Numerical Methods for Nonsmooth Dynamical Systems: Applications in Mechanics and Electronics

Vincent Acary, Bernard Brogliato

Vincent Acary, Bernard Brogliato

Numerical Methods for Nonsmooth Dynamics. Applications in Mechanics and Electronics

SPIN Springer's internal project number, if known

– Monograph –

October 11, 2009

Springer

À Céline et Martin
À Laurence et Bastien

Preface

This book concerns the numerical simulation of dynamical systems whose trajectories may not be differentiable everywhere. They are named *nonsmooth* dynamical systems. They make an important class of systems, first because of the many applications in which nonsmooth models are useful, secondly because they give rise to new problems in various fields of science. Usually nonsmooth dynamical systems are represented as differential inclusions, complementarity systems, evolution variational inequalities, each of these classes itself being split into several subclasses. The book is divided into four parts, the first three parts being sketched in Fig. 0.1. The aim of the first part is to present the main tools from mechanics and applied mathematics which are necessary to understand how nonsmooth dynamical systems may be numerically simulated in a reliable way. Many examples illustrate the theoretical results, and an emphasis is put on mechanical systems, as well as on electrical circuits (the so-called Filippov's systems are also examined in some detail, due to their importance in control applications). The second and third parts are dedicated to a detailed presentation of the numerical schemes. A fourth part is devoted to the presentation of the software platform SICONOS. This book is not a textbook on numerical analysis of nonsmooth systems, in the sense that despite the main results of numerical analysis (convergence, order of consistency, etc.) being presented, their proofs are not provided. Our main concern is rather to present in detail how the algorithms are constructed and what kind of advantages and drawbacks they possess.

Nonsmooth mechanics (resp. nonsmooth electrical circuits) is a topic that has been pioneered and developed in parallel with convex analysis in the 1960s and the 1970s in western Europe by J.J. Moreau, M. Schatzman, and P.D. Panagiotopoulos (resp. by the Dutch school of van Bockhoven and Leenaerts), then followed by several groups of researchers in Montpellier, Munich, Eindhoven, Marseille, Stockholm, Lausanne, Lisbon, Grenoble, Zurich, etc. More recently nonsmooth dynamical systems (especially complementarity systems) emerged in the USA, a country in which, paradoxically, complementarity theory and convex analysis (which are central tools for the study of nonsmooth mechanical and electrical systems) have been developed since a long time. Though nonsmooth mechanics and more generally nonsmooth dynamical systems have long been studied by mechanical engineers (impact mechanics

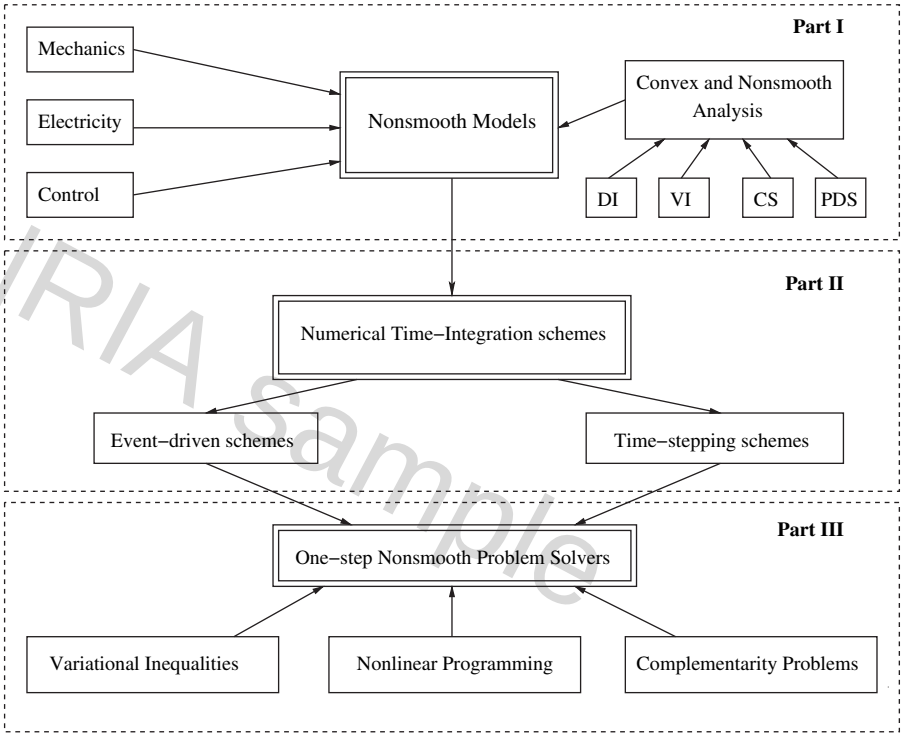


Fig. 0.1. Book Synopsis

can be traced back to ancient Greeks!) and applied mathematicians, their study has more recently attracted researchers of other scientific communities like systems and control, robotics, physics of granular media, civil engineering, virtual reality, haptic systems, image synthesis. We hope that this book will increase its dissemination.

We warmly thank Claude Lemaréchal (INRIA Bipop) for his many comments and discussions on Chap. 12 and Mathieu Renouf (LAMCOS-CNRS, Lyon) whose joint work with the first author contributed to Chap. 13. We also thank Professor F. Pfeiffer (Munich), an ardent promoter of nonsmooth mechanical systems, for his encouragements to us for writing this monograph, and Dr. Ditzinger (Springer Verlag). This work originated from a set of draft notes for a CEA-EdF-INRIA spring school that occurred in Rocquencourt from May 29 to June 02, 2006. The authors thank M. Jean (LMA-CNRS, Marseille, France) for his collaboration to this school and part of the preliminary draft. We would finally like to mention that part of this work was made in the framework of the European project SICONOS IST 2001-37172, from which the software platform SICONOS emerged. In particular the works of F. Pérignon and P. Denoyelle, expert engineers in the INRIA team-project Bipop, are here acknowledged.

Montbonnot,
August 2007

Vincent Acary
Bernard Brogliato

Contents

Part I Formulations of Nonsmooth Dynamical Systems

Part IV The SICONOS Software: Implementation and Examples

List of Acronyms

Absolutely Continuous (AC)
Affine Variational Inequality (AVI)

Complementarity Problem (CP)

Differential Algebraic Equation (DAE)
Dynamical (or Differential) Complementarity System (DCS)
Differential Inclusion (DI)

Evolution Variational Inequality (EVI)

Karush–Kuhn–Tucker (KKT)

Linear Complementarity Problem (LCP)
Linear Complementarity System (LCS)
Linar Independence Constraint Qualification (LICQ)

Mixed Complementarity Problem (MCP)
Mixed Linear Complementarity Problem (MLCP)
Mixed Linear Complementarity System (MLCS)

Nonlinear Complementarity Problem (NCP)
NonLinear Programming (NLP)
Non Smooth Gauss–Seidel (NSGS)

Ordinary Differential Equation (ODE)
Onestep NonSmooth Problem (OSNSP)

Positive Definite (PD)
Positive Semi–Definite (PSD)

Quadratic Program (QP)

Successive Quadratic Program (SQP)

Unilateral Differential Inclusion (UDI)

Variational Inequality (VI)

List of Algorithms

1

Nonsmooth Dynamical Systems: Motivating Examples and Basic Concepts

The aim of this introductory material is to show how one may write the dynamical equations of several physical systems like simple electrical circuits with nonsmooth elements, and simple mechanical systems with unilateral constraints on the position and impacts, Coulomb friction. We start with circuits with ideal diodes, then circuits with ideal Zener diodes. Then a mechanical system with Coulomb friction is analyzed, and the bouncing ball system is presented. These physical examples illustrate gradually how one may construct various mathematical equations, some of which are equivalent (i.e., the same "initial" data produce the same solutions). In each case we also derive the time-discretization of the continuous-time dynamics, and gradually highlight the discrepancy from one system to the next. All the presented tools and algorithms that are briefly presented in this chapter will be more deeply studied further in the book.

1.1 Electrical Circuits with Ideal Diodes

Though this book is mainly concerned with mechanical systems, electrical circuits will also be considered. The reasons are that on one hand electrical circuits with nonsmooth elements are an important class of physical systems, on the other hand their dynamics can nicely be recast in the family of evolution problems like differential inclusions, variational inequalities, complementarity systems, and some piecewise smooth systems. There is therefore a strong analogy between nonsmooth circuits and nonsmooth mechanical systems. This similarity will naturally exist also at the level of numerical simulation, which is the main object of this book.

The objective of this section is to show that electrical circuits containing so-called ideal diodes possess a dynamics which can be interpreted in various ways. It can be written as a complementarity system, a differential inclusion, an evolution variational inequality, or a variable structure system. What these several formalisms really mean will be made clear with simple examples.

1.1.1 Mathematical Modeling Issues

Let us consider the four electrical circuits depicted in Fig. 1.3. The diodes are supposed to be ideal, i.e., the characteristic between the current $i(t)$ and the voltage $v(t)$ (see Fig. 1.1a for the notation) satisfies the *complementarity* conditions:

$$0 \leqslant i(t) \perp v(t) \geqslant 0\,. \tag{1.1}$$

This set of conditions merely means that both the variables current $i(t)$ and voltage $v(t)$ have to remain nonnegative at all times t and that they have to be orthogonal one to each other. So $i(t)$ can be positive only if $v(t) = 0$, and vice versa. The complementarity condition (1.1) between the current across the diode and its voltage certainly represents the most natural way to define the diode characteristic. It is quite similar

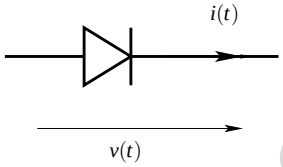


Fig. 1.1a. The diode component

Fig. 1.1b. Characteristics of an ideal diode. A complementarity condition

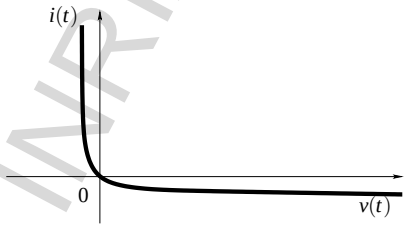


Fig. 1.1c. The graph of the Shockley's law

to the relations between the contact force and the distance between the system and an obstacle, in unilateral mechanics,[1] see Sect. 1.4.

Naturally, other models can be considered for the diode component. The well-known Shockley's law, which is one of the numerous models that can be found in standard simulation software, can be defined as

$$i = i_s \exp(-\frac{v}{\alpha} - 1) \,, \tag{1.2}$$

where the constant α depends mainly on the temperature. This law is depicted in Fig. 1.1c. This model may be considered to be more physical than the ideal one, because the residual saturation current, i_s is taken into account as a function of the voltage across the diode. The same remark applies in mechanics for a compliant contact model with respect to unilateral rigid contact model. Nevertheless, in the numerical practice, the ideal model reveals to be better from the qualitative point of view and also from the quantitative point of view. One of the reasons is that exchanging the highly stiff nonlinear model as in (1.2) by a nonsmooth multivalued model (1.1) leads to more robust numerical schemes. Moreover it is easy to introduce a residual current in the complementarity formalism as follows:

$$0 \leqslant i(t) + \varepsilon_1 \perp v(t) + \varepsilon_2 \geqslant 0 \tag{1.3}$$

for some $\varepsilon_1 \geqslant 0$, $\varepsilon_2 \geqslant 0$. This results in a shift of the characteristic of Fig. 1.1b.

The relation in (1.1) will necessarily enter the dynamics of a circuit containing ideal diodes. It is consequently crucial to clearly understand its meaning. Let us notice that the relation in (1.1) defines the *graph* of a *multivalued function* (or multifunction, or set-valued function), as it is clear that it is satisfied for any $i(t) \geqslant 0$ if $v(t) = 0$. This graph is depicted in Fig. 1.1b.

Using basic convex analysis (which in particular will allow us to accurately define what is meant by the gradient of a function that is not differentiable in the usual way), a nice interpretation of the relation in (1.1) and of its graph in Fig. 1.1b can be obtained with *indicator functions of convex sets*. The indicator of a set K is defined as

$$\psi_K(x) = \begin{cases} 0 & \text{if } x \in K \\ +\infty & \text{if } x \notin K \end{cases} \,. \tag{1.4}$$

This function is highly nonsmooth on the boundary ∂K of K, since it even possesses an infinite jump at such points! It is therefore nondifferentiable at $x \in \partial K$. Nevertheless, if K is a convex set then $\psi_K(\cdot)$ is a convex function, and it is *subdifferentiable* in the sense of convex analysis. Roughly speaking, one will consider *subgradients* instead of the usual gradient of a differentiable function. The subgradients of a convex function are vectors γ defining the directions "under" the graph of the function. More precisely, γ is a subgradient of a convex function $f(\cdot)$ at x if and only if it satisfies

[1] At this stage the similarity between both remains at a pure formal level. Indeed a more physical analogy would lead us to consider that it is rather a relation between a velocity and a force that corresponds to (1.1).

$$f(y) - f(x) \geqslant \gamma^{\mathrm{T}}(y-x) \tag{1.5}$$

for all y. Normally the subdifferential is denoted as $\partial f(\cdot)$, and $\partial f(x)$ can be a set (containing the subgradients γ).

Let us now consider the particular case of the indicator function of $K = \mathbb{R}^+ = \{x \in \mathbb{R} \mid x \geqslant 0\}$. Though this might be at first sight surprising, this function is subdifferentiable at $x = 0$. Its subdifferential is given by

$$\partial \psi_{\mathbb{R}^+}(x) = \begin{cases} \{0\} & \text{if } x > 0 \\ (-\infty, 0] & \text{if } x = 0 \end{cases} . \tag{1.6}$$

Indeed one checks that when $x \geqslant 0$, then $\psi_{\mathbb{R}^+}(y) \geqslant \gamma(y-x)$ for all $y \in \mathbb{R}$ can be satisfied if and only if $\gamma = 0$. Now if $x = 0$, $\psi_{\mathbb{R}^+}(y) \geqslant \gamma y$ is satisfied for all $y \in \mathbb{R}$ if and only if $\gamma \leqslant 0$. One sees that at $x = 0$ the subdifferential is a set, since it is a complete half space. In fact the set $\partial \psi_{\mathbb{R}^+}(x)$ is equal to the so-called normal cone to $\mathbb{R}^+$ at the point x (Fig. 1.2). This can be generalized to convex sets $K \subset \mathbb{R}^n$, so that the subdifferential $\psi_K(x)$ is the normal cone to the set K, computed at the point $x \in K$ and denoted by $N_K(x)$. If the boundary of K is differentiable, this is simply a half-line normal to the tangent plane to K at x, and in the direction outward K.

It becomes apparent that the graph of the subdifferential of the indicator of $\mathbb{R}^+$ resembles a lot the corner law depicted in Fig. 1.1b. Actually, one can now deduce from (1.6) and (1.1) that

$$i(t) \in -\partial \psi_{\mathbb{R}^+}(v(t)) \quad \Longleftrightarrow \quad v(t) \in -\partial \psi_{\mathbb{R}^+}(i(t)) . \tag{1.7}$$

The symmetry between these two inclusions is clear from Fig. 1.1b: if one inverts the multifunction (exchange $i(t)$ and $v(t)$ in Fig. 1.1b), then one obtains exactly the same graph. Actually this is a very particular case of duality between two variables. In a more general setting the graph inversion procedure does not yield the graph of the same multifunction, but the graph of its conjugate. And inverting once again allows

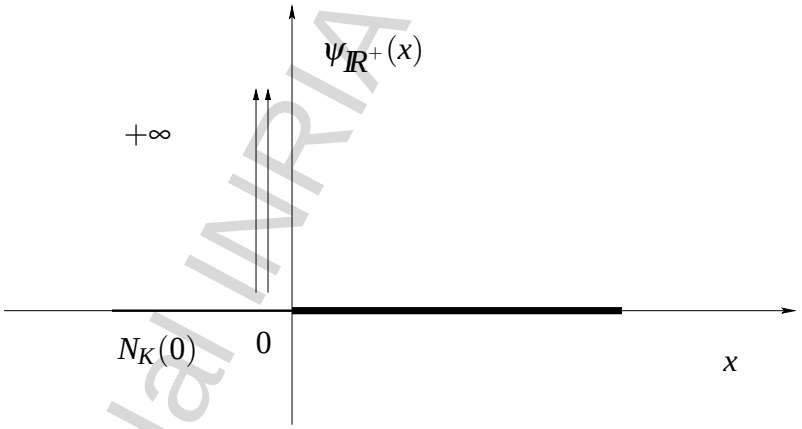


Fig. 1.2. The indicator function of $\mathbb{R}^+$ and the normal cone at $x = 0$, $N_K(0) = \partial \psi_{\mathbb{R}^+}(0) = \mathbb{R}^-$

one to recover the original graph under some convexity and properness assumption: this is the very basic principle of duality (Luenberger, 1992).

