k^+ A_k^{-1} \left(\nabla \phi_k^+\right)^T \right]^{-1} \nabla \phi_k^+ A_k^{-1}.$$

If A_k is a positive definite matrix, so is A_S. Substituting this formula for z_k' into (8.14) gives the LCP

$$0 \le \mu_k^0 \perp \nu_k := A_S \, \mu_k^0 + A_S \, \nabla_x F_k \, f_k \ge 0. \tag{8.15}$$

Thus, if A_k is positive definite, there is a unique solution μ_k^0. If $\left(\mu_k^0\right)_i > 0$, then $i \in \mathcal{I}_{k+1}$ and if $(\nu_k)_i > 0$, then $i \notin \mathcal{I}_{k+1}$. If the solution is strictly complementary (that is, for each i either $\left(\mu_k^0\right)_i > 0$ or $(\nu_k)_i > 0$), then we can write

$$\mathcal{I}_{k+1} = \mathcal{I}_k^+ \cup \left\{ i \mid \left(\mu_k^0\right)_i > 0 \right\}. \tag{8.16}$$

If the solution is not strictly complementarity, then we have some ambiguity in the new active set:

$$\mathcal{I}_k^+ \cup \left\{ i \mid \left(\mu_k^0\right)_i > 0 \right\} \subseteq \mathcal{I}_{k+1} \subseteq \mathcal{I}_k^+ \cup \left\{ i \mid (\nu_k)_i = 0 \right\}.$$

Often this ambiguity can be resolved by looking at higher order derivatives of $\lambda_i(t)$ and $z(t)$.

8.3.4 Switching for index-one problems

An example of how to determine the new active set in the index-one case of discontinuous differential equations can be found in [236, Section 4.2].

In the index-one case, we no longer have continuity of $z(t)$ or $\lambda(t)$. However, the condition of switching from one active set to another depends on $x(t)$, which does depend continuously on t. In particular, we consider the piecewise smooth discontinuous differential equation problem:

$$\begin{aligned} \frac{dx}{dt} &= \sum_{i=1}^{m} z_i(t) \, f_i(x(t)), \qquad x(t_0) = x_0, \\ z(t) &\in \Sigma_m, \\ 0 &\le \langle \widetilde{z} - z(t), h(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in \Sigma_m, \end{aligned}$$

where $\Sigma_m = \left\{ w \in \mathbb{R}^m \mid w \ge 0, \, \sum_{i=1}^m w_i = 1 \right\}$. The active set is

$$\mathcal{I}(x(t)) = \left\{ i \mid h_i(x(t)) = \min_j h_j(x(t)) \right\}.$$

If t_k is a switching time, then let

$$\mathcal{I}_k^* = \left\{ i \mid h_i(x(t_k)) = \min_j h_j(x(t_k)) \right\}.$$

By continuity of $x(\cdot)$ and $h(\cdot)$, the new active set $\mathcal{I}_{k+1} \subseteq \mathcal{I}_k^*$. If $i, j \in \mathcal{I}_{k+1}$, then $h_i(x(t)) = h_j(x(t))$ for all $t \in (t_k, t_{k+1})$, so the forward directional derivatives $\nabla h_i(x(t_k))x'_+(t_k) = \nabla h_j(x(t_k))x'_+(t_k)$. On the other hand, if $\nabla h_i(x(t_k))x'_+(t_k) < \nabla h_j(x(t_k))x'_+(t_k)$ for $i, j \in \mathcal{I}_k^*$, then $j \notin \mathcal{I}_{k+1}$.

Let μ_k be the forward direction derivative of $t \mapsto \min_i h_i(x(t))$ at t_k. Also let $x_k = x(t_k)$. It is not immediately clear that $x'_+(t_k)$ exists, but we will assume that it does. If it does not exist, we can at least consider limits of convergent subsequences $(x(t_k + h) - x(t_k))/h$ as $h \downarrow 0$ and call that $x'_+(t_k)$. Then $x'_+(t_k) \in \operatorname{co}\{ f_i(x_k) \mid i \in \mathcal{I}_{k+1} \} \subseteq \operatorname{co}\left\{ f_i(x_k) \mid i \in \mathcal{I}_k^* \right\}$. Writing $x'_+(t_k) = \sum_{i \in \mathcal{I}_{k+1}} z_{k,i}^* f_i(x_k)$ with $z_k^* \geq 0$ and $\sum_i z_{k,i}^* = 1$, we can extend this vector to indices $i \in \mathcal{I}_k^*$: $z_{k,i}^* = 0$ if $i \notin \mathcal{I}_{k+1}$.

Now

$$\mu_k = \min_{i \in \mathcal{I}_k^*} \nabla h_i(x_k) x'_+(t_k).$$

Let $v_{k,i}^* = \nabla h_i(x_k)x'_+(t_k) - \mu_k \geq 0$. If $v_{k,i}^* > 0$, then $i \notin \mathcal{I}_{k+1}$, so $z_{k,i}^* = 0$. Thus

$$0 \leq v_{k,i}^* \perp z_{k,i}^* \geq 0 \qquad \text{for all } i \in \mathcal{I}_k^*.$$

Substituting for $x'_+(t_k)$ in terms of z_k^* we have

$$v_{k,i}^* = \sum_{j \in \mathcal{I}_k^*} z_{k,j}^* \nabla h_i(x_k) f_j(x_k) - \mu_k.$$

Let $m_{ij} = \nabla h_i(x_k) f_j(x_k)$, forming the matrix $M = \left[m_{ij} \mid i, j \in \mathcal{I}_k^* \right]$, so that $v_k^* = M z_k^* - \mu_k e$. Again e is the vector of ones of the appropriate size. Since $\sum_{j \in \mathcal{I}_k^*} z_{k,j}^* = 1$, if we add $\alpha \geq 0$ to every entry of M, we have

$$v_k^* = \left(M + \alpha e e^T \right) z_k^* - (\mu_k + \alpha) e.$$

If we choose $\alpha > 0$ sufficiently large, then $\mu_k + \alpha > 0$, and we can divide by $\mu_k + \alpha$ to obtain $v_k = v_k^*/(\mu_k + \alpha)$ in terms of $z_k = z_k^*/(\mu_k + \alpha)$:

$$\begin{aligned} v_k &= \left(M + \alpha e e^T \right) z_k - e, \\ 0 \leq v_k &\perp z_k \geq 0. \end{aligned}$$

This is an LCP, and solutions exist if $M + \alpha e e^T$ is strictly copositive, which is true if $m_{ij} + \alpha > 0$ for all $i, j \in \mathcal{I}_k^*$, for example. The new active set can then be identified if the solution is strictly complementary:

$$\mathcal{I}_{k+1} = \left\{ i \in \mathcal{I}_k^* \mid z_{k,i} > 0 \right\}.$$

In the case of solutions that are not strictly complementary, there is some ambiguity in the new active set:

$$\left\{ i \in \mathcal{I}_k^* \mid z_{k,i} > 0 \right\} \subseteq \mathcal{I}_{k+1} \subseteq \left\{ i \in \mathcal{I}_k^* \mid v_{k,i} = 0 \right\}.$$

As with the index-zero case, this ambiguity can often be resolved if we resort to higher order derivatives. Details can be found in [236, Section 4.2].

8.3.5 Algorithm development

By combining all these elements, it is possible to create numerical methods that produce highly accurate solutions. However, there are a number of restrictions that we have made about the structure of the system, particularly as we require a number of nondegeneracy assumptions and that solutions to the associated CPs are strictly complementary. We are also assuming that solutions are at least locally (around each switching point) piecewise smooth. Thus we should also consider the issue of Zeno solutions: solutions which have an infinite number of switching times in a bounded interval. In some situations this can be ruled out. For example, for index-zero DVIs, if the data is *analytic*, then the result of Sussman [254] for continuous piecewise analytic differential equations can be applied to show that Zeno solutions do not exist. But for index-one DVIs, this is not necessarily so [30].

Under the above nondegeneracy and strict complementarity assumptions, we have existence of solutions even if they are Zeno solutions. Proving this requires application of some axiom equivalent to the axiom of choice; see [236, Appendix C]. A more subtle issue relates to *uniqueness* of solutions. Because of the assumption of a constant active set for a suitable interval into the future $(t_k, t_k + \epsilon)$, reverse Zeno solutions, in which any such interval contains infinitely many switches, are effectively invisible. Such solutions can be generic, in the sense that arbitrary small perturbations of the data of the problem typically do not destroy the reverse Zeno solutions. Thus solutions can appear to be unique, while they are in fact not unique.

An issue that can arise in practice with these piecewise smooth methods is that the number of switches, while finite, can be very large. This is particularly true in mechanical impact problems in granular flow. Granular flow problems have a great number of particles in motion and in close proximity to each other. Another example is with the solution of partial differential equations that are DVIs. As the spatial grid is refined, the number of switches in a given time interval can increase quite rapidly. The asymptotic rate at which the number of switches increases with the reduction of the grid spacing depends on the dimension of the problem.

Against these theoretical and practical difficulties of these methods, they have the advantage that high order methods for differential equations and DAEs can be applied. If high accuracy is required, then these piecewise smooth methods are the best methods available for solving these problems. But if only moderate accuracy is required, or there are large numbers of switches in the time interval under consideration, then time-stepping methods are a good alternative.

8.4 Time stepping

Time stepping directly deals with the variational aspects of DVIs. At each time step, a VI or CP is solved for an approximation of the DVI. The solution of the VI is then used for determining the approximate solution at the end of the time step. Unlike the piecewise smooth methods discussed in the previous section, these methods do not require explicit tracking of the active set. Rather the current active set is determined from the solution to the current VI. While piecewise smooth methods have to identify every change of active set, time-stepping methods do not. As a result, time-stepping methods can handle large and frequent changes in the active set.

These methods place more emphasis on good, fast solvers for VIs or CPs. Unlike smoothing or penalty methods which rely on good smooth differential equation solvers and smoothing parameters that are not too extreme, or piecewise smooth methods that rely on nondegeneracy and solution of LCPs at switching times, time-stepping methods require solution of a VI or CP (or perhaps several such problems) at each step. Methods for static VIs, such as nonsmooth Newton methods, are therefore particularly important for time-stepping methods [195, 196, 211].

Time-stepping methods are also useful in a theoretical sense, in that they can be used to show existence of solutions and without requiring nondegeneracy assumptions or strict complementarity assumptions.

Time-stepping methods can be based on various methods for solving differential equations, such as Euler's method, the implicit Euler's method, the midpoint rule, and various Runge–Kutta methods. For index-one or index-two DVIs, the methods must be implicit and satisfy some strong stability properties. The property of B-stability [45, 46, 68, 142] is particularly important. To understand B-stability and its importance, we need to spend some time looking at Runge–Kutta methods in general.

8.4.1 Runge–Kutta methods

Runge–Kutta methods are an important class of methods for solving ordinary differential equations and can be easily adapted to the solution of differential inclusions, DAEs, and partial differential equations. For deeper treatments of Runge–Kutta methods for smooth differential equations, see [15, 18, 46, 119]. The simplest Runge–Kutta methods are the fully explicit and fully implicit Euler methods: to solve the differential equation $dx/dt = f(x(t))$, $x(t_0) = x_0$ to obtain approximate solutions $x_\ell \approx x(t_\ell) = x(t_0 + \ell h)$,

$$\begin{aligned} x_{\ell+1} &= x_\ell + h\, f(x_\ell) && \text{explicit Euler method,} \\ x_{\ell+1} &= x_\ell + h\, f(x_{\ell+1}) && \text{implicit Euler method.} \end{aligned}$$

Higher order Runge–Kutta methods have more stages: $s > 1$. An s-stage Runge–Kutta method has the form

$$y_{\ell,i} = x_\ell + h \sum_{j=1}^{s} a_{ij}\, f(y_{\ell,j}), \qquad i = 1, 2, \ldots, s, \tag{8.17}$$

$$x_{\ell+1} = x_\ell + h \sum_{j=1}^{s} b_j\, f(y_{\ell,j}). \tag{8.18}$$

The constants a_{ij} and b_j together with $c_i := \sum_{j=1}^{s} a_{ij}$ form the *Butcher tableau*:

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array} \qquad \text{or} \qquad \begin{array}{c|c} c & A \\ \hline & b^T \end{array}$$

The most famous Runge–Kutta method is the fourth order method, which can be represented compactly by the tableau (empty entries are zero)

1/2	1/2			
1/2		1/2		
1			1	
	1/6	1/3	1/3	1/6

Alternatively it can be written out as

$$\begin{aligned}
y_{\ell,1} &= x_\ell, \\
y_{\ell,2} &= x_\ell + \frac{1}{2} h\, f(y_{\ell,1}), \\
y_{\ell,3} &= x_\ell + \frac{1}{2} h\, f(y_{\ell,2}), \\
y_{\ell,4} &= x_\ell + h\, f(y_{\ell,3}), \\
x_{\ell+1} &= x_\ell + \frac{1}{6} h \big[f(y_{\ell,1}) + 2 f(y_{\ell,2}) + 2 f(y_{\ell,3}) + f(y_{\ell,4}) \big].
\end{aligned}$$

This is an explicit method. The usual proofs of convergence of Runge–Kutta methods assume that the solution and the function f are both smooth, which is rarely true of solutions of DVIs, at least globally in time. The usual aim in the development of Runge–Kutta methods has been obtaining high order accuracy. With less smooth solutions and for stiff differential equations, the more important issue is stability rather than order of accuracy. Examples of implicit methods, which typically have better stability properties, can be found in Figure 8.1.

Runge–Kutta methods have been used for numerical solution of differential inclusions by a number of authors [31, 81, 141, 143, 142, 144, 158, 159, 190, 191, 257, 258, 259]. Applying a Runge–Kutta method to a differential inclusion $dx/dt \in \Phi(x(t))$ leads to

$$y_{\ell,i} = x_\ell + h \sum_{j=1}^{s} a_{ij}\, v_{\ell,j}, \qquad i = 1, 2, \ldots, s, \tag{8.19}$$

$$v_{\ell,j} \in \Phi(y_{\ell,j}), \qquad i = 1, 2, \ldots, s, \tag{8.20}$$

$$x_{\ell+1} = x_\ell + h \sum_{j=1}^{s} b_j\, v_{\ell,j}. \tag{8.21}$$

This is almost the same as replacing "f" with "Φ" and "$=$" with "$\in$" where appropriate. However, this formulation makes sure that we pick the same element $v_{\ell,j} \in \Phi(y_{\ell,j})$ rather than allowing different elements to be used for different occurrences of $\Phi(y_{\ell,j})$.

For the DVI

$$\begin{gathered}
\frac{dx}{dt} = f(x(t), z(t)), \qquad x(t_0) = x_0, \\
z(t) \in K \quad \& \quad 0 \le \langle \widetilde{z} - z(t),\, F(x(t), z(t)) \rangle \qquad \text{for all } \widetilde{z} \in K,
\end{gathered}$$

we can set

$$\Phi(x) = \{\, f(x,z) \mid z \in K \;\&\; 0 \le \langle \widetilde{z} - z(t),\, F(x,z) \rangle \text{ for all } \widetilde{z} \in K \,\},$$

so that solving the differential inclusion $dx/dt \in \Phi(x)$ is equivalent to solving the DVI.

Consider the case where $x(t) \in \mathbb{R}^n$ and $\Phi(x)$ satisfies Filippov's assumptions:

- $\Phi(x)$ is a closed, convex, and bounded set for all x;
- $x \mapsto \Phi(x)$ is upper semicontinuous;
- there is a constant C where $\langle x, y\rangle \le C\left(1 + \|x\|^2\right)$ for all $y \in \Phi(x)$.

Proofs of convergence can be found for this case in [81, 191]; explicit Euler is covered by [258]. The rate of convergence is much harder to determine and usually requires uniqueness of solutions in order to do so. For example, suppose that Φ satisfies a one-sided Lipschitz condition:

$$\langle x_1 - x_2, y_1 - y_2\rangle \le L\,\|x_1 - x_2\|^2 \qquad \text{whenever } y_i \in \Phi(x_i). \tag{8.22}$$

Under the Filippov assumptions, the one-sided Lipschitz condition implies $x \mapsto -\Phi(x) + Lx$ is maximal monotone, and we can apply the theory of maximal monotone differential inclusions. The implicit Euler method is the simplest method that we can use with general maximal monotone differential inclusions:

$$x_{\ell+1} \in x_\ell + h\,\Phi(x_{\ell+1}), \tag{8.23}$$

which is equivalent to applying the resolvent operator R_h from (2.56) if we identify the space X with its dual X'. The question of the asymptotic size of the error $\|x_\ell - x(t_\ell)\|$ for this method has been investigated by Lippold [161] for the case $\Phi = -A - \partial\phi$ where ϕ is a lower semicontinuous convex function that is Lipschitz on its domain and $A\colon X \to X'$ is linear and monotone, and also by Bastien and Schatzman [27] for general maximal monotone differential inclusions in Gelfand triples. Both obtain

$$\|x_\ell - x(t_\ell)\| = \mathcal{O}(h^{1/2}), \qquad t_\ell \in [0, T],$$

as $h \downarrow 0$. Numerical simulations seem to indicate that

$$\|x_\ell - x(t_\ell)\| = \mathcal{O}(h),$$

but at the time of this writing, there is no proof of this except for the case where $\Phi(x) = f(x) - N_K(x)$ with f Lipschitz and K closed and convex.

For the general case where Φ satisfies the one-sided Lipschitz condition (8.22), we need to restrict h so that $0 < h\,L < 1$ in order to guarantee solutions of the time-stepping problem (8.23).

These results can be extended to more complex Runge–Kutta schemes using (8.19)–(8.21), provided that the Runge–Kutta scheme is *B-stable*, also known as *algebraically stable*. B-stability of a Runge–Kutta method means that whenever $(-f)$ is a continuous monotone function and

$$y_{\ell,i}^{(p)} = x_\ell^{(p)} + h\sum_{j=1}^{s} a_{ij}\, f(y_{\ell,j}^{(p)}), \qquad i = 1, 2, \ldots, s, \tag{8.24}$$

$$x_{\ell+1}^{(p)} = x_\ell^{(p)} + h\sum_{j=1}^{s} b_j\, f(y_{\ell,j}^{(p)}) \tag{8.25}$$

1/2	1/2
	1

(a) mid-point rule

0	0	0
1	1/2	1/2
	1/2	1/2

(b) trapezoidal rule

Figure 8.1: Butcher tableaus for (a) the midpoint and (b) trapezoidal rules.

for $p = 1, 2$, then $\left\| x_{\ell+1}^{(2)} - x_{\ell+1}^{(1)} \right\| \leq \left\| x_{\ell}^{(2)} - x_{\ell}^{(1)} \right\|$. This is a natural nonlinear stability concept, but it appears rather formidable to check if it holds for a given method. Fortunately, there is an easy equivalent algebraic condition [45, 68]:

$$M := \operatorname{diag}(b)A + A^T \operatorname{diag}(b) - b\,b^T \tag{8.26}$$
$$\text{is positive semidefinite,}$$
$$b_i \geq 0 \qquad \text{for all } i\,. \tag{8.27}$$

Note that $\operatorname{diag}(b)$ is the diagonal matrix where the ith diagonal entry is b_i. There are many Runge–Kutta methods that are B-stable. Most of these were developed in order to improve the order of accuracy. Since differential inclusions are often discontinuous and do not have smooth solutions, the order of accuracy is often not particularly important.

One issue that is especially important for handling differential inclusions $dx/dt \in \Phi(x)$ is that the solution remains inside $\operatorname{range}\Phi = \{x \mid \Phi(x) \neq \emptyset\}$: if $x_\ell \in \operatorname{range}\Phi$, then $x_{\ell+1} \in \operatorname{range}\Phi$. The most common way to ensure this is for b^T to be a row of A. This property is known as *stiff accuracy*. In this way $x_{\ell+1} = y_{\ell,i}$ for some i (usually $i = s$) and $v_{\ell,i} \in \Phi(y_{\ell,i}) \neq \emptyset$. Thus, for example, fully implicit Euler is stiffly accurate while neither the explicit Euler nor the midpoint rules are stiffly accurate. However, the trapezoidal method is stiffly accurate. Note that the midpoint and trapezoidal rules are second order methods for smooth differential equations.

If the solution is smooth, then using a higher order Runge–Kutta method can give high order accuracy in the numerical approximations [142], although the order of convergence is often less than the order given for the method. This phenomenon is called order reduction [18, 119], and it is well known for stiff differential equations and DAEs: the effective order of the method is typically the *stage order* of the Runge–Kutta method. This stage order is the largest q where

$$x(t_\ell + c_i\,h) = x(t_\ell) + h\sum_{j=1}^{s} a_{ij}\,x'(t_\ell + c_j\,h) + \mathcal{O}(h^{q+1}) \tag{8.28}$$

for all i. On the other hand, for nonstiff ordinary differential equations, what is often more important is the *quadrature order*, which is the largest p where

$$x(t_\ell + h) = x(t_\ell) + h\sum_{j=1}^{s} b_j\,x'(t_\ell + c_j\,h) + \mathcal{O}(h^{p+1}), \tag{8.29}$$

which is usually significantly larger than q.

To apply these methods to DVIs, we can set up a corresponding VI for the inclusions $v_{\ell,i} \in \Phi(y_{\ell,i})$: $v_{\ell,i} = f(y_{\ell,i}, z_{\ell,i})$ where

$$z_{\ell,i} \in K, \tag{8.30}$$

$$0 \le \langle \widetilde{z} - z_{\ell,i},\, F(y_{\ell,i}, z_{\ell,i}) \rangle \qquad \text{for all } \widetilde{z} \in K, \tag{8.31}$$

$$y_{\ell,i} = x_\ell + h \sum_{j=1}^{s} a_{ij}\, v_{\ell,j}, \qquad i = 1, 2, \ldots, s. \tag{8.32}$$

For an index-zero DVI, if $z \mapsto F(y,z)$ is strongly monotone, uniformly in y, then we can apply the standard theory of Lipschitz differential equations to this Runge–Kutta method to establish the existence of solutions to the Runge–Kutta equations for sufficiently small $h > 0$.

For an index-one DVI

$$\frac{dx}{dt}(t) = f(x(t)) + B(x(t))\, z(t), \tag{8.33}$$

$$z(t) \in K \;\;\&\;\; 0 \le \langle \widetilde{z} - z(t),\, G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K, \tag{8.34}$$

proving the solvability of the Runge–Kutta equations is a bit more complicated. The following treatment follows that of Kastner-Maresch [142], which applies Runge–Kutta methods to differential inclusions with one-sided Lipschitz conditions. The main results of [142] include not only solvability of the Runge–Kutta equations (provided the right-hand side set $\Phi(x)$ satisfies a growth condition) but also that the accuracy of the computed solution is the same as the stiff order of the method, *provided that the solution is smooth*. Of course, we do not expect that the solution will be smooth for DVIs, but we would usually expect it to be piecewise smooth. Thus Runge–Kutta methods can be combined with detect, locate, and restart methods as described in Section 8.3 to accurately compute solutions of DVIs, provided the solutions do not have infinitely many switches in a finite time.

First we show that the Runge–Kutta equations applied to an index-one DVI of the form (8.33)–(8.34) has solutions under conditions on the method and the DVI that are not too restrictive. These Runge–Kutta equations are

$$y_{\ell,i} = x_\ell + h \sum_{j=1}^{s} a_{ij} \left[f(y_{\ell,j}) + B(y_{\ell,j})\, z_{\ell,j} \right], \tag{8.35}$$

$$i = 1, 2, \ldots, s,$$

$$z_{\ell,j} \in K \;\;\&\;\; 0 \le \langle \widetilde{z} - z_{\ell,i},\, G(y_{\ell,i}) \rangle \qquad \text{for all } \widetilde{z} \in K, \tag{8.36}$$

$$x_{\ell+1} = x_\ell + h \sum_{j=1}^{s} b_j \left[f(y_{\ell,j}) + B(y_{\ell,j})\, z_{\ell,j} \right]. \tag{8.37}$$

We will assume that the Runge–Kutta method

(RK1) is algebraically stable (8.26)–(8.27),

(RK2) is stiffly accurate (b^T is a row of A),

(RK3) has A symmetrizable positive definite ($DA + A^T D$ is positive definite for a diagonal matrix D with positive diagonals),

(RK4) satisfies Butcher's simplifying assumptions B(p) and C(p).

1	1
	1

(a) order 1

1/3	5/12	−1/12
1	3/4	1/4
	3/4	1/4

(b) order 3

$(4-\sqrt{6})/10$	$(88-7\sqrt{6})/360$	$(296-169\sqrt{6})/1800$	$(-2+3\sqrt{6})/225$
$(4+\sqrt{6})/10$	$(296+169\sqrt{6})/1800$	$(88+7\sqrt{6})/360$	$(-2-3\sqrt{6})/225$
1	$(16-\sqrt{6})/36$	$(16+\sqrt{6})/36$	1/9
	$(16-\sqrt{6})/36$	$(16+\sqrt{6})/36$	1/9

(c) order 5

Figure 8.2: Radau IIA methods of order 1 ($s = 1$), order 3 ($s = 2$), and order 5 ($s = 3$).

Butcher's simplifying assumptions are

$$\mathrm{B}(p)\colon k\sum_{k=1}^{s} b_j\, c_j^{k-1} = 1 \qquad \text{for } k = 1, 2, \ldots, p,$$

which implies the quadrature order condition (8.29); and

$$\mathrm{C}(q)\colon k\sum_{j=1}^{s} a_{ij}\, c_j^{k-1} = c_i^k \qquad \text{for } i = 1, 2, \ldots, s \text{ and } k = 1, 2, \ldots, q,$$

which implies the stage order condition (8.28).

This might appear to be a formidable list of assumptions; however, important families of methods satisfy these conditions, such as the Radau IIA methods, which have received special attention as powerful methods for solving stiff differential equations. The Radau IIA method with s stages has order $2s-1$; the Radau IIA method with one stage is simply the implicit Euler method. The Radau IIA methods with one, two, and three stages are shown in Figure 8.2.

Kastner-Maresch [142] uses the assumptions (RK1), (RK3), and (RK4). However, Kastner-Maresch does not require that the method be stiffly accurate (RK2). That is because he assumed that the differential inclusion $dx/dt \in \Phi(x)$ had $\Phi(x)$ bounded and satisfying a growth condition. Here we want to include problems of the form $dx/dt \in f(x) - N_K(x)$. Conversely, for $K = C + L$, C bounded, and L a closed convex cone, we require that $G(x(t)) \in L^*$ for all t. Thus we want $G(x^{\ell+1}) \in L^*$ if $G(x^\ell) \in L^*$ using this method. This leads to the requirement of stiffly accurate methods. Without this condition, other methods like the Gauss methods of order $2s$ for s stages can be used, as they can for differential inclusions with a growth condition on $\Phi(x)$.

8.4.2 Existence of solutions to the Runge–Kutta system

The main purpose of this section is to show existence of solutions to the Runge–Kutta conditions (8.35)–(8.37) under some reasonable conditions, at least for sufficiently small $h > 0$. To do this, we set up an iterative sequence of VIs and show that the sequence of

approximate solutions converges. Note that we assume that K is a cone to simplify much of the analysis. Throughout much of the section, it is convenient to use tensor product notation for vectors and matrices:

$$x \otimes y = \left[x_1 y^T, x_2 y^T, \ldots, x_s y^T\right]^T,$$

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1s}B \\ a_{21}B & a_{22}B & \cdots & a_{2s}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{s1}B & a_{s2}B & \cdots & a_{ss}B \end{bmatrix}.$$

As usual, e denotes the vector of ones of the appropriate size.

Let $\mathbf{z}_\ell = [z_{\ell,1}^T, z_{\ell,2}^T, \ldots, z_{\ell,s}^T]^T$ and $\mathbf{y}_\ell = [y_{\ell,1}^T, y_{\ell,2}^T, \ldots, y_{\ell,s}^T]^T$. Then we can write the Runge–Kutta system (8.35)–(8.37) in tensor product form as

$$\mathbf{y}_\ell = e \otimes x_\ell + h(A \otimes I)\left[\mathbf{f}(\mathbf{y}_\ell) + \mathbf{B}(\mathbf{y}_\ell)\mathbf{z}_\ell\right], \tag{8.38}$$

$$\mathbf{z}_\ell \in K^s \;\;\&\;\; 0 \le \langle \widetilde{\mathbf{z}} - \mathbf{z}_\ell, \mathbf{G}(\mathbf{y}_\ell)\rangle \qquad \text{for all } \widetilde{\mathbf{z}} \in K^2, \tag{8.39}$$

$$x_{\ell+1} = x_\ell + h\left(b^T \otimes I\right)\left[\mathbf{f}(\mathbf{y}_\ell) + \mathbf{B}(\mathbf{y}_\ell)\mathbf{z}_\ell\right], \tag{8.40}$$

where

$$\begin{aligned} \mathbf{f}(\mathbf{v})^T &= \left[f(v_1)^T, f(v_2)^T, \ldots, f(v_s)^T\right], \\ \mathbf{G}(\mathbf{v})^T &= \left[G(v_1)^T, G(v_2)^T, \ldots, G(v_s)^T\right], \\ \mathbf{B}(\mathbf{v}) &= \operatorname{diag}(B(v_1), B(v_2), \ldots, B(v_2)), \\ e^T &= [1, 1, , \ldots, 1]. \end{aligned}$$

Theorem 8.1. *Suppose that the functions f, B, G, and ∇G are bounded and Lipschitz with $\nabla G(x)\,B(x)$ symmetric and positive definite (uniformly in x), K is a closed convex cone, and conditions* (RK1) *and* (RK3) *hold. Then provided $G(x_\ell) \in K^*$ there is $h_0 > 0$ such that for $0 < h \le h_0$, the Runge–Kutta system* (8.35)–(8.37) *has a solution. Furthermore, h_0 is independent of x_ℓ, and the solutions are bounded independently of h and x_ℓ for $0 < h \le h_0$.*

Note that given $z_{\ell,i}^{(p)}$, $i = 1, 2, \ldots, s$, and $h > 0$ sufficiently small, we can uniquely solve the equations

$$\mathbf{y}_\ell^{(p)} = e \otimes x_\ell + h(A \otimes I)\left[\mathbf{f}(\mathbf{y}_\ell^{(p)}) + \mathbf{B}(\mathbf{y}_\ell^{(p)})\mathbf{z}_\ell^{(p)}\right]$$

for $\mathbf{y}_\ell^{(p)}$. We suppose that f and B are Lipschitz with constants L_f and L_B, and B is a bounded function with bound β_B. Finally, we assume that $\mathbf{z}_\ell^{(p)}$ is bounded by β_z. Then, a Lipschitz constant of the map $\mathbf{z}_\ell^{(p)} \mapsto \mathbf{y}_\ell^{(p)}$ is

$$h\,\|A\|\,(\beta_B + L_B\beta_z)\,/\,(1 - h\,\|A\|\,(\beta_B + L_B\beta_z)).$$

We approximate the system of VIs (8.35)–(8.36) by a "linearization" around $\mathbf{z}_\ell^{(p)}$:

$$z_{\ell,i} \in K, \tag{8.41}$$

$$0 \le \Big\langle \widetilde{z}_{\ell,i} - z_{\ell,i},\, G(y_{\ell,i}^{(p)}) \tag{8.42}$$

$$+ h \sum_{j=1}^{s} a_{ij}\, \nabla G(x_\ell) B(x_\ell) \left(z_{\ell,j} - z_{\ell,i}^{(p)} \right) \Big\rangle$$

for all $\widetilde{z}_{\ell,j} \in K$, $i = 1, 2, \ldots, s$. The solution $z_{\ell,i}$ to this system of VIs ($i = 1, 2, \ldots, s$) is the new iterate $z_{\ell,i}^{(p+1)}$. Writing $D = \mathrm{diag}(\mathbf{d})$ for the diagonal matrix where $DA + A^T D$ is positive definite, we can multiply the inequality in (8.42) by $d_i > 0$ to get

$$z_{\ell,i} \in K, \tag{8.43}$$

$$0 \le \Big\langle \widetilde{z}_{\ell,i} - z_{\ell,i},\, d_i\, G(y_{\ell,i}^{(p)}) \tag{8.44}$$

$$+ h \sum_{j=1}^{s} d_i\, a_{ij}\, \nabla G(x_\ell) B(x_\ell) \left(z_{\ell,j} - z_{\ell,i}^{(p)} \right) \Big\rangle$$

for all $\widetilde{z}_{\ell,i} \in K$, $i = 1, 2, \ldots, s$. Combining the VIs for $i = 1, 2, \ldots, s$ gives the VI over $L := K \times K \times \cdots \times K$:

$$\mathbf{z}_\ell \in L, \tag{8.45}$$

$$0 \le \Big\langle \widetilde{\mathbf{z}}_\ell - \mathbf{z}_\ell,\, D \otimes I\, \mathbf{G}(\mathbf{y}_\ell^{(p)}) \tag{8.46}$$

$$+ h\, DA \otimes C_\ell \left(\mathbf{z}_\ell - \mathbf{z}_\ell^{(p)} \right) \Big\rangle \qquad \text{for all } \widetilde{\mathbf{z}}_\ell \in L,$$

where $\mathbf{G}(\mathbf{y}_\ell^{(p)}) = [G(y_{\ell,1}^{(p)})^T, \ldots, G(y_{\ell,s}^{(p)})^T]^T$ and $C_\ell = \nabla G(x_\ell) B(x_\ell)$. We want the tensor product $DA \otimes C_\ell$ to be positive definite, as then we can guarantee existence of solutions of (8.45)–(8.46) as well as bounds on these solutions. However, U and V positive definite is not sufficient to guarantee that $U \otimes V$ is also positive definite. Write $U = U_s + U_a$ and $V = V_s + V_a$, where U_s (U_a) is the symmetric (antisymmetric) part of U, and V_s (V_a) is the symmetric (antisymmetric) part of V. Then the symmetric part of $U \otimes V$ is $U_s \otimes V_s + U_a \otimes V_a$. Unless there is some control on U_a and V_a, the effects of the antisymmetric parts can overcome the positive definiteness of $U_s \otimes V_s$. As a simple example, take $U = V = \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}$. The eigenvalues of the symmetric part of $U \otimes V$ are 1 ± 4, so $U \otimes V$ is not positive definite. The assumption of symmetry for $C_\ell = \nabla G(x_\ell)\, B(x_\ell)$ simplifies this part of the proof. This can be weakened to allow for *some* asymmetry in $\nabla G(x)\, B(x)$ as long as $DA \otimes \nabla G(x) B(x)$ is uniformly positive definite.

The main idea of the proof is to show that the iteration $\mathbf{z}_\ell^{(p)} \mapsto \mathbf{z}_\ell = \mathbf{z}_\ell^{(p+1)}$ defined by (8.45)–(8.46) is, for $\|\mathbf{z}_\ell^{(p)}\| \le \beta_z$, a contraction mapping and maintains the property that $\|\mathbf{z}_\ell^{(p+1)}\| \le \beta_z$. A crucial part of the proof is the relationship between $d(D \otimes I\, \mathbf{G}(\mathbf{y}_\ell^{(p)}), (K^s)^*)$ and $\|\mathbf{z}_\ell^{(p+1)}\|$. The VI (8.45)–(8.46) guarantees that $D \otimes I\, \mathbf{G}(\mathbf{y}_\ell^{(p)}) +$

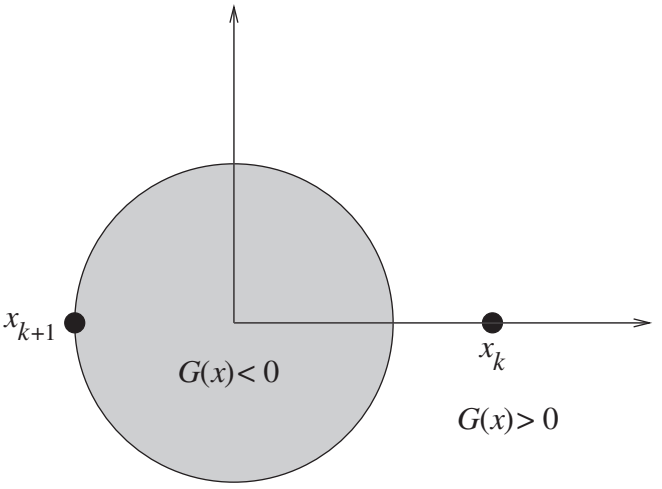


Figure 8.3: Spurious solution to implicit Euler method for a DVI.

$h\,DA \otimes C_\ell\big(\mathbf{z}_\ell^{(p+1)} - \mathbf{z}_\ell^{(p)}\big) \in (K^s)^*$. To bound the distance $d(D \otimes I\,\mathbf{G}(\mathbf{y}_\ell^{(p+1)}), (K^s)^*)$, we must bound the nonlinearities in G by means of the Lipschitz constant of ∇G. This distance bound, in turn, involves $\big\|\mathbf{z}_\ell^{(p+1)} - \mathbf{z}_\ell^{(p)}\big\|$. Fortunately the fact that there is a factor of $\mathcal{O}(h^2)$ in the distance bound is sufficient to show that $\mathbf{z}_\ell^{(p)} \mapsto \mathbf{z}_\ell^{(p+1)}$ is a contraction map for sufficiently small $h > 0$. Details can be found in [246]. The question of whether this result can be extended to $K = C + L$ with C bounded and L a cone (rather than requiring that K itself be a cone) is an open question at the time of this writing.