Let us focus now on the *inclusions* in (1.7). As a matter of fact, one may check that the first one is equivalent to: for any $v(t) \geqslant 0$,

$$\langle i(t), u - v(t) \rangle \geqslant 0, \ \forall\, u \geqslant 0 \tag{1.8}$$

and to: for any $i(t) \geqslant 0$,

$$\langle v(t), u - i(t) \rangle \geqslant 0, \ \forall\, u \geqslant 0\ . \tag{1.9}$$

The objects in (1.8) and (1.9) are called a *Variational Inequality (VI).*

We therefore have three different ways of looking at the ideal diode characteristic: the complementarity relations in (1.1), the inclusion in (1.7), and the variational inequality in (1.8). Our objective now is to show that when introduced into the dynamics of an electrical circuit, these formalisms give rise to various types of dynamical systems as enumerated at the beginning of this section.

Remark 1.1. Another variational inequality can also be written: for all $i(t) \geqslant 0$, $v(t) \geqslant 0$,

$$\langle j - i(t), u - v(t) \rangle \geqslant 0\ , \ \forall\ j, u \geqslant 0\ . \tag{1.10}$$

Having attained this point, the reader might legitimately wonder what is the usefulness of doing such an operation, and what has been gained by rewriting (1.1) as in (1.7) or as in (1.8). Let us answer a bit vaguely: several formalisms are likely to be useful for different tasks which occur in the course of the study of a dynamical system (mathematical analysis, time-discretization and numerical simulation, analysis for control, feedback control design, and so on). In this introductory chapter, we just ask the reader to trust us: all these formalisms are useful and are used. We will see in the sequel that there exists a lot of other ways to write the complementary condition such as zeroes of special functions or extremal points of a functional. All these formulations will lead to specific ways of studying and solving the system.

1.1.2 Four Nonsmooth Electrical Circuits

In order to derive the dynamics of an electrical circuit we need to consider Kirchoff's laws as well as the constitutive relations of devices like resistors, inductors, and capacitors (Chua et al., 1991). The constitutive relation of the ideal diode is the complementarity relation (1.1) while in the case of resistors, inductors, and capacitors we have the classical linear relations between variables like voltages, currents, and charges. Thus, taking into account those constitutive relations and using Kirchoff's laws it follows that the dynamical equations of the four circuits depicted in Fig. 1.3 are given by

$$(\mathbf{a}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC}x_1(t) - \dfrac{\lambda(t)}{R} \\ \dot{x}_2(t) = -\dfrac{1}{LC}x_1(t) - \dfrac{\lambda(t)}{L} \\ 0 \leqslant \lambda(t) \perp w(t) = \dfrac{\lambda(t)}{R} + \dfrac{1}{RC}x_1(t) - x_2(t) \geqslant 0 \end{cases} \tag{1.11}$$

$$(\mathbf{b}) \quad \begin{cases} \dot{x}_1(t) = -x_2(t) + \lambda(t) \\ \dot{x}_2(t) = \dfrac{1}{LC}x_1(t) \\ 0 \leqslant \lambda(t) \perp w(t) = \dfrac{1}{C}x_1(t) + R\lambda(t) \geqslant 0 \end{cases} \tag{1.12}$$

$$(\mathbf{c}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) = -\dfrac{R}{L}x_2(t) - \dfrac{1}{LC}x_1(t) - \dfrac{\lambda(t)}{L} \\ 0 \leqslant \lambda(t) \perp w(t) = -x_2(t) \geqslant 0 \end{cases} \tag{1.13}$$

$$(\mathbf{d}) \quad \begin{cases} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC}x_1(t) \\ \dot{x}_2(t) = -\dfrac{1}{LC}x_1(t) + \dfrac{\lambda(t)}{L} \\ 0 \leqslant \lambda(t) \perp w(t) = x_2(t) \geqslant 0 \end{cases} \tag{1.14}$$

where we considered the current through the inductors for the variable $x_2(t)$, and for the variable $x_1(t)$ the charge on the capacitors as state variables.

Let us now make use of the above equivalent formalisms to express the dynamics in (1.11)–(1.14) in various ways. We will generically call the dynamics in (1.11)–(1.14) a Linear Complementarity System (LCS), a terminology introduced in van der Schaft & Schumacher (1996). An LCS therefore consists of a linear differential equation with state (x_1, x_2), an external signal $\lambda(\cdot)$ entering the differential equation, and a set of complementarity conditions which relate a variable $w(\cdot)$ and $\lambda(\cdot)$. Since $w(\cdot)$ is itself a function of the state and possibly of $\lambda(\cdot)$, the complementarity conditions play a crucial role in the dynamics. The variable λ may be interpreted as a Lagrange multiplier.

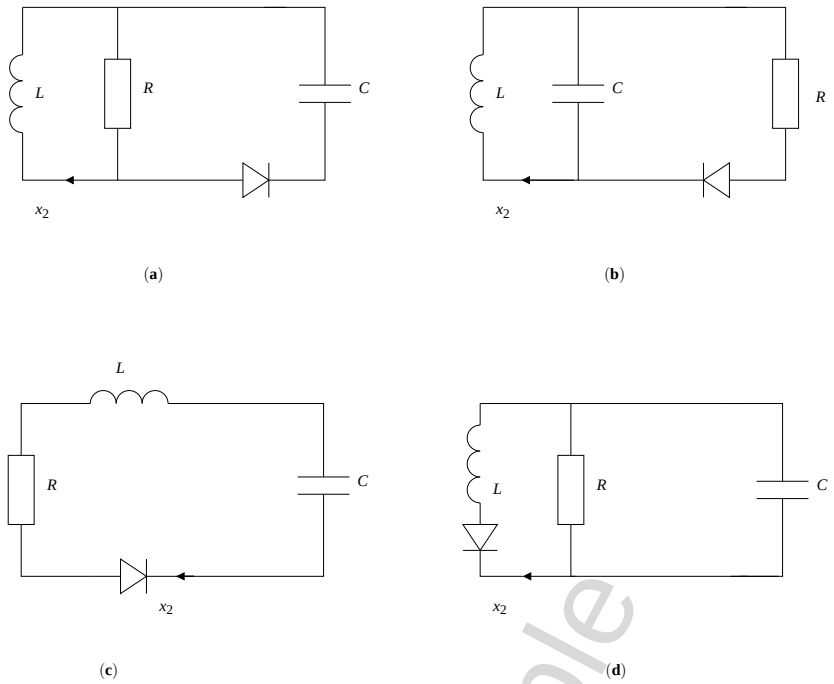


Fig. 1.3. RLC circuits with an ideal diode

1.1.3 Continuous System (Ordinary Differential Equation)

Let us consider for instance the circuit (**a**) whose dynamics is in (1.11). Its complementarity conditions are given by

$$\begin{cases} w(t) = \dfrac{\lambda(t)}{R} + \dfrac{1}{RC}x_1(t) - x_2(t) \\ \\ 0 \leqslant \lambda(t) \perp w(t) \geqslant 0 \end{cases} . \tag{1.15}$$

If we consider $\lambda(t)$ as the unknown of this problem, then the question we have to answer to is: does it possess a solution, and if yes is this solution unique? Here we must introduce a basic tool that is ubiquitous in complementarity systems: the Linear Complementarity Problem (LCP). An LCP is a problem which consists of solving a set of complementarity relations as

$$\begin{cases} w = M\lambda + q \\ \\ 0 \leqslant \lambda \perp w \geqslant 0 \end{cases} , \tag{1.16}$$

where M is a constant matrix and q a constant vector, both of appropriate dimensions. The inequalities have to be understood component-wise and the relation $w \perp \lambda$ means $w^{\mathrm{T}}\lambda = 0$. A fundamental result on LCP (see Sect. 12.4) guarantees that there is a unique λ that solves the LCP in (1.16) for any q if and only if M is a so-called P-matrix (i.e., all its principal minors are positive). In particular, positive definite matrices are P-matrices.

Taking this into account, it is an easy task to see that there is a unique solution $\lambda(t)$ to the LCP in (1.15) given by

$$\lambda(t) = 0 \qquad \text{if } \frac{1}{RC}x_1(t) - x_2(t) \geqslant 0\,, \tag{1.17}$$

$$\lambda(t) = -\frac{1}{C}x_1(t) + Rx_2(t) > 0 \qquad \text{if } \frac{1}{RC}x_1(t) - x_2(t) < 0\,. \tag{1.18}$$

Evidently we could have solved this LCP without resorting to any general result on existence and uniqueness of solutions. However, we will often encounter LCPs with several tenth or even hundreds of variables (i.e., the dimension of M in (1.16) can be very large in many applications). In such cases solving the LCP "with the hands" rapidly becomes intractable. So $\lambda(t)$ in (1.11) considered as the solution at time t of the LCP in (1.15) can take two values, and only two, for all $t \geqslant 0$.

Another way to arrive at the same result for circuit (**a**) is to use once again the equivalence between (1.1) and (1.7). It is straightforward then to see that (1.15) is equivalent to

$$\lambda(t) + \frac{1}{C}x_1(t) - Rx_2(t) \in -\partial\psi_{\mathbb{R}^+}(\lambda(t)) \tag{1.19}$$

(we have multiplied the left-hand side by R and since $\partial\psi_{\mathbb{R}^+}(\lambda(t))$ is a cone $R\partial\psi_{\mathbb{R}^+}(\lambda(t)) = \partial\psi_{\mathbb{R}^+}(\lambda(t))$). It is well known in convex analysis (see Appendix A) that (1.19) is equivalent to

$$\lambda(t) = \mathrm{Proj}_{\mathbb{R}^+}\left[-\frac{1}{C}x_1(t) + Rx_2(t)\right]\,, \tag{1.20}$$

where $\mathrm{Proj}_{\mathbb{R}^+}$ is the projection on $\mathbb{R}^+$. Since $\mathbb{R}^+$ is convex (1.20) possesses a unique solution. Once again we arrive at the same conclusion. The surface that splits the phase space (x_1, x_2) in two parts corresponding to the "switching" of the LCP is the line $-\frac{1}{C}x_1(t) + Rx_2(t) = 0$. On one side of this line $\lambda(t) = 0$, and on the other side $\lambda(t) = -\frac{1}{C}x_1(t) + Rx_2(t) > 0$. We may write (1.11) as

$$\begin{cases} \begin{cases} \dot{x}_1(t) = x_2(t) - \dfrac{1}{RC}x_1(t) \\ \dot{x}_2(t) = -\dfrac{1}{LC}x_1(t) \end{cases} & \text{if } -\dfrac{1}{C}x_1(t) + Rx_2(t) < 0\,, \\ \begin{cases} \dot{x}_1(t) = 0 \\ \dot{x}_2(t) = -\dfrac{R}{L}x_2(t) \end{cases} & \text{if } -\dfrac{1}{C}x_1(t) + Rx_2(t) \geqslant 0\,, \end{cases} \tag{1.21}$$

that is a piecewise linear system, or as

$$\dot{x}(t) - Ax(t) = B\,\mathrm{Proj}_{\mathbb{R}^+}\left[-\frac{1}{C}x_1(t) + Rx_2(t)\right]\,, \tag{1.22}$$

where the matrices A and B can be easily identified.

The fact that the projection operator in (1.20) is a Lipschitz-continuous single-valued function (Goeleven et al., 2003a) shows that the equation (1.22) is an Ordinary Differential Equation (ODE) with a Lipschitz-continuous vector field.[2] We therefore conclude that this complementarity system possesses a global unique and differentiable solution, as a standard result on ODEs (Coddington & Levinson, 1955).

Exactly the same analysis can be done for the circuit (**b**) which is also an ODE.

1.1.4 Hints on the Numerical Simulation of Circuits (a) and (b)

The circuit (**a**) can be simulated with any standard one-step and multistep methods like explicit or implicit (backward) Euler, mid-point, or trapezoidal rules (Hairer et al., 1993, Chap. II.7), which apply to ordinary differential equations with a Lipschitz right-hand side. Nevertheless, all these methods behave globally as a method of order one as the right-hand side is not differentiable everywhere (Hairer et al., 1993; Calvo et al., 2003).

As an illustration, a simple trajectory of the circuit (**a**) is computed with an explicit Euler scheme and a standard Runge–Kutta of order 4 scheme. The results are depicted in Fig. 1.4. With the initial conditions, $x_1(0) = 1$, $x_2(0) = -1$, we observe only one event or switch from one mode to the other. Before the switch, the dynamics is a linear oscillator in x_1 and after the switch, it corresponds to a exponential decay in x_1.

We present in Fig. 1.5 a slightly more rich dynamics with the circuit (**b**), which corresponds to a half-wave rectifier. When the diode blocks the current, $\lambda = 0, w > 0$, the dynamics of the circuit is a pure linear LC oscillator in x_2. When the constraint is active $\lambda > 0, w = 0$ and the diode lets the positive current pass: the dynamics is a damped linear oscillator in x_1. The interest of the circuit (**b**) with respect to the circuit (**a**) is that if R is small other switches are possible in circuit (**b**).

[2] It is also known that the solutions of LCPs as in (1.16) with M a P-matrix are Lipschitz-continuous functions of q (Cottle et al., 1992, Sect. 7.2). So we could have deduced this result from (1.15) and the complementarity formalism of the circuit.

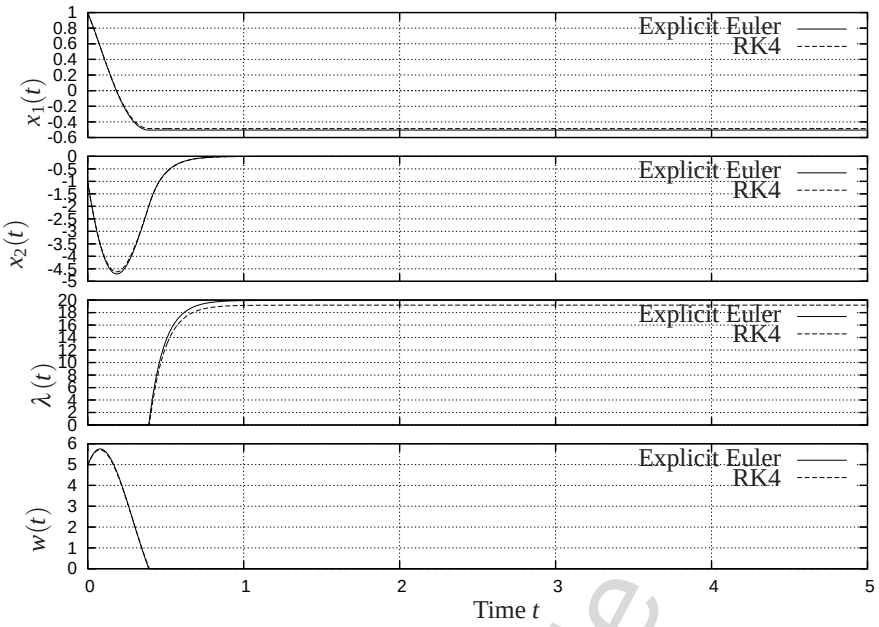


Fig. 1.4. Simulation of the circuit (**a**) with the initial conditions $x_1(0) = 1$, $x_2(0) = -1$ and $R = 10$, $L = 1$, $C = \dfrac{1}{(2\pi)^2}$. Time step $h = 5 \times 10^{-3}$

The Question of the Order

It is noteworthy that even in this simple case, where the "degree" of nonsmoothness is rather low (said otherwise, the system is a gentle nonsmooth system), applying higher order "time-stepping" methods which preserve the order $p \geqslant 2$ is not straightforward. By *time-stepping method*, we mean here a time-discretization method which does not consider explicitly the possible times at which the solution is not differentiable in the process of integration.