Uniqueness of the solution is also not known in general. The proof technique indicates that amongst solutions with $\big\|z_{\ell,i}\big\| \le \beta_z$ (β_z as given in the proof), the solution is unique. This is probably all that is necessary in practice. However, unless there is some control of the nonlinearities in f, B, and G, we cannot guarantee that there are no "spurious" solutions with $\big\|z_{\ell,i}\big\|$ large. In [142] this is dealt with through a one-sided Lipschitz condition for differential inclusions, which is a global condition. But, for DVIs, the following is a counterexample which shows the possibility of unbounded spurious solutions: Consider the DVI

$$\begin{aligned} \frac{dx}{dt} &= \nabla G(x(t))^T z(t), \\ 0 \le z(t) \quad \& \quad 0 &\le (\widetilde{z} - z(t))\, G(x(t)) \qquad \text{for all } \widetilde{z} \ge 0 \end{aligned}$$

with $G\colon \mathbb{R}^2 \to \mathbb{R}$ given by $G(x) = \|x\|_2^2 - 1$. If we applied the implicit Euler method, which is the Radau IIA method with one stage, we have the following CP to solve for step size $h > 0$:

$$\begin{aligned} x_{\ell+1} &= x_\ell + 2h x_{\ell+1} z_{\ell+1}, \\ 0 \le z_{\ell+1} &\perp \|x_{\ell+1}\|_2^2 - 1 \ge 0. \end{aligned}$$

This situation is illustrated by Figure 8.3, where $x_\ell = \alpha \ge 1$. In this case we can have

$x_{\ell+1} = -e_1$ as shown in the figure. The corresponding value of $z_{\ell+1}$ is given by

$$-e_1 = \alpha e_1 - 2h e_1 z_{\ell+1},$$

or equivalently, $z_{\ell+1} = (1+\alpha)/(2h)$. This spurious solution clearly goes to infinity as h goes to zero.

8.4.3 Order of convergence for smooth solutions

Following Kastner-Maresch [142], it can be shown that if the solution of the DVI

$$\frac{dx}{dt} = f(x) + B(x)z(t), \tag{8.47}$$

$$z(t) \in K \quad \& \quad 0 \le \langle \widetilde{z} - z(t), G(x(t)) \rangle \qquad \text{for all } \widetilde{z} \in K \tag{8.48}$$

is smooth on a time interval $[t_0, T]$ with smooth f, B, and G, and $\nabla G(x)\, B(x)$ symmetric positive definite, then a Runge–Kutta method satisfying (RK1)–(RK4) produces numerical solutions that converge with the order of convergence at least equal to the stage order on $[t_0, T]$. Note also that this generalization does not require that the corresponding differential inclusion

$$\frac{dx}{dt}(t) \in f(x) + B(x)\,\{z(t) \mid z(t) \text{ satisfies (8.48)}\}$$

have the one-sided Lipschitz property, as assumed in [142]. The complete proof of high order of convergence is beyond the scope of this book. However, highlights of the proof follow. For details, see [246].

Theorem 8.2. *Under the assumptions of Theorem* 8.1*, if $x(\cdot)$ is smooth on an interval $[t_0, T]$, and $\|x(t_0) - x_0\| = \mathcal{O}(h^q)$, then $\|x(t_\ell) - x_\ell\| = \mathcal{O}(h^q)$, where $t_\ell = t_0 + \ell h \in [t_0, T]$.*

The proof essentially follows [142], although [142] in turn uses a number of results of [74]. In what follows, the hidden constants in "$\mathcal{O}$" do not depend on h for $h > 0$ sufficiently small. Since existence of (bounded) solutions of the Runge–Kutta system has already been established, we start with the Runge–Kutta system in tensor form:

$$\mathbf{y}_\ell = e \otimes x_\ell + h\,(A \otimes I)\big[\mathbf{f}(\mathbf{y}_\ell) + \mathbf{B}(\mathbf{y}_\ell)\mathbf{z}_\ell\big], \tag{8.49}$$

$$\mathbf{z}_\ell \in K^s \quad \& \quad 0 \le \langle \widetilde{\mathbf{z}} - \mathbf{z}_\ell, \mathbf{G}(\mathbf{y}_\ell) \rangle \qquad \text{for all } \widetilde{\mathbf{z}} \in K^s, \tag{8.50}$$

$$x_{\ell+1} = x_\ell + h\left(b^T \otimes I\right)\big[\mathbf{f}(\mathbf{y}_\ell) + \mathbf{B}(\mathbf{y}_\ell)\mathbf{z}_\ell\big]. \tag{8.51}$$

The main task is to prove some perturbation bounds where the Runge–Kutta system is perturbed:

$$\widehat{\mathbf{y}}_\ell = e \otimes \widehat{x}_\ell + h\,(A \otimes I)\big[\mathbf{f}(\widehat{\mathbf{y}}_\ell) + \mathbf{B}(\widehat{\mathbf{y}}_\ell)\widehat{\mathbf{z}}_\ell + \boldsymbol{\eta}_\ell\big], \tag{8.52}$$

$$\widehat{\mathbf{z}}_\ell \in K^s \quad \& \quad 0 \le \langle \widetilde{\mathbf{z}} - \widehat{\mathbf{z}}_\ell, \mathbf{G}(\widehat{\mathbf{y}}_\ell) \rangle \qquad \text{for all } \widetilde{\mathbf{z}} \in K^s, \tag{8.53}$$

$$\widehat{x}_{\ell+1} = \widehat{x}_\ell + h\left(b^T \otimes I\right)\big[\mathbf{f}(\widehat{\mathbf{y}}_\ell) + \mathbf{B}(\widehat{\mathbf{y}}_\ell)\widehat{\mathbf{z}}_\ell\big]. \tag{8.54}$$

Specifically, we want to show that we can bound $\delta\mathbf{y}_\ell := \widehat{\mathbf{y}}_\ell - \mathbf{y}_\ell$ by $C\left(\|\boldsymbol{\eta}_\ell\| + \|\delta x_\ell\|\right)$, where $\delta x_\ell := \widehat{x}_\ell - x_\ell$. From the usual technique for obtaining perturbation bounds in VIs,

$0 \geq \langle \widehat{\mathbf{z}}_\ell - \mathbf{z}_\ell, (D \otimes I)\left[\mathbf{G}(\widehat{\mathbf{y}}_\ell) - \mathbf{G}(\mathbf{y}_\ell)\right]\rangle$. Then we note that $\mathbf{G}(\widehat{\mathbf{y}}_\ell) - \mathbf{G}(\mathbf{y}_\ell) = \nabla \mathbf{G}(\mathbf{y}_\ell)\delta\mathbf{y}_\ell + \mathcal{O}(\|\delta\mathbf{y}_\ell\|^2)$. Following the uniqueness proof for index-one DVIs (Theorem 5.3), we can write $\nabla \mathbf{G}(\mathbf{y})^T = \mathbf{Q}(\mathbf{y})\mathbf{B}(\mathbf{y})$ with $\mathbf{Q}(\mathbf{y})$ block diagonal and symmetric positive definite. Since $\mathbf{Q}(\mathbf{y})$ and $\mathbf{B}(\mathbf{y})$ are block diagonal, they commute with $D \otimes I$. Moving $\nabla \mathbf{G}(\mathbf{y}_\ell)$ to the left of the inner product and transposing then give

$$\mathcal{O}(\|\delta\mathbf{y}_\ell\|^2) \geq \langle \mathbf{Q}(\mathbf{y}_\ell)\,(D \otimes I)\mathbf{B}(\mathbf{y}_\ell)\delta\mathbf{z}_\ell, \delta\mathbf{y}_\ell\rangle\,.$$

Premultiplying the difference of (8.49) and (8.52) by $DA^{-1} \otimes I$ gives

$$\begin{aligned}\left(DA^{-1} \otimes I\right)\delta\mathbf{y}_\ell \qquad\qquad (8.55)\\ = DA^{-1}e \otimes \delta x_\ell + h\,(D \otimes I)\left[\mathcal{O}\left(\|\delta\mathbf{y}_\ell\| + \left\|\boldsymbol{\eta}_\ell\right\|\right) + \mathbf{B}(\mathbf{y}_\ell)\delta\mathbf{z}_\ell\right].\end{aligned}$$

Since $DA + A^T D$ is positive definite, so is $A^{-T}\left(DA + A^T D\right)A^{-1} = DA^{-1} + A^{-T}D$. This means that DA^{-1} is also positive definite and hence strongly monotone. Taking the inner product of (8.55) with $\mathbf{Q}(\mathbf{y}_\ell)\delta\mathbf{y}_\ell$ gives

$$\begin{aligned}&\left\langle\left(DA^{-1} \otimes I\right)\delta\mathbf{y}_\ell, \mathbf{Q}(\mathbf{y}_\ell)\delta\mathbf{y}_\ell\right\rangle\\ &\quad= \left[\mathcal{O}\left(\|\delta x_\ell\|\right) + \mathcal{O}\left(h\left(\|\delta\mathbf{y}_\ell\| + \left\|\boldsymbol{\eta}_\ell\right\|\right)\right)\right]\|\delta\mathbf{y}_\ell\|\\ &\qquad+ h\,\langle \mathbf{Q}(\mathbf{y}_\ell)\,(D \otimes I)\mathbf{B}(\mathbf{y}_\ell)\delta\mathbf{z}_\ell, \delta\mathbf{y}_\ell\rangle\,.\end{aligned}$$

Since $\mathbf{Q}(\mathbf{y}_\ell)\left(DA^{-1} \otimes I\right) = DA^{-1} \otimes Q(x_\ell) + \mathcal{O}(h)$, for sufficiently small $h > 0$ there is an $\alpha > 0$ (independent of h) where

$$\alpha\,\|\delta\mathbf{y}_\ell\|^2 = \mathcal{O}\left(\|\delta x_\ell\| + h\left\|\boldsymbol{\eta}_\ell\right\|\right)\|\delta\mathbf{y}_\ell\| + \mathcal{O}\left(\|\delta\mathbf{y}_\ell\|^2\right).$$

Dividing by $\alpha\,\|\delta\mathbf{y}_\ell\|$ then gives $\|\delta\mathbf{y}_\ell\| = \mathcal{O}\left(\|\delta x_\ell\| + h\left\|\boldsymbol{\eta}_\ell\right\|\right)$.

The second perturbation result we need is that if $\boldsymbol{\eta}_\ell = 0$, then there is a constant C, independent of h, where for sufficiently small $h > 0$ we have

$$\|\delta x_{\ell+1}\|_{Q_{\ell+1}} \leq (1 + C\,h)\,\|\delta x_\ell\|_{Q_\ell}\,,$$

where $Q_\ell = Q(x_\ell)$ and $\|u\|_Q = \langle u, Qu\rangle^{1/2}$ is the norm generated by Q. The method of proof follows Dekker and Verwer [74, Thm. 7.4.2], who considered the case of ordinary differential equations with a one-sided Lipschitz condition. In the case here, there is an additional difficulty, as the natural inner product to use changes with position. It is also crucial that $Q(x)$ be symmetric and that the Runge–Kutta method be algebraically stable.

Note first that with $\boldsymbol{\eta}_\ell = 0$, $\delta\mathbf{y}_\ell = \mathcal{O}(\|\delta x_\ell\|)$ from the previous paragraphs. If we define $\boldsymbol{\xi}_\ell = \mathbf{f}(\widehat{\mathbf{y}}_\ell) - \mathbf{f}(\mathbf{y}_\ell) + \mathbf{B}(\widehat{\mathbf{y}}_\ell)\widehat{\mathbf{z}}_\ell - \mathbf{B}(\mathbf{y}_\ell)\mathbf{z}_\ell$, then $\boldsymbol{\xi}_\ell = h^{-1}\left(A^{-1} \otimes I\right)\left[\delta\mathbf{y}_\ell - e \otimes \delta x_\ell\right]$, so $\left\|\boldsymbol{\xi}_\ell\right\| = \mathcal{O}\left(h^{-1}\,\|\delta x_\ell\|\right)$. Expanding $\delta x_{\ell+1} = \delta x_\ell + h\sum_j b_j\xi_{\ell j}$ in $\|\delta x_{\ell+1}\|^2_{Q_\ell}$ and using $\delta x_\ell = \delta v_{\ell j} - h\sum_k a_{jk}\xi_{\ell k}$ give

$$\begin{aligned}\|\delta x_{\ell+1}\|^2_{Q_\ell} = \|\delta x_\ell\|^2_{Q_\ell} + 2h\sum_{j=1}^{s} b_j\left\langle Q_\ell\,\delta v_{\ell j}, \xi_{\ell j}\right\rangle\\ + h^2\sum_{i,j=1}^{s}\left(b_i b_j - 2b_j a_{ji}\right)\left\langle Q_\ell\xi_{\ell i}, \xi_{\ell j}\right\rangle. \qquad (8.56)\end{aligned}$$

Since Q_ℓ is symmetric positive definite, $\langle Q_\ell \xi_{\ell i}, \xi_{\ell j} \rangle$ forms a symmetric positive semidefinite matrix. On the other hand, from algebraic stability, $\mathrm{diag}(b)A + A^T \mathrm{diag}(b) - bb^T$ is a positive semidefinite matrix, so the last term in (8.56) is nonpositive. Since algebraic stability also requires that $b_j \geq 0$ for all j, we simply have to obtain a suitable upper bound on $\langle Q_\ell \, \delta v_{\ell j}, \xi_{\ell j} \rangle$. Replacing Q_ℓ with $Q(v_{\ell j})$ introduces an error of $\mathcal{O}\left(\|\delta x_\ell\|^2\right)$. Also, $\xi_{\ell j} = B(y_{\ell j})\delta z_{\ell j} + \mathcal{O}\left(\|\delta x_\ell\|^2\right)$. Using the VIs (8.50) and (8.53) along with $\nabla G(y) = B(y)^T Q(y)$ we obtain $\langle Q_\ell \, \delta v_{\ell j}, \xi_{\ell j} \rangle \leq \mathcal{O}\left(\|\delta x_\ell\|^2\right)$. Finally, changing from Q_ℓ to $Q_{\ell+1}$ in the norm gives $\|\delta x_{\ell+1}\|_{Q_{\ell+1}} \leq (1 + Ch)\,\|\delta x_{\ell+1}\|^2_{Q_\ell}$ for some constant C. Combining these results gives

$$\|\delta x_{\ell+1}\|_{Q_{\ell+1}} \leq (1 + Ch)\,\|\delta x_\ell\|_{Q_\ell}$$

for some other constant C.

The first perturbation bound can be applied to the exact solution under Butcher's assumptions B(p) and C(q): for stage order $q \leq p$ we can take $\boldsymbol{\eta}_\ell = \mathcal{O}(h^{q+1})$ for $\widehat{y}_{\ell j} = x(t_\ell + c_j h)$, where $x(\cdot)$ is the exact solution. If $\widetilde{x}_{\ell+1}$ is the result of the Runge–Kutta system with starting value $x(t_\ell)$, then the first perturbation bound gives $\delta \mathbf{y}_\ell = \mathcal{O}\left(\left\|\boldsymbol{\eta}_\ell\right\|\right) = \mathcal{O}(h^{q+1})$. For stiffly accurate methods, then, $\|\widetilde{x}_{\ell+1} - x(t_{\ell+1})\| = \mathcal{O}(h^{q+1})$. Then

$$\begin{aligned}\|x_{\ell+1} - x(t_{\ell+1})\|_{Q_{\ell+1}} &\leq \|x_{\ell+1} - \widetilde{x}_{\ell+1}\|_{Q_{\ell+1}} + \|\widetilde{x}_{\ell+1} - x(t_{\ell+1})\|_{Q_{\ell+1}} \\ &\leq (1 + Ch)\,\|x_\ell - x(t_\ell)\|_{Q_\ell} + \mathcal{O}(h^{q+1}).\end{aligned}$$

Application of a discrete Gronwall lemma (Lemma 5.2) gives the global error bound $\|x_{\ell+1} - x(t_{\ell+1})\|_{Q_{\ell+1}} = \mathcal{O}(h^q)$ for $t_\ell \in [t_0, T]$ for smooth solutions $x(\cdot)$.

8.4.4 Runge–Kutta methods in practice

As noted in Section 8.3 on piecewise smooth solvers, by combining suitable Runge–Kutta methods with techniques for detecting and locating where the smooth pieces join, we can accurately compute piecewise smooth solutions of DVIs. Runge–Kutta methods can even be used to solve the DAEs that arise in Section 8.3.

But the Runge–Kutta methods devised in this section can also be used for problems without identifying the times where smoothness is lost. Of course, we expect only $\mathcal{O}(h)$ accuracy if the derivative of the solution has a jump discontinuity, if this time is not located. In many applications, there are such a large number of points of nonsmoothness that locating them all in order to obtain better than $\mathcal{O}(h)$ accuracy is not worthwhile. In this case, we often settle for the simplest, lowest order method: the one-stage Radau IIA method, which is the fully implicit Euler method.

The fully implicit Euler method when applied to maximal monotone differential inclusion

$$\frac{dx}{dt}(t) \in -\Phi(x(t)), \qquad x(t_0) = x_0$$

becomes a matter of applying resolvents:

$$x_{\ell+1} = R_h(x_\ell), \qquad \ell = 0, 1, 2, \ldots.$$

This problem has been used and studied by various authors [27, 90, 161]. For infinite-dimensional problems the solutions are typically *not* piecewise smooth. Theory so far gives the bound $\mathcal{O}(h^{1/2})$ [27, 161] for the numerical solution of maximal monotone differential inclusions, but in practice the error appears to behave more like $\mathcal{O}(h)$. Future studies should uncover the reason for this state of affairs.

Appendix A

Some Basics of Functional Analysis

When talking about things like vector spaces, the important thing is not how the space is defined or how it is constructed; what is important is how it behaves. This allows us to apply ideas from one area of mathematics to another if the object of discussion behaves in the right way. So we use an abstract definition of what a vector space is, rather than say "a vector is a collection of real numbers x_1, x_2, etc., arranged like this: $x = [x_1, x_2, \ldots, x_n]$." Then we can treat collections of functions as vectors if that gives us insight into the functions.

Readers may wish to turn to texts on mathematical analysis and partial differential equations for discussion of these topics in greater depth, such as [94, 151, 155, 168, 217, 213]. Specialized topics are treated in monographs: for vector-valued measures, see [78, 80]; for Sobolev spaces, see [1, 262]. A short but excellent book on optimization and fixed point theorems is [106].

A.1 Metric spaces

Metric spaces consist of a set of points X together with a *metric* $d_X : X \times X \to \mathbb{R}$ which measures, in some way, the "distance" between the points. We will just use the notation d when X is clear from context. The basic properties of a metric are

$$d(x, y) \geq 0, \tag{A.1}$$

$$d(x, y) = 0 \qquad \text{if and only if } x = y, \tag{A.2}$$

$$d(x, y) = d(y, x), \tag{A.3}$$

$$d(x, y) \leq d(x, z) + d(z, x) \tag{A.4}$$

for all $x, y, z \in X$. The last inequality (A.4) is known as the *triangle inequality*. For real numbers the distance is given by $d(x, y) = |x - y|$. In a metric space we say that a sequence x_k converges $x_k \to x$ as $k \to \infty$ if for any $\epsilon > 0$ there is a K such that $k \geq K$ implies that $d(x_k, x) < \epsilon$. We often write this as $x = \lim_{k\to\infty} x_k$. An *open set* is a set $U \subseteq X$ where for any point $x \in U$ there is a $\delta > 0$ such that if $d(x, y) < \delta$, then $y \in U$ as well. A set $C \subseteq X$ is called *closed* if the complement $X \backslash C$ is open, or equivalently, for any convergent sequence $x_k \to x$ with $x_k \in C$ we also have $x \in C$. This is often described by saying that "C contains

its limit points." For a general set $A \subset X$, the set A together with limits of points in A is the smallest closed set containing A, and is called the *closure* of A, denoted by $\overline{A}$. The largest open subset of A is called the *interior* of A and is denoted by $\operatorname{int} A$.

The collection of all open sets in a space X is called the *topology* of X. A topology has the properties that the empty set $\emptyset$ and X are open sets, arbitrary unions of open sets are also open sets, and *finite* intersections of open sets are also open. A set A is closed if and only if $X \backslash A := \{ x \in X \mid x \notin A \}$ is open. An excellent introduction to the study of topologies, with or without metrics, is [185].

A *neighborhood* of a point x is an open set U containing x. The set A is *compact* (in a metric space) if for every sequence $x_n \in A$ there is a subsequence x_{n_k} ($n_k \to \infty$ if $k \to \infty$) such that $x_{n_k} \to x \in A$. (That is, every sequence in A has a convergent subsequence with a limit in A.) The set A is *precompact* if $\overline{A}$ is compact. Compact sets are particularly important in analysis. For example, any continuous function $f\colon A \to \mathbb{R}$ with A compact has a minimum and a maximum.

There is a definition of compactness based entirely on the notion of open sets: A is compact if whenever $\mathcal{U} = \{ U_\alpha \mid \alpha \in J \}$ is an *open covering* of A (that is, each U_α is open and $A \subseteq \bigcup_{\alpha \in J} U_\alpha$) there is a finite subset $\mathcal{U}' = \left\{ U_{\alpha_1}, U_{\alpha_2}, \ldots, U_{\alpha_m} \right\}$ that is an open covering of A. This concept is equivalent to the one given above for metric spaces. Thus the concept of compactness can be extended to topologies beyond those defined by metrics. The definition of compactness using open coverings leads to a useful theorem.

Lemma A.1. *Suppose $\{ A_\alpha \mid \alpha \in J \}$ is a nonempty collection of compact sets, where any finite subset has nonempty intersection: $A_{\alpha_1} \cap A_{\alpha_2} \cap \cdots \cap A_{\alpha_m} \neq \emptyset$. Then $\bigcap_{\alpha \in J} A_\alpha \neq \emptyset$.*

A function $f\colon X \to Y$ between metric spaces is *continuous* if $x_k \to x$ in X implies that $f(x_k) \to f(x)$ in Y, or equivalently, for any open set U in Y, the set $f^{-1}(U) := \{ x \mid f(x) \in U \}$ is also open. We say f is *Lipschitz continuous* with Lipschitz constant L_f if

$$d_Y(f(x_1), f(x_2)) \leq L_f\, d_X(x_1, x_2) \qquad \text{for all } x_1, x_2 \in X.$$

We say that f is a *homeomorphism* if f is continuous and has an inverse function $f^{-1}\colon Y \to X$ ($f^{-1}(y) = x$ if and only if $f(x) = y$) that is also continuous. If $f\colon X \to \mathbb{R}$ is continuous and X is compact, then f attains both its maximum and minimum on X. In general, if $f\colon X \to Y$ is continuous and X is compact, then $f(X) := \{ f(x) \mid x \in X \}$ is also compact.

A set A is *dense* in X if the closure of A in X is the whole of X: $X = \overline{A}$. A set A is *separable* if there is a countable subset $\{x_1, x_2, x_3, \ldots\}$ that is *dense* in A. This is equivalent to saying that for every $\epsilon > 0$ and $x \in A$ there is an x_k such that $d(x, x_k) < \epsilon$. The real line $\mathbb{R}$ is separable; we can take the rational numbers $\mathbb{Q}$ as a countable dense subset. Simple arguments show that $\mathbb{R}^n$ is separable. Separable spaces are important in numerical analysis, since computers can represent only a countable set of points. Unless we can approximate arbitrary points in a set by a countable set, we cannot expect to do computations in that set.

Many spaces are metric spaces, such as the set of rational numbers $\mathbb{Q}$, the set of real numbers $\mathbb{R}$, the unit circle $\left\{ (x, y) \mid x^2 + y^2 = 1 \right\}$, and the space of bounded functions into a metric space $f, g\colon A \to X$ with the distance between them given by $d(f, g) = \sup_{a \in A} d_X(f(a), g(a))$. An important property for metric spaces to have is *completeness*. That is, all sequences that "should" converge do. As an example, consider the sequence of rational numbers 1, 1.4, 1.41, 1.414, 1.4142, 1.41421, ... These are the truncated decimal expansions of $\sqrt{2}$. Clearly they "ought" to converge. However, they do not converge to

a rational number. We say that $\mathbb{Q}$ is not complete. However, they do converge to a *real* number: $\mathbb{R}$ is complete. When should a sequence x_k converge? Cauchy's answer was that

$$\text{for all } \epsilon > 0 \text{ there is a } K \text{ where } i, j \geq K \text{ implies } d(x_i, x_j) < \epsilon; \tag{A.5}$$

such a sequence is called a *Cauchy sequence*, and in a *complete* metric space, all Cauchy sequences converge to a limit. Every metric space X that is not complete has an extension space $Y \supset X$ with $d_Y(u, v) = d_X(u, v)$ for all $u, v \in X$, and Y is complete. The closure of X in Y is called a completion of X. Almost all spaces that we work with are complete metric spaces. For example, the real numbers $\mathbb{R}$ represent the completion of the rational numbers $\mathbb{Q}$. Note that every bounded increasing (or bounded decreasing) sequence of real numbers x_n has a finite limit. If for every $M \in \mathbb{R}$ we have $x_n \geq M$ for n sufficiently large, we say $\lim_{n\to\infty} x_n = +\infty$ or $x_n \to +\infty$ as $n \to \infty$.

The *supremum* $\sup(A)$ of a set of real numbers A is the smallest α such that $\alpha \geq a$ for every $a \in A$. If A has no such (finite) bound, we say $\sup(A) = +\infty$. The supremum of a function $f\colon X \to \mathbb{R}$ is $\sup f = \sup f(X)$. The *infimum* $\inf(A)$ of a set $A \subseteq \mathbb{R}$ is the largest β such that $\beta \leq a$ for every $a \in A$. The infimum of a function $f\colon X \to \mathbb{R}$ is $\inf f = \inf f(X)$. The *liminf* or limit inferior of a sequence of real numbers x_n, $n = 1, 2, \ldots$, is $\liminf_{n\to\infty} x_n = \lim_{n\to\infty} \inf\{x_n, x_{n+1}, x_{n+2}, \ldots\}$. Similarly, the *limsup* or limit superior of a sequence of real numbers x_n is $\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \sup\{x_n, x_{n+1}, x_{n+2}, \ldots\}$.

An important theorem for complete metric spaces is the *Baire category theorem*.

Theorem A.2. *If X is a complete metric space, and if $\{U_i\}_{i=1}^{\infty}$ is a countable collection of open sets, each of which is dense in X (that is, the closure $\overline{U_i} = X$ for all i), then the intersection $\bigcap_{i=1}^{\infty} U_i$ is also dense in X.*

Note that countable intersections of open sets are called G_δ sets, while the complementary intersections (formed by taking complements) of countable unions of closed sets are called F_σ sets. Since countable intersections of G_δ sets themselves are countable intersections of open sets (and therefore G_δ sets as well), this theorem can be extended to say that countable intersections of dense G_δ sets themselves are dense G_δ sets. The complementary result is that countable unions of F_σ sets that contain no open sets are also F_σ sets that contain no open sets.

A variation on the idea of metric spaces is where the topology (or convergence criterion) is given in terms of an infinite but countable family of metrics: $x_k \to x$ in X if and only if $d_j(x_k, x) \to 0$ as $k \to \infty$ for $j = 1, 2, 3, \ldots$. For example, for each integer $j \geq 0$, the space $C^j(\overline{\Omega})$ of j-times continuously differentiable functions on $\overline{\Omega}$ is a complete metric space with metric

$$d_j(f, g) = \max_{\alpha:|\alpha|\leq j} \max_{x\in\overline{\Omega}} \left|D^\alpha f(x)\right|,$$

where α is a multi-index and $D^\alpha f$ is the appropriate partial derivative of f as described in Section A.5. However, the space of infinitely differentiable functions $C^\infty(\overline{\Omega})$ cannot be given a single (or even a finite) set of metrics to define convergence. Instead we have the infinite family $d_1, d_2, \ldots$ of metrics to define convergence: $f_k \to f$ in $C^\infty(\overline{\Omega})$ if and only if $d_j(f_k, f) \to 0$ as $k \to \infty$ for all j.

A.2 Vector and Banach spaces

Vector spaces are collections of objects called *vectors* (e.g., x, y, etc.) on which there exist the operations of (vector) addition $x + y$ and scalar multiplication αx with a scalar α. We will deal with real vector spaces only, so scalars will be real numbers. Examples of vector spaces include n-dimensional vectors $\mathbb{R}^n$, the set of $m \times n$ matrices, and the continuous real-valued functions on a closed bounded domain $\overline{\Omega}$ (denoted by $C(\overline{\Omega}, \mathbb{R})$ or $C(\overline{\Omega})$).

If there is a finite set $\{z_1, z_2, \ldots, z_r\}$ where every vector x in the space can be written $x = \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_r z_r$ for some scalars α_i, then we say that the set $\{z_1, z_2, \ldots, z_r\}$ is a *generating set* for the vector space, and the vector space is finite dimensional. Otherwise we say the vector space is infinite dimensional. We say the vectors $z_1, z_2, \ldots, z_r$ are *linearly independent* if the only time $\alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_r z_r = 0$ is when $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$. If a generating set is also linearly independent, then we say the set is a *basis* for the vector space, and in any representation of a vector $x = \alpha_1 z_1 + \alpha_2 z_2 + \cdots + \alpha_r z_r$ in terms of the basis, the scalars α_i are unique. Furthermore, we say that the dimension of the vector space is r; this does not depend on the choice of basis.

Normed vector spaces are vector spaces with a *norm* that gives a measure of the size of a vector x: $\|x\|$. Norms must satisfy the following conditions:

- $\|x\| \geq 0$ for all vectors x, and $\|x\| = 0$ implies $x = 0$.
- $\|\alpha x\| = |\alpha| \, \|x\|$ for all vectors x and scalars α.
- $\|x + y\| \leq \|x\| + \|y\|$ for all vectors x and y.

Norms define a metric that is compatible with the vector space structure: $d_X(x, y) = \|x - y\|_X$. Convergence in norm is understood in the sense of this metric as described in the previous section. Examples of norms on $\mathbb{R}^n$ include the following:

- $\|x\|_1 := \sum_{i=1}^n |x_i|$.
- $\|x\|_2 := \left[\sum_{i=1}^n x_i^2\right]^{1/2}$.
- $\|x\|_\infty := \max_{i=1,\ldots,n} |x_i|$.

While there are many different norms we can use on $\mathbb{R}^n$ they are all *equivalent* in the sense that if $\|\cdot\|_a$ and $\|\cdot\|_b$ are equivalent norms, there are constants $C_a, C_b > 0$ such that

$$\frac{1}{C_b} \|x\|_a \leq \|x\|_b \leq C_a \|x\|_a \qquad \text{for all } x.$$

Equivalence of norms means that $x_n \to x$ in $\|\cdot\|_a$ if and only if $x_n \to x$ in $\|\cdot\|_b$; that is, the choice of norm amongst equivalent norms does not affect convergence of sequences. However, *not* all norms are equivalent in infinite-dimensional spaces.

Examples of norms on $C(\overline{\Omega})$ where $\overline{\Omega}$ is a closed and bounded set in $\mathbb{R}^n$ include the following:

- $\|f\|_\infty := \max_{x \in \Omega} |f(x)|$.
- $\|f\|_1 := \int_\Omega |f(x)| \, dx$.
- $\|f\|_2 := \left[\int_\Omega f(x)^2 \, dx\right]^{1/2}$.

None of these three norms is equivalent on $C(\overline{\Omega})$.

Norms are often obtained from *inner products* $(u, v)_X$, which generalize the dot product of vectors in $\mathbb{R}^n$: $(u, v) = u \cdot v = u^T v$. In general, inner products must satisfy the following assumptions:

- $(x, x)_X \geq 0$ for all $x \in X$, and $(x, x)_X = 0$ implies $x = 0$;
- symmetry: $(x, y)_X = (y, x)_X$ for all $x, y \in X$; and
- bilinearity: $(\alpha x + \beta y, z)_X = \alpha (x, z)_X + \beta (y, z)_X$, $(x, \alpha y + \beta z)_X = \alpha (x, y)_X + \beta (y, z)_X$ for any $x, y, z \in X$ and $\alpha, \beta \in \mathbb{R}$.

We will leave off the subscript "X" if the space is clear from context. The norm generated by the inner product is given by

$$\|x\|_X = \sqrt{(x, x)_X}.$$

A *Banach space* is a normed vector space that is also a complete space in the metric defined by the norm. Examples of Banach spaces include

- $\mathbb{R}^n$ for any finite $n \geq 0$,
- $C(\overline{\Omega})$ with the $\|\cdot\|_\infty$ norm for any closed and bounded $\overline{\Omega} \subset \mathbb{R}^d$.

Note that $C(\overline{\Omega})$ with either the $\|\cdot\|_1$ or the $\|\cdot\|_2$ norm is not a complete metric space. For example, if $\overline{\Omega} = [0, 1]$, then the functions $f_k(x) = \min(k, \ln(1/x))$ converge to the limit $f(x) = \ln(1/x)$ in both of these norms, but $f \notin C(\overline{\Omega})$. The ability to construct completions based on a norm enables us to define a number of spaces easily. More explicit constructions which show just what the functions in the space "look like" require more sophisticated tools such as Lebesgue integration theory.

For example, we can define $L^p(\Omega)$ for $1 \leq p < \infty$ and Ω bounded in $\mathbb{R}^d$ as the completion of $C(\overline{\Omega})$ in the norm

$$\|f\|_{L^p(\Omega)} = \left[\int_\Omega |f(x)|^p \, dx \right]^{1/p}.$$

Usually $L^p(\Omega)$ is described as the set of measurable functions for which the Lebesgue integral $\int_\Omega |f(x)|^p \, dx$ is finite, although functions f and g that are equal almost everywhere (that is, the Lebesgue measure of $\{x \in \Omega \mid f(x) \neq g(x)\}$ is zero) are considered to be the same function. The space $L^\infty(\Omega)$ is the space of *essentially bounded* functions from Ω to $\mathbb{R}$. That is, it is the space of functions $f : \Omega \to \mathbb{R}$ where there is an M such that $\{x \in \Omega \mid |f(x)| > M\}$ has Lebesgue measure zero. The norm for $L^\infty(\Omega)$ is given by

$$\|f\|_{L^\infty(\Omega)} = \inf \{M \mid \text{measure}(\{x \in \Omega \mid |f(x)| > M\}) = 0\},$$

where measure (E) is the Lebesgue measure of a set $E \subseteq \mathbb{R}^d$.

An operator is a continuous linear map $A : X \to Y$ between Banach spaces X and Y. For all continuous linear maps, we can assign a norm

$$\|A\|_{\mathcal{L}(X,Y)} = \sup_{0 \neq x \in X} \frac{\|Ax\|_Y}{\|x\|_X}, \tag{A.6}$$

which is finite for continuous linear A, and which makes the space of continuous linear maps $\mathcal{L}(X,Y)$ a Banach space. Differential operators like $\partial/\partial x$ and ∇^2 unfortunately are usually not continuous linear maps, at least not from a space X into itself. We can make them continuous operators $X \to Y$ for suitable spaces X and Y. For example, ∇^2 is a bounded linear operator $H^1(\Omega) \to H^{-1}(\Omega)$ for bounded regions $\Omega \subset \mathbb{R}^d$ (see Section A.5 for more explanation).

An operator $A\colon X \to Y$ is called *compact* if it maps bounded sets to precompact sets. This is equivalent to saying that $\overline{A(B_X)}$ is a compact subset of Y. Solution operators for differential equations are often compact operators, although this can depend on the Banach spaces used. In $\mathbb{R}^n$ all closed bounded sets are compact, so any linear operator $A\colon \mathbb{R}^n \to \mathbb{R}^m$ is a compact operator.

A set S in a vector space X is *absorbing* if for any $x \in X$ there is an $\alpha > 0$ such that $\alpha x \in S$. Alternatively, S is absorbing if $\bigcup_{\alpha>0} \alpha S = X$. The *core* of S is the set of all $x \in S$ where $S - x$ is absorbing. The core of a set S is always a subset of the interior of S.

A.3 Dual spaces, Hilbert spaces, and weak convergence

Hilbert spaces are Banach spaces where the norm is generated by an inner product. Examples include $\mathbb{R}^n$ with the inner product $(u,v) = u^T v$, which generates the usual Euclidean norm: $\|x\| = \left[\sum_{i=1}^n x_i^2\right]^{1/2} = \sqrt{x^T x}$. Another example is the space $L^2(\Omega)$ where the norm $\left[\int_\Omega |f(x)|^2\, dx\right]^{1/2}$ is generated by the inner product

$$(f,g)_{L^2(\Omega)} = \int_\Omega f(x)\, g(x)\, dx.$$

The *dual space* of a normed vector space X is the set of continuous linear functions $X \to \mathbb{R}$ (called linear *functionals*):

$$X' = \{\, f\colon X \to \mathbb{R} \mid f \text{ continuous } \&\ \text{linear}\,\}.$$

We usually denote the application of $f \in X'$ to $x \in X$ by

$$f(x) = \langle f, x\rangle_{X'\times X}$$

or just $\langle f, x\rangle$ when it is clear what X and X' are. If $X = \mathbb{R}^n$, then X' can be represented by $\mathbb{R}^n$: any linear function $\mathbb{R}^n \to \mathbb{R}$ can be represented by $x \mapsto v^T x = \langle v, x\rangle$.

The space X' has a norm

$$\|f\|_{X'} = \sup_{0\neq x\in X} \frac{|\langle f, x\rangle|}{\|x\|_X} = \sup_{x:\|x\|_X=1} |\langle f, x\rangle|.$$

This is finite, as continuity implies $\|f\|_{X'}$ is finite: to see why, suppose $\|f\|_{X'} = +\infty$. Then there would be a sequence x_n with $\|x_n\|_X = 1$ and $|\langle f, x_n\rangle| \to \infty$. By changing the sign of x_n, if necessary, we can make $\langle f, x_n\rangle \geq 0$ for all n; thus $\langle f, x_n\rangle \to +\infty$ as $n \to \infty$. Then setting $y_n = x_n / \langle f, x_n\rangle$ we have $\|y_n\|_X = \|x_n\|_X / \langle f, x_n\rangle \to 0$ as $n \to \infty$, and so $y_n \to 0$ in X. However, $\langle f, y_n\rangle = \langle f, x_n\rangle / \langle f, x_n\rangle = 1 \not\to 0 = \langle f, 0\rangle$, so f is not a continuous function.

Conversely, $\|f\|_{X'}$ being finite means that f is continuous: Suppose $x_n \to x$. Then $|\langle f, x_n\rangle - \langle f, x\rangle| = |\langle f, x_n - x\rangle| \le \|f\|_{X'}\|x_n - x\|_X \to 0$ as $n \to \infty$.

This norm makes X' a normed vector space, and in fact, X' is also a Banach space. If X is a Hilbert space, then there is the *duality map* $J_X : X \to X'$ given by $J_X(x) = (x, \cdot)_X$. Then $\langle J_X(x), y\rangle_{X'\times X} = (x, y)_X$, where the former is a duality pairing and the latter is the inner product on X. Sometimes we identify X with X' by identifying x with $J_X(x)$, but usually we keep these distinct. In general, a map $J_X : X \to X'$ (X not a Hilbert space) is called a duality map if

$$\begin{aligned} &\|J_X(x)\|_{X'} = \|x\|_X, \\ &\langle J_X(x), y\rangle \le \|x\|_X\,\|y\|_X \qquad \text{with equality if and only if } y = x. \end{aligned}$$

It is a standard result of functional analysis that X is a Hilbert space if and only if X has a linear duality map.

The dual space of $L^p(\Omega)$ can be represented by $L^q(\Omega)$, where $1/p + 1/q = 1$ and $1 < p < \infty$: any $f \in L^p(\Omega)'$ can be represented by $h \in L^q(\Omega)$ so that

$$\langle f, g\rangle_{L^p(\Omega)'\times L^p(\Omega)} = \int_\Omega h(x)\,g(x)\,dx.$$

We usually identify the functional $f \in L^p(\Omega)'$ with the function $h \in L^q(\Omega)$. The dual space of $L^1(\Omega)$ is identified with $L^\infty(\Omega)$ in the same way; however, the dual space of $L^\infty(\Omega)$ is *not* $L^1(\Omega)$.

There is a natural map $\natural : X \to X''$ given by $\langle \natural(x), w\rangle_{X''\times X'} = \langle w, x\rangle_{X'\times X}$. (The symbol "$\natural$" in music means that the "natural" note is played rather than the sharp or flat that would usually be played according to the key of the musical piece.) Any space for which $\natural$ is an isomorphism is called *reflexive*. Hilbert spaces are automatically reflexive. Most Banach spaces we deal with are reflexive, such as $L^p(\Omega)$ for $1 < p < \infty$. However, none of $C(\Omega)$, $L^1(\Omega)$, or $L^\infty(\Omega)$ is reflexive.

A.3.1 Adjoints of linear operators

Given a continuous linear operator $A : X \to Y$ between Banach spaces X and Y, there is the *adjoint* operator $A^* : Y' \to X'$ which is defined by

$$\left\langle A^*\eta, x\right\rangle = \langle \eta, Ax\rangle \qquad \text{for all } x \in X \text{ and } \eta \in Y'. \tag{A.7}$$

If A is an $m \times n$ matrix so that $A : \mathbb{R}^n \to \mathbb{R}^m$, then $A^* = A^T$, the transpose of A. There are a number of important properties of adjoint operators:

- if X and Y are reflexive spaces (so we identify X and Y with X'' and Y'', respectively), then $A^{**} = A$;
- the norm $\|A^*\|_{\mathcal{L}(Y',X')} = \|A\|_{\mathcal{L}(X,Y)}$;
- if A is one-to-one, then $A^*(Y')$ is dense in X' (that is, $\overline{A^*(Y')} = X'$);
- if $A(X)$ is dense in Y (that is, $\overline{A(X)} = Y$), then A^* is one-to-one;
- if A is a compact operator, so is A^*.

An operator $A\colon X \to X'$ is called self-adjoint if $\langle Ax, y\rangle = \langle Ay, x\rangle$. If we identify X with X' for a Hilbert space X, this means that $(Ax, y)_X = (Ay, x)_X$. In terms of adjoints, A is self-adjoint means that $A^*\colon X'' \to X'$ and $A^* \circ \natural = A$, where $\natural$ is the natural map $X \to X''$. If X is reflexive and we identify X and X'', then this just means that $A^* = A$.

A.3.2 Weak versus strong topologies

For more general notions of spaces and notions of convergence, the idea of *topology* was invented. In essence it gives us a way of describing when a sequence converges to a limit. In a Banach space, the strong topology is the topology generated by the norm of the space:

$$x_n \to x \quad \text{as } n \to \infty \qquad \text{means} \qquad \|x_n - x\| \to 0 \quad \text{as } n \to \infty.$$

However, often we need less stringent notions of convergence. The notion of *weak convergence* is motivated by the idea that all averages of a sequence of functions might converge, even if the functions in the sequence do not. Consider, for example, the sequence of functions in $L^2(0, 2\pi)$ given by $x_n(t) = \sin(nt)$. The norm of x_n in $L^2(0, 2\pi)$ is $\sqrt{\pi}$ for all n. On the other hand, for *any* smooth or continuous (or even $L^2(0, 2\pi)$ function) ϕ, $\int_0^{2\pi} \phi(t) x_n(t)\, dt \to 0$ as $n \to \infty$. Thus we have a weak limit $x_n \rightharpoonup 0$ but no strong limit: $x_n \not\to 0$.

The general definition of weak convergence is as follows:

$$\begin{aligned} x_n \rightharpoonup x \text{ weakly in } X \qquad &\text{means} \\ \langle y, x_n\rangle_{X'\times X} \to \langle y, x\rangle_{X'\times X} \quad &\text{for all } y \in X'. \end{aligned} \tag{A.8}$$

A consequence of weak convergence is that $\sup_n \|x_n\| < \infty$; that is, weakly convergent sequences are bounded. Important results regarding weak convergence include *Mazur's lemma*.

Lemma A.3 (Mazur). *If $x_n \rightharpoonup x$ (weakly) in a Banach space X as $n \to \infty$, then there is a sequence $y_n \in co\{x_n, x_{n+1}, x_{n+2}, \ldots\}$ where $y_n \to x$ strongly.*

An immediate consequence of this lemma is that if C is a (strongly) closed convex set in X, then C is closed with respect to weak convergence; that is, if $x_n \rightharpoonup x$ (weakly) as $n \to \infty$ and $x_n \in C$ for all n, then the weak limit $x \in C$ as well.

Closely related to weak convergence is *weak* convergence*: if $X = Y'$, and Y a Banach space, then $x_n \overset{*}{\rightharpoonup} x$ as $n \to \infty$ or x_n converges weak* to x means that

$$\langle x_n, y\rangle_{Y'\times Y} \to \langle x, y\rangle_{Y'\times Y} \qquad \text{as } n \to \infty \text{ for all } y \in Y. \tag{A.9}$$

Because of the natural map $\natural\colon Y \to Y'' = X'$, weak convergence implies weak* convergence. If X is a reflexive Banach space, then weak and weak* convergence are identical, as we can then consider $Y'' = Y$. But sometimes weak* convergence is the right form of convergence to consider: Alaoglu's theorem says that if $X = Y'$, Y a Banach space, then every bounded set $A \subset X$ is weak* precompact.

Another consequence of Mazur's lemma is that if X is a reflexive Banach space and $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is a proper ($\inf_x \phi(x) < \infty$), lower semicontinuous ($x_n \to x$ implies $\phi(x) \le \liminf_{n\to\infty} \phi(x_n)$), weakly coercive ($\lim_{\|x\|\to\infty} \phi(x) = \infty$), and convex function,

then ϕ has a minimum value $\inf_x \phi(x) = \phi(x^*)$ for some $x^* \in X$. This is an important result in convex analysis which will be used in the section on convex and nonsmooth analysis.

A.3.3 Compactness in particular spaces

Conditions for sets in certain Banach spaces to be compact have been extremely useful in applications, especially for the Banach spaces $C(0,T)$ of continuous functions $[0,T] \to \mathbb{R}$ and $L^p(0,T)$, and the spaces of vector-valued functions $C(0,T;X)$ and $L^p(0,T;X)$. We are not concerned just about compactness in the strong topology but also in the weak and weak* topologies.

Recall that the norm used for $C(0,T)$ and $C(0,T;X)$ is the supremum norm

$$\|f\|_{C(0,T;X)} = \max_{t\in[0,T]} \|f(t)\|_X ,$$

where for $C(0,T)$, $\|f(t)\|_X$ is just the absolute value of $f(t)$. The main compactness theorem for $C(0,T)$ and $C(0,T;X)$ is the Arzela–Ascoli theorem, which is based on the concept of equicontinuity.

Definition A.4. *A set of functions $F \subset C(0,T;X)$ is* equicontinuous *if for each $t^* \in [0,T]$ and $\epsilon > 0$ there is a $\delta > 0$ where $|t - t^*| < \delta$ implies $\|f(t) - f(t^*)\|_X < \epsilon$ for all $f \in F$.*

Theorem A.5 (Arzela–Ascoli). *A set of functions $F \subset C(0,T;X)$ is compact if and only if it is bounded in the supremum norm and equicontinuous, and for each $t \in [0,T]$ the set $\{ f(t) \mid f \in F \}$ is compact in X.*

A proof of this can be found in, for example, Lang [155]. This reduces for $F \subset C(0,T;\mathbb{R}^n)$ to be precompact if F is bounded and equicontinuous. An important consequence of the Arzela–Ascoli theorem is a compactness theorem of Seidman [224]. Here is Seidman's theorem (following Kuttler [151, pp. 499–501]).

Theorem A.6 (Seidman). *If F is a bounded subset of $L^\infty(a,b;X)$ with f' uniformly bounded in $L^p(a,b;Y)$ $(1 < p \leq \infty)$ for all $f \in F$, then F is precompact in $C(a,b;Z)$ for any Banach spaces $X \subset Z \subseteq Y$ with the imbedding $X \subset Z$ compact and the imbedding $Z \subseteq Y$ continuous.*

Proof. First we show that for every $\epsilon > 0$ there is a constant C_ϵ where

$$\|x\|_Z \leq \epsilon \|x\|_X + C_\epsilon \|x\|_Y \qquad \text{for all } x \in X. \tag{A.10}$$

If this were not true, then for some ϵ there would be a sequence $x_n \in X$ such that $\|x_n\|_Z > \epsilon \|x_n\|_X + n \|x_n\|_Y$. Without loss of generality, we suppose that $\|x_n\|_X = 1$ for all n. Now the imbedding $X \subset Z$ is compact, so there is a convergent subsequence (also denoted by x_n) where $x_n \to x^*$ in Z. Note that $x^* \neq 0$, since $\|x_n\|_Z \geq \epsilon > 0$ for all n. By continuity of the imbedding $Z \subseteq Y$, $x_n \to x^*$ in Y. Dividing the inequality $\|x_n\|_Z > \epsilon \|x_n\|_X + n \|x_n\|_Y$ by n and taking limits, we get $0 \geq \|x^*\|_Y$, which implies that $x^* = 0$, a contradiction. Thus we must conclude that there is indeed a constant C_ϵ making (A.10) true.

Now suppose we have $f \in F$ bounded in $L^\infty(a,b;X)$. Then, for any $\epsilon > 0$ and $s < t$,

$$\begin{aligned}\|f(t)-f(s)\|_Z &\le \epsilon\,\|f(t)-f(s)\|_X + C_\epsilon\,\|f(t)-f(s)\|_Y \\ &\le 2\epsilon \sup_{g\in F}\|g\|_{L^\infty(a,b;X)} + C_\epsilon \int_s^t \left\|f'(\tau)\right\|_Y d\tau \\ &\le 2\epsilon \sup_{g\in F}\|g\|_{L^\infty(a,b;X)} + C_\epsilon \left[\int_s^t \left\|f'(\tau)\right\|_Y^p d\tau\right]^{1/p} |t-s|^{1/q},\end{aligned}$$

where $1/p + 1/q = 1$. Since $p > 1$, we have $1 < q < \infty$, and so for any $\eta > 0$, we can choose

$$\begin{aligned}\epsilon &= \eta \Big/ \left(2 \sup_{g\in F}\|g\|_{L^\infty(a,b;X)}\right), \\ \delta &= \left(\eta \Big/ \left(2C_\epsilon \sup_{f\in F}\left\|g'\right\|_{L^p(a,b;Y)}\right)\right)^q.\end{aligned}$$

Then whenever $|t-s| < \delta$ we have $\|f(t)-f(s)\|_Z < \eta$ for all $f \in F$, so that F is equicontinuous. In addition, the set of values $\{f(t) \mid f \in F, t \in [a,b]\}$ is bounded in X and therefore compact in Z. Thus we can apply the Arzela–Ascoli theorem (Theorem A.5) to see that F is a precompact subset of $C(a,b;Z)$. □

There are also the results of Simon [227] which can be helpful for establishing compactness. See also the textbook of Kuttler [151] for more accessible discussion of these theorems of Seidman and Simon.

Theorem A.7 (Simon). *If F is a bounded subset of $L^q(a,b;X)$ with f' uniformly bounded in $L^1(a,b;Y)$ for all $f \in F$, then F is precompact in $L^p(a,b;Z)$ $(1 \le p < \infty)$ for any Banach spaces $X \subset Z \subseteq Y$ with the imbedding $X \subset Z$ compact and the imbedding $Z \subseteq Y$ continuous.*

A.4 Distributions and measures

Distributions can be constructed using a kind of duality trick, but instead of starting with a Banach space of functions, we start with the nicest space of functions we usually work with: the space of functions that can be differentiated as many times as we please but are zero outside some bounded set. This space of functions is denoted by $C_0^\infty(\mathbb{R}^d)$. The functions $\phi \in C_0^\infty(\mathbb{R}^d)$ are called *test functions*. There is no norm for this space; instead there is a family of *seminorms* $\|\phi\|_{k,R} = \max_{x:\|x\|\le R} \max_{\alpha:|\alpha|\le k} |D^\alpha\phi(x)|$ where for $\alpha = (\alpha_1,\alpha_2,\ldots,\alpha_d)$ (each α_i is a nonnegative integer), $|\alpha| = \sum_{i=1}^d \alpha_i$ and $D^\alpha\phi(x) = \partial^{|\alpha|}\phi/\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\cdots\partial x_n^{\alpha_n}(x)$. Convergence in $C_0^\infty(\mathbb{R}^d)$ works like this: $\phi_n \to \phi$ as $n \to \infty$ if and only if there is a common R such that $\phi_n(x) = 0$ for $\|x\| \ge R$ and all n, and $\|\phi_n - \phi\|_{k,R} \to 0$ as $n \to \infty$ for all k.

Often the space $C_0^\infty(\mathbb{R}^d)$ is denoted by $\mathcal{D}(\mathbb{R}^d)$. The space of *distributions* is the *dual* space to $\mathcal{D}(\mathbb{R}^d)$. An integrable function ψ can be considered to be a distribution by

identifying ψ with the functional $\mathcal{D}(\mathbb{R}^d) \to \mathbb{R}$ given by

$$\phi \mapsto \langle \psi, \phi \rangle := \int_{\mathbb{R}^d} \psi(x)\phi(x)\,dx.$$

Thanks to duality we can extend operations such as differentiation from the test functions to distributions. To see this, suppose that ψ is a function with a continuous derivative. Then in one dimension ($d = 1$) where $\phi(x) = 0$ for $|x| \geq R$,

$$\begin{aligned}
\langle \psi', \phi \rangle &= \int_{-\infty}^{+\infty} \psi'(x)\phi(x)\,dx \\
&= \int_{-R}^{+R} \psi'(x)\phi(x)\,dx \\
&= \psi(x)\phi(x)|_{x=-R}^{x=+R} - \int_{-R}^{+R} \psi(x)\phi'(x)\,dx \\
&= -\int_{-\infty}^{+\infty} \psi(x)\phi'(x)\,dx = -\langle \psi, \phi' \rangle.
\end{aligned}$$

Since differentiation is a continuous operator on $C_0^\infty(\mathbb{R}^d)$, we can extend the differentiation of distributions to general distributions (and not just ψ with continuous derivatives) by defining

$$\langle \psi', \phi \rangle = -\langle \psi, \phi' \rangle$$

for $d = 1$, and generally $\langle \partial\psi/\partial x_i, \phi \rangle = -\langle \psi, \partial\phi/\partial x_i \rangle$. This means that we can differentiate practically any function: if $\psi(x) = |x|$, then $\psi'(x) = \operatorname{sgn}(x)$, and $\psi''(x) = 2\delta(x)$, where δ is the Dirac-δ function (really a distribution) where $\langle \delta, \phi \rangle = \phi(0)$. There are also derivatives of the Dirac-δ function: $\langle \delta', \phi \rangle = -\langle \delta, \phi' \rangle = -\phi'(0)$, $\langle \delta'', \phi \rangle = -\langle \delta', \phi' \rangle = \phi''(0)$, etc.

Most linear operations can be applied to distributions, but many nonlinear operations such as multiplication usually cannot. For example, δ^2 does not exist, although the distribution in two dimensions $g(x, y) = \delta(x)\delta(y)$ does exist: $\langle g, \phi \rangle = \phi(0, 0)$.

For information about tempered distributions in relation to Fourier transforms, see Section C.4.

Since we need to deal with inequalities, we need to understand what "$\psi \geq 0$" means for distributions (whether ordinary or tempered distributions). Since distributions are defined in terms of duality, it is appropriate to define "$\psi \geq 0$" also in terms of duality. Clearly we understand what "$\phi \geq 0$" means for $\phi \in C_0^\infty(\mathbb{R}^d)$ since these are spaces of ordinary functions: "$\phi \geq 0$" means "$\phi(x) \geq 0$ for all $x \in \mathbb{R}^d$." So we define "$\psi \geq 0$" for distributions by

$$\langle \psi, \phi \rangle \geq 0 \qquad \text{for all } \phi \in C_0^\infty(\mathbb{R}^d) \text{ where } \phi \geq 0. \tag{A.11}$$

It turns out that the only such distributions are *measures* [127].

Measures can be considered as functions that take sets as input and return a number, which is a "measure" of the size of the set. The most common measure is the *Lebesgue measure* λ where $\lambda([a,b]) = b - a$, the length of the interval $[a,b]$. The properties of a measure μ are as follows:

1. $\mu(\emptyset) = 0$ (the measure of the empty set is zero); and
2. if $E = \bigcup_{k=1}^{\infty} E_k$ and $E_i \cap E_j = \emptyset$ for $i \neq j$, then $\mu(E) = \sum_{k=1}^{\infty} \mu(E_k)$ (countable additivity).

Note that it is not necessary for $\mu(E)$ to be defined for *all* subsets E. However, the sets E should be closed under countable unions and intersections and also complements. Such a collection of sets is called a σ-algebra. The σ-algebra generated by the collection of open sets in a space X is the collection of *Borel sets* of X; usually $\mu(E)$ is defined for every Borel set E. We can allow $+\infty$ to be a value; for example, the Lebesgue measure of the entire real line is $\lambda(\mathbb{R}) = +\infty$.

Note that a function $f\colon X \to Y$ is called measurable if $f^{-1}(E)$ is a measurable set in X whenever E is a measurable set in Y. Note that this definition depends on what σ-algebras of measurable sets we choose for X and for Y. If X and Y are simply topological spaces, we can use the σ-algebra of Borel sets in X and in Y. In that case, we say f is *Borel measurable* if f is measurable with respect to these σ-algebras. Since $f^{-1}(E \cup F) = f^{-1}(E) \cup f^{-1}(F)$, $f^{-1}(E \cap F) = f^{-1}(E) \cap f^{-1}(F)$, and $f^{-1}(Y \setminus E) = X \setminus f^{-1}(E)$, we can show that f is Borel measurable if $f^{-1}(U)$ is a Borel set in X for every open set U in Y.

Often we deal with nonnegative measures μ where $\mu(E) \geq 0$ for all (appropriate) subsets E. Then condition 2 above for measures implies that if $E_1 \subseteq E_2 \subseteq \cdots \subseteq E_k \subseteq E_{k+1} \subseteq \cdots$, then $\mu(E_k) \uparrow \mu(E)$ as $k \to \infty$, where $E = \bigcup_{k=1}^{\infty} E_k$.

Constructing a measure from a nonnegative distribution involves some technical difficulties, but the basic idea starts with the realization that any test function ϕ can be written as the difference of two nonnegative test functions: $\phi = \phi_1 - \phi_2$. (It is tempting to set $\phi_1(x) = \max(\phi(x), 0)$ and $\phi_2(x) = \max(-\phi(x), 0)$, but this would not give smooth functions in general.) We want to show that we can assign a value for the measure $\mu(E) = \langle \psi, \chi_E \rangle$ where $\chi_E(x) = 1$ if $x \in E$ and $\chi_E(x) = 0$ otherwise for E either closed or open. We do this by finding a decreasing sequence of test functions $\phi_k \downarrow \chi_E$ for E closed and an increasing sequence of test functions $\phi_k \uparrow \chi_E$ for E open. For nonnegative distributions $\langle \psi, \phi_k \rangle$ is a monotone increasing or decreasing sequence that is also bounded. So these sequences converge. The limit is $\mu(E)$. Proving that the μ so constructed is a measure is a standard result of measure theory [61].

There is another way of thinking about measures. The space of bounded measures on the Borel sets in a closed set $A \subseteq \mathbb{R}^d$ (denoted by $\mathcal{M}(A)$) is the dual space to the space of continuous functions $A \to \mathbb{R}$. The duality pairing is represented by integration over a given measure:

$$\langle \mu, \phi \rangle = \int_A \phi(x) \, d\mu(x). \tag{A.12}$$

The integral in (A.12) can be approximated by sums like this:

$$\sum_{k=0}^{N} \mu\left(\phi^{-1}\left([y_k, y_{k+1})\right)\right) \eta_k,$$

where $y_0 \leq \eta_0 < y_1 \leq \eta_1 < \cdots \leq \eta_{N-1} < y_N$. The duality $\mathcal{M}(A) = C(A)'$ for compact sets A can be very useful for applying Alaoglu's theorem, showing the existence of weak* convergent subsequences of a sequence of measures. One way of generating measures on

an interval $[a,b]$ is by functions $g\colon [a,b] \to X$ which have *bounded variation*:

$$\bigvee_a^b g = \sup_{\mathcal{P}} \sum_{i=0}^{N-1} \|g(t_{i+1}) - g(t_i)\|, \tag{A.13}$$

where $\mathcal{P} : a = t_0 < t_1 < t_2 < \cdots < t_N = b$ ranges over partitions of $[a,b]$ (N is not fixed, but rather can go to ∞). Then the *differential measure* dg is given by the *Riemann–Stieltjes integrals* for continuous $f\colon [a,b] \to \mathbb{R}$:

$$\int_{[a,b]} f\,dg = \lim_{|\mathcal{P}|\to 0} \sum_{i=0}^{N-1} f(\tau_i)(g(t_{i+1}) - g(t_i)), \tag{A.14}$$

where $|\mathcal{P}| = \max_i |t_{i+1} - t_i|$ and $\tau_i \in [t_i, t_{i+1}]$ for all i. Note that dg is a measure with values in a Banach space X. If $g(t) = t$, then dg is the Lebesgue measure. Also, if $g\colon [a,b] \to \mathbb{R}$, then the Riemann–Stieltjes integrals are identical with the Bochner integrals for continuous $f\colon [a,b] \to X$ with the measure $\mu = dg$.

Measures have a number of special properties. For all measures μ there is a related nonnegative measure called the *variation measure* $|\mu|$. This is defined by

$$|\mu|(E) = \sup_{\{E_j\}_{j=1}^\infty} \sum_{j=1}^\infty |\mu(E_j)|. \tag{A.15}$$

The collection $\{E_j\}_{j=1}^\infty$ ranges over all collections of disjoint μ-measurable sets whose union is $E = \bigcup_{j=1}^\infty E_j$. Clearly $|\mu|$ is a nonnegative measure and $|\mu(E)| \le |\mu|(E)$ for any set for which $\mu(E)$ is defined. However, it is possible for $|\mu|(E) = +\infty$. Note that if μ is already a nonnegative measure, then $|\mu| = \mu$. If $|\mu|(E)$ is finite for all E, we say that μ is a measure with *bounded variation*. Measures of bounded variation can be written in the form $\mu = \mu_+ - \mu_-$, where μ_+ and μ_- are nonnegative measures. In fact, $\mu_+ = (|\mu| + \mu)/2$ and $\mu_- = (|\mu| - \mu)$.

The integrals $\int_A \phi(x)\,d\mu(x)$, or $\int_A \phi\,d\mu$ for short, are defined not only for continuous ϕ but also for a much larger class of functions. Suppose that ν is a nonnegative measure. A ν-measurable function $\phi \ge 0$ is ν-integrable if there is an increasing sequence of step functions $\phi_k = \sum_{j=1}^{N_k} c_{k,j}\chi_{E_{k,j}}$ that converges pointwise to ϕ, and $\int_A \phi_k(x)\,d\nu(x) = \sum_{j=1}^{N_k} c_{k,j}\,\nu(E_{k,j} \cap A)$ converges. The value of the limit is $\int_A \phi(x)\,d\nu(x)$. A ν-measurable function ϕ is ν-integrable if we can write $\phi = \phi_+ - \phi_-$ with both $\phi_+, \phi_- \ge 0$ and ν-integrable. In the case of functions $\phi\colon A \to X$ with X a separable Banach space, we can define the integrals in the same way: via pointwise limits of step functions $\phi_k = \sum_{j=1}^{N_k} c_{k,j}\chi_{E_{k,j}}$ with $c_{j,k} \in X$, each $E_{k,j}$ ν-measurable. An alternative way of defining these integrals is by subdividing $X = \bigcup_{j=1}^\infty F_j$ (F_j's disjoint) with each F_j a Borel set with $\operatorname{diam} F_j < \delta$ for a given $\delta > 0$. Picking $x_j \in F_k$, we can approximate ϕ by $\phi_\delta = \sum_{j=1}^\infty x_j \chi_{\phi^{-1}(F_j)}$ and $\phi_\delta \to \phi$ pointwise as $\delta \downarrow 0$. Then we take

$$\int_A \phi\,d\nu = \lim_{\delta\to 0} \sum_{j=1}^\infty x_j\,\nu(\phi^{-1}(F_j)).$$

These definitions give the same integral, which is known as the *Bochner integral*.

The space of real-valued functions that are ν-integrable, $\nu \geq 0$ a measure on a measure space A, is written $L^1(\nu)$ and is a Banach space with norm

$$\|\phi\|_{L^1(\nu)} = \int |\phi(a)|\, d\nu(a).$$

The spaces $L^1(\nu)$ and $L^1(\nu; X)$ of integral functions $A \to \mathbb{R}$ and $A \to X$, respectively, have been widely studied. Basic properties include the properties that $f\colon A \to X$ is integrable if and only if it is measurable and $a \mapsto \|f(a)\|_X$ is integrable. Perhaps the most important result is the dominated convergence theorem: if $f_k \to f$ pointwise with each f_k measurable, and $\|f_k(a)\|_X \leq g(a)$, where $g \in L^1(\nu)$ for all k and $a \in A$, then

$$\lim_{k\to\infty} \int_A f_k(a)\, d\nu(a) = \int_A f(a)\, d\nu(a). \tag{A.16}$$

Given a ν-integrable function ϕ, we can define a new measure

$$\nu_\phi(E) = \int_E \phi(a)\, d\nu(a).$$

When can a measure μ be represented like this? The answer is given by the Radon–Nikodym theorem. We say that μ is *absolutely continuous* with respect to a nonnegative ν measure if for $\epsilon > 0$ there is $\delta > 0$ such that for any measurable set E, $\nu(E) < \delta$ implies $|\mu(E)| < \epsilon$. The Radon–Nikodym theorem says that if μ is absolutely continuous with respect to ν, then there is a ν-integrable function h such that for any measurable set E,

$$\mu(E) = \int_E h(x)\, d\nu(x).$$

The function h is called the Radon–Nikodym derivative of μ with respect to ν. This function is unique up to a set of ν-measure zero. It is denoted by $h(x) = d\mu/d\nu(x)$.

Vector measures are measures whose values belong to a vector space, most usually a Banach space. The same additivity properties hold. For vector measures the absolute continuity property becomes the following: for $\epsilon > 0$ there is a $\delta > 0$ such that $\nu(E) < \delta$ implies $\|\mu(E)\|_X < \epsilon$, where the values of μ lie in the Banach space X. However, whether this implies that $\mu(E) = \int_E h(x)\, d\nu(x)$ for all measurable E depends on X *and not the measure* ν. For finite-dimensional Banach spaces $\mathbb{R}^n$, absolute continuity implies a Radon–Nikodym derivative exists since $\mu(E) = [\,\mu_1(E), \mu_2(E), \ldots, \mu_n(E)\,]^T$ and each μ_i is a scalar-valued measure with its own Radon–Nikodym derivative $d\mu_i/d\nu$, so $d\mu/d\nu = [\,d\mu/d\nu_1, d\mu/d\nu_2, \ldots, d\mu_n/d\nu\,]^T$. Any space X for which absolute continuity of a measure with values in X implies a Radon–Nikodym derivative exists is said to have the *Radon–Nikodym property* (RNP). All dual spaces have the RNP, so all Sobolev spaces $W^{s,p}(\Omega)$ with $1 < p \leq \infty$ have the RNP. However, $L^1(\Omega)$ spaces usually do *not* have the RNP.

If a space X has the RNP, then any absolutely continuous *function* $f\colon [a,b] \to X$ has a regular derivative $f'(t) = \lim_{h\to 0} (f(t+h) - f(t))/h$ for Lebesgue almost all t. In particular, if $f\colon [a,b] \to X$ is Lipschitz, then it is differentiable almost everywhere. Here is a simple example to show that $L^1(a,b)$ does *not* have the RNP. Let $f\colon [a,b] \to L^1(a,b)$

be the function $f(t) = \chi_{[a,t)}$. Note that χ_E is the characteristic function for E: $\chi_E(t) = 1$ if $t \in E$ and $\chi_E(t) = 0$ otherwise. This is a Lipschitz function: for $t > s$,

$$\begin{aligned}\|f(t) - f(s)\|_{L^2(a,b)} &= \left\|\chi_{[a,t)} - \chi_{[a,s)}\right\|_{L^1(a,b)} \\ &= \left\|\chi_{[s,t)}\right\|_{L^1(a,b)} = \int_a^b \chi_{[s,t)}(\tau)\,d\tau = |t-s|\,.\end{aligned}$$

But there is no derivative $df/dt(t)$ for *any* t:

$$\begin{aligned}f'(t) &= \lim_{h\to 0} \frac{\chi_{[a,t+h)} - \chi_{[a,t)}}{h} \\ &= \lim_{h\to 0} \frac{1}{h}\chi_{[t,t+h)}.\end{aligned}$$

The limit does exist in the sense of distributions: it is the Dirac-δ function $s \mapsto \delta(s-t)$, not an element of $L^1(a,b)$. Thus there is no derivative of $t \mapsto \chi_{[a,t)}$, and $L^1(a,b)$ does *not* have the RNP.

A useful tool in many situations is the *convolution* of functions: if $f, g\colon \mathbb{R}^d \to \mathbb{R}$,

$$(f * g)(x) = \int_{\mathbb{R}^d} f(x-y)\,g(y)\,dy. \tag{A.17}$$

For functions $f, g\colon [0,\infty) \to \mathbb{R}$ we have the finite time convolution

$$(f * g)(t) = \int_0^t f(t-s)\,g(s)\,ds. \tag{A.18}$$

Note that convolution is a bilinear operation (that is, linear in each argument), and $f * g = g * f$ for either form. There is also *Young's lemma*, where if $p, q, r \geq 1$ and $1/p + 1/q + 1/r = 2$, then

$$\|f * g\|_{L^r} \leq \|f\|_{L^p}\,\|g\|_{L^q}\,. \tag{A.19}$$

If, say, g is actually a measure, then we also have the following bound:

$$\|f * g\|_{L^p} \leq \|f\|_{L^p}\,\|g\|_{\mathcal{M}}\,. \tag{A.20}$$

A.5 Sobolev spaces and partial differential equations

It was noted by Dirichlet that the solution u to the partial differential equation $\nabla^2 u = 0$ in Ω with $u = g$ on $\partial\Omega$ is a minimizer of the integral

$$\int_\Omega \nabla u \cdot \nabla u\,dx \tag{A.21}$$

over all functions u where $u = g$ on $\partial\Omega$. But this integral is not defined for all u, and it is not clear whether it has a true minimizer or just an infimum. Fortunately the theory of Sobolev spaces (being a family of Banach spaces) can help us answer these questions. For more details see, for example, [1].

First we define the space of functions $W^{m,p}(\Omega)$ for m a nonnegative integer, with $1 \le p \le \infty$ and Ω an open subset of $\mathbb{R}^d$. A *multi-index* for d dimensions is a d-tuple $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ where each α_i is a nonnegative integer. We define $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ as the order of α; the partial derivative D^α is given by

$$D^\alpha f(x) = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}(x). \tag{A.22}$$

Also we define $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$. Note that these partial derivatives will be understood as being distributional derivatives where necessary. Then

$$W^{m,p}(\Omega) = \left\{ f \in L^p(\Omega) \mid D^\alpha f \in L^p(\Omega) \text{ for all multi-indexes } \alpha : |\alpha| \le m \right\}, \tag{A.23}$$

which is a Banach space with the norm $\|f\|_{W^{m,p}(\Omega)}$ given by

$$\begin{aligned} \|f\|_{W^{m,p}(\Omega)}^p &= \sum_{\alpha:|\alpha|\le m} \left\| D^\alpha f \right\|_{L^p(\Omega)}^p \\ &= \sum_{\alpha:|\alpha|\le m} \int_\Omega \left| D^\alpha f(x) \right|^p \, dx. \end{aligned} \tag{A.24}$$

There are a number of equivalent norms for $W^{k,p}(\Omega)$, such as

$$\begin{aligned} \|f\|'^p_{W^{m,p}(\Omega)} &= \|f\|^p_{L^p(\Omega)} + \sum_{\alpha:|\alpha|=m} \left\| D^\alpha f \right\|^p_{L^p(\Omega)}, \\ \|f\|''_{W^{m,p}(\Omega)} &= \|f\|_{L^p(\Omega)} + \sum_{\alpha:|\alpha|\le m} \left\| D^\alpha f \right\|_{L^p(\Omega)}, \quad \text{etc.} \end{aligned}$$

Since convergence of sequences and boundedness of norms do not depend on which equivalent norm is used, we will freely trade one norm for an equivalent norm according to circumstance. However, we should be careful when we are considering convergence of a sequence $f_k \to f$ in $W^{m,p}(\Omega)$ that the equivalent norm does not depend on k, or at least that the constants demonstrating the equivalence do not depend on k.

In the special cases where $p = 2$ and $\Omega = \mathbb{R}^d$, equivalent norms can be developed in terms of Fourier transforms:

$$\|f\|^2_{W^{m,2}(\mathbb{R}^d)} = (2\pi)^{-d} \int_{\mathbb{R}^d} \left(1 + |\xi|^2\right)^m |\mathcal{F}f(\xi)|^2 \, d\xi \tag{A.25}$$

with $|\xi| = \left(\xi_1^2 + \xi_2^2 + \cdots + \xi_d^2\right)^{1/2}$. These spaces are often denoted by $H^m(\mathbb{R}^d)$. These are Hilbert spaces, which have the inner product

$$\langle f, g \rangle_{H^m(\mathbb{R}^d)} = \operatorname{Re} (2\pi)^{-d} \int_{\mathbb{R}^d} \left(1 + |\xi|^2\right)^m \overline{\mathcal{F}f(\xi)} \mathcal{F}g(\xi) \, d\xi.$$

The formula can be extended to allow m to be *any* real value, including negative values. While we cannot use the usual properties of Fourier transforms on a general domain $\Omega \subset \mathbb{R}^d$, we can define an equivalent norm for $H^m(\Omega) = W^{m,2}(\Omega)$ using this approach:

$$\|f\|_{H^m(\Omega)} = \inf \left\{ \|g\|_{H^m(\mathbb{R}^d)} \mid g|_\Omega = f \right\}.$$

An alternative approach to constructing a suitable norm for $W^{m,p}(\Omega)$ for $m \in \mathbb{R}$, $m \geq 0$ is to use the *Sobolev–Slobodetskiĭ norm*: for $s = m + \beta$, with m an integer, $0 \leq \beta < 1$, and $1 \leq p < \infty$,

$$\|f\|^p_{W^{m+\beta,p}(\Omega)} = \|f\|^p_{L^p(\Omega)} + \sum_{\alpha:|\alpha|=m} \int_\Omega \int_\Omega \frac{|D^\alpha f(y) - D^\alpha f(x)|^p}{|x-y|^{d+p\beta}}\, dx\, dy.$$

For $p = \infty$ we have

$$\|f\|_{W^{m+\beta,\infty}(\Omega)} = \sum_{\alpha:|\alpha|=m} \sup_{x,y\in\Omega} \frac{|D^\alpha f(y) - D^\alpha f(x)|}{|y-x|^\beta}.$$

There are a number of important relationships between different Sobolev spaces and other well-known spaces. The space of Lipschitz functions on Ω is simply $W^{1,\infty}(\Omega)$; the space of *Hölder continuous* functions of exponent $0 < \beta < 1$ where $|f(y) - f(x)| \leq \text{const}\, |y-x|^\beta$ for all $x, y \in \Omega$ is $W^{\beta,\infty}(\Omega)$.