Let us now quote some ideas from Grüne & Kloeden (2006) which accurately explain the problem of applying standard higher order schemes: *In principle known numerical schemes for ordinary differential equations such as Runge–Kutta schemes can be applied to switching systems, changing the vector field after each switch has occurred. However, in order to maintain the usual consistency order of these schemes, the integration time steps need to be adjusted to the switching times in such a way that switching always occurs at the end of an integration interval. This is impractical in the case of fast switching, because in this case an adjustment of the scheme's integration step size to the switching times would lead to very small time steps causing an undesirably high computational load*. Such a method for the time integration of nonsmooth systems, which consists in locating and adjusting the time step to the events will be called an *event-driven method*. If the location of the events

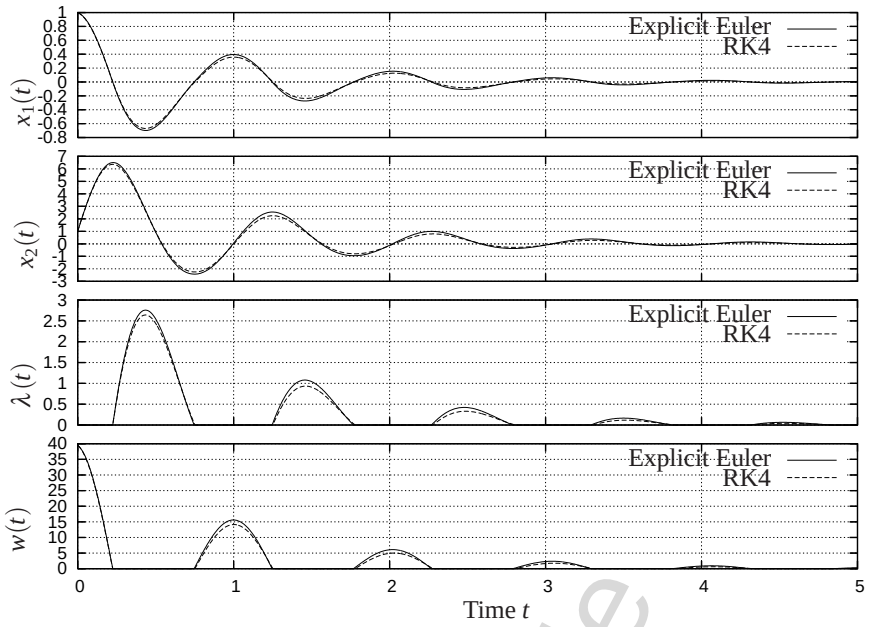


Fig. 1.5. Simulation of the circuit (**b**) with the initial conditions $x_1(0) = 1$, $x_2(0) = 1$ and $R = 10$, $L = 1$, $C = \dfrac{1}{(2\pi)^2}$. Time step $h = 5 \times 10^{-3}$

is sufficiently accurate, the global order of the integration method can be retrieved. If one is not interested in maintaining the order of the scheme larger than one, however, one may apply Runge–Kutta methods directly to an ODE as (1.22).

There are three main conclusions to be retained from this:

1. When the instants of nondifferentiability are not known in advance, or when there are too many such times, then applying an "event-driven" method with order larger than one may not be tractable.
2. We may add another drawback of event-driven methods that may not be present in the system we have just studied, but will frequently occur in the systems studied in this book. Suppose that the events (or times of nondifferentiability, or switching times) possess a finite accumulation point. Then an event-driven scheme will not be able to go further than the accumulation, except at the price of continuing the integration with some ad hoc, physically and mathematically unjustified trick.
3. Finally, there exist higher order standard numerical schemes which continue to perform well for some classes of nonsmooth systems, but at the price of decreasing the global order to one (see Sect. 9.2). However, this global low-order behavior can be compensated by an adaptive time-step strategy which takes benefits from the high accuracy of the time-integration scheme on smooth phases.

It is noteworthy that the events that will be encountered in the systems examined throughout the book usually are not exogenous events but state dependent, hence not known in advance. Therefore, the choice between the event-driven methods or the time-stepping methods depends strongly on the type of systems under study. We will come back later on the difference between time-stepping and event-driven numerical schemes and their respective ranges of applications (especially for mechanical systems).

The Question of the Stability of Explicit Schemes

As we said earlier, the nonsmoothness of the right-hand side destroys the order of convergence of the standard time-stepping integration scheme. Another aspect is the stability, especially for explicit schemes. Most of the results on the stability of numerical integration schemes are based on the assumption of sufficient regularity of the right-hand side.

The question of the simulation of ODEs with discontinuities will be discussed in Sects. 7.2 and 9.1. Some numerical illustrations of troubles in terms of the order of convergence and the stability of the methods are given in Sect. 9.1 where the dynamics of the circuits (**a**) and (**b**) are simulated.

1.1.5 Unilateral Differential Inclusion

Let us now turn our attention to circuit (**c**). This time the complementarity relations are given by

$$0 \leqslant \lambda(t) \perp w(t) = -x_2(t) \geqslant 0\,. \tag{1.23}$$

Contrary to (1.15), it is not possible to calculate $\lambda(t)$ directly from this set of relations. At first sight there is no LCP that can be constructed (indeed now we have a zero matrix M).

Let us, however, imagine that there is a time interval $[\tau, \tau+\varepsilon)$, $\varepsilon > 0$, on which the solution $x_2(t) = 0$ for all $t \in [\tau, \tau+\varepsilon)$. Then on $[\tau, \tau+\varepsilon)$ one has necessarily $-\dot{x}_2(t) \geqslant 0$, otherwise the unilateral constraint $-x_2(t) \geqslant 0$ would be violated. Actually all the derivatives of $x_2(\cdot)$ are identically 0 on $[\tau, \tau+\varepsilon)$. The interesting question is: what happens on the right of $t = \tau+\varepsilon$? Is there one derivative of $x_2(\cdot)$ that becomes positive, so that the system starts to detach from the constraint $x_2 = 0$ at $t = \tau+\varepsilon$? Such a question is important, think for instance of numerical simulation: one will need to implement a correct test to determine whether or not the system keeps evolving on the constraint, or quits it. In fact the test consists of considering the further complementarity condition

$$0 \leqslant \lambda(t^+) \perp -\dot{x}_2(t) = \frac{R}{L}x_2(t^+) + \frac{1}{LC}x_1(t) + \frac{\lambda(t^+)}{L} \geqslant 0 \tag{1.24}$$

which is an LCP to be solved only when $x_2(t) = 0$. The fact that this LCP possesses a solution $\lambda(t) - \dot{x}_2(t) > 0$ is a sufficient condition for the system to change its *mode* of evolution. We can solve for $\lambda(t)$ in (1.24) exactly as we did for (1.15). Both are

LCPs with a unique solution. However, this time the resulting dynamical system is not quite the same, since we have been obliged to follow a different path to get the LCP in (1.24).

In order to better realize this big discrepancy, let us use once again the equivalence between (1.1) and (1.7). We obtain that $\lambda(t) \in -\partial\psi_{\mathbb{R}^+}(-x_2(t))$. Inserting this inclusion in the dynamics (1.13) yields

$$(\mathbf{c}) \quad \begin{cases} \dot{x}_1(t) - x_2(t) = 0 \\ \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) \in \dfrac{1}{L}\partial\psi_{\mathbb{R}^+}(-x_2(t)) \end{cases} \tag{1.25}$$

where it is implicitly assumed that $x_2(0) \leqslant 0$ so that the inequality constraint $x_2(t) \leqslant 0$ will be satisfied for all $t \geqslant 0$.

Passing from the LCP (1.23) to the LCP (1.24) and then from (1.13) to (1.25) can be viewed similarly as the index-reduction operation in a Differential Algebraic Equation (DAE). Indeed, the LCP on x_2 in (1.23) is replaced by the LCP on $\dot{x}_2$ in (1.24).

Unilateral Differential Inclusion

More compactly, (1.25) can be rewritten as

$$-\dot{x}(t) + Ax(t) \in B\partial\psi_{\mathbb{R}^+}(w(t)) \tag{1.26}$$

which we can call a Unilateral Differential Inclusion (UDI) where the matrices A and B can be easily identified. The reason why we employ the word *unilateral* should be obvious. It is noteworthy that the right-hand side of (1.26) is generally a set that is not reduced to a single element, see (1.6). It is also noteworthy that the complementarity conditions are included in the UDI in (1.26). Obviously, the dynamics in (1.26) is not a variable structure or discontinuous vector field system. It is something else.

Evolution Variational Inequality

Using a suitable change of coordinate $z = Rx$, $R = R^{\mathrm{T}} > 0$, it is possible to show (Goeleven & Brogliato, 2004; Brogliato, 2004) that (1.26) can also be seen as an Evolution Variational Inequality (EVI). This time we make use of the equivalence between (1.7) and (1.8) and of a property of electrical circuits composed of resistors, capacitors, and inductances (they are dissipative). Then (1.26) is equivalent to the EVI

$$\begin{cases} \left\langle \dfrac{\mathrm{d}z}{\mathrm{d}t}(t) - RAR^{-1}z(t), v - z(t) \right\rangle \geqslant 0, \forall\, v \in K, \text{ a.e. } t \geqslant 0 \\ \\ z(t) \in K, t \geqslant 0\,, \end{cases} \tag{1.27}$$

where $K = \{(z_1, z_2) | \ -(0\ 1)\, R^{-1} z \geqslant 0\}$ and a.e. means almost everywhere (the solution not being a priori differentiable everywhere). As a consequence of how the set K is constructed, having $z(t) \in K$ is equivalent to having $x_2(t) \leqslant 0$. In fact it can be shown that the EVI in (1.27) possesses unique continuous solutions which are right differentiable (Goeleven & Brogliato, 2004). It is remarkable at this stage to notice that both (1.22) and (1.26) possess unique continuous solutions, however, the solutions of the inclusion (1.26) are less regular.

1.1.6 Hints on the Numerical Simulation of Circuits (c) and (d)

Let us now see how the differential inclusion (1.26) and the LCS in (1.13) may be time-discretized for numerical simulation purpose. Let us start with the LCS in (1.13).

A Direct Backward Euler Scheme

Mimicking the backward Euler discretization for ODEs, a time-discretization of (1.13) is

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} - x_{2,k} = -h\dfrac{R}{L} x_{2,k+1} - \dfrac{h}{LC} x_{1,k+1} - \dfrac{h}{L}\lambda_{k+1} \\ 0 \leqslant \lambda_{k+1} \perp -x_{2,k+1} \geqslant 0 \end{cases}, \tag{1.28}$$

where x_k is the value, at time t_k of a grid $t_0 < t_1 < \cdots < t_N = T$, $N < +\infty$, $h = \dfrac{T - t_0}{N} = t_k - t_{k-1}$, of a step function $x^N(\cdot)$ that approximates the analytical solution $x(\cdot)$.

Let us denote $a(h) = 1 + h\dfrac{R}{L} + h^2 \dfrac{1}{LC}$. Then we can rewrite (1.28) as

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} = (a(h))^{-1} \left\{ x_{2,k} - \dfrac{h}{LC} x_{1,k} - \dfrac{h}{L}\lambda_{k+1} \right\} \\ 0 \leqslant \lambda_{k+1} \perp -(a(h))^{-1} \left\{ x_{2,k} - \dfrac{h}{LC} x_{1,k} \right\} + (a(h))^{-1} \dfrac{h}{L} \lambda_{k+1} \geqslant 0 \end{cases}. \tag{1.29}$$

Remark 1.2. This time-stepping scheme is made of a discretization of the continuous dynamics (the first two lines of (1.29)) and of a LCP whose unknown is λ_{k+1}. We shall call later on the LCP resolution a one-step algorithm. Here the LCP is scalar and can easily be solved by inspection. In higher dimensions specific solvers will be necessary. This is the object of Part III of this book.

Remark 1.3. The LCP matrix M (here a scalar) is equal to $(a(h))^{-1}\frac{h}{L} > 0$ for all $h > 0$, which tends to 0 as $h \to 0$. This is not very good in practice when very small steps are chosen. To cope with this issue, let us choose as the unknown the variable $\bar{\lambda}_{k+1} = h\lambda_{k+1}$. We then solve the LCP

$$0 \leqslant \bar{\lambda}_{k+1} \perp -(a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} + (a(h))^{-1}\frac{1}{L}\bar{\lambda}_{k+1} \geqslant 0\,. \tag{1.30}$$

It is noteworthy that this does not change the result of the algorithm, because the set of nonnegative reals is a cone. This LCP is easily solved:

$$\text{If } x_{2,k} - \frac{h}{LC}x_{1,k} < 0, \text{ then } \bar{\lambda}_{k+1} = 0\,. \tag{1.31}$$

$$\text{If } x_{2,k} - \frac{h}{LC}x_{1,k} \geqslant 0 \text{ then } \bar{\lambda}_{k+1} = L\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} \geqslant 0\,. \tag{1.32}$$

Inserting these values into (1.29) we get:

$$x_{2,k+1} = \begin{cases} (a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} & \text{if } x_{2,k} - \frac{h}{LC}x_{1,k} < 0 \\ \\ 0 & \text{if } x_{2,k} - \frac{h}{LC}x_{1,k} \geqslant 0 \end{cases}. \tag{1.33}$$

A Discretization of the Differential Inclusion (1.26)

Let us now propose an implicit time-discretization of the differential inclusion in (1.26), as follows:

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k+1} \\ \\ x_{2,k+1} - x_{2,k} + \frac{hR}{L}x_{2,k+1} + \frac{h}{LC}x_{1,k+1} \in \frac{1}{L}\partial\psi_{\mathbb{R}^+}(-x_{2,k+1}) \end{cases} \tag{1.34}$$

Notice that we can rewrite the second line of (1.34) as

$$x_{2,k+1} - (a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} \in \partial\psi_{\mathbb{R}^+}(-x_{2,k+1}) \tag{1.35}$$

where we have dropped the factor $\frac{1}{L}$ because $\partial\psi_{\mathbb{R}^+}(-x_{2,k+1})$ is a cone.

Let us now use two properties from convex analysis. Let $K \subset \mathbb{R}^n$ be a convex set, and let x and y be vectors of $\mathbb{R}^n$. Then

$$x - y \in -\partial\psi_K(x) \Longleftrightarrow x = \text{prox}[K;y]\,, \tag{1.36}$$

where "prox" means the closest element of K to y in the Euclidean metric, i.e., $x = \text{argmin}_{z\in K}\frac{1}{2}\parallel z - y \parallel^2$ (see (A.8) for a generalization in a metric M). Moreover using the chain rule of Proposition A.3 one has

$$\partial\psi_{\mathbb{R}^+}(-x) = -\partial\psi_{\mathbb{R}^-}(x)\ . \tag{1.37}$$

Using (1.36) and (1.37) one deduces from (1.35) that

$$x_{2,k+1} - (a(h))^{-1}\left\{x_{2,k} - \frac{h}{LC}x_{1,k}\right\} \in -\partial\psi_{\mathbb{R}^-}(x_{2,k+1})\ , \tag{1.38}$$

so that the algorithm becomes

$$\begin{cases} x_{1,k+1} - x_{1,k} = hx_{2,k+1} \\ \\ x_{2,k+1} = \text{prox}\left[\mathbb{R}^-; (a(h))^{-1}\left\{x_{2,k} - \dfrac{h}{LC}x_{1,k}\right\}\right] \end{cases}\ . \tag{1.39}$$

We therefore have proved the following:

Proposition 1.4. *The algorithm (1.28) is equivalent to the algorithm (1.34). They both allow one to advance from step k to step $k+1$, solving the proximation in (1.39).*

In Figs. 1.6 and 1.7, simulation results of the presented algorithm are given.

1.1.6.1 Approximating the Measure of an Interval

It is worthy to come back on the trick presented in Remark 1.3 that has been used to calculate the solution of the LCP in (1.29), i.e., to calculate $\tilde{\lambda}_{k+1} = h\lambda_{k+1}$ rather than λ_{k+1}.