Duality of Sobolev spaces is straightforward for $H^m(\Omega)$: we can identify $H^m(\Omega)'$ with $H^{-m}(\Omega)$. An explicit identification can be done using Fourier transforms for $\Omega = \mathbb{R}^d$.

Often we can show that $W^{r,p}(\Omega)$ can be imbedded in $W^{s,t}(\Omega)$ with $\Omega \subset \mathbb{R}^d$. Usually we assume that the domain Ω has a *Lipschitz boundary*; that is, in a neighborhood of any point on the boundary, the boundary can be represented locally as the graph of a Lipschitz function. Some results require that the domain have a smooth boundary, where the boundary can be represented locally as the graph of a smooth function. The imbedding map is often compact (mapping bounded sets to sets whose closure is compact), which is extremely useful in proving existence results. The theorem that shows most of these connections is the standard *Sobolev imbedding theorem*.

Theorem A.8. *Suppose that Ω is a bounded domain in $\mathbb{R}^d$ with a smooth boundary; then the following imbeddings are compact:*

- $W^{m+\ell,p}(\Omega) \to W^{m,p}(\Omega)$ *for* $\ell > 0$, $m \geq 0$, *and* $1 \leq p \leq \infty$;
- $W^{m+\ell,p}(\Omega) \to W^{m,r}(\Omega)$ *for* $1 \leq r < dp/(d-\ell p)$, *provided* $\ell p < d$ *and* $\ell \geq 0$ *is an integer;*
- $W^{m+\ell,p}(\Omega) \to W^{m,p}(\Omega)$ *for* $\ell p = d$, *and* $1 \leq p < \infty$, *and* $\ell \geq 0$ *is an integer;*
- $W^{m+\ell,p}(\Omega) \to C^m(\Omega)$ *for* $\ell p > d$, *and* $\ell \geq 0$ *is an integer.*

Differentiation has a straightforward effect on Sobolev spaces: provided $m \geq |\alpha|$, the partial derivative operator $D^\alpha : W^{m,p}(\Omega) \to W^{m-|\alpha|,p}(\Omega)$ is continuous. In fact, for *any* real m, $D^\alpha : H^m(\Omega) \to H^{m-|\alpha|}(\Omega)$ is continuous.

There are also results for what happens when we restrict the domain of the functions to subsets of $\overline{\Omega}$. First, if $\Omega' \subset \Omega$ and Ω' is also a domain in $\mathbb{R}^d$ so that it has strictly positive volume, then the imbeddings $W^{m,p}(\Omega) \to W^{m,p}(\Omega')$ are continuous (but not compact). The really interesting part is when we restrict ourselves to a subset $\Gamma \subset \Omega$ which has a different dimension. This is most used in dealing with $\Gamma = \partial\Omega$, the boundary of the domain. It can seem rather odd that this is possible since a function in $W^{m,p}(\Omega)$ is in general not even

defined on $\partial\Omega$. Nevertheless, we can consider limits of values to define $u(x)$ for almost all $x \in \partial\Omega$ where $u \in W^{m,p}(\Omega)$, provided m and p are sufficiently large. These results are called *trace theorems*. The most useful example of these results is as follows.

Theorem A.9. *Suppose that Ω is a domain in $\mathbb{R}^d$ with smooth boundary; then the trace operator (extending restriction of the domain) $\gamma\colon W^{m,p}(\Omega) \to W^{m-1/p,p}(\partial\Omega)$ is continuous and surjective for $m > 1/p$. Furthermore, for $p = 2$ there is an extension operator $\rho\colon W^{m-1/p,p}(\partial\Omega) \to W^{m,p}(\Omega)$ so that $\gamma \circ \rho$ is the identity operator on $W^{m-1/p,p}(\partial\Omega)$. If $m = 1$, then the boundary need only be Lipschitz.*

Note that the trace operator on $H^m(\Omega)$ is $\gamma\colon H^m(\Omega) \to H^{m-1/2}(\partial\Omega)$ for $m > 1/2$.

Modifications of the basic Sobolev spaces are typically defined and used as needed. A common example is the Sobolev space of functions zero on the boundary $\partial\Omega$ of the given domain Ω. This is important for partial differential equations such as $\nabla^2 u = 0$ in Ω with boundary conditions $u = g$ on $\partial\Omega$. If $g \in H^{1/2}(\Omega)$, then we can use the extension operator $H^{1/2}(\partial\Omega) \to H^1(\Omega)$ to obtain $\widetilde{g} \in H^1(\Omega)$, where $\widetilde{g}$ on $\partial\Omega$ is equal to g. Let $w = u - \widetilde{g}$. Then $\nabla^2 w = -\nabla^2 \widetilde{g} \in H^{-1}(\Omega) = H^1(\Omega)'$. But w on $\partial\Omega$ is $g - g = 0$, so $w \in H_0^1(\Omega) = \{ z \in H^1(\Omega) \mid \gamma u = 0 \}$, where $\gamma\colon H^1(\Omega) \to H^{1/2}(\partial\Omega)$ is the trace operator. Now $H_0^1(\Omega)$ is a Banach space in its own right, and in fact a Hilbert space, since it is a closed subspace of $H^1(\Omega)$. Since $-\nabla^2 w$ is a distributional derivative, for any smooth function ϕ zero on $\partial\Omega$, we can use integration by parts:

$$\int_\Omega \phi \left[-\nabla^2 w \right] dx = \int_\Omega \nabla\phi \cdot \nabla w \, dx,$$

which is defined and continuous in ϕ also for any $\phi \in H_0^1(\Omega)$. In fact, finding w solving $\nabla^2 w = -\nabla^2 \widetilde{g}$ is equivalent to minimizing

$$\frac{1}{2} \int_\Omega |\nabla w|^2 \, dx + \left\langle -\nabla^2 \widetilde{g}, w \right\rangle_{H^{-1} \times H^1}$$

over $w \in H_0^1(\Omega)$. This function of w is closed, lower semicontinuous, and finite for any $w \in H_0^1(\Omega)$ and, as we shall see in the next section, has a minimizer. This gives the solution $w \in H_0^1(\Omega)$ for $\nabla^2 w = -\nabla^2 \widetilde{g}$, and so $u = w + \widetilde{g}$ solves the Dirichlet problem with $u \in H^1(\Omega)$.

A.6 Principles of nonlinear analysis

There are a number of principles of nonlinear analysis that are applicable to a wide range of situations. Amongst these are fixed point theorems and variational principles. The context for these is typically Banach spaces, although they can often be applied to more general situations. Nonlinear analysis is a very broad subject, but there are some excellent works on the topic, such as Aubin and Ekeland [20] and Zeidler [275]. Zeidler also has a four-volume treatment of nonlinear functional analysis which goes into many more topics in considerable depth.

Fixed point theorems are theorems of the following form: if $f\colon A \to A$ is a function with certain properties (for example, f is continuous) and A has certain properties (for

example, A is a bounded closed convex set in $\mathbb{R}^n$), then there is a point $x^* \in A$ where $f(x^*) = x^*$; that is, $x^*$is a *fixed point* of f.

The first is due to Banach and is known as the *contraction mapping theorem*.

Proposition A.10. *Suppose that $f: X \to X$, where X is a complete metric space, and f is a contraction; that is, there is an $\alpha < 1$ such that $d(f(x), f(y)) \le \alpha\, f(x, y)$ for all $x, y \in X$. Then f has a unique fixed point $x^* \in X$ and for any $x_0 \in X$, the sequence $x_{k+1} = f(x_k)$ converges to x^*.*

This theorem does not require any finite-dimensionality or compactness condition, which makes it particularly useful when other (apparently more powerful) theorems do not apply. On the other hand, if we are dealing with finite-dimensional spaces, then we can use continuity alone for f. The most celebrated such result is the *Brouwer fixed point theorem*.

Proposition A.11. *Suppose $f: A \to A$, where A is a closed convex set in $\mathbb{R}^n$, is continuous. Then f has a fixed point in A.*

There are many sources as well as proofs of this theorem. Often it is given as an easy consequence of homology theory in algebraic topology via the no-retraction theorem of balls to spheres; see, for example, [215, pp. 3–5] or [232, pp. 193–194]. However, there are many proofs that do not use algebraic topology, and even fairly easy proofs. A good source for a number of these is [106].

A straightforward corollary of Brouwer's theorem is that weakly coercive functions (that is, functions where $\lim_{\|x\|\to\infty} \langle f(x), x\rangle = +\infty$) that map $\mathbb{R}^n \to \mathbb{R}^n$ and are continuous have zeros.

Corollary A.12. *Suppose that $f: \mathbb{R}^n \to \mathbb{R}^n$ is continuous and weakly coercive. Then there is an x where $f(x) = 0$.*

Proof. Choose $R > 0$ so that $\langle f(x), x\rangle > 0$ for all x with $\|x\| \ge R$. Let B be the unit ball in $\mathbb{R}^n$. Then set $g(x) = x - \alpha(x) f(x)$, where

$$\alpha(x) = \frac{\langle f(x), x\rangle + \sqrt{\langle f(x), x\rangle^2 + \|f(x)\|^2 \left(R^2 - \|x\|^2\right)}}{\|f(x)\|^2 + 1}.$$

Calculations show that $\|g(x)\| \le R$ for all x with $\|x\| \le R$, and that $\alpha(x) > 0$ for all x. Both $\alpha(x)$ and $g(x)$ are continuous in x. Then by Brouwer's theorem there is a fixed point $x^* \in R\overline{B}$: $g(x^*) = x^*$. That is, $x^* - \alpha(x^*) f(x^*) = x^*$, from which we observe that $f(x^*) = 0$, as desired. □

Because only continuity is required for this theorem, Brouwer's fixed point theorem is essentially a topological result; A does not need to be convex, but rather need only be topologically equivalent to a closed convex set in finite dimensions. This result can be extended to infinite-dimensional spaces, provided we add a compactness condition. This gives the *Leray–Schauder fixed point theorem*.

Proposition A.13. *Suppose $f: A \to A$, where A is a compact convex set in a Banach space X, is continuous. Then f has a fixed point in A.*

There is another generalization of Brouwer's theorem due to Kakutani (see [106]) for convex set-valued functions that are upper semicontinuous (that is, for every $x \in A$ and $\epsilon > 0$ there is a $\delta > 0$ such that $\Phi(y) \subseteq \Phi(x) + \epsilon B$ whenever $\|y - x\| < \delta$).

Proposition A.14. *Suppose $\Phi\colon A \to \mathcal{P}(A)$, where A is a compact convex set in a Banach space X, is upper semicontinuous, where $\Phi(x)$ is a nonempty closed convex set for each $x \in A$. Then Φ has a fixed point in the sense that there is an $x^* \in A$ where $x^* \in \Phi(x^*)$.*

The proof of Kakutani's theorem can be reduced to the Leray–Schauder or Brouwer's theorem by using piecewise linear approximations to Φ.

Other approaches to nonlinear analysis include minimax-type theorems, of which the Ky Fan theorem is perhaps the archetype.

Proposition A.15. *Let K be a compact convex subset of a Banach space, and suppose that $f\colon K \times K \to \mathbb{R}$ is lower semicontinuous in the first variable and concave in the second variable. Then there exists $x^* \in K$ such that* $\sup_{y \in K} f(x^*, y) \leq \sup_{y \in K} f(y, y)$.

Apart from the contraction mapping theorem, these theorems all rely on compactness in some way or another in order to obtain existence of certain points (zeros, fixed points, or solutions of certain inequalities). Without compactness, there are a few alternatives that have been widely successful. One is to use variational methods, where we seek the minimum of a certain function. Convex analysis (see the following chapter) provides much of the theory for this approach, although extensions to nonsmooth nonconvex functions are an active area of interest for many researchers. Another approach is to use monotonicity or pseudomonotonicity to obtain existence of solutions. For more details on these approaches, see Sections 2.5 and 4.2.

Appendix B

Convex and Nonsmooth Analysis

B.1 Convex sets and functions

For a Banach space X, a function $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is *convex function* if for all $x, y \in X$ and $0 \le \theta \le 1$,

$$\phi(\theta x + (1-\theta)y) \le \theta\,\phi(x) + (1-\theta)\,\phi(y). \tag{B.1}$$

Note that we take $0 \cdot \infty = 0$, $r \cdot \infty = \infty$ if $r > 0$, and $r < \infty$ for any $r \in \mathbb{R}$. The *domain* of a convex function is $\operatorname{dom}\phi = \{x \mid \phi(x) < \infty\}$. A set $C \subseteq X$ is a *convex set* if whenever $x, y \in C$ and $0 \le \theta \le 1$ we have $\theta x + (1-\theta)y \in C$. For a given convex set C there is the indicator function

$$I_C(x) = \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{if } x \notin C. \end{cases} \tag{B.2}$$

If C is convex, so is I_C; if C is closed, then I_C is lower semicontinuous; if $C \neq \emptyset$, then I_C is proper.

A function $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is convex if and only if the *epigraph*

$$\operatorname{epi}\phi = \{(x,r) \in X \times \mathbb{R} \mid r \ge \phi(x)\} \tag{B.3}$$

is a convex set in $X \times \mathbb{R}$. If ϕ is a *lower semicontinuous function*, that is, $x_k \to x$ as $k \to \infty$ implies $\liminf_{k\to\infty} \phi(x_k) \ge \phi(x)$, then $\operatorname{epi}\phi$ is a closed set. If ϕ is *proper*, that is, $\phi(x_0) < \infty$ for some $x_0 \in X$, then $\operatorname{epi}\phi$ and $\operatorname{dom}\phi$ are nonempty.

A powerful tool for studying closed convex sets is the *separating hyperplane theorem*: if K is a closed convex set in a Banach space X, and $z \notin K$, then there are a $y \in X'$ and $\beta \in \mathbb{R}$ such that

$$\begin{aligned} &0 > \langle y, z\rangle + \beta, \\ &0 \le \langle y, x\rangle + \beta \qquad \text{for all } x \in K. \end{aligned}$$

That is, the hyperplane $\{x \in X \mid 0 = \langle y, x\rangle + \beta\}$ separates K from z. We will use this as a basic theorem, although it can be proved in terms of the Hahn–Banach theorem.

A convex set C is called *solid* if $\operatorname{int} C \neq \emptyset$. Note that for a solid convex set, $C = \overline{\operatorname{int} C}$.

A *cone* $C \subseteq X$ is a set where $x \in C$ and $\alpha \geq 0$ implies $\alpha x \in C$. A convex cone is a set that is both convex and a cone. This can be expressed by the following conditions: if $x, y \in C$ and $\alpha \geq 0$, then $\alpha x \in C$ and $x + y \in C$. Properties of convex cones are discussed in Section B.1.3.

B.1.1 Support functions

An additional topic that often arises is that of *support functions* of convex sets. The support function of a convex set $K \subseteq X$, with X a Banach space, is the function $\sigma_K : X' \to \mathbb{R} \cup \{\infty\}$ given by

$$\sigma_K(\xi) = \sup_{x \in K} \langle \xi, x \rangle . \tag{B.4}$$

These are duals of the corresponding indicator functions: $\sigma_K = I_K^*$. (Dual convex functions are defined in Section B.2.) Support functions are also *not* to be confused with the support *of* a function.

Lemma B.1. *Support functions have the following properties:*

1. *σ_K is a convex lower semicontinuous positively homogeneous function;*
2. *$\overline{K} = \left\{ x \mid \langle \xi, x \rangle \leq \sigma_K(\xi) \text{ for all } \xi \in X' \right\}$;*
3. *if K is also a cone, then $K^\circ = -K^* = \operatorname{dom} \sigma_K := \{\xi \mid \sigma_K(\xi) < \infty\}$.*

Proof.

1. To show that σ_K is convex, note that it is the supremum of a family of linear (and therefore convex) functions.

   To show that σ_K is lower semicontinuous, suppose that $\xi_\ell \to \xi$ in X' as $\ell \to \infty$. For any $\epsilon > 0$ there is an $x \in K$ such that $\langle \xi, x \rangle + \epsilon \geq \sigma_K(\xi) \geq \langle \xi, x \rangle$. Then $\liminf_{\ell \to \infty} \sigma_K(\xi_\ell) \geq \liminf_{\ell \to \infty} \langle \xi_\ell, x \rangle = \langle \xi, x \rangle \geq \sigma_K(\xi) - \epsilon$. Since this is true for all $\epsilon > 0$, we see that $\liminf_{\ell \to \infty} \sigma_K(\xi_\ell) \geq \sigma_K(\xi)$, and that σ_K is lower semicontinuous.

   To show that σ_K is positively homogeneous, note that for $\alpha \geq 0$,

   $$\sigma_K(\alpha \xi) = \sup_{x \in K} \langle \alpha \xi, x \rangle = \sup_{x \in K} \alpha \langle \xi, x \rangle = \alpha \sup_{x \in K} \langle \xi, x \rangle = \alpha \sigma_K(\xi).$$

2. It is easy to show that $\sigma_K = \sigma_{\overline{K}}$, so we assume at the outset that K is closed. Then $z \in K$ implies that $\langle \xi, z \rangle \leq \sup_{x \in K} \langle \xi, x \rangle = \sigma_K(\xi)$. This shows that $K \subseteq \left\{ x \mid \langle \xi, x \rangle \leq \sigma_K(\xi) \text{ for all } \xi \in X' \right\}$. Suppose $z \notin K$. Then there is a separating hyperplane: $\zeta \in X'$ and $\alpha \in \mathbb{R}$ where

   $$\begin{aligned} &\langle \zeta, z \rangle - \alpha > 0, \\ &\langle \zeta, x \rangle - \alpha \leq 0 \qquad \text{for all } x \in K. \end{aligned}$$

   Taking the supremum over $x \in K$, we see that $\alpha \geq \sigma_K(\zeta)$. Thus $\sigma_K(\zeta) \leq \alpha < \langle \zeta, z \rangle$. Thus the reverse inclusion holds, and the two sets are equal, as desired.

3. Suppose that K is a convex cone. Then, if $\xi \in K^\circ$, we have $\langle \xi, x\rangle \le 0$ for all x, and so $\sigma_K(\xi) = \sup_{x\in K}\langle \xi, x\rangle \le 0$. In fact, since $0 \in K$, we have $\sigma_K(\xi) = 0$ for all $\xi \in K^\circ$, so $K^\circ \subseteq \mathrm{dom}\,\sigma_K$. Conversely, suppose that $\xi \in \mathrm{dom}\,\sigma_K$. If $\langle \xi, x\rangle > 0$ for any $x \in K$, noting that $\alpha x \in K$ as well, we see that

$$\sigma_K(\xi) \ge \sup_{\alpha\ge 0}\langle \xi, \alpha x\rangle = +\infty,$$

contradicting our assumption that $\xi \in \mathrm{dom}\,\sigma_K$. Thus $\mathrm{dom}\,\sigma_K \subseteq K^\circ$. Hence the two sets are equal. □

B.1.2 Convex projections in Hilbert spaces

Here we discuss properties of the convex projection (or "nearest point map") Π_K for a closed convex set K.

Lemma B.2. *If X is a Hilbert space and K is a closed convex set, then $\Pi_K : X \to K$, where $\Pi_K(x)$ is the nearest point in K to x, is a well-defined Lipschitz continuous function with Lipschitz constant one. The function Π_K is characterized by the property that*

$$(x - \Pi_K(x), z - \Pi_K(x)) \le 0 \qquad \textit{for all } z \in K. \tag{B.5}$$

Furthermore, Π_K is a monotone function.

Proof. First we show that there is a nearest point in K to x. Suppose $z_m \in K$ is an infimizing sequence so that $\|z_m - x\| \to \inf_{z\in K}\|z - x\|$ as $m \to \infty$. If we set $R = \inf_{z\in K}\|z - x\| + 1$, for m sufficiently large, $\|z_m - x\| \le R$. Thus the sequence z_m, $m = 1, 2, \ldots$, is bounded. By Alaoglu's theorem there is a weak* convergent subsequence (also denoted by z_m) with weak* limit z^*. Since X is a Hilbert space, the subsequence is also weakly convergent. Now $z^* \in K$. If this were not true, by the separating hyperplane theorem there would be $w \in X'$ and $\beta \in \mathbb{R}$ such that $\langle w, z\rangle + \beta \ge 0$ for all $z \in K$ but $\langle w, z^*\rangle + \beta < 0$. Thus $\langle w, z_m\rangle + \beta \ge 0$ for all m, and since $z_m \rightharpoonup z^*$ weakly we have $\langle w, z^*\rangle + \beta = \lim_{m\to\infty}\langle w, z_m\rangle + \beta \ge 0$, contradicting $\langle w, z^*\rangle + \beta < 0$. Thus $z^* \in K$.

Now we show that there is only *one* nearest point in K to x. Suppose there were more: z_1 and z_2. Then, as K is convex, $\theta z_1 + (1-\theta)z_2 \in K$ for any $0 \le \theta \le 1$. Now, as X is a Hilbert space,

$$\begin{aligned}\|x - (\theta z_1 + (1-\theta)z_2)\|^2 &= \|x - z_2 - \theta(z_1 - z_2)\|^2 \\ &= \|x - z_2\|^2 - 2\theta(x - z_2, z_1 - z_2) + \theta^2\|z_1 - z_2\|^2.\end{aligned}$$

Taking derivatives with respect to θ at $\theta = 0$, we see that

$$(x - z_2, z_1 - z_2) \le 0$$

since z_2 is a nearest point. We get a similar result for z_1: $(x - z_1, z_2 - z_1) \le 0$ after swapping the roles of z_1 and z_2. Adding these last two inequalities we get

$$\|z_1 - z_2\|^2 = (z_1 - z_2, z_1 - z_2) \le 0,$$

which can occur only if $z_1 = z_2$. Thus $\Pi_K(x)$ is well defined.

Repeating the argument of the previous paragraph with $z_2 = z^*$ (which is the nearest point) and $z_1 = z$ a given point in K, we see that $(x - z^*, z - z^*) \le 0$. That is,

$$(x - \Pi_K(x), z - \Pi_K(x)) \le 0 \qquad \text{for all } z \in K.$$

Note that the conditions $w \in K$ and $(x - w, y - w) \le 0$ for all $y \in K$ are sufficient to imply that $w = \Pi_K(x)$ in a Hilbert space. To see this, note that

$$\begin{aligned} \|x - (w + \theta(y - w))\|^2 &= \|x - w\|^2 - 2\theta(x - w, y - w) + \theta^2 \|y - w\|^2 \\ &\ge \|x - w\|^2 - 2\theta(x - w, y - w) \ge \|x - w\|^2 \end{aligned} \tag{B.6}$$

for $0 \le \theta \le 1$, and in particular, $\|x - y\| \ge \|x - w\|$ for all $y \in K$. Thus $w = \Pi_K(x)$.

From (B.6) we also have $(x - \Pi_K(x), \Pi_K(y) - \Pi_K(x)) \le 0$ for all x, y. Swapping the roles of x and y we get $\langle y - \Pi_K(y), \Pi_K(x) - \Pi_K(y)\rangle \le 0$. Adding these inequalities we get

$$(x - y - \Pi_K(x) + \Pi_K(y), \Pi_K(y) - \Pi_K(x)) \le 0,$$

so

$$\begin{aligned} 0 \le (\Pi_K(y) - \Pi_K(x), \Pi_K(y) - \Pi_K(x)) &\le (y - x, \Pi_K(y) - \Pi_K(x)); \\ \text{and thus} \qquad \|\Pi_K(y) - \Pi_K(x)\|^2 &\le \|y - x\| \, \|\Pi_K(y) - \Pi_K(x)\|. \end{aligned}$$

The second to last inequality implies that Π_K is a monotone function. The last inequality shows (after division by $\|\Pi_K(y) - \Pi_K(x)\|$) that Π_K is Lipschitz with Lipschitz constant one. □

B.1.3 Convex cones

Convex cones play a particularly important role in our theory: a set K is a *cone* if $x \in K$ and $\alpha \ge 0$ imply that $\alpha x \in K$. In the next section we will meet some important cones such as the tangent and normal cones. For now we will look at some of the important cones and properties of cones.

One of the most common cones is the cone of nonnegative vectors, or the nonnegative orthant:

$$\mathbb{R}^n_+ = \left\{ \mathbf{x} \in \mathbb{R}^n \mid x_i \ge 0 \text{ for all } i \right\}.$$

Another is the cone of nonnegative functions on a set Ω:

$$\left\{ f \in L^2(\Omega) \mid f(x) \ge 0 \text{ for all } x \in \Omega \right\}.$$

Half-spaces are also cones,

$$H_n := \mathbb{R}_+ \times \mathbb{R}^{n-1},$$

as is the full-space $\mathbb{R}^n$. Clearly these are different kinds of cones. For example, a *pointed cone* K is a cone where

$$K \cap (-K) = \{0\}. \tag{B.7}$$

The half-space H_n is *not* a pointed cone, but $\mathbb{R}^n_+$ is pointed. Note that for any closed convex cone, $V := K \cap (-K)$ must be a vector space. If we find a complementary subspace W so that $V + W = X$ and $V \cap W = \{0\}$, then $(K \cap W) \cap (-K \cap W) = \{0\}$, so that $K \cap W$ is a pointed cone.

A closed convex cone K has a *dual cone* K^* given by

$$K^* = \left\{ w \in X' \mid \langle w, z \rangle \geq 0 \text{ for all } z \in K \right\}. \tag{B.8}$$

Closely related is the *polar cone* to a cone K, which is $K^\circ = -K^*$. If X is a reflexive Banach space (such as a Hilbert space or $\mathbb{R}^n$), then we can identify the space X with its second dual X''. Under this identification $K^{**} = K$, or equivalently $K^{\circ\circ} = K$, whenever K is a closed convex cone.

Topology can interact with convexity in important ways. For example, a cone K is *solid* if it has nonempty interior; that is, there are an $x \in K$ and an $r > 0$ where $x + r\,B \subset K$, with B the unit ball in X. A solid closed convex cone K is the closure of its interior. In $\mathbb{R}^n$, any closed convex cone K is solid if and only if K^* is pointed. However, this is not true in infinite dimensions.

Consider, for example,

$$K = \left\{ \mathbf{x} \in \ell^2 \mid x_i \geq 0 \text{ for all } i = 1, 2, 3, \ldots \right\}.$$

Identifying ℓ^2 with its dual space $\left(\ell^2\right)'$, which we can do since it is a Hilbert space, we have $K = K^*$. That is, K is a *self-dual cone*. It is also clear that K is a pointed cone, as $\mathbf{x} = -\mathbf{y}$, with $x_i, y_i \geq 0$ for all i, can occur only if $x_i = y_i = 0$. Yet K does not contain any open ball. Let $\mathbf{x} \in K$, and take $r > 0$. We can find a $\mathbf{y} \in \mathbf{x} + r\,B \notin K$. Since $\mathbf{x} \in \ell^2$, we have from the definition of ℓ^2, $\sum_i x_i^2 < +\infty$; thus $x_i \to 0$ as $i \to \infty$. Choose i, where $|x_i| < r/2$. Let $\mathbf{e}_i$ be the ith unit basis vector in ℓ^2. Then $\mathbf{y} = \mathbf{x} - r\,\mathbf{e}_i/2 \in \mathbf{x} + r\,B$ but $y_i = x_i - r/2 < 0$, so $\mathbf{y} \notin K$.

Sometimes we need a stronger concept than pointedness in infinite-dimensional problems: We say that a closed convex cone K is *strongly pointed* if K^* is a solid cone; that is, it has nonempty interior. The most important property of strongly pointed cones is that norms can be bounded below and above in terms of inner products or duality pairings.

Lemma B.3. *If K is a strongly pointed cone in a Banach space X, then there is a $\nu \in K^*$ where*

$$\|x\|_X \leq \langle \nu, x \rangle \leq \|\nu\|_{X'}\, \|x\|_X \qquad \textit{for all } x \in K.$$

Proof. The second inequality is clear. Since K is strongly pointed, K^* is a solid cone. Pick μ in the interior of K^*. Then there is a $\delta > 0$ such that $\mu + \delta\, B_{X'} \subset K^*$. This means that for every $x \in K$ and $\xi \in X'$ with $\|\xi\|_{X'} < \delta$, we have $\langle \mu + \xi, x \rangle \geq 0$. Taking the infimum over all $\xi \in X'$ with $\|\xi\|_{X'} < \delta$ gives $\langle \mu, x \rangle - \delta\, \|x\|_X \geq 0$. Setting $\nu = \mu/\delta$ and rearranging give the desired result. □

In finite dimensions there is no difference between a cone being pointed and being strongly pointed.

Lemma B.4. *If $K \subseteq \mathbb{R}^n$ is a closed convex cone, then it is pointed if and only if it is strongly pointed.*

Proof. ($\Rightarrow$) Suppose that K is pointed, so that $K \cap (-K) = \{0\}$. Then $0 \notin \operatorname{co}(K \cap S)$, where S is the unit sphere in $\mathbb{R}^n$: $S = \{x \in \mathbb{R}^n \mid \|x\| = 1\}$. To see this, suppose $0 = \sum_{i=1}^m \theta_i x_i$, where $\theta_i \geq 0$ for all i, $\sum_{i=1}^m \theta_i = 1$, and $x_i \in K \cap S$. Choose j, where $\theta_j \neq 0$; then $0 \neq \theta_j x_j \in K$ and $-\theta_j x_j = \sum_{i:i\neq j} \theta_i x_i \in K$. That is, $0 \neq \theta_j x_j \in K \cap (-K)$, contradicting pointedness of K. So we see that $0 \notin \operatorname{co}(K \cap S)$. Since $K \cap S$ is a closed and bounded subset of $\mathbb{R}^n$, it is compact. By Carathéodory's theorem on convex sets in $\mathbb{R}^n$, every element of $\operatorname{co}(K \cap S)$ can be written as $\sum_{i=1}^{n+1} \theta_i x_i$, where $\theta_i \geq 0$ for all i, $\sum_{i=1}^{n+1} \theta_i = 1$, and $x_i \in K \cap S$. That is, $\operatorname{co}(K \cap S)$ is the image under a continuous map of the Cartesian product $\Sigma_{n+1} \times (K \cap S)^{n+1}$; hence $\operatorname{co}(K \cap S)$ is compact and also closed. By the separating hyperplane theorem, there is a $\nu \in \mathbb{R}^n$ and an α where $\langle \nu, x\rangle - \alpha \geq 0$ for all $x \in K \cap S$, and $\langle \nu, 0\rangle - \alpha < 0$. Thus $\alpha > 0$. For any $0 \neq y \in K$, we have $\langle \nu, y\rangle = \|y\| \langle \nu, y/\|y\|\rangle \geq \alpha \|y\| > 0$. This shows that $\nu \in K^*$. Suppose that $\mu \in X'$ and $\|\mu\|_{X'} < \alpha$. Then $\nu + \mu \in K^*$ since for any $0 \neq y \in K$ we have $\langle \nu + \mu, y\rangle \geq \alpha \|y\| - \|\mu\|_{X'} \|y\| = (\alpha - \|\mu\|) \|y\| \geq 0$. Thus ν lies in the interior of K^*, and so K is strongly pointed.

($\Leftarrow$) Suppose that K is strongly pointed. By Lemma B.3, there is $\nu \in K^*$ such that $\langle \nu, x\rangle \geq \|x\|$ for all $x \in K$. If $x \in K \cap (-K)$, then both $x, -x \in K$, and so $\|x\| \leq \langle \nu, x\rangle$ and $\|x\| \leq -\langle \nu, x\rangle$; that is, $\langle \nu, x\rangle \leq -\|x\| \leq \|x\| \leq \langle \nu, x\rangle$, which can be true only if $x = 0$. That is, $K \cap (-K) = \{0\}$, and K is a pointed cone. □

Note that the dual of the dual cone K^{**} is the original cone K, provided K is a closed convex cone in a reflexive space X that can be identified with X''. The proof is an exercise in using the separating hyperplane theorem. First we show that K^* is a closed convex cone.

Lemma B.5. *If $K \subseteq X$, with X a Banach space, then K^* is a closed convex cone.*

Proof.
***K^** is closed:** Suppose that $y_n \in K^*$ for all n and that $y_n \to y$ as $n \to \infty$. Then, for every $x \in K$, $\langle y_n, x\rangle \geq 0$. Taking the limit as $n \to \infty$ we see that $\langle y, x\rangle \geq 0$ for every $x \in K$. Thus $y \in K^*$ as well. So K^* is closed.

***K^** is convex:** Suppose $y_1, y_2 \in K^*$ and $0 \leq \theta \leq 1$. Note that $\langle y_1, x\rangle$ and $\langle y_2, x\rangle \geq 0$ for every $x \in K$. So

$$\langle \theta y_1 + (1-\theta) y_2, x\rangle = \theta \langle y_1, x\rangle + (1-\theta)\langle y_2, x\rangle \geq 0.$$

Since this is true for all $x \in K$, $\theta y_1 + (1-\theta) y_2 \in K^*$ and K^* is convex.

***K^** is a cone:** Suppose that $y \in K^*$ and $\alpha \geq 0$. Then, for every $x \in K$,

$$\langle \alpha y, x\rangle = \alpha \langle y, x\rangle \geq 0,$$

so $\alpha y \in K^*$. □

Now we can go on to show that $K^{**} = K$. Since $K^{**} \subseteq X''$ we need to identify X with X'' using the *natural map* $\natural$: $X \to X''$ (see (2.1)). Identifying X with X'' means that we identify $\natural(x)$ with x.

Theorem B.6. *If K is a closed convex cone in a reflexive Banach space X, with X identified with X'', then $K^{**} = K$. (If we do not identify X with X'', then we have $\natural(K) = K^{**}$.)*

Proof. First we show that $K \subseteq K^{**}$. Suppose that $x \in K$. Then, for any $\eta \in K^*$, $\langle x, \eta\rangle_{X\times X'} = \langle \eta, x\rangle_{X'\times X''} \geq 0$ (identifying X with X'' via the natural map). So $x \in (K^*)^* = K^{**}$. Thus $K \subseteq K^{**}$.

Now we show that $K^{**} \subseteq K$. If this were not true, then there would be a $w \in K^{**}\backslash K$. Since K is a closed convex set, by the separating hyperplane theorem there are $\eta \in X'$ and $\beta \in \mathbb{R}$ such that $\langle \eta, x\rangle + \beta \geq 0$ for every $x \in K$, but $\langle \eta, w\rangle + \beta < 0$. Since K is a cone, we can set $x = 0$, so $\beta \geq 0$. Thus $\langle \eta, w\rangle \leq \langle \eta, w\rangle + \beta < 0$. Also, since K is a cone, if $x \in K$, so is $\alpha x \in K$ for any $\alpha \geq 0$. Hence $0 \leq \langle \eta, \alpha x\rangle + \beta = \alpha(\langle \eta, x\rangle + \beta/\alpha)$, provided $\alpha > 0$, so $\langle \eta, x\rangle + \beta/\alpha \geq 0$ for any $\alpha > 0$. Taking $\alpha \to \infty$ gives $\langle \eta, x\rangle \geq 0$ for any $x \in K$. Thus $\eta \in K^*$. But as $w \in K^{**}$, $\langle w, \eta\rangle \geq 0$, contradicting our above result that $\langle w, y\rangle < 0$. Thus our assumption that $w \in K^{**}\backslash K$ is false. This shows that $K^{**} \subseteq K$.

Combining the two inclusions gives $K = K^{**}$. □

Vectors in a Hilbert space can be "split" into K and K° much like vectors can be split into components parallel and orthogonal to a vector space [176].

Lemma B.7 (Moreau). *Let K be a closed convex cone in a Hilbert space X with X' identified with X. For all $x \in X$,*

$$x = \Pi_K(x) + \Pi_{K^\circ}(x). \tag{B.9}$$

Furthermore, $\langle \Pi_K(x), \Pi_{K^\circ}(x)\rangle = 0$.

Proof. From (B.5), $(x - \Pi_K(x), z - \Pi_K(x))_X \leq 0$ for all $z \in K$. Let $w = x - \Pi_K(x)$. First we take $z = 0$ so that $(w, -\Pi_K(x))_X \leq 0$; then we take $z = 2\Pi_K(x)$ so that $(w, \Pi_K(x))_X \leq 0$; thus $(w, \Pi_K(x))_X = 0$. Now, to show that $w = \Pi_{K^\circ}(x)$, we need $0 \geq (x - w, u - w)_X = (\Pi_K(x), u - w)_X$ for all $u \in K^\circ$. But $(\Pi_K(x), u)_X \leq 0$ because $u \in K^\circ$, and $(\Pi_K(x), w)_X = 0$, so the inequality holds, as desired. □

For some cones $K \subseteq \mathbb{R}^n$, $K^* = K$. These are *self-dual* cones. The simplest example is $K = \mathbb{R}_+$, the nonnegative real numbers. Since $(K_1 \times K_2)^* = K_1^* \times K_2^*$, the Cartesian product of self-dual cones is self-dual; in particular, the nonnegative orthant consisting of componentwise nonnegative vectors $\mathbb{R}^n_+$ is self-dual. Thus GCPs are truly a generalization of CPs: just use $K = \mathbb{R}^n_+$ in a GCP to get the corresponding CP.

Here are some structural properties of dual cones.

Lemma B.8. *Suppose K, $K_1 \subseteq X$, and $K_2 \subseteq Y$ are closed convex cones. Then the following hold:*

1. $(K_1 \times K_2)^* = K_1^* \times K_2^*$.
2. *For $A\colon X \to Y$ linear, continuous, and invertible, $(AK)^* = \left(A^{-1}\right)^* K^*$, where $\left(A^{-1}\right)^* = (A^*)^{-1} : X' \to Y'$, is the adjoint of the inverse $A^{-1}\colon Y \to X$.*
3. *If $K \subseteq K_1$, then $K_1^* \subseteq K^*$.*
4. *If $X = Y$, then $(K_1 + K_2)^* = K_1^* \cap K_2^*$.*
5. *If $X = Y$, then $K_1^* + K_2^* \subseteq (K_1 \cap K_2)^* = \overline{K_1^* + K_2^*}$.*

Proof.

1. Note that

$$\begin{aligned}(K_1 \times K_2)^* &= \left\{ (\xi, \eta) \in X' \times Y' \mid \langle (\xi, \eta), (x, y) \rangle \geq 0 \text{ for all } (x, y) \in K_1 \times K_2 \right\} \\ &= \left\{ (\xi, \eta) \in X' \times Y' \mid \langle \xi, x \rangle + \langle \eta, y \rangle \geq 0 \text{ for all } x \in K_1,\ y \in K_2 \right\}.\end{aligned}$$

   Taking $x = 0$ shows that $\eta \in K_2^*$; taking $y = 0$ shows that $\xi \in K_1^*$. So $(K_1 \times K_2)^* \subseteq K_1^* \times K_2^*$. To show the reverse inclusion, if $\xi \in K_1^*$and $\eta \in K_2^*$, then for any $(x, y) \in K_1 \times K_2$, $\langle (\xi, \eta), (x, y) \rangle = \langle \xi, x \rangle + \langle \eta, y \rangle \geq 0$, so $K_1^* \times K_2^* \subseteq (K_1 \times K_2)^*$. Combining the two inclusions shows the two sets are equal.

2. Note that

$$\begin{aligned}(A\,K)^* &= \left\{ \eta \in Y' \mid \langle \eta, Az \rangle \geq 0 \text{ for all } z \in K \right\} \\ &= \left\{ \eta \in Y' \mid \left\langle A^* \eta, z \right\rangle \geq 0 \text{ for all } z \in K \right\} \\ &= \left\{ \left(A^*\right)^{-1} \xi \mid \langle \xi, z \rangle \geq 0 \ \text{ for all } z \in K \right\} \\ &= \left(A^*\right)^{-1} K^*.\end{aligned}$$

3. If $K \subset K_1$ and $\xi \in K_1^*$, then $\langle \xi, x \rangle \geq 0$ for all $x \in K_1$, so $\langle \xi, x \rangle \geq 0$ for all $x \in K$ and thus $\xi \in K^*$. Note that this result does not rely on either K or K_1 being closed.

4. We calculate

$$\begin{aligned}(K_1 + K_2)^* &= \left\{ \zeta \in X' \mid \langle \zeta, x + y \rangle \geq 0 \text{ for all } x \in K_1,\ y \in K_2 \right\} \\ &= \left\{ \zeta \in X' \mid \langle \zeta, x \rangle + \langle \zeta, y \rangle \geq 0 \text{ for all } x \in K_1,\ y \in K_2 \right\}.\end{aligned}$$

   Clearly $(K_1 + K_2)^* \supseteq K_1^* \cap K_2^*$. To show the reverse inequality, set $x = 0$, so that $\zeta \in (K_1 + K_2)^*$ implies $\langle \zeta, y \rangle \geq 0$ for all $y \in K_2$. Similarly, setting $y = 0$ gives $\langle \zeta, x \rangle \geq 0$ for all $x \in K_1$. Thus $z \in K_1^* \cap K_2^*$, and $(K_1 + K_2)^* \subseteq K_1^* \cap K_2^*$. Combining the two inclusions gives equality: $(K_1 + K_2)^* = K_1^* \cap K_2^*$.

5. Now we show that $K_1^* + K_2^* \subseteq (K_1 \cap K_2)^*$. Suppose that $\zeta = \xi + \eta$ with $\xi \in K_1^*$and $\eta \in K_2^*$. For any $w \in K_1 \cap K_2$, $\langle \zeta, w \rangle = \langle \xi + \eta, w \rangle = \langle \xi, w \rangle + \langle \eta, w \rangle \geq 0$; thus $\eta \in (K_1 \cap K_2)^*$.

   Finally, we show that $\overline{K_1^* + K_2^*} = (K_1 \cap K_2)^*$. If L is a convex cone, but not necessarily closed, then $L \subseteq \overline{L}$ so by item 3, $\overline{L}^* \subseteq L^*$, and again $L^{**} \subseteq \overline{L}^{**} = \overline{L}$ by Theorem B.6. Now $K_1^* + K_2^*$ is a convex cone, but not necessarily closed, as we will see later. Then

$$\begin{aligned}\overline{K_1^* + K_2^*} &= \left(K_1^* + K_2^*\right)^{**} = \left(K_1^{**} \cap K_2^{**}\right)^* \\ &= (K_1 \cap K_2)^{**}.\end{aligned}$$

   Thus $K_1^* + K_2^* \subseteq (K_1 \cap K_2)^* = \overline{K_1^* + K_2^*}$, as desired. □

Remark B.9. Note that we do not get equality in the last inclusion of Lemma B.8 (item 5). Here is an example in $\mathbb{R}^3$: Take

$$K_1 = \left\{ \alpha \mathbf{e}_1 + \mathbf{y} \mid \|\mathbf{y}\|_2 \le \alpha, \mathbf{e}_1^T \mathbf{y} = 0 \right\},$$
$$K_2 = \left\{ \alpha \mathbf{e}_2 + \mathbf{y} \mid \|\mathbf{y}\|_2 \le \alpha, \mathbf{e}_2^T \mathbf{y} = 0 \right\}.$$

Both of these cones are self-dual since they are both orthogonal transforms of the Lorentz cone L_2 (see (2.34)): $(Q\,L_2)^* = Q^{-T} L_2^* = Q\,L_2$.

Now $K_1 + K_2 = K_1^* + K_2^*$ is not a closed set, and so it cannot be a dual cone: the plane $\left\{ \mathbf{x} \mid (\mathbf{e}_1 + \mathbf{e}_2)^T \mathbf{x} = 0 \right\}$ is in $\overline{K_1 + K_2}$, but the only part of this plane in $K_1 + K_2$ is the line generated by $\mathbf{e}_1 - \mathbf{e}_2$: let $P = \{ x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 \mid x_1, x_2 \in \mathbb{R} \}$ be the $(\mathbf{e}_1, \mathbf{e}_2)$ plane. Then $K_1 \cap P = \{ x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 \mid |x_2| \le x_1 \}$ and $K_2 \cap P = \{ y_1 \mathbf{e}_1 + y_2 \mathbf{e}_2 \mid |y_1| \le y_2 \}$. Setting $0 \le x_1 = -x_2$ we have $x_1 (\mathbf{e}_1 - \mathbf{e}_2) \in K_1 \cap P$; setting $0 \le y_2 = -y_1$ we have $-y_2 (\mathbf{e}_1 - \mathbf{e}_2) \in K_2 \cap P$. Thus the entire line generated by $\mathbf{e}_1 - \mathbf{e}_2$ is in $K_1 \cup K_2 \subset K_1 + K_2$.

But now consider $\mathbf{z} = \rho (\mathbf{e}_1 - \mathbf{e}_2) + \sigma \mathbf{e}_3$, $\sigma \ne 0$. Suppose $\mathbf{z} = \mathbf{x} + \mathbf{y}$, with $\mathbf{x} \in K_1$ and $\mathbf{y} \in K_2$. Write $\mathbf{x} = \alpha \mathbf{e}_1 + \mathbf{u}$, $\mathbf{y} = \beta \mathbf{e}_2 + \mathbf{v}$ with $\mathbf{e}_1^T \mathbf{u} = 0$, $\mathbf{e}_2^T \mathbf{v} = 0$, $\|\mathbf{u}\| \le \alpha$, and $\|\mathbf{v}\| \le \beta$. Put $\mathbf{u} = u_2 \mathbf{e}_2 + u_3 \mathbf{e}_3$ and $\mathbf{v} = v_1 \mathbf{e}_1 + v_3 \mathbf{e}_3$; $u_2^2 + u_3^2 \le \alpha^2$ and $v_1^2 + v_3^2 \le \beta^2$. So $\rho (\mathbf{e}_1 - \mathbf{e}_2) = (\alpha + v_1) \mathbf{e}_1 + (\beta + u_2) \mathbf{e}_2$ and therefore $\rho = \alpha + v_1 = -\beta - u_2$. This implies that $\alpha + \beta = -v_1 - u_2$. Since $|v_1| \le \beta$ and $|u_2| \le \alpha$, this means that $|v_1| = \beta$ and $|u_2| = \alpha$; thus $v_3 = u_3 = 0$ and so $\sigma = 0$. Thus $\mathbf{z}$ is not in $K_1 + K_2$.

However, we want to show that $\mathbf{z} \in \overline{K_1 + K_2}$. For $\alpha > |\rho|$, put $\mathbf{x}(\alpha) = (\alpha + \rho/2) \mathbf{e}_1 - \sqrt{(\alpha + \rho/2)^2 - \sigma^2/4}\, \mathbf{e}_2 + \frac{1}{2} \sigma \mathbf{e}_3$ and $\mathbf{y}(\alpha) = (\alpha - \rho/2) \mathbf{e}_2 - \sqrt{(\alpha - \rho/2)^2 - \sigma^2/4}\, \mathbf{e}_1 + \frac{1}{2} \sigma \mathbf{e}_3$. Simple calculations show that $\mathbf{x}(\alpha) \in K_1$ and $\mathbf{y}(\alpha) \in K_2$, provided $\alpha \ge (|\rho| + |\sigma|)/2$. Since

$$\sqrt{(\alpha + \rho/2)^2 - \sigma^2/4} = \alpha + \rho/2 + \mathcal{O}(1/\alpha) \qquad \text{as } \alpha \to \infty,$$

we have

$$\mathbf{x}(\alpha) + \mathbf{y}(\alpha) = \big[(\alpha + \rho/2) - (\alpha - \rho/2) \big] \mathbf{e}_1 + \big[(\alpha - \rho/2) - (\alpha + \rho/2) \big] \mathbf{e}_2 + \sigma \mathbf{e}_3 + \mathcal{O}(1/\alpha).$$

Taking $\alpha \to \infty$ we see that $\mathbf{x}(\alpha) + \mathbf{y}(\alpha) \to \rho (\mathbf{e}_1 - \mathbf{e}_2) + \sigma \mathbf{e}_3$, so that $\mathbf{z} \in \overline{K_1 + K_2}$, as desired.

B.1.4 Tangent cones and normal cones

We can define the *tangent cone* to K at a point $x \in K$ as

$$T_K(x) = \left\{ \lim_{j \to \infty} \frac{x_j - x}{t_j} \mid x_j \in K \text{ for all } j,\ t_j \downarrow 0 \text{ as } j \to \infty \right\}. \tag{B.10}$$

This represents the shape of K close to x.

Lemma B.10. *Suppose that K is a closed convex set and $x \in K$. Then $T_K(x)$ is a closed convex cone. Also, if $x \notin K$, then $T_K(x) = \emptyset$.*

Proof. First we show that $T_K(x)$ is a cone: Suppose $y \in T_K(x)$ and $\alpha \ge 0$. We show that $\alpha y \in T_K(x)$. If $\alpha = 0$, then $\alpha y = 0$; $0 \in T_K(x)$ since we can take $x_j = x$ for all j. If

$\alpha > 0$, then if $y = \lim_{j\to\infty} (x_j - x)/t_j$ with $x_j \in K$ for all j and $t_j \downarrow 0$ as $j \to \infty$, then $\alpha y = \lim_{j\to\infty} (x_j - x)/(t_j/\alpha) \in T_K(x)$.

Now we show that $T_K(x)$ is convex. Suppose that $y, z \in T_K(x)$ and $0 \le \theta \le 1$. We wish to show that $\theta y + (1-\theta)z \in T_K(x)$. Let $y = \lim_{j\to\infty} (y_j - x)/s_j$ and $z = \lim_{j\to\infty} (z_j - x)/t_j$ with $y_j, z_j \in K$ and $s_j, t_j \downarrow 0$ as $j \to \infty$. Let $r_j = \min(s_j, t_j) > 0$; $r_j \downarrow 0$ as $j \to \infty$. Since $x \in K$ and K is convex, for any $0 \le \sigma_j \le 1$, $\widehat{y}_j = \sigma_j y_j + (1-\sigma_j)x \in K$. Note that $\widehat{y}_j - x = \sigma_j (y_j - x)$. If we choose $0 \le \sigma_j = r_j/s_j \le 1$, then $(\widehat{y}_j - x)/r_j = \sigma_j (y_j - x)/r_j = (y_j - x)/s_j$. Similarly we can find $\widehat{z}_j = \tau_j z_j + (1-\tau_j)x \in K$ with $\tau_j = r_j/t_j$ so that $(\widehat{z}_j - x)/r_j = \tau_j (z_j - x)/r_j = (z_j - x)/t_j$. Then

$$\begin{aligned}\theta y + (1-\theta)z &= \theta \lim_{j\to\infty} \frac{\widehat{y}_j - x}{r_j} + (1-\theta) \lim_{j\to\infty} \frac{\widehat{z}_j - x}{r_j} \\ &= \lim_{j\to\infty} \frac{\theta\widehat{y}_j + (1-\theta)\widehat{z}_j - x}{r_j} \in T_K(x).\end{aligned}$$

To show that $T_K(x)$ is closed, suppose that $y^{(j)} \to y$, $y^{(j)} \in T_K(x)$, so that $y^{(j)} = \lim_{l\to\infty} (y_l^{(j)} - x)/t_l^{(j)}$ with $y_l^{(j)} \in K$ and $t_l^{(j)} \downarrow 0$ as $l \to \infty$. We can construct a sequence l_j as follows: $l_1 = 1$;

$$l_{j+1} = \min\left\{ l \ge l_j + 1 \mid t_l^{(j+1)} \le \frac{1}{2} t_{l_j}^{(j)} \text{ and } \left\| y^{(j)} - \left(y_l^{(j)} - x \right)/t_l^{(j)} \right\| < 2^{-j} \right\}.$$

Note that $t_{l_j}^{(j)} \downarrow 0$ as $j \to \infty$. Then, setting $z_j = y_{l_j}^{(j)} \in K$,

$$\begin{aligned}\left\| y - \left(z^{(j)} - x \right)/t_{l_j} \right\| &\le \left\| y - y^{(j)} \right\| + \left\| y^{(j)} - \left(z^{(j)} - x \right)/t_{l_j} \right\| \\ &\le \left\| y - y^{(j)} \right\| + 2^{-j} \to 0 \quad \text{as } j \to \infty.\end{aligned}$$

Thus $y \in T_K(x)$.

To show that for $x \notin K$, $T_K(x) = \emptyset$, note that the distance between x and K is positive, as K is closed. Then, for any sequence $x_j \in K$ and $t_j \downarrow 0$ as $j \to \infty$, $\| (x_j - x)/t_j \| \to +\infty$ as $j \to \infty$, and so $(x_j - x)/t_j$ cannot have a limit as $j \to \infty$. □

The tangent cone $T_K(x)$ is essentially the result of "blowing up" K around x and taking the result to the limit. Note that the tangent cone can also be defined as

$$T_K(x) = \overline{\bigcup_{t>0} \frac{1}{t}(K - x)} \qquad \text{for } x \in K. \tag{B.11}$$

To see why, note that if $r > s > 0$, then for $x \in K$,

$$\frac{1}{r}(K - x) \subseteq \frac{1}{s}(K - x)$$

since K is convex. Alternatively, we can write

$$T_K(x) = \overline{\text{cone}(K - x)} \qquad \text{for } x \in K, \tag{B.12}$$

where

$$\text{cone}(A) = \{\alpha x \mid \alpha \geq 0,\, x \in A\} \tag{B.13}$$

is the cone generated by a set $A \subseteq X$. Note that neither (B.11) nor (B.12) applies if $x \notin K$; if $x \notin K$, then $T_K(x) = \emptyset$.

Closely related to the tangent cone is the *normal cone*:

$$N_K(x) = -T_K(x)^* = T_K(x)^\circ \subset X' \qquad \text{for } x \in K, \tag{B.14}$$

the negative of the dual to the tangent cone. If $x \notin K$, then we define $N_K(x) = \emptyset$. An equivalent definition is that for $x \in K$,

$$N_K(x) = \{\eta \mid \langle \eta, w - x\rangle \leq 0 \text{ for all } w \in K\}. \tag{B.15}$$

Lemma B.11. *Definition* (B.14) *is equivalent to* (B.15).

Proof. Suppose $x \in K$ and $\eta \in T_K(x)^\circ$, so that $\langle \eta, w\rangle \leq 0$ for all $w \in T_K(x)$. Pick $z \in K$. As K is convex, for all $0 \leq \theta \leq 1$, $\theta z + (1-\theta)x \in K$. Then, taking $\theta \downarrow 0$, we see that $z - x \in T_K(x)$. So $\langle \eta, z - x\rangle \leq 0$ for all $z \in K$.

Suppose that $x \in K$ and $\langle \eta, w - x\rangle \leq 0$ for all $w \in K$. Then, if $z \in T_K(x)$, then $z = \lim_{j\to\infty} \big(w_j - x\big)/t_j$, where $w_j \in K$ and $t_j \downarrow 0$. Thus

$$\langle \eta, z\rangle = \lim_{j\to\infty} \big\langle \eta, w_j - x\big\rangle/t_j \leq 0.$$

Since this inequality holds for all $z \in T_K(x)$, it follows that $\eta \in T_K(x)^\circ$. □

The normal cone is closely related to the projection operator since for all x,

$$J_X\,(x - \Pi_K(x)) \in N_K\,(\Pi_K(x)). \tag{B.16}$$

The *recession cone* of a closed convex set $K \subseteq X$ is denoted by K_∞ and given by the formula

$$K_\infty := \left\{ \lim_{i\to\infty} t_i x_i \mid x_i \in K,\, t_i \downarrow 0 \text{ as } i \to \infty \right\}.$$

An equivalent expression for any $x \in K$ is

$$K_\infty = \bigcap_{t>0} t\,(K - x). \tag{B.17}$$

The recession cone is the set of all $v \in X$ where for any $x \in K$ we have $x + \alpha v \in K$ for all $\alpha \geq 0$. Thus, for any $x \in K$,

$$x + K_\infty \subseteq K.$$

If K is a cone as well as being closed and convex, then $K = K_\infty$. We can think of K_∞ as the set of directions in K "at infinity." The recession cone can clearly be seen to be a closed

convex cone (being the nested intersection of a family of closed convex sets in (B.17)). It is also nonempty (provided, of course, that $K \neq \emptyset$), as then $0 \in K_\infty$.

It is tempting to think that if K is a closed convex set, then K is bounded if and only if $K_\infty = \{0\}$. It is certainly true that if K is bounded, then $K_\infty = \{0\}$. The converse is also true in finite dimensions. However, in infinite-dimensional spaces "$K_\infty = \{0\}$" does not necessarily imply that K is bounded. Consider, for example,

$$K = \left\{ \mathbf{x} \in \ell^2 \mid \sum_{j=1}^{\infty} \frac{1}{j} x_j^2 \leq 1 \right\}.$$

This is not a bounded set, as $\sqrt{j}\,\mathbf{e}_j \in K$ for $j = 1, 2, \ldots$. However, no ray belongs to K: $0 \in K$, and for any $\mathbf{x} \neq 0$, $0 + \alpha \mathbf{x} \in K$ means that $\alpha^2 \sum_{j=1}^{\infty} x_j^2/j \leq 1$. This gives an upper bound on α, and so we cannot take $\alpha \to +\infty$. Thus $K_\infty = \{0\}$.

In infinite dimensions, we also need to consider weak convergence. It turns out that

$$K_\infty = \{\, v \in X \mid t_k x_k \rightharpoonup v \text{ weakly}, t_k \downarrow 0, x_k \in K \text{ for all } k \,\}.$$

To show this, since strong convergence implies weak convergence,

$$K_\infty \subseteq \{\, v \in X \mid t_k x_k \rightharpoonup v \text{ weakly}, t_k \downarrow 0, x_k \in K \text{ for all } k \,\}.$$

To see the reverse inclusion, we use (B.17): $K_\infty = \bigcap_{t>0} t(K - x)$ for some (or indeed, any) $x \in K$. Then, if $t_k x_k \rightharpoonup v$, for any $s > 0$ and sufficiently large k, $0 < t_k < s$, so $t_k (x_k - x) \in s(K - x)$. Now $s(K - x)$ is a closed convex set, and so it is weakly closed, and therefore the weak limit of $t_k (x_k - x)$, which is v, must belong to $s(K - x)$. Thus $v \in \bigcap_{s>0} s(K - x) = K_\infty$.

Recession cones are not just of interest for their own sake. They can be important for determining if certain properties hold. Consider, for example, that K is a closed convex set, and L_j, $j = 1, 2, \ldots$, is a nested family of closed convex sets with $L_{j+1} \subseteq L_j$ for all j. Is

$$\bigcap_{j=1}^{\infty} \big(K + L_j\big) = K + \bigcap_{j=1}^{\infty} L_j \,?$$

This turns out to depend on recession cones—at least for $\mathbb{R}^n$!

Lemma B.12. *Suppose K and L_j, $j = 1, 2, 3, \ldots$, are all convex and closed in $\mathbb{R}^n$ with $L_1 \supset L_2 \supset \cdots$, and $K_\infty \cap (-\bigcap_{j=1}^{\infty} (L_j)_\infty) = \{0\}$. Then*

$$\bigcap_{j=1}^{\infty} (K + L_j) = K + \bigcap_{j=1}^{\infty} L_j.$$

Proof. Clearly

$$K + \bigcap_{j=1}^{\infty} L_j \subseteq \bigcap_{j=1}^{\infty} (K + L_j).$$

To show the reverse inclusion, suppose that $z \in K + L_j$ for all j. Then $z = x_j + y_j$, where $x_j \in K$ and $y_j \in L_j$. If the sequence x_j is bounded, then so is the sequence y_j, and by taking convergent subsequences, there are limits $x^* \in K$ and $y^* \in \bigcap_{j=1}^{\infty} L_j$ where $z = x^* + y^*$.

Now suppose that $\|x_j\| \to \infty$ as $j \to \infty$, possibly by taking a subsequence. By restriction to a further subsequence, $x_j/\|x_j\| = (z - y_j)/\|x_j\| \to \widehat{x} \in K_\infty$, and so $y_j/\|y_j\| \to \widehat{y} = -\widehat{x} \in (L_k)_\infty$ for all k. Thus $0 \neq \widehat{x} \in K_\infty \cap (-\bigcap_{k=1}^{\infty}(L_k)_\infty)$, contradicting the above assumption. Thus the sequences x_j and y_j are bounded, and so $z \in K + \bigcap_{j=1}^{\infty} L_j$ for all $z \in \bigcap_{j=1}^{\infty}(K + L_j)$.

Hence $\bigcap_{j=1}^{\infty}(K + L_j) = K + \bigcap_{j=1}^{\infty} L_j$, as desired. □

The "pointedness" assumption, $K_\infty \cap (-\bigcap_{j=1}^{\infty}(L_j)_\infty) = \{0\}$, is necessary. Consider the example $K = \{(x,0) \mid x \leq 0\}$ and $L_j = \{(x,y) \mid x \geq j\,|y|\}$ for $j = 1,2,3,\ldots$. Then $K + L_j = \mathbb{R}^2$ for all j, so $\bigcap_{j=1}^{\infty}(K + L_j) = \mathbb{R}^2$. But $K + \bigcap_{j=1}^{\infty} L_j = K + (\mathbb{R}_+ \times \{0\}) = \mathbb{R} \times \{0\}$. Note that in this case, $K_\infty \cap (-\bigcap_{j=1}^{\infty}\left(L_j\right)_\infty) = -\mathbb{R}_+ \times \{0\} \neq \{(0,0)\}$.

B.1.5 Existence of minimizers

In this section we show the existence of global minimizers of proper convex lower semicontinuous functions $\phi\colon X \to \mathbb{R} \cup \{\infty\}$, provided that they are *coercive*, $\phi(x) \to +\infty$ if $\|x\| \to \infty$, and X is a reflexive Banach space. We can show that if $\psi : \mathbb{R}^n \to \mathbb{R}$ is continuous and coercive, then ψ has a global minimizer: the level sets $\{x \mid \psi(x) \leq \psi(x_0)\}$ for some (any) point $x_0 \in \mathbb{R}^n$ are closed and bounded and therefore compact. Thus we have a global minimizer and a global maximizer. But this does not work in infinite dimensions. If we take $X = \ell^2 = \left\{(x_1, x_2, x_3, \ldots) \mid \sum_{i=1}^{\infty} x_i^2 < \infty\right\}$, we can set $\psi(x) = \sum_{i=1}^{\infty} x_i^2/i + (1 - \|x\|_{\ell^2}^2)^2$, which is continuous on ℓ^2 and has infimum zero, but we can never reach this infimum. Thus convexity may be necessary.

On the other hand, we need coercivity even for nice convex functions. For example, $f(x) = e^x$ is a convex function of x which is bounded below, with infimum zero. However, we cannot reach zero. We can approach zero only by taking $x \to -\infty$. Thus coercivity may be necessary.

Theorem B.13. *If $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is proper, convex, lower semicontinuous, and coercive, with X a reflexive Banach space, then there is a global minimizer x^*.*

Proof. Suppose that $\phi(x_0) < \infty$. If we apply the separating hyperplane theorem to $\operatorname{epi}\phi$ and the point $(x_0, \phi(x_0) - 1)$, then we get an affine lower bound for ϕ: $\phi(x) \geq \langle \xi, x \rangle + \beta$. So ϕ is bounded below on bounded sets. Now the level set $L_0 := \{x \mid \phi(x) \leq \phi(x_0)\}$ is a closed (since ϕ is lower semicontinuous) and bounded (since ϕ is coercive) convex set. So ϕ is bounded below on L_0 and has an infimum. Let x_k be an infimizing sequence: $\phi(x_k) \to \inf_{x \in L_0} \phi(x) = \inf_{x \in X} \phi(x)$. Since X is reflexive, by Alaoglu's theorem there is a weakly converging subsequence (also denoted by x_k), so $x_k \rightharpoonup x^*$ as $k \to \infty$. By Mazur's lemma there are $z_k \in \operatorname{co}\{x_k, x_{k+1}, \ldots\}$ such that $z_k \to x^*$ strongly. Now z_k is a convex combination of $x_k, x_{k+1}, \ldots$, so $\inf_{x \in X} \phi(x) \leq \phi(z_k) \leq \sup\{\phi(x_k), \phi(x_{k+1}), \ldots\} \to \inf_{x \in X} \phi(x)$ as $k \to \infty$. Since ϕ is lower semicontinuous, $\phi(x^*) \leq \liminf_{k \to \infty} \phi(z_k) \leq \inf_{x \in X} \phi(x)$. Adding

the inequality $\inf_{x\in X}\phi(x) \le \phi(x^*)$, we see that $\phi(x^*) = \inf_{x\in X}\phi(x)$, and x^* is a global minimizer. □

B.2 Subdifferentials and generalized gradients

The *subdifferential* of a convex function $\phi\colon X \to \mathbb{R}\cup\{\infty\}$ at $x \in X$ is the set

$$\partial\phi(z) = \left\{ g \in X' \mid \phi(z) \ge \phi(x) + \langle g, z-x\rangle \text{ for all } z \in X \right\}.$$

Subdifferentials $\partial\phi(x)$ are closed convex sets. As is shown in Section 4.2, the subdifferential of a proper lower semicontinuous convex function is a maximal monotone operator. Also, $\partial\phi(x) \ne \emptyset$ for all $x \in \operatorname{int}\operatorname{dom}\phi$. If ϕ is differentiable at x, then $\partial\phi(x) = \{\nabla\phi(x)\}$. For proper lower semicontinuous convex functions, x^* minimizes ϕ if and only if $0 \in \partial\phi(x^*)$.

The following theorem summarizes the basic properties of subdifferentials of convex functions.

Theorem B.14. *Let $\phi, \psi : X \to \mathbb{R}\cup\{\infty\}$ be proper, convex, and lower semicontinuous and X be a reflexive Banach space. Then the following hold:*

1. *The directional derivatives $\phi'(x;v) = \lim_{s\downarrow 0}(\phi(x+sv)-\phi(x))/s$ exist (possibly with the value $+\infty$). Furthermore, $\phi(z) \ge \phi(x) + \phi'(x; z-x)$, and*

$$\partial\phi(x) = \left\{ g \in X' \mid \langle g, v\rangle \le \phi'(x;v) \text{ for all } v \in X \right\}.$$

2. *$\partial\phi(x)$ is a closed convex set,*

3. *Graph $\partial\phi$ is a closed set in $X \times X'$.*

4. *$\partial\phi(x) \ne \emptyset$ whenever $x \in \operatorname{int}\operatorname{dom}\phi$ (and ϕ is Lipschitz on some neighborhood of x).*

5. *If ϕ is Gateaux differentiable at x, then $\partial\phi(x) = \{\nabla\phi(x)\}$.*

6. *x is a global minimizer of ϕ if and only if $0 \in \partial\phi(x)$.*

7. *Provided $\phi + \psi$ is proper, then $\partial(\phi+\psi)(x) \supseteq \partial\phi(x) + \partial\psi(x)$ with equality if either $x \in \operatorname{int}\operatorname{dom}\phi$ or $x \in \operatorname{int}\operatorname{dom}\psi$.*

Proof.

1. Now, for any $0 < \theta < 1$, $\phi(x+\theta sv) = \phi((1-\theta)x + \theta(x+sv)) \le (1-\theta)\phi(x) + \theta\phi(x+sv)$, and so $(\phi(x+\theta sv)-\phi(x))/(\theta s) \le (\phi(x+sv)-\phi(x))/s$. That is, $s \mapsto (\phi(x+sv)-\phi(x))/s$ is a nondecreasing function of $s > 0$. Thus the limit as $s \downarrow 0$ exists, although it can possibly have the value $+\infty$. Hence $\phi(x+sv) \ge \phi(x) + s\phi'(x;v)$. Putting $s = 1$ and $v = z - x$ for a given z shows that $\phi(z) \ge \phi(x) + \phi'(x; z-x)$, as desired. For the characterization of the subdifferential, if $\langle g, v\rangle \le \phi'(x;v)$ for all v, then clearly $g \in \partial\phi(x)$. Conversely, suppose that $g \in \partial\phi(x)$. Then, for any v, $(\phi(x+sv)-\phi(x))/s \ge (\phi(x) + s\langle g, v\rangle - \phi(x))/s = \langle g, v\rangle$. Taking the limit as $s \downarrow 0$ shows that $\phi'(x;v) \ge \langle g, v\rangle$ for any v.

2. First, $\partial\phi(x)$ is closed: Suppose $g_k \to g$ and $g_k \in \partial\phi(x)$. Then, for all $y \in X$, $\phi(y) - \phi(x) - \langle g_k, y - x\rangle \geq 0$. Taking $k \to \infty$ we get $\phi(y) - \phi(x) - \langle g, y - x\rangle \geq 0$ for all $y \in X$, and so $g \in \partial\phi(x)$.

   Second, $\partial\phi(x)$ is convex: Suppose $g_1, g_2 \in \partial\phi(x)$ and $0 \leq \theta \leq 1$. Then, for all $y \in X$ and $i = 1, 2$, $\phi(y) - \phi(x) - \langle g_i, y - x\rangle \geq 0$. Taking convex combinations of these inequalities gives $\phi(y) - \phi(x) - \langle \theta g_1 + (1-\theta) g_2, y - x\rangle \geq 0$. Since this is true for all $y \in X$, $\theta g_1 + (1-\theta) g_2 \in \partial\phi(x)$, and $\partial\phi(x)$ is convex.

3. Suppose $g_k \in \partial\phi(x_k)$ and $x_k \to x$ and $g_k \to g$ as $k \to \infty$. Then, for every $z \in X$, $\phi(z) \geq \phi(x_k) + \langle g_k, z - x_k\rangle$. Taking liminfs, noting that $\liminf_{k\to\infty} \phi(x_k) \geq \phi(x)$ as ϕ is lower semicontinuous, we obtain $\phi(z) \geq \phi(x) + \langle g, z - x\rangle$. Thus $g \in \partial\phi(x)$ and $\partial\phi$ has a closed graph.

4. We show that $\partial\phi(x) \neq \emptyset$ if $x \in \operatorname{int}\operatorname{dom}\phi$. Our first task is to show that ϕ is bounded on some open set containing x. Let $E_k = \{ z \in X \mid \phi(z) \leq k \}$. Clearly $\operatorname{dom}\phi = \bigcup_{k=1}^{\infty} E_k$. Note that $E_k \subseteq E_{k+1}$. Since ϕ is lower semicontinuous, each E_k is closed. The intersection $\bigcap_{k=1}^{\infty} \left[\operatorname{dom}\phi \backslash E_k\right] = \emptyset$. Thus, by the Baire category theorem, some set $\operatorname{dom}\phi \backslash E_k$ is *not* dense in $\operatorname{dom}\phi$, and so E_k must contain an open set for some k. So pick a point $z \in \operatorname{dom}\phi$ and an $\eta > 0$ such that $z + \eta B_X \subset \operatorname{dom}\phi$. Now, since $x \in \operatorname{int}\operatorname{dom}\phi$, $z + s(x - z) \in \operatorname{dom}\phi$ for $s \in [0, s^*]$ for some $s^* > 1$. Put $w = z + s^*(x - z)$. Now, for any point y in the convex hull of $z + \eta B_X$ and w, $y = \theta w + (1-\theta) z'$ with $z' \in z + \eta B_X \subset E_k$, so $\phi(y) \leq \theta\phi(w) + (1-\theta)\phi(z') \leq \max\{k, \phi(w)\}$. This convex hull contains an open set around x. Thus ϕ is bounded on an open ball $x + \epsilon B_X$ for some $\epsilon > 0$.

   We now show that $|\phi(y) - \phi(x)| \leq L \|x - y\|$ for $y \in x + \epsilon B_X$. Now we can choose M so that $\phi(y) \leq M$ for all $y \in x + \epsilon B_X$. Note that ϕ must be bounded below on $x + \epsilon B_X$; without loss of generality, let us suppose that $\phi(y) \geq M$ for all $y \in x + \epsilon B_X$. Thus, for any $y \in x + \epsilon B_X$, there is the point $w = x + \epsilon(x - y)/(2\|x - y\|) \in x + \epsilon B_X$: $x = \theta y + (1-\theta) w$ where $\theta = \epsilon/(\epsilon + 2\|x - y\|)$. Thus $\phi(x) - \phi(y) \leq (\theta - 1)\phi(y) + (1-\theta)\phi(w) \leq 2M\, 2\|x - y\| / (\epsilon + 2\|x - y\|) \leq (4M/\epsilon)\|x - y\|$. To get an inequality in the reverse direction, note that if we set $w = y + s(y - x)$ for $s > 0$, $\phi(y) \leq s\,\phi(x)/(1+s) + \phi(w)/(1+s)$ and so $\phi(y) - \phi(x) \leq (\phi(w) - \phi(x))/(1+s) \leq 2M/(1+s)$. Since we can take any $s > 0$ where $\|w - x\| = (1+s)\|y - x\| < \epsilon$, taking $1 + s$ to the limiting value of $\epsilon/\|y - x\|$ we get the bound $\phi(y) - \phi(x) \leq 2M\|y - x\|/\epsilon$. Either way, we get $|\phi(x) - \phi(y)| \leq (4M/\epsilon)\|y - x\|$. From this we can see that the directional derivative at x, $\phi'(x; \cdot)$ is Lipschitz and convex. Then, by the Hahn–Banach theorem, there is a $g \in X'$ where $\langle g, v\rangle \leq \phi'(x; v)$ for all $v \in X$, so $g \in \partial\phi(x)$ and $\partial\phi(x) \neq \emptyset$.