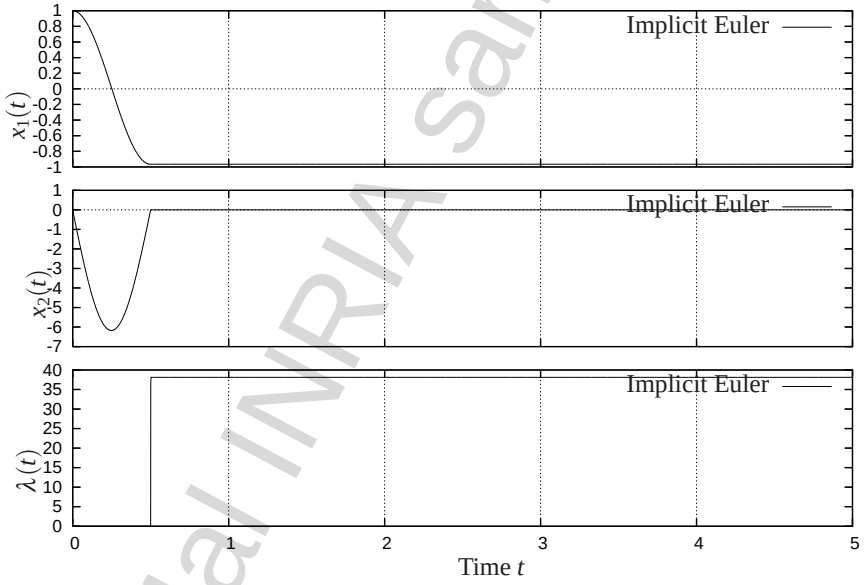


Fig. 1.6. Simulation of the circuit (**c**) with the initial conditions $x_1(0) = 1,\ x_2(0) = 0$ and $R = 0.1,\ L = 1,\ C = \frac{1}{(2\pi)^2}$. Time step $h = 1 \times 10^{-3}$

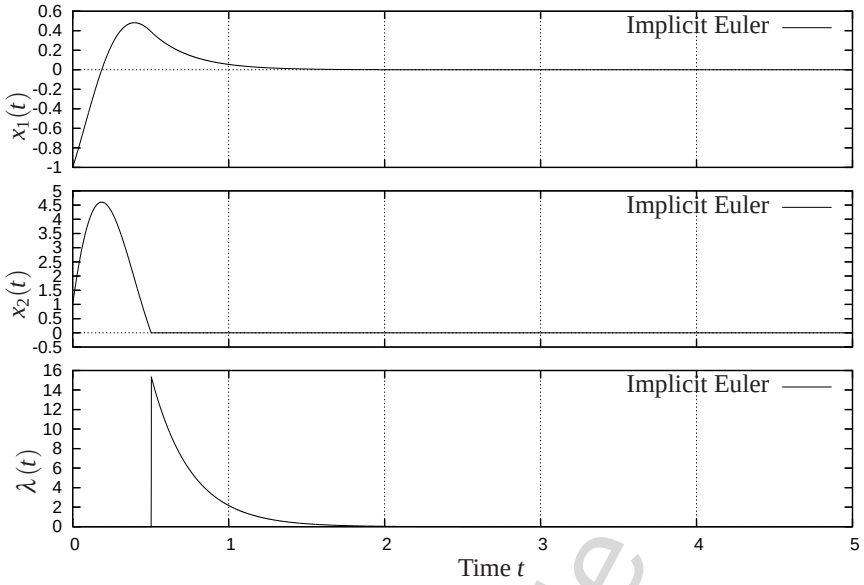


Fig. 1.7. Simulation of the circuit (**d**) with the initial conditions $x_1(0) = 1,\ x_2(0) = -1$ and $R = 10,\ L = 1,\ C = \frac{1}{(2\pi)^2}$. Time step $h = 1 \times 10^{-3}$

First of all, it follows from (1.34) and (1.28) that the element of the set $\partial\psi_{\mathbb{R}^+}(-x_{2,k+1})$ is not λ_{k+1}, but $\bar{\lambda}_{k+1}$. Retrospectively, our "trick" therefore appears not to be a trick, but a natural thing to do. Second, this means that the primary variables which are used in the integration are not $(x_{1,k}, x_{2,k}, \lambda_{k+1})$, but $(x_{1,k}, x_{2,k}, \bar{\lambda}_{k+1})$. Suppose that the initial value for the variable $x_2(\cdot)$ is negative. Then its right limit (supposed at this stage of the study to exist) has to satisfy $x_2(0^+) \geqslant 0$. Thus a jump occurs initially in $x_2(\cdot)$, so that the multiplier λ is at $t = 0$ a Dirac measure:[3]

$$\lambda = -L(x_2(0^+) - x_2(0^-))\delta_0 \tag{1.40}$$

The numerical scheme has to be able to approximate this measure! It is not possible numerically to achieve such a task, because this would mean approximating some kind of infinitely large value over one integration interval. However, what is quite possible is to calculate the value of

$$\mathrm{d}\lambda([t_k, t_{k+1}]) = \int_{[t_k, t_{k+1}]} \mathrm{d}\lambda\ , \tag{1.41}$$

i.e., the measure of the interval $[t_k, t_{k+1}]$.

[3] Throughout the book, right and left limits of a function F(·) will be denoted as $F(t^+)$ or $F^+(t)$, and $F(t^-)$ or $F^-(t)$, respectively.

Outside atoms of λ this is easy as λ is simply the Lebesgue measure. At atoms of λ this is again a bounded value. In fact, $\bar{\lambda}_{k+1} = h\lambda_{k+1}$ is an approximation of the measure of the interval by $\mathrm{d}\lambda$ i.e.,

$$\bar{\lambda}_{k+1} = h\lambda_{k+1} \approx \int_{[t_k, t_{k+1}]} \mathrm{d}\lambda \tag{1.42}$$

for each time-step interval.

Such an algorithm is therefore guaranteed to compute only *bounded* values, even if state jumps occur. Such a situation is common when we consider mechanical systems (see Sect. 1.4), dynamical complementarity systems (see Chap. 4), or higher relative degree systems (see Chap. 5).

Remark 1.5. A noticeable discrepancy between the equations (1.11) of the circuit (**a**) and the equations (1.13) of the circuit (**c**) is as follows. The complementarity relations in (1.11) are such that for any initial value of $x_1(\cdot)$ and $x_2(\cdot)$, there always exist a bounded value of the multiplier λ (which is a function of time and of the states) such that the integration proceeds. Such is not the case for (1.13), as pointed out just above. The *relative degree r* between w and λ plays a significant role in the dynamics (the relative degree is the number of times one needs to differentiate w in order to make λ appear explicitly: in (1.11) one has $r = 0$, but in (1.13) one has $r = 1$). A comprehensible presentation of the notion of relative degree is given in Chap. 4.

1.1.6.2 The Necessity of an Implicit Discretization

Another reason why considering the discretization of the inclusion in (1.25) is important is the following. Suppose one writes an explicit right-hand side $\partial\psi_{\mathbb{R}^+}(-x_{2,k})$ in (1.34) instead of the implicit form $\partial\psi_{\mathbb{R}^+}(-x_{2,k+1})$. Then after few manipulations and using (1.36) one obtains

$$\begin{array}{c} a(h)x_{2,k+1} + \dfrac{h}{LC}x_{1,k} - x_{2,k} \in \partial\psi_{\mathbb{R}^+}(-x_{2,k}) \\ \Updownarrow \\ x_{2,k} = \mathrm{prox}\left[\mathbb{R}^+; -a(h)x_{2,k+1} - \dfrac{h}{LC}x_{1,k}\right] \end{array} \tag{1.43}$$

which is absurd.

The implicit way of discretizing the inclusion is thus the only way that leads to a sound algorithm. This will still be the case with more general inclusions with right-hand sides of the form $\partial\psi_K(x)$ for some domain $K \subset \mathbb{R}^n$.

Let us now start from the complementarity formalism (1.28), with an explicit form

$$0 \leqslant \lambda_{k+1} \perp -x_{2,k} \geqslant 0\,. \tag{1.44}$$

Then we get the complementarity problem

$$0 \leqslant \lambda_{k+1} \perp -\left(1+\frac{h}{R}L\right) x_{2,k+1} - \frac{h}{LC} x_{1,k+1} - \frac{h}{L}\lambda_{k+1} \geqslant 0\,. \tag{1.45}$$

Clearly this complementarity problem cannot be used to advance the algorithm from step k to step $k+1$. This intrinsic implicit form of the discretization of the Differential Inclusion (DI) we work with here is not present in other types of inclusions, where explicit discretizations are possible, see Chap. 9.

1.1.7 Calculation of the Equilibrium Points

It is expected that studying the equilibrium points of complementarity systems as in (1.13) and (1.11) will lead either to a Complementarity Problem (CP) (like LCPs), or inclusions (see (1.7)), or variational inequalities (see (1.8)). Let us point out briefly the usefulness of the tools that have been introduced above, for the characterization of the equilibria of the class of nonsmooth systems we are dealing with.

In general one cannot expect that even simple complementarity systems possess a unique equilibrium. Consider for instance circuit (**c**) in (1.13). It is not difficult to see that the set of equilibria is given by $\{(x_1^*, x_2^*) |\, x_1^* \leqslant 0, x_2^* = 0\}$.

Let us consider now (1.26) and its equivalent (1.27). The fixed points z^* of the EVI in (1.27) have to satisfy

$$\langle -RAR^{-1}z^*, v - z^* \rangle \geqslant 0, \forall\, v \in K\,. \tag{1.46}$$

This is a variational inequality, and the studies concerning existence and uniqueness of solutions of a Variational Inequality (VI)are numerous. We may for instance use results in Yao (1994) which relate the set of solutions of (1.46) to the monotonicity of the operator $x \mapsto -RAR^{-1}x$. In this case, monotonicity is equivalent to semi-positive definiteness of $-RAR^{-1}$ and strong monotonicity is equivalent to positive definiteness of $-RAR^{-1}$ (Facchinei & Pang, 2003, p. 155). If the matrix $-RAR^{-1}$ is semi-positive definite, then Yao (1994, theorem 3.3) guarantees that the set of equilibria is nonempty, compact, and convex. If $-RAR^{-1}$ is positive definite, then from Yao (1994, theorem 3.5) there is a unique solution to (1.46), consequently a unique equilibrium for the system (1.26).

The monotonicity is of course a sufficient condition only. In order to see this, let us consider a linear complementarity system

$$\begin{cases} \dot{x}(t) = Ax(t) + B\lambda(t) \\ 0 \leqslant Cx(t) + D \perp \lambda(t) \geqslant 0 \end{cases}. \tag{1.47}$$

The fixed points of this LCS are the solutions of the problem

$$\begin{cases} 0 = Ax^* + B\lambda \\ 0 \leqslant Cx^* + D \perp \lambda \geqslant 0 \end{cases}. \tag{1.48}$$

If we assume that A is invertible, then we can construct the following LCP

$$0 \leqslant -CA^{-1}B\lambda + D \perp \lambda \geqslant 0 \tag{1.49}$$

which is not to be confused with the LCP in (1.24). If the matrix $-CA^{-1}B$ is a P-matrix then this LCP has a unique solution λ^* and we conclude that there is a unique equilibrium state $x^* = -A^{-1}B\lambda^*$. Clearly there is no monotonicity argument in this reasoning as the set of P-matrices contains that of positive definite matrices (i.e., a P-matrix is not necessarily positive definite).

As an illustration we may consider once again the circuits and (**c**) and (**d**). In the case of (1.13) we have

$$A = \begin{pmatrix} 0 & 1 \\ -\dfrac{1}{LC} & -\dfrac{R}{L} \end{pmatrix}, -CA^{-1}B = 0, \text{ and } D = 0\,. \tag{1.50}$$

There is an infinity of solutions for the LCP in (1.49), as pointed out above. In the case of (1.14) we have

$$A = \begin{pmatrix} -\dfrac{1}{RC} & 1 \\ -\dfrac{1}{LC} & 0 \end{pmatrix}, -CA^{-1}B = \frac{1}{R} > 0, \text{ and } D = 0\,. \tag{1.51}$$

There is a unique solution. We leave it to the reader to calculate explicitly the solutions (or the set of solutions). It is easily checked that no one of the two matrices $-A$ is semi-positive definite and they therefore do not define monotone operators. The sufficient criterion alluded to above is therefore not applicable.

In the case of circuit (**a**) with dynamics in (1.11), the fixed points are given as the solutions of a complementarity problem of the form

$$\begin{cases} 0 = Ax^* + B\lambda \\ 0 \leqslant Cx^* + D\lambda \perp \lambda \geqslant 0 \end{cases}, \tag{1.52}$$

where

$$A = \begin{pmatrix} -\dfrac{1}{RC} & 1 \\ -\dfrac{1}{LC} & 0 \end{pmatrix}, B = -\begin{pmatrix} \dfrac{1}{R} \\ \dfrac{1}{L} \end{pmatrix}, C = \begin{pmatrix} \dfrac{1}{RC} & -1 \end{pmatrix}, \text{ and } D = \frac{1}{R} \tag{1.53}$$

Since A is invertible, with inverse

$$A^{-1} = \begin{pmatrix} 0 & -1 \\ \dfrac{1}{LC} & -\dfrac{1}{RC} \end{pmatrix}$$

one can express x^* as $x^* = -A^{-1}B\lambda$. Therefore $Cx^* + D\lambda = (D - CA^{-1}B)\lambda$, and the LCP is: $0 \leqslant \lambda \perp (D - CA^{-1}B)\lambda \geqslant 0$. The solution is $\lambda = 0$ independently of

the sign of the scalar $D - CA^{-1}B$. This can also be seen from the inclusion $\lambda \in -\partial\psi_{\mathbb{R}^+}((D - CA^{-1}B)\lambda)$, taking into account (1.6).

It is noteworthy that computing the fixed points of our circuits may be done by solving LCPs. In dimension 1 or 2, this may be done by checking the two or four possible cases, respectively. In higher dimensions, such enumerative procedures become impossible, and specific algorithms for solving LCPs (or other kinds of CPs) have to be used. Such algorithms will be described later in Part III.

1.2 Electrical Circuits with Ideal Zener Diodes

1.2.1 The Zener Diode

Let us consider, now, a further electrical device: the ideal Zener diode whose schematic symbol is depicted in Fig. 1.8a. A Zener diode is a type of diode that permits current to flow in the forward direction like a normal diode, but also in the reverse direction if the voltage is larger than the rated breakdown voltage known as "Zener knee voltage" or "Zener voltage" denoted by $V_z > 0$. The ideal characteristic between the current $i(t)$ and the voltage $v(t)$ can be seen in Fig. 1.8b.

Let us seek an analytical representation of the current–voltage characteristic of the ideal Zener diode. For this we are going to use some convex analysis tools and make some manipulations: subdifferentiate, conjugate, invert. Let us see how this works, with Fig. 1.9 as a guide.