5. If ϕ is Gateaux differentiable at x, then for any $v \in X$,

$$\frac{\phi(x + sv) - \phi(x)}{s} \to \phi'(x; v) = \langle \nabla\phi(x), v\rangle \qquad \text{as } s \to 0.$$

   Thus, for any z, putting $s = 1$ and $v = z - x$ we get $\phi(z) \geq \phi(x) + \langle \nabla\phi(x), z - x\rangle$ for all $z \in X$. Hence $\nabla\phi(x) \in \partial\phi(x)$. To show that nothing else is in $\partial\phi(x)$, suppose that $g \in \partial\phi(x)$. Then $\phi(x + s\,v) \geq \phi(x) + s\,\langle g, v\rangle$ for all $s > 0$ and $v \in X$. Taking limits $s \downarrow 0$ gives $\phi'(x; v) = \langle \nabla\phi(x), v\rangle \geq \langle g, v\rangle$ for all $v \in X$. Replacing v with $-v$ shows

that the reverse inequality holds, and so $\langle \nabla\phi(x), v\rangle = \langle g, v\rangle$ for all $v \in X$. The only way this can happen is if $g = \nabla\phi(x)$.

6. Now suppose that x is a global minimizer of ϕ. Then clearly $\phi(z) \geq \phi(x) = \phi(x) + \langle 0, z - x\rangle$ for all z, so $0 \in \partial\phi(x)$. Conversely, suppose that $0 \in \partial\phi(x)$. Then, for all z, $\phi(z) \geq \phi(x) + \langle 0, z - x\rangle = \phi(x)$, as desired.

7. Suppose $g \in \partial\phi(x)$ and $h \in \partial\psi(x)$. Then, for any $z \in X$,

$$\phi(z) + \psi(z) \geq \phi(x) + \psi(x) + \langle g + h, z - x\rangle,$$

and so $g + h \in \partial(\phi + \psi)$. Conversely, suppose that $x \in \operatorname{int}\operatorname{dom}\phi$. Then, by item 4, ϕ is Lipschitz in a neighborhood of x. Then the directional derivative $\phi'(x; v)$ not only exists but also is finite (bounded by the local Lipschitz constant of ϕ times $\|v\|$). The directional derivative of ψ also exists, but it may be infinite. Then, for any $v \in X$, the directional derivative $(\phi + \psi)'(x; v) = \phi'(x; v) + \psi'(x; v)$ exists, but it may be infinite. The subdifferential of a convex function f can be written as $\partial f(x) = \left\{ g \in X' \mid f'(x; v) \geq \langle g, v\rangle \text{ for all } v \right\}$. Since we have $(\phi + \psi)'(x; v) = \phi'(x; v) + \psi'(x; v)$ for all $v \in X$, we have equality of $\partial(\phi + \psi)(x)$ and $\partial\phi(x) + \partial\psi(x)$. □

The Baire category argument for item 4 is given in Borwein and Zhu (see [36, Thm. 4.1.3]). Item 7 can be shown to hold if $0 \in \operatorname{int}[\operatorname{dom}\partial\phi - \operatorname{dom}\partial\psi]$ via Lemma 2.32.

B.2.1 Fenchel duality

An important concept in convex analysis is that of *Fenchel dual* of proper convex functions. The Fenchel dual $\phi^*\colon X' \to \mathbb{R} \cup \{\infty\}$ of a convex function $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is

$$\phi^*(y) = \sup_{x \in X} \langle y, x\rangle - \phi(x). \tag{B.18}$$

The main properties of Fenchel dual functions are given in the following theorem.

Theorem B.15. *Suppose that X is a reflexive Banach space, and that $\phi\colon X \to \mathbb{R} \cup \{\infty\}$ is proper, convex, and lower semicontinuous. Then the Fenchel dual $\phi^*\colon X' \to \mathbb{R} \cup \{\infty\}$ is also proper, convex, and lower semicontinuous. Also $\phi^{**} = \phi$. Furthermore, for any $x \in X$ and $y \in X'$, we have*

$$\phi(x) + \phi^*(y) \geq \langle x, y\rangle$$

with equality if and only if $y \in \partial\phi(x)$, or equivalently, $x \in \partial\phi^*(y)$.

Proof. From the definition (B.18), ϕ^* is the supremum of linear functions and is therefore convex.

To show that it is lower semicontinuous, suppose that $y_k \to y$ in X'. Then

$$\phi^*(y_k) = \sup_{x \in X} \langle y_k, x\rangle - \phi(x).$$

For any $\epsilon > 0$ choose x_ϵ such that $\langle y, x_\epsilon\rangle - \phi(x_\epsilon) \le \phi^*(y) \le \langle y, x_\epsilon\rangle - \phi(x_\epsilon) + \epsilon$. Then

$$\begin{aligned}\phi^*(y_k) &\ge \langle y_k, x_\epsilon\rangle - \phi(x_\epsilon)\\ &= \langle y, x_\epsilon\rangle + \langle y_k - y, x_\epsilon\rangle - \phi(x_\epsilon)\\ &\ge \phi^*(y) - \epsilon - \|y_k - y\|\,\|x_\epsilon\|\\ &\to \phi^*(y) - \epsilon \qquad \text{as } k \to \infty.\end{aligned}$$

Thus $\liminf_{k\to\infty} \phi^*(y_k) \ge \phi^*(y) - \epsilon$. Since $\epsilon > 0$ is arbitrary, $\liminf_{k\to\infty} \phi^*(y_k) \ge \phi^*(y)$ and ϕ^* is lower semicontinuous.

To show that ϕ^* is proper, note that ϕ is proper, so that there is a point $x \in X$ where $\phi(x) < \infty$. Then, since $\operatorname{epi}\phi$ is a closed, convex set, and $(x, \phi(x) - 1) \notin \operatorname{epi}\phi$, by the separating hyperplane theorem, there must be $(\eta, \gamma) \in X' \times \mathbb{R}$ and $\beta \in \mathbb{R}$ such that

$$\begin{aligned}\left\langle \begin{bmatrix} \eta \\ \gamma \end{bmatrix}, \begin{bmatrix} z \\ s \end{bmatrix} \right\rangle + \beta \ge 0 \qquad &\text{for all } z \in X,\ s \ge \phi(z),\\ \left\langle \begin{bmatrix} \eta \\ \gamma \end{bmatrix}, \begin{bmatrix} x \\ \phi(x) - 1 \end{bmatrix} \right\rangle + \beta < 0.&\end{aligned}$$

In particular, taking $z = x$ and $s = \phi(x)$ in the first inequality gives $\langle \eta, x\rangle + \gamma\,\phi(x) + \beta \ge 0$. On the other hand, the second inequality gives $\langle \eta, x\rangle + \gamma\,\phi(x) - \gamma + \beta < 0$, so $\gamma > 0$. Again using the first inequality but with $z \in X$ and $s = \phi(z)$, we get

$$\langle \eta, z\rangle + \gamma\,\phi(z) + \beta \ge 0.$$

Negating and dividing by $\gamma > 0$ give $\langle -\eta/\gamma, z\rangle - \phi(z) \le -\beta/\gamma$; taking the supremum gives $\phi^*(-\eta/\gamma) \le -\beta/\gamma < +\infty$, and so ϕ^* is proper.

Suppose that $x \in X$ and $y \in X'$. Then

$$\begin{aligned}\phi(x) + \phi^*(y) &= \phi(x) + \sup_{z\in X} \langle y, z\rangle - \phi(z)\\ &\ge \phi(x) + \langle y, x\rangle - \phi(x) \;=\; \langle y, x\rangle.\end{aligned}$$

This is known as *weak duality*.

We can now justify the term duality by showing that $\phi^{**} = \phi$. First we show that $\phi^{**} \le \phi$:

$$\begin{aligned}\phi^{**}(x) &= \sup_{y\in X'} \langle x, y\rangle - \phi^*(x)\\ &= \sup_{y\in X'} \left[\langle x, y\rangle - \sup_{z\in X} \langle z, y\rangle + \phi(x) \right]\\ &= \sup_y \inf_z \langle x - z, y\rangle + \phi(z).\end{aligned}$$

Taking $z = x$ in place of the infimum shows that

$$\phi^{**}(x) \le \sup_y \langle x - x, y\rangle + \phi(x) \;=\; \phi(x) \qquad \text{for all } x \in X.$$

On the other hand, we can show that $\phi \le \phi^{**}$: Suppose that $\phi(x) < \infty$. For $\epsilon > 0$, since $(x, \phi(x) - \epsilon) \notin \operatorname{epi}\phi$, there are $\eta_\epsilon \in X'$ and $\beta_\epsilon, \gamma_\epsilon \in \mathbb{R}$ such that

$$\left\langle \begin{bmatrix} \eta_\epsilon \\ \gamma_\epsilon \end{bmatrix}, \begin{bmatrix} z \\ s \end{bmatrix} \right\rangle + \beta_\epsilon \ge 0 \qquad \text{for all } z \in X,\ s \ge \phi(z),$$
$$\left\langle \begin{bmatrix} \eta_\epsilon \\ \gamma_\epsilon \end{bmatrix}, \begin{bmatrix} x \\ \phi(x) - \epsilon \end{bmatrix} \right\rangle + \beta_\epsilon < 0.$$

That is, $\langle \eta_\epsilon, z\rangle + \gamma_\epsilon \phi(z) + \beta_\epsilon \ge 0$ for all $z \in X$ while $\langle \eta_\epsilon, x\rangle + \gamma_\epsilon \phi(x) - \gamma_\epsilon \epsilon + \beta_\epsilon < 0$. Note that $\gamma_\epsilon > 0$. Dividing by γ_ϵ and combining these inequalities give $\phi(x) < \langle -\eta_\epsilon/\gamma_\epsilon, x - z\rangle + \phi(z) + \epsilon$. Then

$$\begin{aligned} \phi^{**}(x) &= \sup_y \inf_z \langle x - z, y\rangle + \phi(z) \\ &\ge \inf_z \langle x - z, -\eta_\epsilon/\gamma_\epsilon\rangle + \phi(z) \\ &\ge \phi(x) - \epsilon. \end{aligned}$$

In the case where $\phi(x) = \infty$, we use the separating hyperplane theorem again, but using $(x, M) \notin \operatorname{epi}\phi$ for arbitrary $M \in \mathbb{R}$. In either case we get $\phi^{**}(x) \ge \phi(x)$ for all x. Combining the results shows that $\phi^{**} = \phi$.

Now suppose that $\phi(x) + \phi^*(y) = \langle y, x\rangle$. Then, for any $z \in X$, $\phi(x) + \langle y, z\rangle - \phi(z) \le \langle y, x\rangle$; rearranging gives $\phi(z) \ge \phi(x) + \langle y, z - x\rangle$. In other words, $y \in \partial\phi(x)$.

To show the reverse implication, suppose that $y \in \partial\phi(x)$, so that $\phi(z) \ge \phi(x) + \langle y, z - x\rangle$ for all $z \in X$. Then $z \mapsto \phi(z) - \langle y, z - x\rangle$ has a global minimum at $z = x$. So

$$\begin{aligned} \phi^*(y) &= \sup_z \langle z, y\rangle - \phi(z) \\ &= \sup_z \langle z - x, y\rangle - \phi(z) + \langle x, y\rangle \\ &= -\phi(x) + \langle x, y\rangle, \end{aligned}$$

and $\phi(x) + \phi^*(y) = \langle x, y\rangle$.

To show that $\phi(x) + \phi^*(y) = \langle x, y\rangle$ is equivalent to $x \in \partial\phi^*(y)$, apply the previous two paragraphs to $\psi = \phi^*$ and use $\psi^* = \phi^{**} = \phi$. Then $\psi(y) + \psi^*(x) = \langle x, y\rangle$ is equivalent to $x \in \partial\psi(y)$; unwrapping the substitutions gives the result. □

The difference $\phi(x) + \phi^*(y) - \langle y, x\rangle$ is called the *duality gap*. If the duality gap is zero, then $\phi^*(y) = \langle y, x\rangle - \phi(x) = \max_{z\in X} \langle y, z\rangle - \phi(z)$. If ϕ^* is a computable function, then the duality gap is a way of determining how close a feasible point is to being optimal for convex optimization problems.

B.2.2 Constrained convex optimization and KKT conditions

Minimizing a convex function subject to convex inequality constraints is a standard problem in nonlinear programming. Usually we have Lagrange multipliers and KKT (Karush–Kuhn–Tucker or just Kuhn–Tucker) conditions. But this depends on the holding of some constraint qualification.

Consider the problem

$$\min_x f(x) \qquad \text{subject to} \tag{B.19}$$

$$g_i(x) \le 0, \qquad i = 1, 2, \ldots, m, \tag{B.20}$$

with f and the g_i's convex finite-valued functions. The feasible set is

$$K = \{x \mid g_i(x) \le 0,\ i = 1, 2, \ldots, m\}, \tag{B.21}$$

which is a closed convex set. Then (B.19)–(B.20) is equivalent to minimizing $f + I_K$, where I_K is the indicator function for K. The optimality condition then becomes $0 \in \partial(f + I_K)(x)$. As long as f is finite valued on a neighborhood of K, this is equivalent to $0 \in \partial f(x) + \partial I_K(x) = \partial f(x) + N_K(x)$ by Theorem B.14(7). If f is differentiable throughout K, we have the optimality condition

$$\nabla f(x) \in -N_K(x).$$

The usual rule for $N_K(x)$ where K is given by (B.21) with smooth g_i is

$$N_K(x) = \text{cone}\,\{\nabla g_i(x) \mid g_i(x) = 0\}.$$

However, this can fail, even for just one constraint. Consider, for example, $g(x, y) = x^2 + y^2$ and $K = \{(x, y) \mid g(x, y) \le 0\}$. This rule would give $N_K(0,0) = \{(0,0)\}$, whereas $K = \{(0,0)\}$, and so in fact $N_K(0,0) = \mathbb{R}^2$. The rule $N_K(x) = \text{cone}\,\{\nabla g_i(x) \mid g_i(x) = 0\}$ applies as long as the *Slater constraint qualification* holds, that is, if there is a point $\widehat{x}$ where

$$g_i(\widehat{x}) < 0 \qquad \text{for all } i = 1, 2, \ldots, m. \tag{B.22}$$

We handle this in two steps. The first is to show that the Slater constraint qualification for a general locally Lipschitz function gives a nice formula for the normal cone.

Lemma B.16. *Let $K = \{x \in X \mid \phi(x) \le 0\}$, where ϕ is a locally Lipschitz convex function and X a reflexive Banach space, and suppose that there is an x_0 where $\phi(x_0) < 0$. Then, for any x^* with $\phi(x^*) = 0$,*

$$N_K(x^*) = \text{cone}\,\partial\phi(x^*).$$

Proof. Note that the directional derivative $\phi'(x^*; v)$ is finite and defined for all v, positively homogeneous, and convex in v. First, we show that $T_K(x^*) \subseteq \left\{v \mid \phi'(x; v) \le 0\right\}$. Let $v = \lim_{k\to\infty}(x_k - x^*)/t_k$ with $x_k \in K$ and $t_k \downarrow 0$. If L is a local Lipschitz constant for ϕ, then

$$\begin{aligned}\phi'(x^*; v) &= \lim_{k\to\infty} \frac{\phi(x^* + t_k v) - \phi(x^*)}{t_k} \\ &\le \lim_{k\to\infty} \frac{\phi(x_k) - \phi(x^*)}{t_k} + L \left\| \frac{x_k - x^*}{t_k} - v \right\| \le 0,\end{aligned}$$

so $T_K(x^*) \subseteq \left\{v \mid \phi'(x^*; v) \le 0\right\}$. To show the reverse inclusion, we need the Slater constraint qualification: let $v_0 = x_0 - x^*$. Now $\phi'(x^*; v_0) \le \phi(x^* + v_0) - \phi(x^*) < 0$. So, by convexity of $\phi'(x^*; \cdot)$, for $0 < \theta < 1$ and $v \in T_K(x^*)$,

$$\phi'(x^*; (1-\theta)v + \theta v_0) \le (1-\theta)\phi'(x^*; v) + \theta\,\phi'(x^*; v_0) < 0.$$

Thus, for sufficiently small $t > 0$, $\phi(x^* + t[(1-\theta)v + \theta v_0]) < 0$, so taking $t \downarrow 0$, $(1-\theta)v + \theta v_0 \in T_K(x^*)$. Since $T_K(x^*)$ is a closed set, taking $\theta \downarrow 0$ gives $v \in T_K(x^*)$, as desired.

Then

$$\begin{aligned} T_K(x^*) &= \{\, v \mid \phi'(x^*; v) \le 0 \,\} \\ &= \left\{ v \mid \sup_{\xi \in \partial\phi(x^*)} \langle \xi, v \rangle \le 0 \right\}. \end{aligned} \tag{B.23}$$

So $N_K(x^*) = T_K(x^*)^\circ = \{\eta \mid \langle \eta, v \rangle \le 0 \text{ for all } v \in T_K(x^*)\} \supseteq \partial\phi(x^*)$. Since $N_K(x^*)$ is a cone, $N_K(x^*) \supseteq \operatorname{cone} \partial\phi(x^*)$. To show the reverse inclusion, suppose that $\eta \in N_K(x^*) \setminus \operatorname{cone} \partial\phi(x^*)$. By the separating hyperplane theorem, there is a $z \in X$ where

$$\begin{aligned} &\langle \eta, z \rangle + \beta > 0, \\ &\langle \alpha\xi, z \rangle + \beta \le 0 \qquad \text{for all } \alpha \ge 0, \quad \xi \in \partial\phi(x^*). \end{aligned}$$

Taking $\alpha = 0$ gives $\beta \le 0$; taking $\alpha \to +\infty$ gives $\langle \xi, z \rangle \le 0$ for all $\xi \in \partial\phi(x^*)$. Thus $z \in T_K(x^*)$ by (B.23). On the other hand, $\langle \eta, z \rangle > 0$, contradicting $N_K(x^*) = T_K(x^*)^\circ$. Thus we must conclude that $N_K(x^*) \subseteq \operatorname{cone} \partial\phi(x^*)$. Combining the two inclusions gives $N_K(x^*) = \operatorname{cone} \partial\phi(x^*)$. □

To get

$$\begin{aligned} N_K(x) = \operatorname{cone} \{\, \nabla g_i(x) \mid g_i(x) = 0 \,\} \qquad &\text{for} \\ K = \{\, x \mid g_i(x) \le 0,\ i = 1, 2, \ldots, m \,\} & \end{aligned}$$

requires an additional step. If we set $g_{max}(x) = \max_{i=1,2,\ldots,m} g_i(x)$, then g_{max} is a locally Lipschitz convex function and $K = \{x \mid g_{max}(x) \le 0\}$. So $N_K(x) = \operatorname{cone} \partial g_{max}(x)$. What we now need is a formula for ∂g_{max}.

Lemma B.17. *If $\phi_{max}(x) = \max_{i=1,2,\ldots,m} \phi_i(x)$, where each $\phi\colon X \to \mathbb{R}$ is a locally Lipschitz convex function, then*

$$\partial\phi_{max}(x^*) = co \bigcup_{i:\phi_i(x^*)=\phi_{max}(x^*)} \partial\phi_i(x^*). \tag{B.24}$$

Proof. If $\phi_i(x^*) < \phi_{max}(x^*)$, then $\phi_i(x) < \phi_{max}(x)$ for all x in a neighborhood of x^*, so we can ignore ϕ_i if this is so. So we assume without loss of generality that $\phi_i(x^*) = \phi_{max}(x^*)$ for all i. Then

$$\begin{aligned} \phi'_{max}(x^*; v) &= \lim_{t \downarrow 0} \frac{\phi_{max}(x^* + tv) - \phi_{max}(x^*)}{t} \\ &= \lim_{t \downarrow 0} \max_i \frac{\phi_i(x^* + tv) - \phi_i(x^*)}{t} \\ &= \max_i \lim_{t \downarrow 0} \frac{\phi_i(x^* + tv) - \phi_i(x^*)}{t} \;=\; \max_i \phi'_i(x^*; v). \end{aligned}$$

So

$$\begin{aligned} \partial\phi_{max}(x^*) &= \{\, \xi \mid \langle \xi, v \rangle \le \phi'_{max}(x^*; v) \text{ for all } v \,\} \\ &= \{\, \xi \mid \langle \xi, v \rangle \le \phi'_i(x^*; v) \text{ for all } v \text{ and } i \,\} \;\supseteq\; \partial\phi_i(x^*) \quad \text{for all } i. \end{aligned}$$

Thus $\partial\phi_{max}(x^*) \supseteq \text{co}\bigcup_i \partial\phi_i(x^*)$. Since each $\partial\phi_i(x^*)$ is closed, convex, and bounded, it is also weakly compact (as X is a reflexive space), and so $\text{co}\bigcup_i \partial\phi_i(x^*)$ is also weakly compact and thus strongly closed.

To show the reverse inclusion, suppose that $\eta \in \partial\phi_{max}(x^*)\backslash\text{co}\bigcup_i \partial\phi_i(x^*)$. Then, by the separating hyperplane theorem, there are $z \in X$ and $\beta \in \mathbb{R}$ such that

$$\begin{aligned} &\langle \eta, z\rangle + \beta > 0, \\ &\langle \xi, z\rangle + \beta \le 0 \qquad \text{for all } \xi \in \text{co}\bigcup_i \partial\phi_i(x^*). \end{aligned}$$

The latter inequality reduces to $\langle \xi, z\rangle + \beta \le 0$ for all $\xi \in \partial\phi_i(x^*)$ for some i. Taking the supremum over all such ξ's shows that $\max_i \phi_i'(x; z) \le -\beta < \langle \eta, z\rangle \le \phi'_{max}(x; z)$, which is a contradiction. Thus $\partial\phi_{max}(x^*) = \text{co}\bigcup_{i:\phi_i(x^*)=\phi_{max}(x^*)} \partial\phi_i(x^*)$. □

This result can be extended to $\phi\colon X \times A \to \mathbb{R}$ continuous, with $\phi(x,a)$ convex and Lipschitz in x. If $\phi_{max}(x) = \max_{a\in A} \phi(x,a)$, then

$$\partial\phi_{max}(x^*) = \overline{\text{co}} \bigcup_{a:\phi(x^*,a)=\phi_{max}(x^*)} \partial_x\phi(x,a). \tag{B.25}$$

Such formulas are proved in even more generality in, for example, Clarke [55] and are very important for certain optimization problems.

Now if $K = \{x \mid g_i(x) \le 0, i = 1, 2, \ldots, m\}$ with g_i convex and Slater's constraint qualification holds (for some x_0, $g_i(x_0) < 0$ for all i), then if $x^* \in K$,

$$\begin{aligned} N_K(x^*) &= \text{cone co} \bigcup_{i:g_i(x^*)=0} \partial g_i(x^*) \\ &= \left\{ \sum_{i=1}^m \lambda_i \xi_i \mid \lambda_i \ge 0, \xi_i \in \partial g_i(x^*), \lambda_i \cdot g_i(x^*) = 0 \text{ for all } i \right\}. \end{aligned}$$

The optimization criterion for (B.19)–(B.20) then becomes the existence of Lagrange multipliers $\lambda_i \ge 0$ such that $\lambda_i\, g_i(x) = 0$ for all i and

$$0 \in \partial f(x^*) + \sum_{i=1}^m \lambda_i\, \partial g_i(x^*), \tag{B.26}$$

the convex subdifferential version of the KKT conditions.

What if the Slater constraint qualification fails? What should we use for the necessary conditions? Fritz John came up with an answer for smooth (but generally nonconvex constraints) [135]. In the convex case it works like this: let $\phi_{max}(x) = \max_i \phi_i(x)$. If Slater's constraint qualification fails, then $\phi'_{max}(x; v) \ge 0$ for all v, so $0 \in \partial\phi_{max}(x)$. Thus there are $\theta_i \ge 0$, $\sum_{i=1}^m \theta_i = 1$, where $0 \in \sum_{i=1}^m \theta_i\, \partial\phi_i(x)$. We can combine this with (B.26) to give the condition that

$$0 \in \lambda_0\, \partial f(x) + \sum_{i=1}^m \lambda_i\, \partial g_i(x^*), \tag{B.27}$$

where all $\lambda_i \ge 0$ (including $i = 0$) and at least one of the λ_i's is strictly positive. This is known as the *Fritz John condition* for optimality, and it holds even if constraint qualifications fail.

B.2.3 Inf-convolutions

The inf-convolution of two proper convex lower semicontinuous functions $f, g: X \to \mathbb{R} \cup \{\infty\}$ is the function

$$(f \boxdot g)(x) = \inf_y f(y) + g(x-y) \tag{B.28}$$

$$= \inf_{y,z:x=y+z} f(y) + g(z). \tag{B.29}$$

Note that $\operatorname{dom} f \boxdot g = \operatorname{dom} f + \operatorname{dom} g$. Now $f \boxdot g$ is a convex function, but it is not necessarily lower semicontinuous.

Lemma B.18. *If f and g are convex functions, so is $f \boxdot g$.*

Proof. Fix $x_1, x_2 \in X$ and $0 \le \theta \le 1$; let $x = \theta x_1 + (1-\theta)x_2$. For any $\epsilon > 0$, choose $y_{1,\epsilon}$, $y_{2,\epsilon}$ such that $f \boxdot g(x_i) + \epsilon \ge f(y_{i,\epsilon}) + g(x_i - y_{i,\epsilon})$, $i = 1, 2$. Set $y_\epsilon = \theta y_{1,\epsilon} + (1-\theta) y_{2,\epsilon}$. Then

$$\begin{aligned} f \boxdot g(x) &\le f(y_\epsilon) + g(x - y_\epsilon) \\ &\le \theta f(y_{1,\epsilon}) + (1-\theta) f(y_{2,\epsilon}) + \theta g(x_1 - y_{1,\epsilon}) + (1-\theta) g(x_2 - y_{2,\epsilon}) \\ &\le \theta \left[f \boxdot g(x_1) + \epsilon \right] + (1-\theta) \left[f \boxdot g(x_2) + \epsilon \right]. \end{aligned}$$

Since $\epsilon > 0$ is arbitrary, we see that $f \boxdot g$ is indeed convex. □

Inf-convolutions are closely related to Fenchel duals: $(f \boxdot g)^* = f^* + g^*$. However, we will need to use some results about inf-convolutions (and related functions) assuming some properties of the domains of f and g. In particular, we want to give conditions under which $f \boxdot g$ is locally Lipschitz (which implies that it *is* locally lower semicontinuous). Following the proof of Theorem B.14(4), if we can show that a convex function is locally bounded above, then it is locally Lipschitz. The conditions under which we can prove that $f \boxdot g$ is locally Lipschitz are variously called *constraint qualifications*, or *transversality conditions*. These kinds of conditions generalize what are ordinarily referred to as constraint qualifications or transversality conditions well beyond their readily recognizable forms.

Our basic result is the following.

Lemma B.19. *If the domain of $f \boxdot g$ contains an open set, then $f \boxdot g$ is bounded in a neighborhood of z_0 for any $z_0 \in \operatorname{int}(\operatorname{dom} f + \operatorname{dom} g)$.*

Note that it is sufficient to show that $f \boxdot g$ is Lipschitz on a neighborhood of each point in the interior of $\operatorname{dom} f + \operatorname{dom} g$. Before we give the proof of this lemma, we need some preliminary results.

A set S is *convex series closed* if for $\theta_i \ge 0$, $\sum_{i=1}^\infty \theta_i = 1$ and $x_i \in S$, $\widehat{x} = \sum_{i=1}^\infty \theta_i x_i$ implies $\widehat{x} \in S$.

The importance of these concepts is that they relate to the interior of convex series closed sets. The proof follows Borwein and Zhu [36].

Lemma B.20. *If S is convex series closed, then* $\operatorname{int} S = \operatorname{int} \overline{S}$.

Proof. We know that $S \subseteq \overline{S}$, so $\operatorname{int} S \subseteq \operatorname{int} \overline{S}$. We want to show the reverse conclusion, so suppose that $z \in \operatorname{int} \overline{S}$. Choose $\delta > 0$ such that $z + \delta B_X \subset \overline{S}$. Now, translating by $-z$,

$$\delta B_X \subset \overline{S - z} \subset S - z + \frac{1}{2}\delta B_X.$$

Multiplying by 2^{-k} gives $2^{-k}\delta B_X \subset 2^{-k}(S-z) + 2^{-k-1}\delta B_X$. Expanding this gives

$$\frac{1}{2}\delta B_X \subset \frac{1}{2}(S-z) + \frac{1}{2^2}(S-z) + \cdots + \frac{1}{2^{k+1}}(S-z) + \frac{1}{2^{k+2}}\delta B_X.$$

Then, for any $w \in \frac{1}{2}\delta B_X$,

$$w \in \sum_{i=1}^{k+1} 2^{-i}(s_i - z) + 2^{-k-1}\delta B_X. \tag{B.30}$$

Since S is convex series closed, so is the translate $S - z$, and $\sum_{i=1}^{\infty} 2^{-i}(s_i - z) \in S - z$; that is, $\sum_{i=1}^{\infty} 2^{-i} s_i \in S$. On the other hand, taking the limit as $k \to \infty$ of (B.30) gives

$$w = \sum_{i=1}^{\infty} 2^{-i}(s_i - z) \in S - z.$$

Since this is true for all $w \in \frac{1}{2}\delta B_X$, it follows that $z + \frac{1}{2}\delta B_X \subset S$, and so $z \in \operatorname{int} S$. Hence $\operatorname{int} \overline{S} \subset \operatorname{int} S$, and the equality of the two interiors follows. □

This has to be combined with the following result regarding closed convex absorbing sets: A set A is *absorbing* if for any $x \in X$ there is an $\alpha > 0$ such that $\alpha x \in A$.

Lemma B.21. *If S is a closed convex absorbing set in a Banach space X, then* $0 \in \operatorname{int} S$.

Proof. Note that if S is absorbing, so is $S \cap (-S)$, so we assume without loss of generality that S is *balanced* (that is, if $x \in S$, then $-x \in S$). Since S is absorbing, $\bigcup_{k=1}^{\infty} k\, S = X$. Since this is a countable union of G_δ sets whose union is open, some $k\, S$ contains an open set by the Baire category theorem. Thus S contains an open set $x_0 + \delta\, B_X$. So $-x_0 + \delta\, B_X \subset S$ as well; taking convex combinations with $\theta = 1/2$ gives $\delta\, B_X \subset S$. In other words, $0 \in \operatorname{int} S$. □

Now we can return to the local Lipschitz property of inf-convolutions.

***Proof of Lemma* B.19.** Suppose that $z_0 \in \operatorname{int} \operatorname{dom} f \,\square\, g$, and pick $x_0 \in \operatorname{dom} f$ and $y_0 \in \operatorname{dom} g$, where $z_0 = x_0 + y_0$. Without loss of generality, shift the values of f and g so that $f(x_0) = g(y_0) = 0$. Let $\widetilde{f} = f + I_{x_0 + \overline{B_X}}$ and $\widetilde{g} = g + I_{y_0 + \overline{B_X}}$. Now $\widetilde{f}$ and $\widetilde{g}$ are convex proper lower semicontinuous functions with $f \le \widetilde{f}$ and $g \le \widetilde{g}$, so $f \,\square\, g \le \widetilde{f} \,\square\, \widetilde{g}$, and we have to show just that $\widetilde{f} \,\square\, \widetilde{g}$ is locally bounded. Let L_α be the level set $L_\alpha := \left\{ z \mid \left(\widetilde{f} \,\square\, \widetilde{g}\right)(z) < \alpha \right\}$ for $\alpha > 0$. First we show that $L_\alpha - z_0$ is absorbing; that is, for any z, there is a $\theta > 0$ such that $z_0 + \theta(z - z_0) \in L_\alpha$. To see this, we note that since $\operatorname{dom} f + \operatorname{dom} g$ contains a neighborhood of z_0, $z_0 + \theta(z - z_0) \in \operatorname{dom} f + \operatorname{dom} g$ for sufficiently small $\theta > 0$.

For such a θ there must be an $x \in \operatorname{dom} f$ and a $y \in \operatorname{dom} g$ where $z_0 + \theta(z - z_0) = x + y$. Now $\operatorname{dom} f$ and $\operatorname{dom} g$ are convex sets, so for any $0 \le \eta \le 1$, $x_0 + \eta(x - x_0) \in \operatorname{dom} f$ and $y_0 + \eta(y - y_0) \in \operatorname{dom} g$. Pick $\eta > 0$ sufficiently small so that $\eta < 1/\max(\|x - x_0\|, \|y - y_0\|)$; thus $x_0 + \eta(x - x_0) \in \operatorname{dom} \widetilde{f}$ and $y_0 + \eta(y - y_0) \in \operatorname{dom} \widetilde{g}$. In fact, $\widetilde{f}(x_0 + \eta(x - x_0)) \le \eta f(x)$, $\widetilde{g}(y_0 + \eta(y - y_0)) \le \eta g(y)$, so $\big(\widetilde{f} \,\square\, \widetilde{g}\big)(z_0 + \theta\eta(z - z_0)) \le \eta(f(x) + g(y))$. If we also make $0 < \eta < \alpha/(f(x) + g(y))$, then $z_0 + \theta\eta(z - z_0) \in L_\alpha$.

We now want to show that L_α is convex series closed. First, L_α is bounded, as $\operatorname{dom}\big(\widetilde{f} \,\square\, \widetilde{g}\big) = \operatorname{dom} \widetilde{f} + \operatorname{dom} \widetilde{g} \subseteq x_0 + y_0 + 2\overline{B_X}$. So consider $z := \sum_{i=1}^{\infty} \theta_i z_i$ with $\theta_i \ge 0$ for all i and $\sum_{i=1}^{\infty} \theta_i = 1$ and $z_i \in L_\alpha$. This series certainly converges. We want to show that it converges to an element of L_α. To do this, note that for any $z_i \in L_\alpha$ there must be $x_i \in \operatorname{dom} \widetilde{f}$ and $y_i \in \operatorname{dom} \widetilde{g}$ where $x_i + y_i = z_i$ and $\widetilde{f} \,\square\, \widetilde{g}(z_i) \le \widetilde{f}(x_i) + \widetilde{g}(y_i) \le \big(\widetilde{f} \,\square\, \widetilde{g}(z_i) + \alpha\big)/2 < \alpha$. Since the $x_i \in x_0 + \overline{B_X}$ and $y_i \in y_0 + \overline{B_X}$, both x_i and y_i are bounded, and so $x := \sum_{i=1}^{\infty} \theta_i x_i$ and $y := \sum_{i=1}^{\infty} \theta_i y_i$ converge. Since $\widetilde{f}$ and $\widetilde{g}$ are both lower semicontinuous and convex, it follows that

$$\begin{aligned} \widetilde{f} \,\square\, \widetilde{g}(z) &\le \widetilde{f}(x) + \widetilde{g}(y) \\ &\le \sum_{i=1}^{\infty} \theta_i \big(\widetilde{f}(x_i) + \widetilde{g}(y_i)\big) \;<\; \alpha, \end{aligned}$$

and so $z \in L_\alpha$. Thus L_α is convex series closed.

Now $L_\alpha - z_0$ is a convex absorbing set, and so $\overline{L_\alpha} - z_0$ is a closed convex absorbing set. Hence, $\overline{L_\alpha}$ contains an open neighborhood of z_0 by Lemma B.20, and $z_0 \in \operatorname{int} \overline{L_\alpha}$. But for convex series closed sets, $\operatorname{int} \overline{L_\alpha} = \operatorname{int} L_\alpha$, so $z_0 \in \operatorname{int} L_\alpha$. Thus $\widetilde{f} \,\square\, \widetilde{g}$ is locally bounded above, and so $f \,\square\, g$ is also locally bounded above; hence $f \,\square\, g$ is locally Lipschitz around every point in its domain. □

B.2.4 Nonsmooth analysis: Beyond convex analysis

First we need some terminology for *smooth* functions. In particular, we say $f\colon X \to Y$ (where X and Y are Banach spaces) is *Fréchet differentiable* at $x \in X$ with derivative $\nabla f(x)$, a continuous linear map $X \to Y$, if

$$\lim_{\|h\| \to 0} \frac{f(x + h) - f(x) - \nabla f(x) h}{\|h\|} = 0. \tag{B.31}$$

We say that $f\colon X \to Y$ is *Gateaux differentiable* at $x \in X$ with derivative $\nabla f(x)$, a continuous linear map $X \to Y$, if for all $v \in X$,

$$\lim_{h \to 0} \frac{f(x + tv) - f(x)}{t} = \nabla f(x) v. \tag{B.32}$$

There are "weak" versions of these concepts where the limit is understood as a weak, rather than strong, limit.

For nonsmooth and nonconvex functions there are generalizations of subdifferentials. The best known are the *generalized gradients* of Clarke [53, 54, 55]. The definition of these is a little involved, requiring a two-step definition. Given a locally Lipschitz function

$\phi\colon X \to \mathbb{R}$, we define the Clarke directional derivative

$$\phi^\circ(x;d) = \limsup_{x'\to x;\, d'\to d;\, h\downarrow 0} \frac{\phi(x'+hd') - \phi(x')}{h}. \tag{B.33}$$

Since ϕ is Lipschitz near x this is well defined and finite. Furthermore, $\phi^\circ(x;d)$ is a positively homogeneous ($\phi^\circ(x;\alpha d) = \alpha\,\phi^\circ(x;d)$ for $\alpha \geq 0$) and *convex* function of d. Then we can define the Clarke generalized gradient

$$\partial\phi(x) = \left\{ g \in X' \mid \langle g, d\rangle \leq \phi^\circ(x,;d) \text{ for all } d \in X \right\}. \tag{B.34}$$

If $X = \mathbb{R}^n$, then

$$\partial\phi(x) = \overline{\text{co}}\left\{ \lim_{k\to\infty} \nabla\phi(x_k) \mid x_k \to x \text{ as } k \to \infty \right\}. \tag{B.35}$$

This relies on *Rademacher's theorem*, which says that a Lipschitz function $\mathbb{R}^n \to \mathbb{R}$ is differentiable almost everywhere. This can be generalized to certain Banach spaces called *Asplund spaces*. Asplund spaces X have the property that any convex function $\phi\colon X \to \mathbb{R}$ is Frechét differentiable on a G_δ set that is dense in $\text{dom}\,\phi$ (see [16]). In fact, every locally Lipschitz function $\phi\colon X \to \mathbb{R}$ is differentiable except on a dense subset of X (see [210]), and (B.35) can be used for the Clarke generalized gradient. A Banach space X is an Asplund space if and only if its dual X' has the RNP.