The inversion consists of expressing $v(t)$ as a function of $-i(t)$: this is done in Fig. 1.9b. Computing the subderivative of the function $f(\cdot)$ of Fig. 1.9c, one gets the multivalued mapping of Fig. 1.9b. Indeed we have

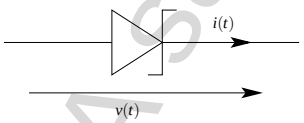


Fig. 1.8a. The Zener diode schematic symbol

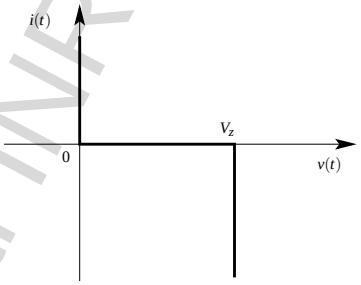


Fig. 1.8b. The ideal characteristic of a Zener diode

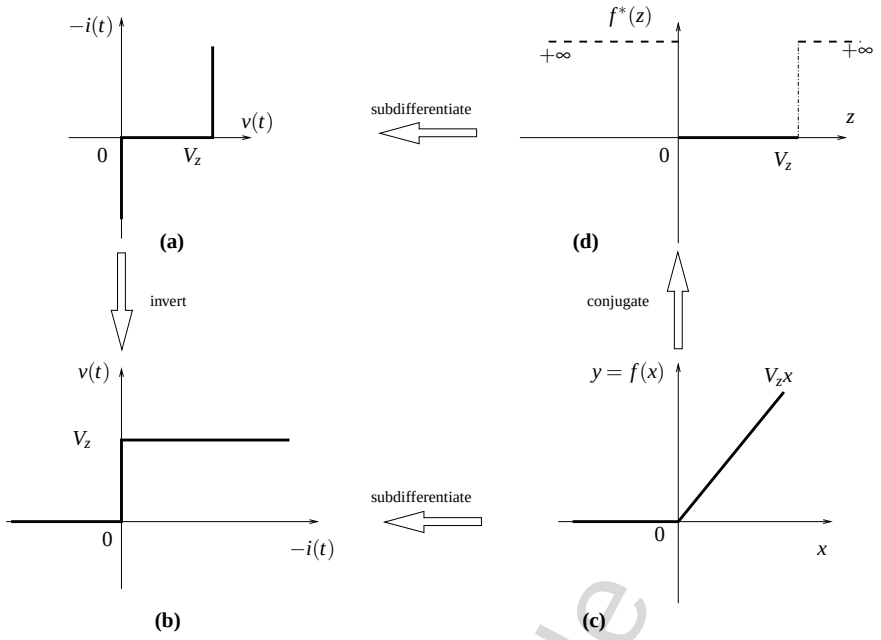


Fig. 1.9. The Zener diode characteristic

$$f(x) = \begin{cases} V_z x \text{ if } x \geqslant 0 \\ 0 \quad \text{if } x < 0 \end{cases} \tag{1.54}$$

from which it follows that the subdifferential of $f(\cdot)$ is

$$\partial f(x) = \begin{cases} V_z & \text{if } x > 0 \\ [0, V_z] & \text{if } x = 0 \\ 0 & \text{if } x < 0 \end{cases} . \tag{1.55}$$

Notice that the function $f(\cdot)$ is convex, proper, continuous, and that the graphs of the multivalued mappings of Fig. 1.9a and b are maximal monotone. *Monotonicity* means that if you pick any two points $-i_1$ and $-i_2$ on the abscissa of Fig. 1.9b, and the corresponding v_1 and v_2, then it is always true that

$$\langle -i_1 - (-i_2), v_1 - v_2 \rangle \geqslant 0 \tag{1.56}$$

Similarly for Fig. 1.9a *maximality* means that it is not possible to add any new branch to the graphs of these mappings, without destroying the monotonicity. This is indeed the case for the graphs of Fig. 1.9a and b.

Let us now introduce the notion of the conjugate of a convex function $f(\cdot)$ that is defined as

$$f^*(z) = \sup_{x\in\mathbb{R}}(\langle x,z\rangle - f(x)) \,. \tag{1.57}$$

Let us calculate the conjugate of the function $f(\cdot)$ above:

$$\begin{aligned} f^*(z) &= \sup_{x\in\mathbb{R}} \begin{cases} xz - V_z x \text{ if } x \geqslant 0 \\ xz \qquad\quad \text{if } x < 0 \end{cases} = \sup_{x\in\mathbb{R}} \begin{cases} x(z-V_z) \text{ if } x \geqslant 0 \\ xz \qquad\quad \text{if } x < 0 \end{cases} \\ &= \begin{cases} \begin{cases} +\infty \text{ if } z > V_z \\ 0 \quad \text{if } z \leqslant V_z \end{cases} \\ \\ \begin{cases} 0 \quad \text{if } z \geqslant 0 \\ +\infty \text{ if } z < 0 \end{cases} \end{cases} = \begin{cases} +\infty \text{ if } z < 0 \text{ and } z > V_z \\ \\ 0 \quad \text{if } 0 \leqslant z \leqslant V_z \end{cases} \\ &= \psi_{[0,V_z]}(z) \,, \end{aligned} \tag{1.58}$$

where we retrieve the indicator function that was already met when we considered the ideal diode, see Sect. 1.1.1.

We therefore deduce from Fig. 1.9 that

$$-i(t) \in \partial\psi_{[0,V_z]}(v(t)), \text{ whereas } v(t) \in \partial f(-i(t)) \,. \tag{1.59}$$

The function $f(\cdot) = \psi^*_{[0,V_z]}(\cdot)$ is called in convex analysis the *support* function of the set $[0,V_z]$. It is known that the support function and the indicator function of a convex set are conjugate to one another.

We saw earlier that the subderivative of the indicator function of a convex set is also the normal cone to this convex set. Here we obtain that $\partial\psi_{[0,V_z]}(v(t))$ is the normal cone $N_{[0,V_z]}(v(t))$, that is $\mathbb{R}^-$ when $v(t) = 0$ and $\mathbb{R}^+$ when $v(t) = V_z$. It is the singleton $\{0\}$ when $0 < v(t) < V_z$.

1.2.2 The Dynamics of a Simple Circuit

Differential Inclusions and Filippov's Systems

Now that these calculations have been led, let us consider the dynamics of the circuit in Fig. 1.3c, where we replace the ideal diode by an ideal Zener diode. Choosing the same state variables (x_1 is the capacitor charge, x_2 is the current through the circuit), we obtain:

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) = \dfrac{1}{L}v(t) \end{cases}, \tag{1.60}$$

where $v(\cdot)$ is the voltage of the Zener diode. We saw that $v(t) \in \partial f(-i(t))$, thus we get

$$\begin{cases} \dot{x}_1(t) - x_2(t) = 0 \\ \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) \in \dfrac{1}{L}\partial f(-x_2(t)) \end{cases}, \tag{1.61}$$

which is a differential inclusion.

Compare the inclusions in (1.25) and in (1.61). They look quite similar, however, the sets in their right-hand sides are quite different. Indeed the set in the right-hand side of (1.25) is unbounded, whereas the set in the right-hand side of (1.61) is bounded, as it is included in $[0, V_z]$. More precisely, the set-valued mapping $\partial f(\cdot)$ is nonempty, compact, convex, upper semi-continuous, and satisfies a linear growth condition: for all $v \in \partial f(x)$ there exists constants k and a such that $\| v \| \leqslant k \| x \| + a$.

The differential inclusion (1.61) possesses an absolutely continuous solution, and we may even assert here that this solution is unique for each initial condition, because in addition the considered set-valued mapping is maximal monotone, see Lemma 2.13, Theorem 2.41. This is also sometimes called a Filippov's system or a Filippov's DI, associated with the switching surface $\Sigma = \{x \in I\!R^2 \mid x_2 = 0\}$. See Sect. 2.1 for a precise definition of Filippov's systems. Simple calculations yield that the vector field in the neighborhood of Σ is as depicted in Fig. 1.10. The surface Σ is crossed transversally by the trajectories when $x_1(t) < 0$ and $x_1(t) > CV_z$. It is an attracting surface when $x_1(t) \in [0, V_z]$ (where t means the time when the trajectory attains Σ). According to Filippov's definition of the solution, Σ is a sliding surface in the latter case, which means that $x_2(t) = 0$ after the trajectory has reached this portion of Σ. Notice that we may rewrite the second line in (1.61) as

$$\dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) = \lambda(t), \;\; \lambda(t) \in \frac{1}{L}\partial f(-x_2(t)) \,. \tag{1.62}$$

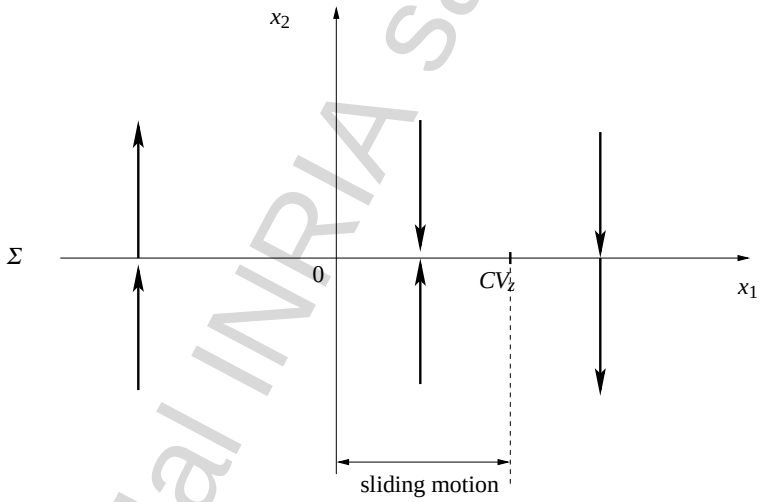


Fig. 1.10. The vector field on the switching surface Σ

Despite passing from (1.61) to (1.62) looks like wasted effort, it means that the inclusion in (1.61) is equivalent to integrate its left-hand side by looking for an element of the set in its right-hand side, at each time instant. This is in fact the case for *all* the differential inclusions that we shall deal with in this book. In other words the integration proceeds along Σ with an element $\lambda \in \partial f(0)$ such that $\lambda(t) = \frac{x_1(t)}{C}$, where t is the "entry" time of the trajectory in Σ (notice that as long as $x_2 = 0$ then x_1 remains constant).

Remark 1.6. The fact that the switching surface Σ is attracting in $x_1(t) \in [0, V_z]$, is intimately linked with the maximal monotonicity of the set-valued mapping $\partial f(\cdot)$. This mapping is sometimes called a *relay* function in the systems and control community (Fig. 1.11).

A First Complementarity System Formulation

Let us now seek a complementarity formulation of the multivalued mapping $\partial f(\cdot) = \partial \psi^*_{[0,V_z]}(\cdot)$ whose graph is in Fig. 1.9a. Let us introduce two slack variables (or multipliers) λ_1 and λ_2, and the set of conditions:

$$\begin{cases} 0 \leqslant \lambda_1(t) \perp -i(t) + |i(t)| \geqslant 0 \\ 0 \leqslant \lambda_2(t) \perp i(t) + |i(t)| \geqslant 0 \\ \lambda_1(t) + \lambda_2(t) = V_z \\ v(t) = \lambda_2(t) \end{cases} . \tag{1.63}$$

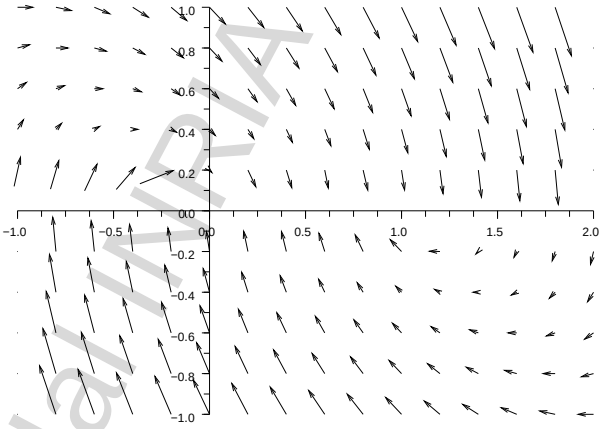


Fig. 1.11. Example of the vector field on the switching surface Σ for $R = C = L = V_z = 1$

Let us check by inspection that indeed (1.63) represents the mapping of Fig. 1.9a. If $-i(t) > 0$, then $-i(t) + |i(t)| > 0$, so $\lambda_1(t) = 0$ and $\lambda_2(t) = V_z = v(t)$ (and $i(t) + |i(t)| = 0$). If $-i(t) < 0$ then $i(t) + |i(t)| > 0$, so $\lambda_2(t) = 0$, and $\lambda_1(t) = V_z$ (and $-i(t) + |i(t)| = 0$) and $v(t) = \lambda_2(t) = 0$. Now if $i(t) = 0$, then one easily calculates that $0 \leqslant \lambda_1(t) \leqslant V_z$, $0 \leqslant \lambda_2(t) \leqslant V_z$. Thus $0 \leqslant v(t) \leqslant V_z$.

Thanks to the complementary formulation (1.63), the inclusion (1.61) can be formulated as a Dynamical (or Differential) Complementarity System (DCS)

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) = \dfrac{1}{L}v(t) \\ 0 \leqslant \lambda_1(t) \perp -x_2(t) + |x_2(t)| \geqslant 0 \\ 0 \leqslant \lambda_2(t) \perp \quad x_2(t) + |x_2(t)| \geqslant 0 \\ \lambda_1(t) + \lambda_2(t) = V_z \\ v(t) = \lambda_2(t) \end{cases} . \tag{1.64}$$

This DCS is not an LCS due to the presence of the absolute value function in the complementarity condition and the two last algebraic equations. We notice that the variables $\lambda_1(t)$ and $\lambda_2(t)$ can be eliminated from (1.64) using the last two equalities, leading to another formulation of the DCS:

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) = \dfrac{1}{L}v(t) \\ 0 \leqslant V_z - v(t) \perp -x_2(t) + |x_2(t)| \geqslant 0 \\ 0 \leqslant v(t) \perp x_2(t) + |x_2(t)| \geqslant 0 \end{cases} \tag{1.65}$$

which is neither an LCS.

A Mixed Linear Complementarity Formulation

It is possible from (1.63) to obtain a so-called Mixed Linear Complementarity System (MLCS) which is a generalization of an LCS with an additional system of linear equations. The goal is to obtain after discretization a so-called Mixed Linear Complementarity Problem (MLCP) which is a generalization of an LCP with an additional system of linear equations, such that

$$\begin{cases} Au + Cw + a = 0 \\ 0 \leqslant w \perp Du + Bw + d \geqslant 0 \end{cases} . \tag{1.66}$$

To obtain an MLCS formulation, let us introduce the positive part and the negative part of the current $i(t)$ as

$$i^+(t) = \frac{1}{2}(i(t) + |i(t)|) = \max(0, i(t)) \geqslant 0\,, \tag{1.67}$$

$$i^-(t) = \frac{1}{2}(i(t) - |i(t)|) = \min(0, i(t)) \leqslant 0\,. \tag{1.68}$$

The system (1.63) can be rewritten equivalently as

$$\begin{cases} 0 \leqslant \lambda_1(t) \perp i^+(t) - i(t) \geqslant 0 \\ 0 \leqslant \lambda_2(t) \perp i^+(t) \geqslant 0 \\ i(t) = i^-(t) + i^+(t) \\ \lambda_1(t) + \lambda_2(t) = V_z \\ v(t) = \lambda_2(t) \end{cases}, \tag{1.69}$$

where the absolute value has disappeared, but a linear equation has been added. Substitution of two of the last three equations into the complementarity conditions leads to an intermediate complementarity formulation of the relay function as

$$\begin{cases} 0 \leqslant \lambda_1(t) \perp i^+(t) - i(t) \geqslant 0 \\ 0 \leqslant v(t) \perp i^+(t) \geqslant 0 \\ \lambda_1(t) + v(t) = V_z \end{cases} \tag{1.70}$$

or as

$$\begin{cases} 0 \leqslant V_z - v(t) \perp i^+(t) - i(t) \geqslant 0 \\ 0 \leqslant v(t) \perp i^+(t) \geqslant 0 \end{cases}. \tag{1.71}$$

The linear dynamical system (1.60) together with one of the reformulations (1.69), (1.70), or (1.71) leads to an MLCS formulation. Nevertheless, the complete substitution of the equation into the complementarity condition yields a DCS

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) = \dfrac{1}{L}v(t) \\ 0 \leqslant V_z - v(t) \perp x_2^+(t) - x_2(t) \geqslant 0 \\ 0 \leqslant v(t) \perp x_2^+(t) \geqslant 0 \end{cases}, \tag{1.72}$$

which is neither an LCS nor an MLCS.

A Linear Complementarity Formulation

Due to the simplicity of the equations involved in the MLCS formulation (1.71), it is possible to find an LCS formulation of the dynamics. Indeed, the following system

$$\begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) = \dfrac{1}{L}\lambda_2(t) \\ x_2^+(t) = x_2(t) - x_2^-(t) \\ \lambda_1(t) = V_z - \lambda_2(t) \\ 0 \leqslant \begin{pmatrix} x_2^+(t) \\ \lambda_1(t) \end{pmatrix} \perp \begin{pmatrix} \lambda_2(t) \\ -x_2^-(t) \end{pmatrix} \geqslant 0 \end{cases} \tag{1.73}$$

can be recast into the following LCS form

$$\begin{cases} \dot{x}(t) = Ax(t) + B\tilde{\lambda}(t) \\ w(t) = Cx(t) + D\tilde{\lambda}(t) + g \\ 0 \leqslant w(t) \perp \tilde{\lambda}(t) \geqslant 0 \end{cases} \tag{1.74}$$

with

$$A = \begin{bmatrix} 0 & 1 \\ -\dfrac{1}{LC} & -\dfrac{R}{L} \end{bmatrix}, B = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, C = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, D = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, g = \begin{bmatrix} 0 \\ V_z \end{bmatrix}. \tag{1.75}$$

The reformulation appears to be a special case for more general reformulations of relay systems or two-dimensional friction problems into LCS. For more details, we refer to Pfeiffer & Glocker (1996) and to Sect. 9.3.3. In the more general framework of ODE with discontinuous right-hand side, an LCS reformulation can be found in Chap. 7.

1.2.3 Numerical Simulation by Means of Time-Stepping Schemes

In view of this preliminary material, we may consider now the time-discretization of our system. Clearly our objective here is still to introduce the topic, and the reader should not expect an exhaustive description of the numerical simulation of the system.

1.2.3.1 Explicit Time-Stepping Schemes Based on ODE with Discontinuities Formulations

A forward Euler scheme may be applied on an ODE with discontinuities of the form, $\dot{x} = f(x,t)$, where the right-hand side may possess discontinuities (see Sect. 2.8). For

the right-hand side of the circuit with the Zener diode, a switched model may be given by

$$f(x,t) = \begin{cases} \begin{bmatrix} x_2 \\ -\dfrac{R}{L}x_2 - \dfrac{1}{LC}x_1 \end{bmatrix} & \text{for } -x_2 < 0 \quad (1.76a) \\ \begin{bmatrix} 0 \\ 0 \end{bmatrix} & \text{for } -x_2 = 0 \quad (1.76b) \\ \begin{bmatrix} x_2 \\ -\dfrac{R}{L}x_2 - \dfrac{1}{LC}x_1 - V_z \end{bmatrix} & \text{for } -x_2 > 0. \quad (1.76c) \end{cases}$$

The simulation for this choice of the right-hand side is illustrated in Fig. 1.12. We can observe that some "chattering" effects due to the fact that the sliding mode given by (1.76b) cannot be reached due to the numerical approximation on x_2. This artifact results in spurious oscillations of the diode voltage $v(t) = \lambda(t)$ and the diode current $x_2(t) = \omega(t)$ as we can observe on the zoom in Fig. 1.13.

One way to circumvent the spurious oscillations is to introduce a "sliding band", i.e., an interval where the variable x_2 is small in order to approximate the sliding mode. This interval can be for instance chosen as $|x_2| \leqslant \eta$ such that the new right-hand side is given by

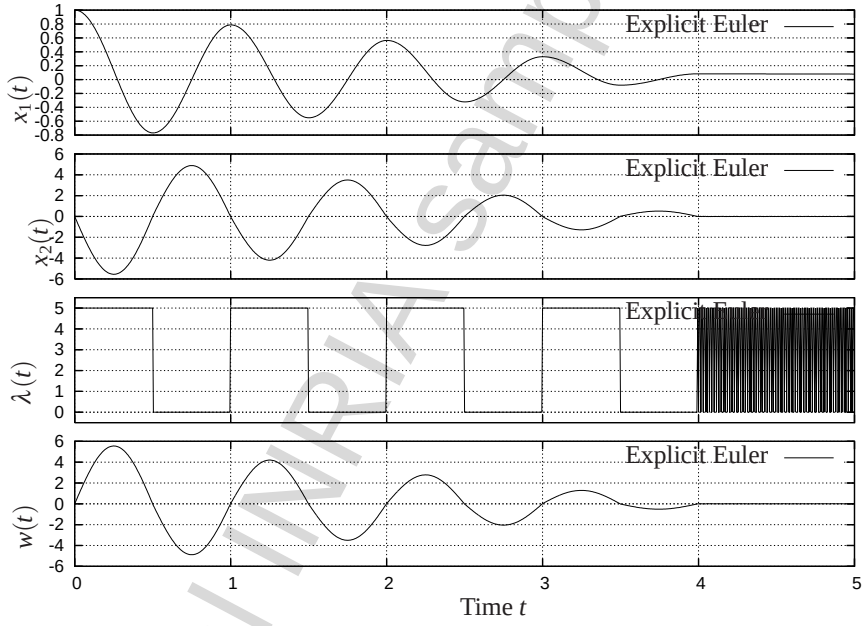


Fig. 1.12. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Explicit Euler scheme with the right-hand side defined by (1.76). Time step $h = 5 \times 10^{-3}$

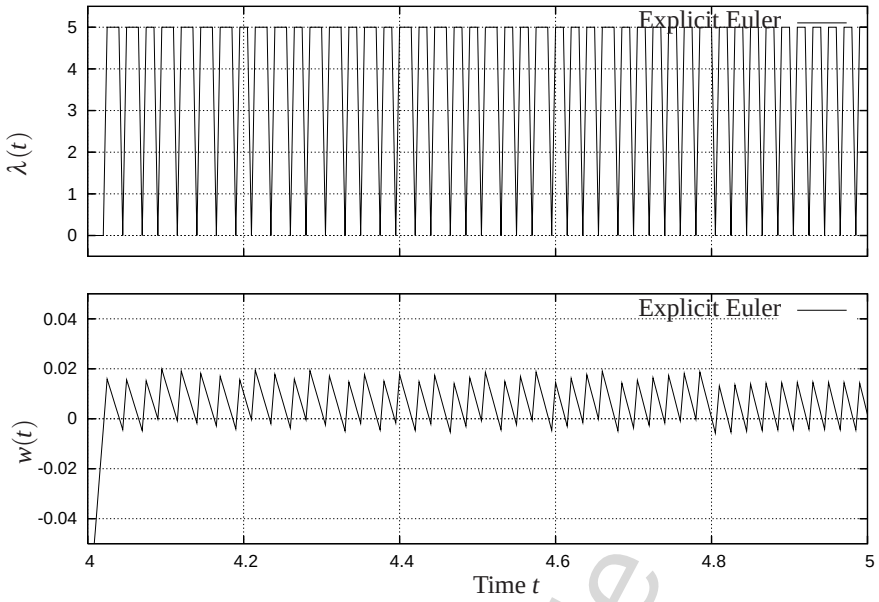


Fig. 1.13. Zoom on the "chattering" behavior simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Explicit Euler scheme with the right-hand side defined by (1.76). Time step $h = 5 \times 10^{-3}$

$$f(x,t) = \begin{cases} \begin{bmatrix} x_2 \\ -\dfrac{R}{L}x_2 - \dfrac{1}{LC}x_1 \end{bmatrix} & \text{for } -x_2 < -\eta \qquad (1.77a) \\ \begin{bmatrix} x_2 \\ -\dfrac{R}{L}x_2 \end{bmatrix} & \text{for } |x_2| \leqslant \eta \qquad (1.77b) \\ \begin{bmatrix} x_2 \\ -\dfrac{R}{L}x_2 - \dfrac{1}{LC}x_1 - V_z \end{bmatrix} & \text{for } -x_2 > \eta \qquad (1.77c) \end{cases}$$

Simulation results depicted in the Figs. 1.14 and 1.15 show that the spurious oscillations have been cancelled.

The switched models (1.76) and (1.77) are incomplete models. In more general situations they may fail due the lack of conditions for the transition from the sliding mode to the other modes. Clearly, the value of the dual variable $\lambda(t) = v(t)$ has to be checked to know if the system stays in the sliding mode. We will see in Sect. 9.3.3 that all these conditional statements can be in numerous cases replaced by an LCP formulation.

It is noteworthy that the previous numerical trick is not an universal solution for the problem of chattering. Indeed, the switched model given by the right-hand side (1.77) allows the solution to stay near the boundary of the sliding band. The new

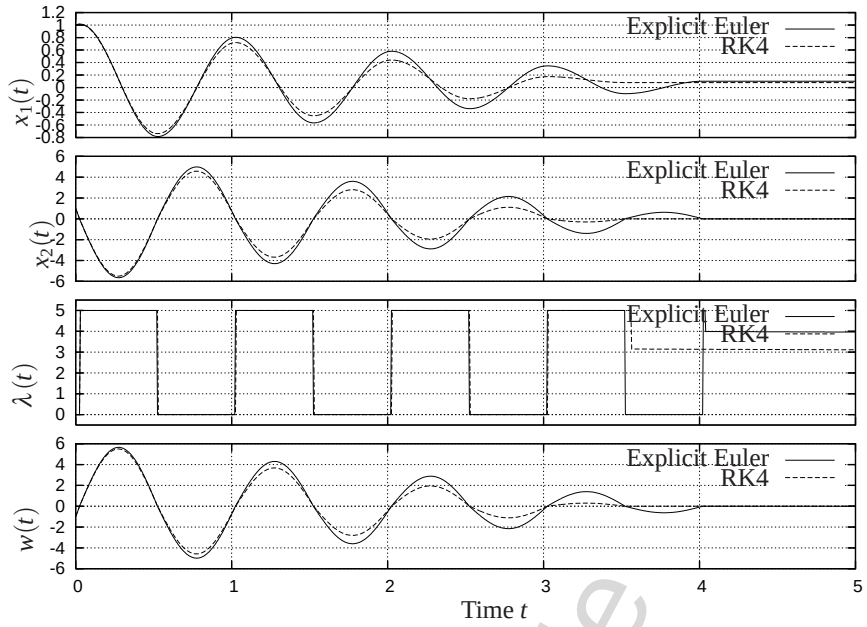


Fig. 1.14. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Euler and four order Runge–Kutta explicit scheme with the right-hand side defined by (1.77). Time step $h = 5 \times 10^{-3}$

model is still a discontinuous system and therefore some numerical instabilities of the ODE solver can appear. More smart approaches for the choice of the right-hand side in the sliding band can be found in Karnopp (1985), Leine et al. (1998), Leine & Nijmeijer (2004) and will be described in Sect. 9.3.2.

The fact that we are able to express the Filippov's DI as an equivalent model of ODE with a switched right-hand side allows one to use any other explicit schemes such as explicit Runge–Kutta methods. In Figs. 1.15 and 1.16, the results of the simulation with the right-hand side (1.76) and (1.77) are depicted. The conclusions are the same as above. One notices also that two different methods provide different results (see Figs. 1.14 and 1.15). We will discuss in Sect. 9.2 the question of the order and the stability of such a higher order method for Filippov's DIs.

1.2.3.2 Explicit Discretization of the Differential Inclusion and the Complementarity Systems

Explicit Discretization of the Differential Inclusion (1.61)

Consider the forward Euler method

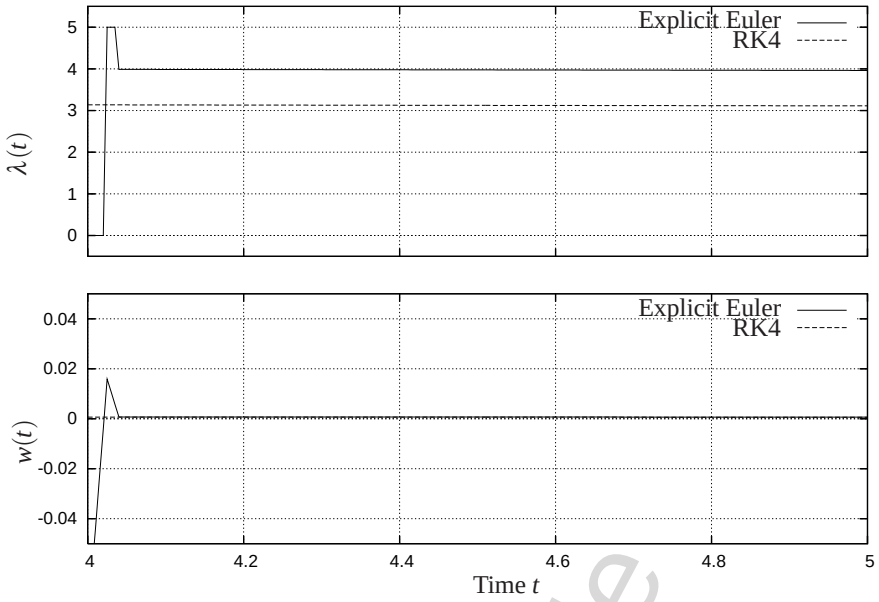


Fig. 1.15. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Euler and four order Runge–Kutta explicit scheme with the right-hand side defined by (1.77). Time step $h = 5 \times 10^{-3}$

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L} x_{2,k} + \dfrac{h}{LC} x_{1,k} \in \dfrac{h}{L} \partial f(-x_{2,k}) \,, \end{cases} \tag{1.78}$$

where x_k is the value, at time t_k of a grid $t_0 < t_1 < \cdots < t_N = T$, $N < +\infty$, $h = \dfrac{T - t_0}{N} = t_k - t_{k-1}$, of a step function $x^N(\cdot)$ that approximates the analytical solution $x(\cdot)$.

Compare with the time-discretization of the inclusion (1.25) that is proposed in Sect. 1.1.5. This time considering an implicit scheme is not mandatory (this may improve the overall quality of the numerical integration especially from the stability point of view, but is not a consequence of the dynamics contrary to what happens with (1.25)). One of the major discrepancies with the circuit (1.25) is that the values of x_2 are no longer constrained to stay in a set by the inclusion (1.78).

Explicit Discretization of the Complementarity Systems (1.65)

Let us investigate how the complementarity system (1.65) may be discretized. We get

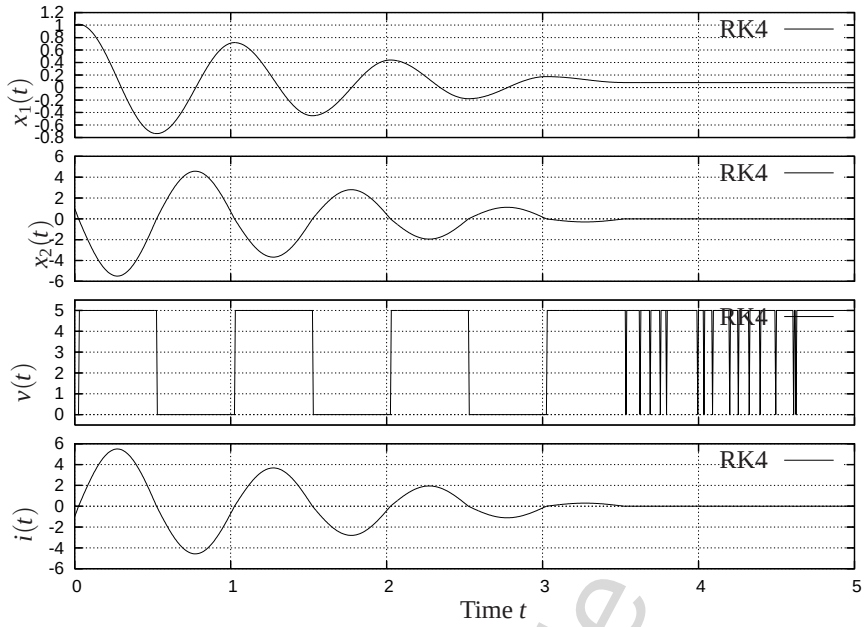


Fig. 1.16. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \frac{1}{(2\pi)^2}$. Four order Runge–Kutta explicit scheme with the right-hand side defined by (1.76). Time step $h = 5 \times 10^{-3}$

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L} x_{2,k} + \dfrac{h}{LC} x_{1,k} = \dfrac{h}{L} \lambda_{2,k} \\ 0 \leqslant V_z - \lambda_{2,k} \perp -x_{2,k} + |x_{2,k}| \geqslant 0 \\ 0 \leqslant \lambda_{2,k} \perp \quad x_{2,k} + |x_{2,k}| \geqslant 0 \end{cases} . \tag{1.79}$$

One computes that if $x_{2,k} > 0$ then $\lambda_{2,k} = 0$, while $x_{2,k} < 0$ implies $\lambda_{2,k} = V_z$. Moreover $x_{2,k} = 0$ implies that $\lambda_{2,k} \in [0, V_z]$. We conclude that the two schemes in (1.78) and (1.79) are the same.