Clarke generalized gradients can be used for optimization: $0 \in \partial\phi(x^*)$ is now a *necessary* but not sufficient condition for x^* to be a minimizer of a locally Lipschitz function ϕ. The condition $0 \in \partial\phi(x^*)$ does not even mean that the directional derivatives $\phi'(x^*;d) \geq 0$ for all d. The example of $\phi(x) = -|x|$ has $0 \in \partial\phi(0)$, but $\phi'(x^*;\pm 1) = -1$. If $\phi'(x;d) = \phi^\circ(x;d)$ for all d, we say that ϕ is *Clarke regular* at x. If, for all z in a neighborhood of x, we have $\phi(z) \geq \phi(x) + \phi^\circ(x;d) - r\,\|x-z\|^2$ for a fixed r, we say ϕ is *r-prox regular*, or just *prox-regular*, if we do not wish to specify a particular r.

An alternative to Clarke's generalized gradients for locally Lipschitz functions is the *Bouligand generalized gradient* given by

$$\partial_B\phi(x) = \left\{ \lim_{k\to\infty} \nabla\phi(x_k) \mid x_k \to x \text{ as } k \to \infty \right\}. \tag{B.36}$$

From (B.35) and (B.36) we can develop Bouligand and Clarke *generalized Jacobian* matrices for locally Lipschitz $\psi\colon \mathbb{R}^n \to \mathbb{R}^m$:

$$\partial_B\psi(x) = \left\{ \lim_{k\to\infty} \nabla\psi(x_k) \mid x_k \to x \text{ as } k \to \infty \right\},$$

$$\partial\psi(x) = \overline{\text{co}}\left\{ \lim_{k\to\infty} \nabla\phi(x_k) \mid x_k \to x \text{ as } k \to \infty \right\}.$$

The Clarke generalized Jacobian allows us to generalize the inverse and implicit functions to locally Lipschitz functions: if $\psi\colon \mathbb{R}^n \to \mathbb{R}^n$ and $\partial\psi(x^*)$ contains no singular matrix, then ψ is a local homeomorphism; in fact, there is a ball $y^* + r\,B$ with $y^* = \psi(x^*)$ where there is a Lipschitz inverse function ψ^{-1} defined on $y^* + r\,B$ with $r > 0$ and $\psi^{-1}(y^*) = x^*$ (see [113]).

Other versions of nonsmooth variational analysis and "generalized gradients" or "generalized subdifferentials" can be found in, for example, the books of Mordukhovich [175], Rockafellar and Wets [214], and Borwein and Zhu [36]. These different versions of variational analysis can operate in different domains (finite-dimensional versus infinite-dimensional spaces), different kinds of functions ("smooth + convex," Lipschitz, or lower semicontinuous), or with different applications in mind (control theory, differential equations, optimization, or solution of equations). We will not venture further into these different notions of variational analysis and generalized differentiation except to say that while many of the concepts developed elsewhere are sharper than, say, the Clarke generalized gradient, they are often more difficult to compute and have less regularity. Nevertheless, they often have applicability to numerous practical problems.

Appendix C

Differential Equations

We will consider here differential equation initial value problems of the form

$$\frac{dx}{dt}(t) = f(t, x(t)), \qquad x(t_0) = x_0 \in X, \tag{C.1}$$

where X is a given Banach space. We take this to mean that dx/dt exists almost everywhere, $x(\cdot)$ is an absolutely continuous function, $dx/dt(t) = f(t, x(t))$ for almost all t, and finally that $x(t_0) = x_0$. This understanding of what a solution of a differential equation is derives from the work of *Carathéodory*; we say that such a function is a solution in the sense of Carathéodory.

C.1 Existence theory for Lipschitz ordinary differential equations

The most basic result in the theory of differential equation initial value problems is that if $f(t,x)$ is a Lipschitz function of x, then a solution to (C.1) exists and is unique. A refinement of this result is given below.

Theorem C.1. *If $\|f(t,x) - f(t,z)\| \leq L(t)\,\|x - y\|$ and $\|f(t,0)\| \leq k(t)$ for all t, x, and y, and if L, k are locally integrable functions, then solutions* (C.1) *exist to in the sense of Carathéodory. Furthermore, if $z(\cdot)$ is a solution of $dz/dt = f(t, z(t))$, $z(t_0) = z_0$ in the same sense, then $\|x(t) - z(t)\| \leq \exp(\int_{t_0}^{t} L(\tau)\,d\tau)\,\|x_0 - z_0\|$.*

Proof. We prove this using *Picard iteration*. First we write the problem in integral form:

$$x(t) = x_0 + \int_{t_0}^{t} f(\tau, x(\tau))\,d\tau.$$

Next we create a fixed point iteration:

$$x^{(k+1)}(t) = x_0 + \int_{t_0}^{t} f(\tau, x^{(k)}(\tau))\,d\tau, \qquad x^{(0)}(t) = x_0 \quad \text{for all } t.$$

The map $z(\cdot) \mapsto y(\cdot)$ given by

$$y(t) = x_0 + \int_{t_0}^{t} f(\tau, z(\tau))\, d\tau$$

is a Lipschitz continuous map $C[t_0, t_1] \to C[t_0, t_1]$ with Lipschitz constant $\int_{t_0}^{t_1} L(\tau)\, d\tau$. To see this, if $z_1 \mapsto y_1$ and $z_2 \mapsto y_2$, then

$$\begin{aligned} \|y_2(t) - y_1(t)\| &\le \int_{t_0}^{t} \|f(\tau, z_2(\tau)) - f(\tau, z_1(\tau))\|\, d\tau \\ &\le \int_{t_0}^{t} L(\tau)\, \|z_2(\tau) - z_1(\tau)\|\, d\tau \\ &\le \left(\int_{t_0}^{t_1} L(\tau)\, d\tau \right) \sup_{t_0 \le t \le t_1} \|z_2(t) - z_1(t)\|. \end{aligned}$$

Thus if $t_1 > t_0$ is chosen sufficiently small, then $\int_{t_0}^{t_1} L(\tau)\, d\tau < 1$, and so the iteration is a contraction map. Applying the contraction mapping theorem, there is one and only one fixed point of the iteration which is the limit $x(t) = \lim_{k\to\infty} x^{(k)}(t)$, which is the solution of (C.1) on $[t_0, t_1]$.

The solution can be extended to an interval $[t_1, t_2]$ where $\int_{t_1}^{t_2} L(\tau)\, d\tau < 1$ with the "initial" value $x(t_1)$ given from the solution on $[t_0, t_1]$. This can be repeated; suppose we choose t_{k+1} at stage k so that $\int_{t_k}^{t_{k+1}} L(\tau)\, d\tau = 1/2$. Then $t_\infty := \lim_{k\to\infty} t_k = +\infty$; otherwise $\int_{t_0}^{t_\infty} L(\tau)\, d\tau = +\infty$, which contradicts the assumption that L is locally integrable. Thus a solution exists for all $t \ge t_0$.

The bound on $\|x(t) - z(t)\|$ for different initial conditions follows by the Gronwall lemmas of the following section. □

Note that this theorem works in infinite dimensions as well as finite dimensions. There is no need to impose compactness conditions. However, there are situations where existence can be shown for initial value problems which rely on some compactness properties.

C.2 Gronwall-type lemmas

To carry out proofs for more general situations, we need a lemma originally due to Gronwall [116], which in Bellman's integral form [28] is something like this.

Lemma C.2. *Suppose that $r\colon [a,b] \to \mathbb{R}$ is an absolutely continuous function such that $r'(t) \le \beta(t) r(t)$ for almost all t with $\beta\colon [a,b] \to \mathbb{R}$ integrable; then*

$$r(t) \le r(a) \exp\left(\int_a^t \beta(\tau)\, d\tau \right).$$

This result can be easily derived from the integral version of the lemma.

Lemma C.3. *Suppose that $r\colon [a,b] \to \mathbb{R}$ is an integrable function satisfying*

$$r(t) \le \alpha(t) + \int_a^t \beta(\tau) r(\tau)\, d\tau, \qquad \alpha, \beta \in L^1(a,b).$$

Then

$$r(t) \le \alpha(t) + \int_a^t \alpha(\tau)\beta(\tau)\exp\left(\int_\tau^t \beta(s)\,ds\right)d\tau$$

for all t. If $\alpha(t)$ is a constant α, this implies that

$$r(t) \le \alpha \exp\left(\int_a^t \beta(\tau)\,d\tau\right) \qquad \text{for all } t.$$

In fact, the above results hold if β is a measure on $[a,b]$ with $\beta(\{a\}) = 0$ and α is β-integrable.

There are numerous variations and generalizations of these results, many of which require positivity or monotonicity. Here is an example of a nonlinear Gronwall lemma, which in integral form is due to Bihari [32].

Lemma C.4. *Suppose that $r\colon [a,b] \to \mathbb{R}$ is an absolutely continuous function and that $\theta\colon \mathbb{R} \to \mathbb{R}$ is a positive and nondecreasing function. Then, if $r'(t) \le \theta(r(t))$ for almost all t, then $r(t) \le \rho(t)$ for all $t \in [a,b]$, where ρ is the unique solution of the differential equation*

$$\rho'(t) = \theta(\rho(t)), \qquad \rho(a) = r(a),$$

provided that $\int_{r(a)}^{\infty} ds/\theta(s) > b - a$.

In this version, θ might be a function that grows superlinearly (for example, $\theta(r) = 1 + r^2$), and so it can be used to give short-time existence bounds. The existence and uniqueness of the solution, at least for a sufficiently short time interval, are not obvious a priori, so this must be shown in the proof.

Proof. To show the existence of a solution to $\rho' = \theta(\rho)$, we define ρ to be the inverse function to $\psi(s) := a + \int_{r(a)}^{s} ds'/\theta(s')$. Now $s' \mapsto 1/\theta(s')$ is a locally bounded, positive, nonincreasing function, and so it is integrable. Thus ψ is absolutely continuous, and $\psi'(s) = 1/\theta(s)$ for almost all s. Hence, on finite intervals $[s_1, s_2]$, $1/\theta(s_2) \le \psi'(s) \le 1/\theta(s_1)$ for all $s \in [s_1, s_2]$. This implies that the inverse function ρ exists and is absolutely continuous on $[a, b] \subset \operatorname{range}\psi$. The usual differentiation rules show that $\rho'(t) = \theta(\rho(t))$ for almost all t; $\rho(a) = r(a)$ as $\psi(r(a)) = a$. Thus there is a solution of the differential equation for ρ.

To show that there is only one solution to the differential equation, suppose that $\rho' = \theta(\rho)$ and $\rho(a) = r(a)$. Then since θ is positive and bounded on bounded intervals, $(d/dt)(\psi(\rho(t)) = (1/\theta(\rho(t)))\,\rho'(t) = 1$ for almost all t, where ψ is the function defined in the previous paragraph. This together with $\psi(r(a)) = a$ implies that ψ and ρ are inverse functions. Therefore ρ is unique, as ψ is unique.

Now, to show the bound, we first note that $(d/dt)\psi(r(t)) = r'(t)/\theta(r(t)) \le 1$ for almost all t. Since $\psi(r(a)) = a$, it follows that $\psi(r(t)) \le t$. Since ψ is strictly increasing, this means that $r(t) \le \rho(t)$, as desired. □

This result cannot only be used to show local boundedness when $\theta(r)$ grows superlinearly in r, but it can also be used to get subexponential bounds when $\theta(r)$ grows sublinearly

in r. An example of this is in mechanical systems with bounded external forces $\mathbf{f}(t)$:

$$m\frac{d^2\mathbf{x}}{dt^2} = -\nabla V(\mathbf{x}) + \mathbf{f}(t).$$

The energy is $E(t) = \frac{1}{2}m\,\|\dot{\mathbf{x}}\|^2 + V(\mathbf{x}) \geq \inf_{\mathbf{x}} V(\mathbf{x}) =: E_{min}$. The rate of change of the energy can be bounded by

$$\begin{aligned}\frac{dE}{dt} &\leq \mathbf{f}(t)^T\dot{\mathbf{x}}(t)\\ &\leq \|\mathbf{f}(t)\|\,\|\dot{\mathbf{x}}(t)\|\\ &\leq C\,(E - E_{min})^{1/2}.\end{aligned}$$

Then the energy is bounded by the solution of $d\rho/dt = C\,(\rho - E_{min})^{1/2}$, $\rho(t_0) = E(t_0) > E_{min}$, which is $\rho(t) = E_{min} + (2C\,t + \sqrt{E(t_0) - E_{min}})^2$, showing that the energy grows at most quadratically in time for bounded external forces.

Discrete Gronwall lemmas have also been developed for handling time discretizations of differential equations and related systems. The simplest of these starts with

$$r_{k+1} \leq r_k + h\,\beta\,r_k \qquad \text{for all } k$$

and obtains a bound

$$r_k \leq e^{\beta\,h\,k}\,r_0,$$

which depends only on the product $h\,k$ and r_0. Nonlinear versions have also been developed. One of these is Lemma 5.2 in Section 5.2.1.

C.3 Carathéodory's existence theorem for continuous ordinary differential equations

Gronwall lemmas (both continuous and discrete) can be used to show existence of solutions to differential equations and inclusions. For example, consider Carathéodory's existence theorem for ordinary differential equations in finite dimensions with merely *continuous* right-hand side (in the state variable). More formally, we have the following theorem.

Theorem C.5 (Carathéodory). *Consider the differential equation*

$$\frac{dx}{dt} = f(t,x), \qquad x(t_0) = x_0 \in \mathbb{R}^n,$$

where $f(t,x)$ is continuous in x and measurable in t, with bounds $\|f(t,x)\| \leq \psi(t)\,\theta(\|x\|)$ where θ is continuous and monotone, and $0 < \psi \in L^1_{loc}(\mathbb{R})$. Then solutions exist for the above differential equation on a sufficiently small time interval $\left[0, T^*\right]$, $T^* > 0$.

Proof. Note that we can assume without loss of generality that $\psi(t) \geq 1$ for all t. We start by using a modified explicit Euler method:

$$x_{h,k+1} = x_{h,k} + \int_{t_k}^{t_{k+1}} f(\tau, x_{h,k})\,d\tau,$$

where $\int_{t_k}^{t_{k+1}} \psi(\tau)\,d\tau = h > 0$. Then

$$\begin{aligned}\|x_{h,k+1}\| &\le \|x_{h,k}\| + \int_{t_k}^{t_{k+1}} \|f(\tau, x_{h,k})\|\,d\tau \\ &\le \|x_{h,k}\| + \int_{t_k}^{t_{k+1}} \psi(\tau)\,d\tau\,\theta\left(\|x_{h,k}\|\right) \\ &= \|x_{h,k}\| + h\,\theta\left(\|x_{h,k}\|\right).\end{aligned}$$

Applying Lemma 5.2, then for sufficiently small $h > 0$ we have the bound

$$\|x_{h,k}\| \le \phi\left(\int_{t_0}^{t_k} \psi(\tau)\,d\tau\right) + 1 \qquad \text{where}$$

$$\frac{d\phi}{ds}(s) = \theta(\phi(s)), \qquad \phi(0) = \|x_0\|.$$

Since ϕ is bounded on a sufficiently small interval $[0, s^*]$, $s^* > 0$, then we have a uniform bound on $\|x_{h,k}\| \le \phi^* := \phi(s^*) + 1$ for all h.

Choose T^* so that $\int_{t_0}^{T^*} \psi(\tau)\,d\tau \le s^*$. Let $x_h(t)$ be given by

$$x_h(t) = x_{h,k} + \int_{t_k}^{t} f(\tau, x_{h,k})\,d\tau \qquad \text{for } t_k \le t \le t_{k+1}.$$

This is well defined on $[t_0, T^*]$ and absolutely continuous there since it is the indefinite integrable of a locally integrable function. Furthermore, the functions x_h are equicontinuous since

$$\|x_h(t) - x_h(s)\| \le (t - s)\,\theta(\phi^*) \qquad \text{for } t_0 \le s \le t \le T^*.$$

Since the discrete-time trajectories $x_h(t)$ are bounded by ϕ^* in $\mathbb{R}^n$, we can apply the Arzela–Ascoli theorem to conclude that there is a uniformly convergent subsequence (also denoted by x_h) with a limit $\widehat{x}$. The limit is clearly continuous. It is also a solution of the differential equation. To see this, let $\widetilde{x}_h(t) = x_{h,k}$ for $t_k \le t < t_{k+1}$. Then, for $s \ge t$,

$$\begin{aligned}\widehat{x}(s) - \widehat{x}(t) &= \lim_{h\to 0} x_h(s) - x_h(t) \\ &= \lim_{h\to 0} \int_t^s f(\tau, \widetilde{x}_h(\tau))\,d\tau \\ &= \int_t^s \lim_{h\to 0} f(\tau, \widetilde{x}_h(\tau))\,d\tau\end{aligned}$$

by the dominated convergence theorem. By continuity of $f(t, x)$ in x,

$$\lim_{h\to 0} f(\tau, \widetilde{x}_h(\tau)) = f(\tau, \lim_{h\to 0} \widetilde{x}_h(\tau)).$$

Now, for $t_k \le t < t_{k+1}$,

$$\begin{aligned}\|x_h(t) - \widetilde{x}_h(t)\| &= \left\| x_{h,k} + \int_{t_k}^{t} f(\tau, x_{h,k})\,d\tau - x_{h,k} \right\| \\ &\le \int_{t_k}^{t} \psi(\tau)\,d\tau\,\theta(\phi^*) \le h\,\theta(\phi^*),\end{aligned}$$

which goes to zero as $h \to 0$. Thus $\lim_{h\to 0} \widetilde{x}_h(t) = \widehat{x}(t)$ (taking the limit in the subsequence). So

$$\widehat{x}(t) - \widehat{x}(s) = \int_s^t f(\tau, \widehat{x}(\tau))\, d\tau$$

for all $t_0 \le s \le t \le T^*$ by the dominated convergence theorem. This immediately implies that $\widehat{x}\colon [t_0, T^*] \to \mathbb{R}^n$ is absolutely continuous, and so it is differentiable almost everywhere and its derivative is $d\widehat{x}/dt(t) = f(t, \widehat{x}(t))$ for almost all t. Finally, the initial conditions are correct, since $x_h(t_0) = x_0$ for all $h > 0$, so the limit is $\widehat{x}(t_0) = x_0$, as desired. □

Unlike the case of Lipschitz continuous right-hand sides, we cannot guarantee uniqueness of solutions. A simple counterexample is $dx/dt = \sqrt{|x|}$, $x(0) = 0$. Then $x(t) = 0$ and $x(t) = \frac{1}{4}t^2$ are both solutions.

C.4 Laplace and Fourier transforms

For dealing with linear differential equations, especially with constant coefficients, there are few tools better than the Laplace and Fourier transforms. Laplace transforms can usually be applied only to one variable. The Laplace transform of a measurable function $f\colon [0,\infty) \to X$ with X a Banach space is given by

$$\mathcal{L}f(s) = \int_0^\infty e^{-st} f(t)\, dt, \tag{C.2}$$

provided $t \mapsto e^{-st} f(t)$ is integrable. Essentially, f needs to just be measurable and have a growth rate that is at most exponential. Then for $\operatorname{Re} s$ sufficiently positive, $\mathcal{L}f(s)$ is well defined and analytic in s. Provided the relevant Laplace transforms are well defined, the following rules hold for constants α, β:

$$\begin{aligned}
\mathcal{L}\big[\alpha f + \beta g\big](s) &= \alpha\, \mathcal{L}f(s) + \beta\, \mathcal{L}g(s),\\
\mathcal{L}\big[f'\big](s) &= s\, \mathcal{L}f(s) - f(0),\\
\mathcal{L}\big[t\, f(t)\big](s) &= -\frac{d}{ds}\mathcal{L}f(s),\\
\mathcal{L}\big[e^{at} f(t)\big](s) &= \mathcal{L}f(s-a),\\
\mathcal{L}\big[f * g\big](s) &= \mathcal{L}f(s)\, \mathcal{L}g(s),
\end{aligned}$$

where $(f * g)(t) = \int_0^t f(\tau)\, g(t-\tau)\, d\tau$ is the convolution of two functions with domains $[0,\infty)$. There is the Laplace inversion formula: if $\mathcal{L}f(s) = g(s)$, then

$$f(t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{st}\, g(s)\, ds, \qquad i = \sqrt{-1}. \tag{C.3}$$

The number c should be chosen so that g is analytic on the half-plane $\{s \in \mathbb{C} \mid \operatorname{Re} s > c\}$ and $g(s) \to 0$ if $|s| \to \infty$ in this half-plane.

If $f : \mathbb{R}^d \to \mathbb{R}$, then the Fourier transform of f is given by

$$\mathcal{F}f(\xi) = \int_{\mathbb{R}^d} e^{-i\langle x,\xi\rangle} f(x)\,dx. \tag{C.4}$$

Now $\mathcal{F}f$ is defined for all f that are "sufficiently regular" (e.g., if f is integrable). If $\mathcal{F}f$ is also "sufficiently regular," then f can be recovered by the formula

$$\mathcal{F}^{-1}f(x) = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} e^{i\langle x,\xi\rangle}\,\mathcal{F}f(\xi)\,d\xi. \tag{C.5}$$

The main properties of the Fourier transform are, provided all integrals and derivatives are well defined,

$$\begin{aligned}
\mathcal{F}\big[\alpha f + \beta g\big](\xi) &= \alpha\,\mathcal{F}f(\xi) + \beta\,\mathcal{F}g(\xi),\\
\mathcal{F}\left[\frac{\partial f}{\partial x_k}(x)\right](\xi) &= i\xi_k\,\mathcal{F}f(\xi),\\
\mathcal{F}\big[x_k\,f(x)\big](\xi) &= -i\frac{\partial}{\partial \xi_k}\mathcal{F}f(\xi),\\
\mathcal{F}\big[f * g\big](\xi) &= \mathcal{F}f(\xi)\,\mathcal{F}g(\xi),
\end{aligned}$$

where

$$(f * g)(x) = \int_{\mathbb{R}^d} f(y)\,g(x-y)\,dy.$$

Perhaps the most important property is Plancherel's theorem. A modern version of this theorem follows.

Theorem C.6. *If $f, g \in L^2(\mathbb{R}^d)$, then*

$$\int_{\mathbb{R}^d} \overline{f(x)}\,g(x)\,dx = (2\pi)^{-d}\int_{\mathbb{R}^d} \overline{\mathcal{F}f(\xi)}\,\mathcal{F}g(\xi)\,d\xi,$$

where $\overline{(\cdot)}$ is the complex conjugate of $(\cdot)$.

This can be used to extend the Fourier transforms to tempered distributions. Let $\mathcal{S}(\mathbb{R}^d)$ be the set of functions ϕ where

$$x \mapsto x^\beta\,D^\alpha\phi(x) := x_1^{\beta_1}x_2^{\beta_2}\cdots x_d^{\beta_d}\,\frac{\partial^{\alpha_1+\cdots+\alpha_d}\phi}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\cdots\partial x_d^{\alpha_d}}$$

is bounded for all multi-indices α and β. Using the maximum values of these functions as seminorms, we have $\phi_k \to \phi$ in $\mathcal{S}(\mathbb{R}^d)$ if $\max_x \left|x^\beta\,D^\alpha(\phi_k - \phi)(x)\right| \to 0$ as $k \to \infty$ for all multi-indices α and β. It can be easily shown using the above rules that the Fourier transform maps $\mathcal{S}(\mathbb{R}^d)$ into itself. Tempered distributions are the dual space $\mathcal{S}(\mathbb{R}^d)'$ of functionals $\mathcal{S}(\mathbb{R}^d) \to \mathbb{R}$. For a tempered distribution ψ we define its Fourier transform via

$$\langle \mathcal{F}\psi, \mathcal{F}\phi\rangle = (2\pi)^d\,\langle \psi, \phi\rangle \qquad \text{for all } \phi \in \mathcal{S}(\mathbb{R}^d).$$

The same rules apply to the Fourier transform for tempered distributions, but the operations (such as differentiation) must be considered in a distributional sense.

Bibliography

[1] Robert A. Adams and John J. F. Fournier, *Sobolev spaces*, second ed., Pure and Applied Mathematics (Amsterdam), vol. 140, Elsevier/Academic Press, Amsterdam, 2003.

[2] Jeongho Ahn, *A vibrating string with dynamic frictionless impact*, Appl. Numer. Math. **57** (2007), no. 8, 861–884.

[3] Jeongho Ahn and David E. Stewart, *Existence of solutions for a class of impact problems without viscosity*, SIAM J. Math. Anal. **38** (2006), no. 1, 37–63.

[4] Charalambos D. Aliprantis and Kim C. Border, *Infinite dimensional analysis*, third ed., Springer, Berlin, 2006.

[5] Eugene Allgower and Kurt Georg, *Simplicial and continuation methods for approximating fixed points and solutions to systems of equations*, SIAM Rev. **22** (1980), no. 1, 28–85.

[6] ———, *Numerical continuation methods: An introduction*, Springer-Verlag, Berlin, Heidelberg, New York, 1990.

[7] Eugene L. Allgower and Kurt Georg (eds.), *Computational solution of nonlinear systems of equations*, Lectures in Appl. Math. 26, Amer. Math. Soc., 1990, Proceedings of the 1988 SIAM–AMS summer seminar on Computational Solution of Nonlinear Systems of Equations at Colorado.

[8] Luigi Amerio, *Continuous solutions of the problem of a string vibrating against an obstacle*, Rend. Sem. Mat. Univ. Padova **59** (1978), 67–96 (1979).

[9] Guillame Amontons, *De la resistance causée dans les machines*, Memoires de l'Academie Royale A (1699), 257–282, (Chez Gerard Kuyper, Amsterdam, 1706).

[10] Lars-Erik Andersson, *A global existence result for a quasistatic contact problem with friction*, Adv. Math. Sci. Appl. **5** (1995), no. 1, 249–286.

[11] Mihai Anitescu, James Cremer, and Florian A. Potra, *Formulating three-dimensional contact dynamics problems*, Mech. Structures Mach. **24** (1996), no. 4, 405–437.

[12] Mihai Anitescu and Florian A. Potra, *Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems*, Nonlinear Dynam. **14** (1997), no. 3, 231–247.

[13] ———, *Time-stepping schemes for stiff multi-rigid-body dynamics with contact and friction*, submitted to *Internat. J. Numer. Methods Eng.*.

[14] Mihai Anitescu, Florian A. Potra, and David E. Stewart, *Time-stepping for three-dimensional rigid body dynamics*, Comput. Methods Appl. Mech. Engrg. **177** (1999), 183–197.

[15] Uri M. Ascher and Linda R. Petzold, *Computer methods for ordinary differential equations and differential-algebraic equations*, SIAM, Philadelphia, 1998.

[16] Edgar Asplund, *Fréchet differentiability of convex functions*, Acta Math. **121** (1968), 31–47.

[17] Kendall E. Atkinson, *An introduction to numerical analysis*, first ed., J. Wiley and Sons, 1978.

[18] Kendall E. Atkinson, Weimin Han, and David E. Stewart, *Numerical solution of ordinary differential equations*, Pure and Applied Mathematics, J. Wiley and Sons, Hoboken, NJ, 2009.

[19] Jean-Pierre Aubin and Arigo Cellina, *Differential inclusions: Set-valued maps and viability theory*, Springer-Verlag, Berlin, New York, 1984.

[20] Jean-Pierre Aubin and Ivar Ekeland, *Applied nonlinear analysis*, Dover Publications Inc., Mineola, NY, 2006, Reprint of the 1984 original.

[21] Jean-Pierre Aubin and Hélène Frankowska, *Set-valued analysis*, Progress in Systems and Control, no. 2, Birkhäuser, Boston, Basel, Berlin, 1990.

[22] Robert J. Aumann, *Integrals of set-valued functions*, J. Math. Anal. Appl. **12** (1965), 1–12.

[23] Claudio Baiocchi and António Capelo, *Variational and quasivariational inequalities: Applications to free boundary problems*, Wiley, Chichester, New York, 1984.

[24] Patrick Ballard, *The dynamics of discrete mechanical systems with perfect unilateral constraints*, Arch. Ration. Mech. Anal. **154** (2000), no. 3, 199–274.

[25] Viorel Barbu, *Optimal control of variational inequalities*, Research Notes in Mathematics, vol. 100, Pitman (Advanced Publishing Program), Boston, MA, 1984.

[26] F. Bashforth and J. C. Adams, *An attempt to test the theories of capillary action*, Cambridge University Press, London, 1883.

[27] Jérôme Bastien and Michelle Schatzman, *Numerical precision for differential inclusions with uniqueness*, M2AN Math. Model. Numer. Anal. **36** (2002), no. 3, 427–460.

[28] Richard Bellman, *The stability of solutions of linear differential equations*, Duke Math. J. **10** (1943), 643–647.

[29] Jöran Bergh and Jörgen Löfström, *Interpolation spaces*, Grundlehren der mathmatischen Wissenschaften, vol. 223, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[30] Alain Bernard and Ahmed el Kharroubi, *Régulations déterministes et stochastiques dans le premier "orthant" de* $\mathbf{R}^n$, Stochastics Stochastics Rep. **34** (1991), no. 3-4, 149–167.

[31] Wolf-Jürgen Beyn and Janosch Rieger, *Numerical fixed grid methods for differential inclusions*, Computing **81** (2007), no. 1, 91–106.

[32] Imre Bihari, *A generalization of a lemma of Bellman and its application to uniqueness problems of differential equations*, Acta Math. Acad. Sci. Hungar. **7** (1956), 81–94.

[33] Béla Bollobás, *Graph theory*, Graduate Texts in Mathematics, vol. 63, Springer-Verlag, New York, 1979.

[34] Jonathan M. Borwein, *Maximality of sums of two maximal monotone operators*, Proc. Amer. Math. Soc. **134** (2006), no. 10, 2951–2955.

[35] ———, *Maximality of sums of two maximal monotone operators in general Banach space*, Proc. Amer. Math. Soc. **135** (2007), no. 12, 3917–3924.

[36] Jonathan M. Borwein and Qiji J. Zhu, *Techniques of variational analysis*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, vol. 20, Springer-Verlag, New York, 2005.

[37] Frank Philip Bowden and David Tabor, *Friction: An introduction to tribology*, Anchor Press/Doubleday, Garden City, NY, 1973.

[38] Kathryn E. Brenan, Stephen L. Campbell, and Linda R. Petzold, *Numerical solution of initial-value problems in differential-algebraic equations*, Classics in Applied Mathematics, vol. 14, SIAM, Philadelphia, 1996, Originally published by North–Holland, 1989.

[39] Richard P. Brent, *Algorithms for minimization without derivatives*, Prentice–Hall Series in Automatic Computation, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[40] Haïm Brézis, *Équations et inéquations non linéaires dans les espaces vectoriels en dualité*, Ann. Inst. Fourier (Grenoble) **18** (1968), no. 1, 115–175.

[41] ———, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North–Holland Mathematics Studies, no. 5, Notas de Matemática (50), North–Holland, Amsterdam, 1973.

[42] Bernard Brogliato, *Nonsmooth impact mechanics: Models, dynamics and control*, Lecture Notes in Control and Information Sciences, no. 220, Springer, Berlin, Heidelberg, New York, 1996.

[43] Felix E. Browder, *Nonlinear operators and nonlinear equations of evolution in Banach spaces*, Nonlinear Functional Analysis (Proc. Sympos. Pure Math., Vol. XVIII, Part 2, Chicago, Ill., 1968), Amer. Math. Soc., Providence, RI, 1976, pp. 1–308.

[44] Richard L. Burden and J. Douglas Faires, *Numerical analysis*, 7th ed., Wadsworth Group, Pacific Grove, CA, 2001.

[45] Kevin Burrage and John C. Butcher, *Stability criteria for implicit Runge–Kutta methods*, SIAM J. Numer. Anal. **16** (1979), no. 1, 46–57.

[46] John C. Butcher, *Numerical methods for ordinary differential equations*, second ed., John Wiley & Sons Ltd., Chichester, 2008.

[47] M. Kanat Çamlıbel, W. P. M. H. Heemels, and J. M. Schumacher, *Consistency of a time-stepping method for a class of piecewise-linear networks*, IEEE Trans. Circuits Syst. I Fund. Theory Appl. **49** (2002), no. 3, 349–357.

[48] M. Kanat Çamlıbel, W. P. M. H. Heemels, and J. M. Schumacher, *On linear passive complementarity systems*, Eur. J. Control **8** (2002), no. 3, 220–237.

[49] Charles Castaing and Manuel D. P. Monteiro Marques, *Sweeping processes by nonconvex closed moving sets with perturbation*, C. R. Acad. Sci. Paris Sér. I Math. **319** (1994), no. 2, 127–132.

[50] Anindya Chatterjee, *On the realism of complementarity conditions in rigid-body collisions*, Nonlinear Dynam. **20** (1999), no. 2, 159–168.

[51] T. J. Chung, *General continuum mechanics*, Cambridge University Press, Cambridge, UK, 2007.

[52] Claudio Citrini and Clelia Marchionna, *Some unilateral problems for the vibrating string equation*, Eur. J. Mech. A Solids **8** (1989), no. 1, 73–85.

[53] Frank H. Clarke, *Generalized gradients and applications*, Trans. Amer. Math. Soc. **205** (1975), 247–262.

[54] ———, *Methods of dynamic and nonsmooth optimization*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 57, SIAM, Philadelphia, 1989.

[55] ———, *Optimization and nonsmooth analysis*, SIAM, Philadelphia, 1990, Originally published by the Canadian Math. Soc., 1983.

[56] Marius Cocou, *Existence of solutions of a dynamic Signorini's problem with nonlocal friction in viscoelasticity*, Z. Angew. Math. Phys. **53** (2002), no. 6, 1099–1109.

[57] Marius Cocou, Elaine Pratt, and Michel Raous, *Formulation and approximation of quasistatic frictional contact*, Internat. J. Engrg. Sci. **34** (1996), no. 7, 783–798.

[58] Marius Cocou and Jean-Marc Ricaud, *Existence results for a class of implicit evolution inequalities and applications to dynamic unilateral contact problems with friction*, C. R. Acad. Sci. Paris Sér. I Math. **329** (1999), no. 9, 839–844.

[59] ———, *Analysis of a class of implicit evolution inequalities associated to viscoelastic dynamic contact problems with friction*, Internat. J. Engrg. Sci. **38** (2000), no. 14, 1535–1552.

[60] Marius Cocou and Rémi Rocca, *Existence results for unilateral quasistatic contact problems with friction and adhesion*, M2AN Math. Model. Numer. Anal. **34** (2000), no. 5, 981–1001.

[61] Donald L. Cohn, *Measure theory*, Birkhäuser, Boston, 1980.

[62] Monica-Gabriela Cojocaru, Patrizia Daniele, and Anna Nagurney, *Projected dynamical systems and evolutionary variational inequalities via Hilbert spaces with applications*, J. Optim. Theory Appl. **127** (2005), no. 3, 549–563.

[63] Monica-Gabriela Cojocaru and Leo B. Jonker, *Existence of solutions to projected differential equations in Hilbert spaces*, Proc. Amer. Math. Soc. **132** (2004), no. 1, 183–193.

[64] Thomas H. Cormen, Charles E. Leiserson, and Ronald L. Rivest, *Introduction to algorithms*, 2nd ed., MIT Press, Cambridge, MA, 2001.

[65] Bernard Cornet, *Existence of slow solutions for a class of differential inclusions*, J. Math. Anal. Appl. **96** (1983), no. 1, 130–147.

[66] Richard W. Cottle and George B. Dantzig, *Complementary pivot theory of mathematical programming*, Linear Algebra Appl. **1** (1968), no. 1, 103–125.

[67] Richard W. Cottle, Jong-Shi Pang, and Richard E. Stone, *The linear complementarity problem*, Series on Computer Science and Scientific Computing, Academic Press, Boston, San Diego, New York, 1992.

[68] Michel Crouzeix, *Sur la B-stabilité des méthodes de Runge-Kutta*, Numer. Math. **32** (1979), no. 1, 75–82.

[69] C. F. Curtiss and J. O. Hirschfelder, *Integration of stiff equations*, Proc. Nat. Acad. Sci. U. S. A. **38** (1952), 235–243.

[70] Leonardo da Vinci, *Codex arundel (ff.* 40*v,* 41*r)*, Available via *Turning the Pages* at http://www.bl.uk/onlinegallery/ttp/ttpbooks.html, prepared in the late 1400s.