However, the complementarity formalism does not bring any advantage over the inclusion formalism, as it does not yield neither an LCP nor an MLCP, even with the reformulation proposed in the preceding section. The main reason for that is not the presence of absolute values in the complementarity formalism which can be avoided by adding an equality, but the fact that $\lambda_{2,k}$ has to be complementary to the positive part of $x_{2,k}$ which is not an unknown at the beginning of the step.

For instance, if we choose the MLCS formulation given by the dynamical system (1.60) and the formulation (1.70), we get the following complementarity problem

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L} x_{2,k} + \dfrac{h}{LC} x_{1,k} = \dfrac{h}{L} \lambda_{2,k} \\ 0 \leqslant \lambda_{1,k} \perp x_{2,k}^{+} - x_{2,k} \geqslant 0 \\ 0 \leqslant \lambda_{2,k} \perp x_{2,k}^{+} \geqslant 0 \\ \lambda_{1,k} + \lambda_{2,k} = V_{\mathrm{z}} \end{cases} . \tag{1.80}$$

In such a "fake" complementarity problem, one has to perform the procedure described in the Remark 1.7, which implies to choose a threshold on the value of $x_{2,k}$.

To conclude this part, whatever the mathematical formalism which is used to formulate the dynamics, explicit discretizations lead to algorithms without any sense.

Remark 1.7. One has to choose a value for $\lambda_{2,k}$ in the interval $[0, V_{\mathrm{z}}]$ when $x_{2,k} = 0$. More concretely when implementing the algorithm on a computer, one has to choose a threshold $\eta > 0$ such that $x_{2,k}$ is considered to be null when $|x_{2,k}| \leqslant \eta$. One possibility is to choose the Filippov's solution that makes the trajectory slide on the surface $\Sigma = \{x \in \mathbb{R}^2 \mid x_2 = 0\}$. If $x_{1,k} \notin [0, CV_{\mathrm{z}}]$ we have seen that the trajectories cross transversally Σ. Thus the chosen value of $\lambda_{2,k}$ is not important. If $x_{1,k} \in [0, CV_{\mathrm{z}}]$ one may simply choose $\lambda_{2,k} = \frac{x_{1,k}}{C}$ or $\lambda_{2,k} = -\frac{L}{h} x_{2,k} + R x_{2,k} + \frac{x_{1,k}}{C}$ to keep $x_{2,k+1}$ in the required neighborhood of Σ. With the solution, we have also to check the value of the dual variable $v(t) = \lambda_2(t)$ to know when the application of this rule has to be stopped.

1.2.3.3 An Implicit Time-Stepping Scheme

Implicit Discretization of the Differential Inclusion (1.61)

Let us try the following implicit scheme[4]:

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L} x_{2,k+1} + \dfrac{h}{LC} x_{1,k+1} \in \dfrac{h}{L} \partial f(-x_{2,k+1}) \, . \end{cases} \tag{1.81}$$

After some manipulations this may be rewritten as

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} + a(h) \left[\dfrac{h}{LC} x_{1,k} - x_{2,k} \right] \in a(h) \dfrac{h}{L} \partial f(-x_{2,k+1}) \, , \end{cases} \tag{1.82}$$

[4] The scheme chosen here is fully implicit for the sake of simplicity.

where $a(h) = \left(1 + \dfrac{hR}{L} + \dfrac{h^2}{LC}\right)^{-1}$.

Denoting

$$b = a(h)\left[\frac{h}{LC}x_{1,k} - x_{2,k}\right]$$

the second line of (1.82) may be rewritten as

$$x_{2,k+1} + b \in a(h)\frac{h}{L}\partial f(-x_{2,k+1})\ . \tag{1.83}$$

It is this inclusion that we are going to examine now. This will allow us to illustrate graphically why the monotonicity is a crucial property. In Fig. 1.17 the graph of the linear function

$$\mathscr{D}_b = \left\{(\lambda_{2,k+1}, x_{2,k+1}) \in \mathbb{R}^2 \mid \quad \lambda_{2,k+1} = x_{2,k+1} + b\right\}$$

is depicted for three values of b, together with the graph of the set-valued function,

$$\mathscr{G} = \left\{(\lambda_{2,k+1}, x_{2,k+1}) \in \mathbb{R}^2 \mid \quad \lambda_{2,k+1} \in a(h)\frac{h}{L}\partial f(-x_{2,k+1})\right\}\ .$$

It is apparent that for any value of b, there is always a single intersection between the two graphs. One concludes that the generalized equation (1.83) with unknown $x_{2,k+1}$ has a unique solution, which allows one to advance the algorithm from k to $k+1$.

If there is an exogenous input $u(t)$ that acts on the system so that the dynamics is changed to

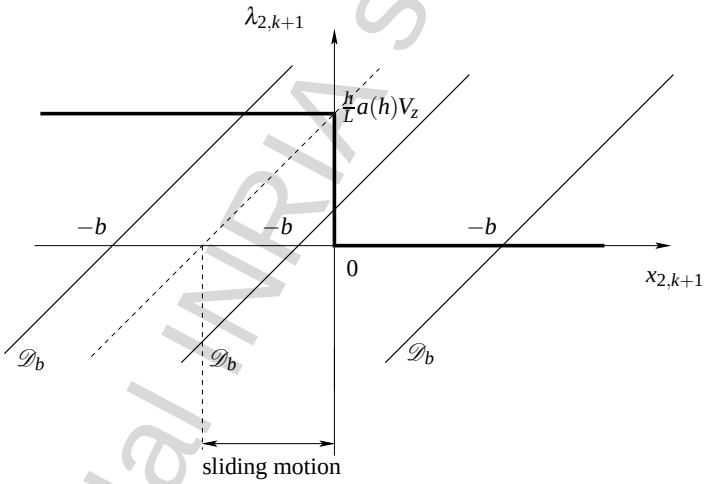


Fig. 1.17. Implicit scheme: the maximal monotone case

$$\dot{x}_2(t) + \frac{R}{L}x_2(t) + \frac{1}{LC}x_1(t) + \frac{u(t)}{L} \in \frac{1}{L}\partial f(-x_2(t)) \tag{1.84}$$

then the variable b is changed to $b + a(h)\dfrac{u_k}{L}$. Varying u_{k+1} corresponds to a horizontal translations of the straight lines in Fig. 1.17.

Remark 1.8 (A nonmonotone example). Suppose now that the dynamics is

$$x_{2,k+1} + b \in -a(h)\frac{h}{L}\partial f(-x_{2,k+1})\,. \tag{1.85}$$

We know this is not possible with the circuit we are studying. For the sake of the reasoning we are leading let us imagine this is the case. Then we get the situation depicted in Fig. 1.18. There exist values of b for which the generalized equation has two or three solutions. Uniqueness is lost.

Remark 1.9 (Comparison with the procedure in Remark 1.7). Coming back to Fig. 1.17, one sees that the values of b that yield a sliding motion along the surface Σ, correspond to all the values such that the graph of the linear function intersects the vertical segment of the graph of the multifunction. Contrarily to what happens with the explicit scheme where a threshold has to be introduced, "detecting" the sliding motion is now the result of a resolution of the intersection problem. No artificial threshold is needed due to the fact that we have to verify the inclusion of a value into a set of nonempty interior.

Implicit Discretization of the Complementarity Systems

Let us choose one of the LCS formulations described in the previous section given by the dynamics (1.73). An implicit time-discretization is given by

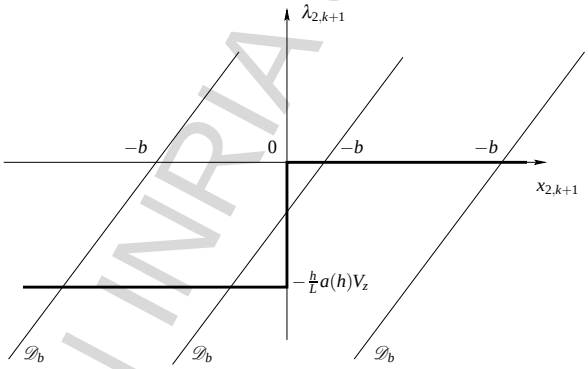


Fig. 1.18. Implicit scheme: the nonmonotone case

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} - x_{2,k} + \dfrac{hR}{L} x_{2,k+1} + \dfrac{h}{LC} x_{1,k+1} = \dfrac{h}{L} \lambda_{2,k+1} \\ x^+_{2,k+1} = x_{2,k+1} - x^-_{2,k+1} \\ \lambda_{1,k+1} = V_z - \lambda_{2,k+1} \\ 0 \leqslant \begin{pmatrix} x^+_{2,k+1} \\ \lambda_{1,k+1} \end{pmatrix} \perp \begin{pmatrix} \lambda_{2,k+1} \\ -x^-_{2,k+1} \end{pmatrix} \geqslant 0 \end{cases} . \tag{1.86}$$

Using the previous notations for $a(h)$ and b, we get the following system

$$\begin{cases} x_{1,k+1} - x_{1,k} = h x_{2,k+1} \\ x_{2,k+1} + b = a(h) \dfrac{h}{L} \lambda_{2,k+1} \\ x^+_{2,k+1} = x_{2,k+1} - x^-_{2,k+1} \\ \lambda_{1,k+1} = V_z - \lambda_{2,k+1} \\ 0 \leqslant \begin{pmatrix} x^+_{2,k+1} \\ \lambda_{1,k+1} \end{pmatrix} \perp \begin{pmatrix} \lambda_{2,k+1} \\ -x^-_{2,k+1} \end{pmatrix} \geqslant 0 \end{cases} . \tag{1.87}$$

The value of $\lambda_{2,k+1}$ is obtained at each time step by the following LCP

$$\begin{cases} w = \begin{bmatrix} \dfrac{a(h)h}{L} & 1 \\ -1 & 0 \end{bmatrix} z + \begin{bmatrix} -b \\ V_z \end{bmatrix} \\ 0 \leqslant w \perp z \geqslant 0 \end{cases} \tag{1.88}$$

with $w = [x^+_{2,k+1}, \lambda_{1,k+1}]^T$ and $z = [\lambda_{2,k+1}, x^-_{2,k+1}]^T$. We see in this case that the interest of the LCS formulation is to open the door to LCP solvers instead of having to check the modes.

Simulation Results

The simulation results are presented in Fig. 1.19. We can notice that the spurious oscillations in Figs. 1.12, 1.13 and 1.16 have disappeared due to the fact that the sliding is correctly modeled with the implicit approach.

1.2.3.4 Convergence Properties

Consider the explicit Euler scheme in (1.78). Then there exists a subsequence of the sequence $\{x^n(\cdot)\}_n$ that converges uniformly as $n \to +\infty$ to some (the) solution of the

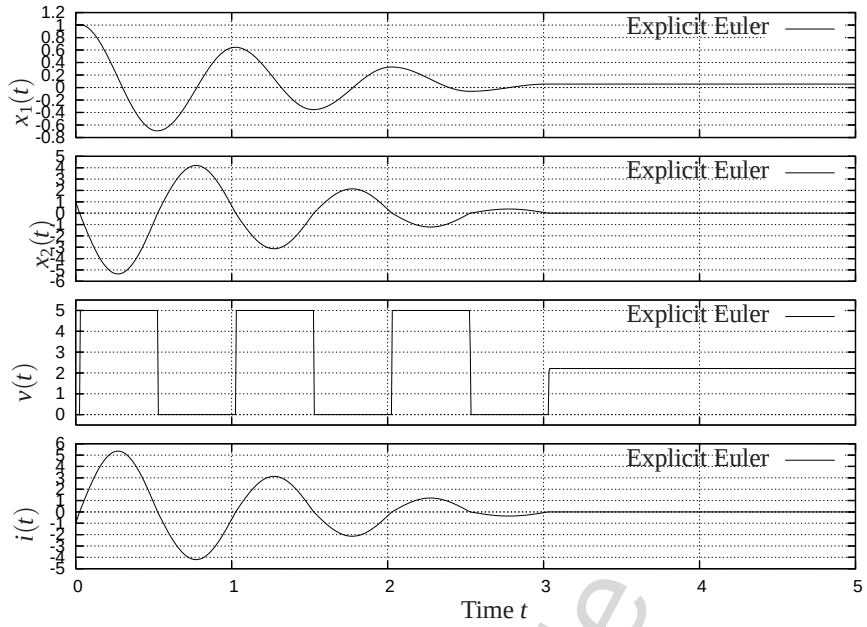


Fig. 1.19. Simulation of the RLC circuit with a Zener diode with the initial conditions $x_1(0) = 1, x_2(0) = 1$ and $R = 0.1, L = 1, C = \dfrac{1}{(2\pi)^2}$. Implicit Euler scheme. Time step $h = 5 \times 10^{-3}$

inclusion in (1.78). This is a consequence of Theorem 9.5. A similar result applies to the implicit scheme in (1.81), considered as a particular case of a linear multistep algorithm.

More details will be given in Chap. 9 on one-step and multistep time-stepping methods for differential inclusion with absolutely continuous solutions such as Filippov's DI. When uniqueness of solutions holds, more can be said on the convergence of the scheme, see Theorems 9.8, 9.9 and 9.11.

1.2.4 Numerical Simulation by Means of Event-Driven Schemes

The Filippov's DI (1.61) may also be simulated by means of *event-driven schemes*. We recall that the event-driven approach is based on a time integration of an ODE or a DAE between two nonsmooth *events*. At events, if the evolution of the system is nonsmooth, then a reinitialization is applied. From the numerical point of view, the time integration on smooth phases is performed by any standard one-step or multistep ODE or DAE solvers. This approach needs an accurate location of the events in time which is based on some root-finding procedure.

In order to illustrate a little bit more what can be an event-driven approach for a Filippov's differential inclusion with an exogenous signal $u(t)$, we introduce the notion of *modes*, where the system evolves smoothly. Three modes can be defined as follows,

$$\begin{aligned}
\text{mode} - \quad &: \begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) + \dfrac{u(t)}{L} = 0 \end{cases} && \text{if } x_2 \in \mathscr{I}_- \,, \\
\text{mode } 0 \quad &: \begin{cases} \dot{x}_1(t) = 0 \\ \dot{x}_2(t) = 0 \end{cases} && \text{if } x_2 \in \mathscr{I}_0 \,, \\
\text{mode} + \quad &: \begin{cases} \dot{x}_1(t) = x_2(t) \\ \dot{x}_2(t) + \dfrac{R}{L}x_2(t) + \dfrac{1}{LC}x_1(t) + \dfrac{u(t)}{L} = V_z \end{cases} && \text{if } x_2 \in \mathscr{I}_+ \,,
\end{aligned} \tag{1.89}$$

respectively associated with the three sets,

$$\begin{aligned}
\mathscr{I}_-(t) &= \{i \in I\!R \mid i < 0\} \\
\mathscr{I}_0(t) &= \{i \in I\!R \mid i = 0\} \\
\mathscr{I}_+(t) &= \{i \in I\!R \mid i > 0\}
\end{aligned} \,. \tag{1.90}$$

In each mode, the dynamical system is represented by an ODE that can be integrated by any ODE solver. The transition between two modes is activated when the sign of a guard function changes, i.e., when an event is detected.

For the modes, "$-$" and "+", it suffices to check that the sign of i is changing to detect an event. A naive approach is to check when the variable x_2 is crossing a threshold $\varepsilon > 0$ sufficiently small. This naive approach may lead to numerical troubles, such as chattering due to the possible drift from the constraint $x_2 = 0$ in the mode when we integrate $\dot{x}_2(t) = 0$. To avoid this artifact, it is better to check the guard functions $v(t)$ and $V_z - v$, which are dual to the current x_2^+ and x_2^- in the complementarity formalism, see (1.71) with $x_2 = i$. We will see in Chap. 7 that considering a complementarity formulation, or more generally, a formulation that exhibits a duality leads to powerful event-driven schemes.