[71] Charles A. de Coulomb, *Théorie des machines simples, en ayant égard au frottement de leurs parties, et la roideur des cordages. Piéce qui a reporté le Prix double de l'Académie des Sciences pour l'année* 1781, Mémoires des Savans Etrangers **X** (1785), 163–332, Réimprimé 1809 chez Bachelier, Paris.

[72] Djairo G. de Figueiredo, *An existence theorem for pseudo-monotone operator equations in Banach spaces*, J. Math. Anal. Appl. **34** (1971), 151–156.

[73] Klaus Deimling, *Multivalued differential equations*, Series on Nonlinear Analysis and Applications 1, Walter de Gruyter, Berlin, New York, 1992.

[74] Kees Dekker and Jan G. Verwer, *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, CWI Monographs, vol. 2, North–Holland, Amsterdam, 1984.

[75] T. J. Dekker, *Finding a zero by means of successive linear interpolation*, Constructive Aspects of the Fundamental Theorem of Algebra (Proc. Sympos., Zürich-Rüschlikon, 1967), Wiley-Interscience, New York, 1969, pp. 37–48.

[76] Étienne Delassus, *Considérations sur le frottement de glissement*, Nouv. Ann. de Math. (4ème série) **20** (1920), 485–496.

[77] ———, *Sur les lois du frottement de glissement*, Bull. Soc. Math. France **51** (1923), 22–33.

[78] J. Diestel and J. J. Uhl, Jr., *Vector measures*, Mathematical Surveys, vol. 15, Amer. Math. Soc., Providence, RI, 1977.

[79] Reinhard Diestel, *Graph theory*, third ed., Graduate Texts in Mathematics, vol. 173, Springer-Verlag, Berlin, 2005.

[80] Nicolae Dinculeanu, *Vector measures*, Pergamon Press, London, New York, 1967.

[81] Asen Dontchev and Frank Lempio, *Difference methods for differential inclusions: A survey*, SIAM Rev. **34** (1992), 263–294.

[82] D. Dowson, *History of tribology*, 2nd ed., Professional Engineering Publishers, 1997, First edition, 1975.

[83] Paul Dupuis and Anna Nagurney, *Dynamical systems and variational inequalities*, Ann. Oper. Res. **44** (1993), no. 1-4, 9–42.

[84] Georges Duvaut and Jacques-Louis Lions, *Inequalities in mechanics and physics*, Grundlehren der mathematischen Wissenschaften, vol. 219, Springer-Verlag, Berlin, Heidelberg, New York, 1976, Original in French: "Les inéquations en mécanique et en physique" (1972).

[85] Christof Eck and Jiří Jarušek, *Existence results for the static contact problem with Coulomb friction*, Math. Models Methods Appl. Sci. **8** (1998), no. 3, 445–468.

[86] ———, *Existence results for the semicoercive static contact problem with Coulomb friction*, Nonlinear Anal. **42** (2000), no. 6, 961–976.

[87] ———, *A survey on dynamic contact problems with Coulomb friction*, Multifield Problems, Springer, Berlin, 2000, pp. 254–261.

[88] ———, *Existence of solutions for the dynamic frictional contact problem of isotropic viscoelastic bodies*, Nonlinear Anal. **53** (2003), no. 2, 157–181.

[89] Christof Eck, Jiří Jarušek, and Miroslav Krbec, *Unilateral contact problems*, Pure and Applied Mathematics (Boca Raton), vol. 270, Chapman & Hall/CRC, Boca Raton, FL, 2005.

[90] Charles M. Elliott, *On the convergence of a one-step method for the numerical solution of ordinary differential inclusions*, IMA J. Numer. Anal. **5** (1985), 3–21.

[91] Michael Erdmann, *On a representation of friction in configuration space*, Int. J. Robotics Research **13** (1994), no. 3, 240–271.

[92] Leonhard Euler, *Sue le frottement des corps solides*, Mémoires Acad. Sci. Berlin **4** (1748), 122–132.

[93] ———, *Sur la diminution de la resistance du frottement*, Mémoires Acad. Sci. Berlin **4** (1748), 133–148.

[94] Lawrence C. Evans, *Partial differential equations*, second ed., Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 2010.

[95] Francisco Facchinei and Jong-Shi Pang, *Finite-dimensional variational inequalities and complementarity problems I*, Springer Series in Operations Research, Springer, New York, 2003.

[96] ———, *Finite-dimensional variational inequalities and complementarity problems* II, Springer Series in Operations Research, Springer, New York, 2003.

[97] Jacques Faraut and Adam Korányi, *Analysis on symmetric cones*, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 1994.

[98] Gaetano Fichera, *Sul problema elastostatico di Signorini con ambigue condizioni al contorno*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8) **34** (1963), 138–142.

[99] ———, *Problemi elastostatici con vincoli unilaterali: Il problema di Signorini con ambigue condizioni al contorno*, Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. I (8) **7** (1963/1964), 91–140.

[100] ———, *La nascita della teoria delle disequazioni variazionali ricordata dopo trent'anni*, Incontro scientifico italo-spagnolo: organizzato in collaborazione con la Real Academia de Ciencias Exactas, Fisicas y Naturales e la Real Academia Nacional de Medicinia (Roma, 21 ottobre 1993), Atti dei Convegni Lincei, vol. 114, Accademia nazionale dei Lincei, 1995, pp. 47–53.

[101] Alekseĭ F. Filippov, *On some questions in the theory of optimal regulation: Existence of a solution of the problem of optimal regulation in the class of bounded measurable functions*, Vestnik Moskov. Univ. Ser. Mat. Meh. Astr. Fiz. Him. **1959** (1959), no. 2, 25–32.

[102] ———, *Differential equations with discontinuous right-hand side*, Amer. Math. Soc. Transl. **42** (1964), 199–231, Original in Russian in Math. Sbornik **5** (1960), 99–127.

[103] ———, *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control **5** (1967), no. 4, 609–621.

[104] ———, *Differential equations with discontinuous right-hand side*, Kluwer Academic Publishers, Dordrecht, 1988.

[105] Simon Fitzpatrick, *Representing monotone operators by convex functions*, Workshop/Miniconference on Functional Analysis and Optimization (Canberra, 1988), Proc. Centre Math. Anal. Austral. Nat. Univ., vol. 20, Austral. Nat. Univ., Canberra, 1988, pp. 59–65.

[106] Joel N. Franklin, *Methods of mathematical economics*, Classics in Applied Mathematics, vol. 37, SIAM, Philadelphia, 2002, Linear and nonlinear programming, fixed-point theorems, Reprint of the 1980 original.

[107] Terry L. Friesz, David Bernstein, Tony E. Smith, Roger L. Tobin, and B.-W. Wie, *A variational inequality formulation of the dynamic network user equilibrium problem*, Oper. Res. **41** (1993), no. 1, 179–191.

[108] C. B. Garcia, F. J. Gould, and T. R. Turnbull, *Relations between PL maps, complementary cones, and degree in linear complementarity problems*, Homotopy Methods and Global Convergence (Porto Cervo, 1981), NATO Conf. Ser. II: Systems Sci., vol. 13, Plenum, New York, 1983, pp. 91–144.

[109] C. B. Garcia and Willard I. Zangwill, *Pathways to solutions, fixed points and equilibria*, Computational Mathematics, Prentice–Hall, Englewood Cliffs, NJ, 1981.

[110] C. William Gear and Linda R. Petzold, *ODE methods for the solution of differential/algebraic systems*, SIAM J. Numer. Anal. **21** (1984), no. 4, 716–728.

[111] Christoph Glocker, *Set-valued force laws*, Lecture Notes in Applied Mechanics, vol. 1, Springer-Verlag, Berlin, 2001.

[112] Oscar Gonzalez and Andrew M. Stuart, *A first course in continuum mechanics*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, UK, 2008.

[113] M. Seetharama Gowda, *Inverse and implicit function theorems for H-differentiable and semismooth functions*, Optim. Methods Softw. **19** (2004), 443–461.

[114] M. Seetharama Gowda and Roman Sznajder, *Automorphism invariance of P- and GUS-properties of linear transformations on Euclidean Jordan algebras*, Math. Oper. Res. **31** (2006), no. 1, 109–123.

[115] M. Seetharama Gowda, Roman Sznajder, and J. Tao, *Some P-properties for linear transformations on Euclidean Jordan algebras*, Linear Algebra Appl. **393** (2004), 203–232.

[116] Thomas H. Gronwall, *Note on the derivatives with respect to a parameter of the solutions of a system of differential equations*, Ann. of Math. (2) **20** (1919), no. 4, 292–296.

[117] Jonathan L. Gross and Jay Yellen (eds.), *Handbook of graph theory*, Discrete Mathematics and Its Applications (Boca Raton), CRC Press, Boca Raton, FL, 2004.

[118] Richard Haberman, *Mathematical models*, Classics in Applied Mathematics, vol. 21, SIAM, Philadelphia, 1998, Mechanical vibrations, population dynamics, and traffic flow, Reprint of the 1977 original.

[119] Ernst Hairer and Gerhard Wanner, *Solving ordinary differential equations* II: *Stiff and differential–algebraic problems*, Ser. in Comp. Math. 14, Springer-Verlag, New York, Berlin, Heidelberg, 1991.

[120] Weimin Han and Mircea Sofonea, *Time-dependent variational inequalities for viscoelastic contact problems*, J. Comput. Appl. Math. **136** (2001), no. 1-2, 369–387.

[121] ———, *Quasistatic contact problems in viscoelasticity and viscoplasticity*, AMS/IP Studies in Advanced Mathematics, vol. 30, American Mathematical Society, Providence, RI, 2002.

[122] W. P. M. H. Heemels, M. Kanat Çamlıbel, and J. M. Schumacher, *On the dynamic analysis of piecewise-linear networks*, IEEE Trans. Circuits Syst. I Fund. Theory Appl. **49** (2002), no. 3, 315–327.

[123] W. P. M. H. Heemels, J. M. Schumacher, and S. Weiland, *The rational complementarity problem*, Linear Algebra Appl. **294** (1999), no. 1-3, 93–135.

[124] ———, *Linear complementarity systems*, SIAM J. Appl. Math **60** (2000), no. 4, 1234–1269.

[125] Claude Henry, *An existence theorem for a class of differential equations with multivalued right-hand side*, J. Math. Anal. Appl. **41** (1973), 179–186.

[126] Karl Heun, *Neue methoden zur approximativen integration der differentialgleichungen einer unabhängigen veränderlichen*, Z. Math. Phys. **45** (1900), 23–38.

[127] Lars Hörmander, *The analysis of linear partial differential operators* I: *Distribution theory and Fourier analysis*, Grundlehren der mathematischen Wissenschaften, no. 256, Springer-Verlag, Berlin, Heidelberg, New York, 1983.

[128] ———, *The analysis of linear partial differential operators* III: *Pseudodifferential operators*, Grundlehren der mathematischen Wissenschaften, no. 274, Springer-Verlag, Berlin, Heidelberg, New York, 1983.

[129] Shouchuan Hu and Nikolas S. Papageorgiou, *Handbook of multivalued analysis. Vol.* I, Mathematics and Its Applications, vol. 419, Kluwer Academic Publishers, Dordrecht, 1997.

[130] A.-G. Ibrahim and A. M. Gomaa, *Strong and weak solutions for differential inclusions with moving constraints in Banach spaces*, Pure Math. Appl. **8** (1997), no. 1, 53–65.

[131] Laurent O. Jay, *Convergence of a class of Runge-Kutta methods for differential-algebraic systems of index 2*, BIT **33** (1993), no. 1, 137–150.

[132] ———, *Convergence of Runge-Kutta methods for differential-algebraic systems of index* 3, Appl. Numer. Math. **17** (1995), no. 2, 97–118.

[133] ———, *Solution of index 2 implicit differential-algebraic equations by Lobatto Runge-Kutta methods*, BIT **43** (2003), no. 1, 93–106.

[134] ———, *Specialized Runge-Kutta methods for index* 2 *differential-algebraic equations*, Math. Comp. **75** (2006), no. 254, 641–654.

[135] Fritz John, *Extremum problems with inequalities as subsidiary conditions*, Studies and Essays, Interscience, New York, 1948, Presented to R. Courant on his 60th birthday, January 8, 1948, pp. 187–204.

[136] David E. Johnson, Johnny R. Johnson, John L. Hilburn, and Peter D. Scott, *Electric circuit analysis*, third ed., John Wiley & Sons, New York, 1997.

[137] S. Karamardian, *Generalized complementarity problem*, J. Optim. Theory Appl. **8** (1971), 161–168.

[138] _____, *The complementarity problem*, Math. Programming **2** (1972), no. 1, 107–129.

[139] _____, *Complementarity problems over cones with monotone and pseudomonotone maps*, J. Optim. Theory Appl. **18** (1976), no. 4, 445–454.

[140] _____, *An existence theorem for the complementarity problem*, J. Optim. Theory Appl. **19** (1976), no. 2, 227–232.

[141] Alois E. Kastner-Maresch, *Diskretisierungsverfahren zur Lösung von Differentialinklusionen*, Ph.D. thesis, Universität Bayreuth, 1990.

[142] _____, *Implicit Runge–Kutta methods for differential inclusions*, Numer. Funct. Anal. Optim. **11** (1990), 937–958.

[143] _____, *The implicit midpoint rule applied to discontinuous differential equations*, Computing **49** (1992), no. 1, 45–62.

[144] Alois E. Kastner-Maresch and Frank Lempio, *Difference methods with selection strategies for differential inclusions*, Numer. Funct. Anal. Optim. **14** (1993), no. 5-6, 555–572.

[145] Anatoly A. Kilbas, Hari M. Srivastava, and Juan J. Trujillo, *Theory and applications of fractional differential equations*, North–Holland Mathematics Studies, vol. 204, Elsevier Science B.V., Amsterdam, 2006.

[146] Jong Uhn Kim, *A boundary thin obstacle problem for a wave equation*, Comm. Partial Differential Equations **14** (1989), no. 8&9, 1011–1026.

[147] Virginia Kiryakova, *Generalized fractional calculus and applications*, Pitman Research Notes in Mathematics Series, vol. 301, Longman Scientific & Technical, Harlow, 1994.

[148] Anders Klarbring, Andro Mikelić, and Meir Shillor, *A global existence result for the quasistatic frictional contact problem with normal compliance*, Unilateral Problems in Structural Analysis, IV (Capri, 1989), Internat. Ser. Numer. Math., vol. 101, Birkhäuser, Basel, 1991, pp. 85–111.

[149] Markus Kunze and Manuel D. P. Monteiro Marques, *BV solutions to evolution problems with time-dependent domains*, Set-Valued Anal. **5** (1997), no. 1, 57–72.

[150] Martin W. Kutta, *Beitrag zur näherunsweisen Integration oder Differentialgleichungen*, Zeit. Math. Physik **46** (1901), 435–453.

[151] Kenneth L. Kuttler, *Modern analysis*, CRC Press, Boca Raton, FL, 1998.

[152] Kenneth L. Kuttler, A. Park, Meir Shillor, and Wen Zhang, *Unilateral dynamic contact of two beams*, Math. Comput. Modelling **34** (2001), no. 3-4, 365–384.

[153] Kenneth L. Kuttler and Meir Shillor, *Set-valued pseudomonotone maps and degenerate evolution inclusions*, Commun. Contemp. Math. **1** (1999), no. 1, 87–123.

[154] ———, *Dynamic contact with Signorini's condition and slip rate dependent friction*, Electron. J. Differential Equations (2004), no. 83, 21 pp.

[155] Serge Lang, *Real and functional analysis*, second ed., Graduate Texts in Mathematics, vol. 142, Springer, Berlin, Heidelberg, New York, 1993.

[156] Gilles Lebeau and Michelle Schatzman, *A wave problem in a half-space with a unilateral constraint at the boundary*, J. Differential Equations **53** (1984), 309–361.

[157] Carlton E. Lemke and Joseph T. Howson, Jr., *Equilibrium points of bimatrix games*, J. Soc. Indust. Math. **12** (1964), 413–423.

[158] Frank Lempio, *Modified Euler methods for differential inclusions*, Set-Valued Analysis and Differential Inclusions (Boston, Basel, Berlin) (A. B. Kurzhansky and V. M. Veliov, eds.), Birkhäuser Boston, Boston, MA, 1993.

[159] ———, *Set-valued interpolation, differential inclusions, and sensitivity in optimization*, Recent Developments in Well-posed Variational Problems, Math. Appl., vol. 331, Kluwer Academic Publishers, Dordrecht, 1995, pp. 137–169.

[160] Jacques-Louis Lions and Guido Stampacchia, *Variational inequalities*, Comm. Pure Appl. Math. **20** (1967), 493–519.

[161] Günter Lippold, *Error estimates for the implicit Euler approximation of an evolution inequality*, Nonlinear Anal. **15** (1990), no. 11, 1077–1089.

[162] Noel G. Lloyd, *Degree theory*, Cambridge Tracts in Mathematics, no. 73, Cambridge University Press, Cambridge, UK, 1978.

[163] Mongi Mabrouk, *Liaisons unilatérales et chocs élastiques quelconques: Un résultat d'existence*, C. R. Acad. Sci. Sér. Paris I Math. **326** (1998), 1353–1357.

[164] Avishai Mandelbaum, *The dynamic complementarity problem*, Unpublished manuscript, 1989.

[165] J. A. C. Martins and J. T. Oden, *Existence and uniqueness results for dynamics contact problems with nonlinear normal and friction interface laws*, Nonlinear Anal. **11** (1987), no. 3, 407–428.

[166] J. A. C. Martins, F. M. F. Simões, Fabio Gastaldi, and Manuel D. P. Monteiro Marques, *Dissipative graph solutions for a* 2 *degree-of-freedom quasistatic frictional contact problem*, Internat. J. Engrg. Sci. **33** (1995), no. 13, 1959–1986.

[167] Edward J. McShane and Robert B. Warfield, Jr., *On Filippov's implicit functions lemma*, Proc. Amer. Math. Soc. **18** (1967), 41–47.

[168] Robert E. Megginson, *An introduction to Banach space theory*, Graduate Texts in Mathematics, vol. 183, Springer, Berlin, New York, 1998.

[169] Nimrod Megiddo and Masakazu Kojima, *On the existence and uniqueness of solutions in nonlinear complementarity theory*, Math. Programming **12** (1977), no. 1, 110–130.

[170] Fulbert Mignot and Jean-Pierre Puel, *Inéquations d'évolution paraboliques avec convexes dépendant du temps. Applications aux inéquations quasi variationnelles d'évolution*, Arch. Rational Mech. Anal. **64** (1977), no. 1, 59–91.

[171] George J. Minty, *Monotone networks*, Proc. Roy. Soc. London. Ser. A **257** (1960), 194–212.

[172] ———, *On the maximal domain of a "monotone" function*, Michigan Math. J. **8** (1961), 135–137.

[173] ———, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J. **29** (1962), 341–346.

[174] Manuel D. P. Monteiro Marques, *Differential inclusions in nonsmooth mechanical problems: Shocks and dry friction*, Progress in Nonlinear Differential Equations and Their Applications, vol. 9, Birkhäuser Verlag, Basel, Boston, Berlin, 1993.

[175] Boris S. Mordukhovich, *Variational analysis and generalized differentiation.* I, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 330, Springer-Verlag, Berlin, 2006.

[176] Jean-Jacques Moreau, *Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires*, C. R. Acad. Sci. Paris **255** (1962), 238–240.

[177] ———, *Sur les mesures différentielles de fonctions vectorielles et certain problèmes d'évolution*, C. R. Acad. Sci. Paris Sér. A-B **282** (1976), 837–840.

[178] ———, *Evolution problem associated with a moving convex set in a Hilbert space*, J. Differential Equations **26** (1977), no. 3, 347–374.

[179] ———, *Application of convex analysis to some problems of dry friction*, Trends in Applications of Pure Mathematics to Mechanics, Vol. II (Second Sympos., Kozubnik, 1977), Monographs Stud. Math., vol. 5, Pitman, Boston, MA, 1979, pp. 263–280.

[180] ———, *Bounded variation in time*, Topics in Nonsmooth Mechanics, Birkhäuser, Basel, Boston, 1988, pp. 1–74.

[181] ———, *Unilateral contact and dry friction in finite freedom dynamics*, Nonsmooth Mechanics and Applications (J.-J. Moreau and P. D. Panagiotopoulos, eds.), International Centre for Mechanical Sciences, Courses and Lectures 302, Springer-Verlag, Vienna, New York, 1988, pp. 1–82.

[182] ———, *Numerical experiments in granular dynamics: Vibration-induced size segregation*, Contact Mechanics (New York) (M. Raous, M. Jean, and J.-J. Moreau, eds.), Proceedings of the 2nd Contact Mechanics International Symposium, September 19–23, 1994 in Carry-Le-Rouet, France, Plenum Press, 1995, pp. 347–158.

[183] Dumitru Motreanu and Mircea Sofonea, *Evolutionary variational inequalities arising in quasistatic frictional contact problems for elastic materials*, Abstr. Appl. Anal. **4** (1999), no. 4, 255–279.

[184] F. R. Moulton, *An introduction to celestial mechanics*, second ed., Macmillan, New York, 1914.

[185] James R. Munkres, *Topology: A first course*, Prentice–Hall, Englewood Cliffs, NJ, 1975.

[186] Anna Nagurney, *Network economics: A variational inequality approach*, third ed., Advances in Computational Economics, vol. 1, Springer, Heidelberg, New York, 1999.

[187] Anna Nagurney and Ding Zhang, *Projected dynamical systems and variational inequalities with applications*, International Series in Operations Research & Management Science, Springer, New York, Heidelberg, Berlin, 1995.

[188] Zdzisław Naniewicz and Panagiotos D. Panagiotopoulos, *Mathematical theory of hemivariational inequalities and applications*, Monographs and Textbooks in Pure and Applied Mathematics, vol. 188, Marcel Dekker, New York, 1995.

[189] Joachim Naumann, *Einführung in die Theorie parabolischer Variationsungleichungen*, Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], vol. 64, BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1984, With English, French, and Russian summaries.

[190] Hans-Dietrich Niepage, *Numerical methods for discontinuous ODEs and differential inclusions*, Computational Ordinary Differential Equations, Inst. Math. Appl. Conf. Ser. New Ser., no. 39, Oxford University Press, New York, 1992.

[191] Hans-Dietrich Niepage and Werner Wendt, *On the discrete convergence of multistep methods for differential inclusions*, Numer. Funct. Anal. Optim. **9** (1987), 591–617.

[192] Jorge Nocedal and Stephen J. Wright, *Numerical optimization*, second ed., Springer Series in Operations Research and Financial Engineering, Springer, New York, 2006.

[193] J. T. Oden and J. A. C. Martins, *Models and computational methods for dynamic friction phenomena*, Comput. Methods Appl. Mech. Engrg. **52** (1985), no. 1-3, 527–634.

[194] Paul Painlevé, *Sur le lois du frottement de glissemment*, C. R. Acad. Sci. Paris **121** (1895), 112–115, Following articles under the same title appeared in this journal, vol. 141, pp. 401–405 and 546–552 (1905).

[195] Jong-Shi Pang, *Newton's method for B-differentiable equations*, Math. Oper. Res. **15** (1990), no. 2, 311–341.

[196] Jong-Shi Pang and Liqun Qi, *Nonsmooth equations: Motivation and Algorithms*, SIAM J. Optim. **3** (1993), no. 3, 443–465.

[197] Jong-Shi Pang and David E. Stewart, *A unified approach to discrete frictional contact problems*, Internat. J. Engrg. Sci. **37** (1999), no. 13, 1747–1768.

[198] _____, *A unified approach to frictional contact problems*, Internat. J. Engrg. Sci. **37** (1999), no. 13, 1747–1768.

[199] _____, *Differential variational inequalities*, Math. Program. **113** (2008), no. 2, 345–424.

[200] Laetitia Paoli and Michelle Schatzman, *Schéma numérique pour un modèle de vibrations avec contraintes unilatérales et perte d'énergie aux impacts, en dimension finie.*, C. R. Acad. Sci. Paris Sér. I Math **317** (1993), no. 2, 211–215.

[201] _____, *A numerical scheme for impact problems.* I. *The one-dimensional case*, SIAM J. Numer. Anal. **40** (2002), no. 2, 702–733.

[202] _____, *A numerical scheme for impact problems.* II. *The multidimensional case*, SIAM J. Numer. Anal. **40** (2002), no. 2, 734–768.

[203] _____, *Numerical simulation of the dynamics of an impacting bar*, Comput. Methods Appl. Mech. Engrg. **196** (2007), no. 29-30, 2839–2851.

[204] _____, *Mouvement à un nombre fini de degrés de liberté avec contraintes unilatérales: Cas avec perte d'énergie*, Math. Model. Numer. Anal. **27** (1993), no. 6, 673–717.

[205] _____, *Vibrations avec contraintes unilatérales et perte d'énergie aux impacts, en dimension finie*, C. R. Acad. Sci. Paris, Sér. I Math. **317** (1993), 97–101.

[206] Cosmin Petra, Bogdan Gavrea, Mihai Anitescu, and Florian A. Potra, *A computational study of the use of an optimization-based method for simulating large multibody systems*, Optim. Methods Softw. **24** (2009), no. 6, 871–894.

[207] Adrien Petrov and Michelle Schatzman, *Viscoélastodynamique monodimensionnelle avec conditions de Signorini*, C. R. Math. Acad. Sci. Paris **334** (2002), no. 11, 983–988.

[208] _____, *Mathematical results on existence for viscoelastodynamic problems with unilateral constraints*, SIAM J. Math. Anal. **40** (2009), no. 5, 1882–1904.

[209] Siméon D. Poisson, *Traité de mécanique*, second ed., Bachelier, Paris, 1833.

[210] David Preiss, *Differentiability of Lipschitz functions on Banach spaces*, J. Funct. Anal. **91** (1990), no. 2, 312–345.

[211] Liqun Qi, *Convergence analysis of some algorithms for solving nonsmooth equations*, Math. Oper. Res. **18** (1993), no. 1, 227–244.

[212] Michael Renardy, *Ill-posedness at the boundary for elastic solids sliding under Coulomb friction*, J. Elasticity **27** (1992), no. 3, 281–287.

[213] Michael Renardy and Robert C. Rogers, *An introduction to partial differential equations*, Texts in Applied Mathematics, vol. 13, Springer-Verlag, New York, Berlin, Heidelberg, 1993.

[214] R. Tyrell Rockafellar and Roger J.-B. Wets, *Variational analysis*, Grundlehren der mathematischen Wissenschaften, vol. 317, Springer, Berlin, Heidelberg, New York, 1998.

[215] Joseph J. Rotman, *An introduction to algebraic topology*, Graduate Texts in Mathematics, vol. 119, Springer-Verlag, New York, 1988.

[216] Edward J. Routh, *A treatise on the dynamics of a system of rigid bodies*, MacMillan, London, 1860.

[217] Walter Rudin, *Principles of mathematical analysis*, third ed., International Series in Pure and Applied Mathematics, McGraw–Hill, New York, 1976.

[218] Carle D.T. Runge, *Über die numerische Auflösung von Differentialgleichungen*, Math. Ann. **46** (1895), 167–178.

[219] Michelle Schatzman, *Sur une classe de problèmes hyperboliques non linéaires*, C. R. Acad. Sci. Paris Sér. A-B **277** (1973), A671–A674.

[220] ———, *Le système différentiel* $(d^2u/dt^2) + \partial\varphi(u) \ni f$ *avec conditions initiales*, C. R. Acad. Sci. Paris Sér. A-B **284** (1977), no. 11, A603–A606.

[221] ———, *A hyperbolic problem of second order with unilateral constraints: The vibrating string with a concave obstacle*, J. Math. Anal. Appl. **73** (1980), no. 1, 138–191.

[222] ———, *Un problème hyperbolique du* 2*ème ordre avec contrainte unilatérale: La corde vibrante avec obstacle ponctuel*, J. Differential Equations **36** (1980), no. 2, 295–334.

[223] Hans Schumacher, *Between mathematical programming and systems theory: Linear complementarity systems*, Advances in Mathematical Systems Theory, Birkhäuser Boston, Boston, MA, 2001, pp. 185–219.

[224] Thomas I. Seidman, *The transient semiconductor problem with generation terms.* II, Nonlinear Semigroups, Partial Differential Equations and Attractors (Washington, DC, 1987), Lecture Notes in Math., vol. 1394, Springer, Berlin, 1989, pp. 185–198.

[225] Antonio Signorini, *Sopra alcune questioni di elastostatica*, Atti della Societá Italiana per il Progresso delle Scienze, 1933.

[226] ———, *Questioni di elasticità non linearizzata e semilinearizzata*, Rend. Mat. e Appl. (5) **18** (1959), 95–139.

[227] Jacques Simon, *Compact sets in the space* $L^p(0,T;B)$, Ann. Mat. Pura Appl. (4) **146** (1987), 65–96.

[228] Georgi V. Smirnov, *Introduction to the theory of differential inclusions*, American Mathematical Society, Providence, RI, 2002.

[229] Mircea Sofonea, Weimin Han, and Meir Shillor, *Analysis and approximation of contact problems with adhesion or damage*, Pure and Applied Mathematics (Boca Raton), vol. 276, Chapman & Hall/CRC, Boca Raton, FL, 2006.

[230] Mircea Sofonea and Meir Shillor, *Variational analysis of quasistatic viscoplastic contact problems with friction*, Commun. Appl. Anal. **5** (2001), no. 1, 135–151.

[231] Masha Sosonkina, Layne T. Watson, and David E. Stewart, *Note on the end game in homotopy zero curve tracking*, ACM Trans. Math. Software **22** (1996), no. 3, 281–287.

[232] Edwin H. Spanier, *Algebraic topology*, Ser. Higher Math. 7, McGraw–Hill, New York, 1966.

[233] A. J. M. Spencer, *Continuum mechanics*, Dover Publications, Mineola, NY, 2004, Reprint of the 1980 edition [Longman, London; MR0597343].

[234] Ulisse Stefanelli, *Nonlocal quasivariational evolution problems*, J. Differential Equations **229** (2006), no. 1, 204–228.

[235] David E. Stewart, *A high accuracy method for solving ODEs with discontinuous right-hand side*, Numer. Math. **58** (1990), no. 3, 299–328.

[236] ———, *High accuracy numerical methods for ordinary differential equations with discontinuous right-hand side*, Ph.D. thesis, University of Queensland, St. Lucia, Queensland 4072, Australia, 1990.

[237] ———, *Existence of solutions to rigid body dynamics and the paradoxes of Painlevé*, C. R. Acad. Sci. Paris Sér. I Math. **325** (1997), 689–693.

[238] ———, *Convergence of a time-stepping scheme for rigid body dynamics and resolution of Painlevé's problems*, Arch. Ration. Mech. Anal. **145** (1998), no. 3, 215–260.

[239] ———, *Rigid-body dynamics with friction and impact*, SIAM Rev. **42** (2000), no. 1, 3–39.

[240] ———, *Formulating measure differential inclusions in infinite dimensions*, J. Set-Valued Anal. **8** (2000), no. 3, 273–293.

[241] ———, *Reformulations of measure differential inclusions and their closed graph property*, J. Differential Equations **175** (2001), no. 1, 108–129.

[242] ———, *Convolution complementarity problems with application to impact problems*, IMA J. Appl. Math. **71** (2006), no. 1, 92–119, First published Advance Access May 2005; DOI 10.1093/imamat/hxh087.

[243] ———, *Differentiating complementarity problems and fractional index convolution complementarity problems*, Houston J. Math. **33** (2006), no. 1, 301–322.

[244] ———, *Uniqueness for index-one differential variational inequalities*, Nonlinear Anal. Hybrid Syst. **2** (2008), no. 3, 812–818.

[245] ———, *Uniqueness for solutions of differential complementarity problems*, Math. Program. **118** (2009), no. 2, 327–345.

[246] ———, *Runge–Kutta methods for index one differential variational inequalities*, Tech. Report, Department of Mathematics, University of Iowa, 2010, TBA.

[247] David E. Stewart and Jeffrey C. Trinkle, *Dynamics, friction and complementarity problems*, Complementarity and Variational Problems: State of the Art (Philadelphia, PA) (M. C. Ferris and J.-S. Pang, eds.), SIAM, Philadelphia, 1997, pp. 425–439.

[248] ———, *An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and Coulomb friction*, Internat. J. Numer. Methods Engrg. **39** (1996), no. 15, 2673–2691.

[249] David E. Stewart and Theodore J. Wendt, *Fractional index convolution complementarity problems*, Nonlinear Anal. Hybrid Syst. **1** (2007), no. 1, 124–134.

[250] D. Stoianovici and Yildirim Hurmuzlu, *A critical study of the applicability of rigid-body collision theory*, ASME J. Appl. Mech. **63** (1996), no. 2, 307–316.

[251] Ben G. Streetman and Sanjay Banerjee, *Solid state electronic devices*, sixth ed., Ser. Solid State Physical Electronics, Prentice–Hall, Englewood Cliffs, NJ, 2005.

[252] William J. Stronge, *Rigid body collisions with friction*, Proc. Roy. Soc. London Ser. A **431** (1990), no. 1881, 169–181.

[253] ———, *Impact mechanics*, first ed., Cambridge University Press, Cambridge, UK, 2000.

[254] Hector J. Sussmann, *Bounds on the number of switchings for trajectories of piecewise analytic vector fields*, J. Differential Equations **43** (1982), 399–418.

[255] Akira Takayama, *Mathematical economics*, second ed., Cambridge University Press, Cambridge, UK, 1985.

[256] Alessandro Tasora and Mihai Anitescu, *A fast NCP solver for large rigid-body problems with contacts, friction, and joints*, Multibody Dynamics, Comput. Methods Appl. Sci., vol. 12, Springer, Berlin, 2009, pp. 45–55.

[257] Klaus Taubert, *Differenzverfahren für gewöhnliche Anfangswertaufgaben mit unstetiger rechte Seite*, Numerische Behandlung Nichtlineare Integrodifferential- und Differentialgleichungen (R. Ansorge and W. Törnig, eds.), Lecture Notes in Mathematics 395, Springer-Verlag, Berlin, New York, 1974, pp. 137–148.

[258] ———, *Differenzverfahren für Schwingungen mit trockener und zäher Reibung und für Regelungssysteme*, Numer. Math. **26** (1976), 379–395.

[259] ———, *Converging multistep methods for initial value problems involving multivalued maps*, Computing **27** (1981), 123–136.

[260] Michael E. Taylor, *Partial differential equations* I*: Basic theory*, Applied Mathematical Sciences, vol. 115, Springer, New York, 1996.

[261] Roger Temam and Alain Miranville, *Mathematical modeling in continuum mechanics*, second ed., Cambridge University Press, Cambridge, UK, 2005.

[262] Hans Triebel, *Interpolation theory, function spaces, differential operators*, North–Holland, Amsterdam, New York, 1978.

[263] William T. Tutte, *Graph theory*, Encyclopedia of Mathematics and Its Applications, vol. 21, Cambridge University Press, Cambridge, UK, 2001, With a foreword by Crispin St. J. A. Nash-Williams, Reprint of the 1984 original.

[264] Arjan J. van der Schaft and J. M. Schumacher, *Complementarity modeling of hybrid systems*, IEEE Trans. Automat. Control **43** (1998), no. 4, 483–489.

[265] Lieven Vandenberghe, Bart L. de Moor, and Joos Vandewalle, *The generalized linear complementarity problem applied to the complete analysis of resistive piecewise-linear circuits*, IEEE Trans. Circuits and Systems **36** (1989), no. 11, 1382–1391.

[266] Walter D. Wallis, *A beginner's guide to graph theory*, second ed., Birkhäuser Boston, Boston, MA, 2007.

[267] John G. Wardrop, *Some theoretical aspects of road traffic research*, Proceedings, Institute of Civil Engineers, Part II **1** (1952), 325–378.

[268] Layne T. Watson, Stephen C. Billups, and Alexander P. Morgan, *Algorithm* 652: *HOMPACK: A suite of codes for globally convergent homotopy algorithms*, ACM Trans. Math. Software **13** (1987), 281–310.

[269] Layne T. Watson, J. Patrick Bixler, and Aubrey B. Poore, *Continuous homotopies for the linear complementarity problem*, SIAM J. Matrix Anal. Appl. **10** (1989), no. 2, 259–277.

[270] Joseph Wloka, *Partial differential equations*, Cambridge University Press, Cambridge, UK, 1987, Originally published as *Partielle Differentialgleichungen* by B.G. Teubner, Stuttgart, 1982.

[271] Philip Wolfe, *The simplex method for quadratic programming*, Econometrica **27** (1959), 382–398.

[272] Yosef Yomdin, *Sard's theorem and its improved versions in numerical analysis*, Computational Solution of Nonlinear Systems of Equations (Fort Collins, CO, 1988), Lectures in Appl. Math., vol. 26, American Mathematical Society, Providence, RI, 1990, pp. 701–706.

[273] Eberhard Zeidler, *Nonlinear functional analysis and its applications.* II*/A*, Springer-Verlag, New York, 1990, Linear Monotone Operators, Translated from the German by the author and Leo F. Boron.

[274] ———, *Nonlinear functional analysis and its applications.* II*/B*, Springer-Verlag, New York, 1990, Nonlinear Monotone Operators, Translated from the German by the author and Leo F. Boron.

[275] ———, *Applied functional analysis*, Applied Mathematical Sciences, vol. 109, Springer-Verlag, New York, 1995.

Index