Once the event is detected, a mode transition has to be performed to provide the time integrator with the new next mode. The operation is made by inspecting the sign of $\dot{x}_2(t)$ at the event by solving for instance the inclusion. We will see also in Chap. 7 that a good manner to perform this task is to relay the mode transition onto a CP resolution.

1.2.5 Conclusions

The message of Sects. 1.2.3 and 1.2.4 is the following: explicit schemes, when applied to Filippov's systems like (1.60), yield poor results. One should prefer implicit schemes. More details on the properties of various methods are provided in Chap. 9. The picture is similar for event-driven algorithms, where one has to be careful with the choice of the variable to check mode transitions. Mode transitions should preferably be steered by the multiplier λ rather than by the state $x(\cdot)$. In mechanics with Coulomb friction, this is equivalent to decide between sticking and sliding, watching whether or not the contact force lies strictly inside the friction cone or on its boundary. For Filippov's inclusion Stewart's method is described in Sect. 7.1.2.

1.3 Mechanical Systems with Coulomb Friction

In this section we treat the case of a one-degree-of-freedom mechanical system subject to Coulomb friction with a bilateral constraint and a constant normal force, as depicted in Fig. 1.20. Its dynamics is given by

$$m\ddot{q}(t) + f(t) \in -mg\mu \ \mathrm{sgn}(\dot{q}(t)) \ , \tag{1.91}$$

where $q(\cdot)$ is the position of the mass, $f(\cdot)$ is some force acting on the mass, g is the gravity, $\mu > 0$ is the friction coefficient. The sign multifunction is defined as

$$\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ [-1,1] & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases} \ . \tag{1.92}$$

In view of the foregoing developments one deduces that

$$\mathrm{sgn}(x) = \partial |x| \ , \tag{1.93}$$

i.e., the subdifferential of the absolute value function. It is easy to see that this system is quite similar to the circuit with an ideal Zener diode in (1.61). It can also be expressed using a complementarity formalism as follows:

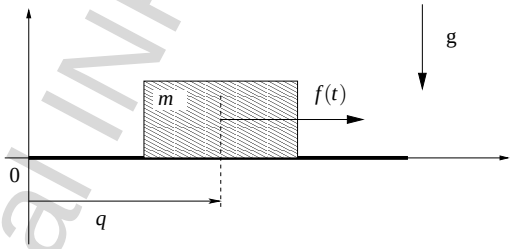


Fig. 1.20. A one-degree-of-freedom mechanical system with Coulomb friction

$$\begin{cases} 0 \leqslant \lambda_1 \perp -x + |x| \geqslant 0 \\ 0 \leqslant \lambda_2 \perp x + |x| \geqslant 0 \\ \lambda_1 + \lambda_2 = 2 \\ \operatorname{sgn}(x) = \dfrac{\lambda_1 - \lambda_2}{2} \end{cases} \tag{1.94}$$

which is quite similar to the set of relations in (1.63). Consequently what has been done for the Zener diode can be redone for such a simple system with Coulomb friction, which is a Filippov's DI.

Similarly to the Zener circuit, the one-degree-of-freedom mechanical system with Coulomb friction can be formulated as an LCS, introducing the positive and the negative parts of the velocity:

$$\begin{cases} \dot{q}(t) = v(t) \\ m\dot{v}(t) + f(t) = -\lambda(t) = \dfrac{1}{2}(\lambda_2(t) + \lambda_1(t)) \\ v^+(t) = v(t) - v^-(t) \\ \lambda_1(t) = 2mg\mu - \lambda_2(t) \\ 0 \leqslant \begin{pmatrix} \lambda_1(t) \\ v^+(t) \end{pmatrix} \perp \begin{pmatrix} -v^-(t) \\ \lambda_2(t) \end{pmatrix} \geqslant 0 \end{cases} . \tag{1.95}$$

1.4 Mechanical Systems with Impacts: The Bouncing Ball Paradigm

In this section some new notions are used, which are all defined later in the book.

1.4.1 The Dynamics

Let us write down the dynamics of a ball with mass m, subjected to gravity and to a unilateral constraint on its position, depicted in Fig. 1.21:

$$\begin{cases} m\ddot{q}(t) + f(t) = -mg + \lambda \\ 0 \leqslant q(t) \perp \lambda \geqslant 0 \\ \dot{q}(t^+) = -e\dot{q}(t^-) \text{ if } q(t) = 0 \text{ and } \dot{q}(t^-) \leqslant 0 \\ q(0) = q_0 \geqslant 0, \dot{q}(0^-) = \dot{q}_0 \end{cases} . \tag{1.96}$$

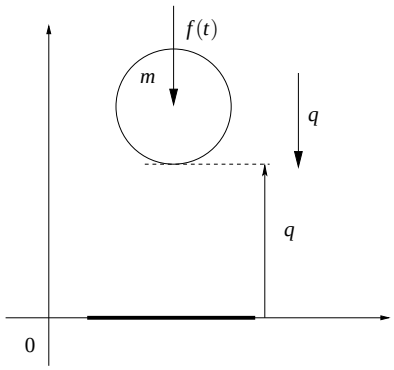


Fig. 1.21. The one-dimensional bouncing ball

The variable λ is a Lagrange multiplier that represents the contact force: it has to remain nonnegative. The complementarity condition between $q(t)$ and λ implies that when $q(t) > 0$ then $\lambda = 0$, while $\lambda > 0$ is possible only if $q(t) = 0$. This is a particular contact model which excludes effects like magnetism (nonzero contact force with $q(t) > 0$) or gluing (negative contact force). This relationship between q and λ is a set-valued function whose graph is as in Fig. 1.1b. The third ingredient in (1.96) is an impact law, which reinitializes the velocity when the trajectory tends to violate the inequality constraint.

Let us analyze the dynamics (1.96) on phases of smooth motion, i.e., either $q(t) > 0$ or $q(t) = 0$ for all $t \in [a,b]$, for some $0 \leqslant a < b$. As seen above the complementarity condition implies that $\lambda(t) = 0$ in the first case. In the second case it allows for $\lambda(t) \geqslant 0$. Let us investigate how the multiplier may be calculated, employing a reasoning similar to the one in Sect. 1.1.5 to get the LCP in (1.24). On $[a,b)$ one has $q(t) = 0$ and $\dot{q}(t) = 0$. So a necessary condition for the inequality constraint not to be violated in a right neighborhood of b is that $\ddot{q}(t^+) \geqslant 0$ on [a, b].

Actually as shown by Glocker (2001, Chap. 7) it is possible to reformulate the contact force law in (1.96), i.e.,

$$-\lambda(t) \in \partial \psi_{\mathbb{R}^+}(q(t)) \;\Leftrightarrow\; 0 \leqslant \lambda(t) \perp q(t) \geqslant 0 \tag{1.97}$$

(compare with (1.1), (1.6), (1.7)) at the acceleration level as follows:

$$-\lambda(t^+) \in \begin{cases} 0 & \text{if } q(t) > 0 \\ 0 & \text{if } q(t) = 0 \text{ and } \dot{q}(t^+) > 0 \\ 0 & \text{if } q(t) = 0 \text{ if } \dot{q}(t^+) = 0 \text{ and } \ddot{q}(t^+) > 0 \\ [-\infty, 0] & \text{if } q(t) = 0 \text{ if } \dot{q}(t^+) = 0 \text{ and } \ddot{q}(t^+) = 0 \end{cases} . \tag{1.98}$$

All the functions are expressed as their right limits, since given the state of the system at some instant of time, one is interested to know what happens in the very near future of this time.

Let us now focus on the calculation of $\lambda(t^+)$ in the latter case. Using the dynamics one has

$$m\ddot{q}(t^+) + f(t^+) = -mg + \lambda(t^+)\,. \tag{1.99}$$

From the third and fourth lines of (1.98) we deduce that

$$\begin{cases} 0 \leqslant \ddot{q}(t^+) \perp \lambda(t^+) \geqslant 0 \text{ if } q(t) = 0 \text{ and } \dot{q}(t^+) = 0 \\ \\ \lambda(t^+) = 0 \qquad\qquad \text{if } (q(t), \dot{q}(t^+)) > 0 \end{cases}. \tag{1.100}$$

where the lexicographical inequality means that the first non zero element has to be positive. Inserting (1.99) into the first line of (1.100) yields

$$0 \leqslant -\frac{1}{m} f(t^+) - g + \frac{1}{m}\lambda(t^+) \perp \lambda(t^+) \geqslant 0 \tag{1.101}$$

which is an LCP with unknown $\lambda(t^+)$. We therefore have derived an LCP allowing us to compute the multiplier. However, this time two differentiations have been needed, when only one differentiation was sufficient to get (1.24).

Remark 1.10. One can rewrite (1.100) as

$$\begin{cases} -\lambda(t^+) \in \partial\psi_{\mathbb{R}^+}(\ddot{q}(t^+)) \text{ if } q(t) = 0 \text{ and } \dot{q}(t^+) = 0 \\ \\ \lambda(t^+) = 0 \qquad\qquad \text{if } (q(t), \dot{q}(t^+)) > 0 \end{cases}. \tag{1.102}$$

Similarly a contact force law at the velocity level can be written as

$$\begin{cases} -\lambda(t^+) \in \partial\psi_{\mathbb{R}^+}(\dot{q}(t^+)) \text{ if } q(t) = 0 \\ \\ \lambda(t^+) = 0 \qquad\qquad \text{if } q(t) > 0 \end{cases}. \tag{1.103}$$

Such various formulations of the contact law strongly rely on Glocker's Proposition C.8 in Appendix C. Notice that inserting (1.97) into (1.96) allows us to express the first and second lines of (1.96) as an inclusion in the cone $\partial\psi_{\mathbb{R}^+}(q(t))$

To complete this remark, the whole system (1.103) can be rewritten as a single inclusion as

$$-\lambda(t^+) \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}\left(\dot{q}(t^+)\right)\,, \tag{1.104}$$

where $T_{\mathbb{R}^+}(q(t))$ is the tangent cone to $\mathbb{R}^+$ at $q(t)$: it is equal to $\mathbb{R}$ if $q(t) > 0$, and equal to $\mathbb{R}^+$ if $q(t) \leqslant 0$. In the same way the whole system (1.102) can be rewritten as

$$-\lambda(t^+) \in \partial\psi_{T_{T_{\mathbb{R}^+}(q(t))}(\dot{q}(t^+))}\left(\ddot{q}(t^+)\right)\,, \tag{1.105}$$

where $T_{T_{\mathbb{R}^+}(q(t))}(\dot{q}(t^+))$ is the tangent cone at $\dot{q}(t^+)$ to the tangent cone at $q(t)$ to $\mathbb{R}^+$. We will see again such cones in Sect. 5.4.2, for higher order systems.

1.4.2 A Measure Differential Inclusion

Suppose that the velocity is a function of local bounded variation (LBV). This implies that the discontinuity instants are countable, and that for any $t \geqslant 0$ there exists an $\varepsilon > 0$ such that on $(t, t+\varepsilon)$ the velocity is smooth. This also implies that at jump instants the acceleration is a Dirac measure. In fact, the acceleration is the *Stieltjes measure*, or the *differential measure* of the velocity (see Definition C.4).

If we assume that the position $q(\cdot)$ is an Absolutely Continuous (AC) function, we may say that the velocity is equal to some Lebesgue integrable and LBV function $v(\cdot)$ such that

$$q(t) = q(0) + \int_0^t v(s)\mathrm{d}s\,. \tag{1.106}$$

We denote the acceleration as the differential measure $\mathrm{d}v$ associated with $v(\cdot)$.

With this material in mind, let us rewrite the system (1.96) as the following DI involving measures:

$$-m\,\mathrm{d}v - f(t)\mathrm{d}t - mg\,\mathrm{d}t \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right)\,. \tag{1.107}$$

We recall that $T_{\mathbb{R}^+}(q(t))$ is the tangent cone to $\mathbb{R}^+$ at $q(t)$. Therefore the right-hand side of the inclusion in (1.107) is the normal cone to the tangent cone $T_{\mathbb{R}^+}(q(t))$, calculated at the "averaged" velocity $\dfrac{v(t^+)+ev(t^-)}{1+e}$, where $v(t^+)$ is the right limit of $v(\cdot)$ at t, and $v(t^-)$ is the left limit.

Let us check that (1.96) and (1.107) represent the same dynamics. On an interval $(t, t+\varepsilon)$ on which the solution is smooth (infinitely differentiable) then

$$v(t) = \dot{q}(t), \quad \mathrm{d}v = \ddot{q}(t)\mathrm{d}t, \quad \frac{v(t^+)+ev(t^-)}{1+e} = \dot{q}(t)\,. \tag{1.108}$$

Thus we obtain

$$-m\ddot{q}(t) - f(t) - mg \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}(\dot{q}(t))\,. \tag{1.109}$$

We considered intervals of time on which no impact occur, i.e., either $q(t) > 0$ (free motion) or $q(t) = 0$ (constrained motion). In the first case $T_{\mathbb{R}^+}(q(t)) = \mathbb{R}$ so that $\partial\psi_{T_{\mathbb{R}^+}(q(t))}(\dot{q}(t)) = \{0\}$. In the second case $T_{\mathbb{R}^+}(q(t)) = \mathbb{R}^+$. The right-hand side is therefore equal to the normal cone $\partial\psi_{\mathbb{R}^+}(\dot{q}(t))$. So if $\dot{q}(t) = 0$ we get $\partial\psi_{\mathbb{R}^+}(0) = \mathbb{R}^-$. If $\dot{q}(t) > 0$ we get $\partial\psi_{\mathbb{R}^+}(\dot{q}(t)) = \{0\}$. In other words either the velocity is tangential to the constraint (in this simple case zero) and we get the inclusion $-m\ddot{q}(t) - f(t) - mg \in \mathbb{R}^-$, or the velocity points inside the admissible domain and $-m\ddot{q}(t) - f(t) - mg = 0$. One may see the cone in the right-hand side of (1.107) as a way to represent in one shot the contact force law both at the position and the velocity levels.

Let us now consider an impact time t. Then $\mathrm{d}v = (v(t^+) - v(t^-))\delta_t$. Since the Lebesgue measure has no atoms, the terms $-f(t)\mathrm{d}t - mg\,\mathrm{d}t$ disappear and we get

$$-m(v(t^+) - v(t^-)) \in \partial\psi_{\mathbb{R}^+}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right)\,. \tag{1.110}$$

The fact that the inclusion of the measure $m\,\mathrm{d}v$ into a cone can be written as in (1.110) is proved rigorously in Monteiro Marques (1993) and Acary et al. (in press). Since the right-hand side is a cone we can simplify the m and we finally obtain

$$-\frac{v(t^+)+ev(t^-)}{1+e}+v(t^-)\in\partial\psi_{\mathbb{R}^+}\left(\frac{v(t^+)+ev(t^-)}{1+e}\right). \tag{1.111}$$

Now using (1.36) and the fact that $v(t^-)\leqslant 0$ it follows that $v(t^+)+ev(t^-)=0$, which is the impact rule in (1.96).

The measure differential inclusion in (1.107) therefore encompasses all the phases of motion in one compact formulation. It is a particular case of the so-called *Moreau's sweeping process*.

1.4.3 Hints on the Numerical Simulation of the Bouncing Ball

Let us provide now some insights on the consequences of the dynamics in (1.96) and in (1.107) in terms of numerical algorithms.

1.4.3.1 Event-Driven Schemes

One notices that (1.96) contains in its intrinsic formulation some kind of conditional statements ("if...then" test procedure). Such a formalism is close to event-driven schemes. Therefore, we may name it an event-driven-like formalism. Two smooth dynamical modes can be defined from the dynamics in (1.96):

$$\left[\begin{array}{ll}\text{Mode 1 "free flight":} & \begin{cases} m\ddot{q}(t^+)+f(t)=-mg \\ \lambda=0 \end{cases} \quad \text{if } (q(t),\dot{q}(t^+))>0 \\ \\ \text{Mode 2 "contact":} & \begin{cases} m\ddot{q}(t^+)+f(t)=-mg+\lambda \\ 0\leqslant \ddot{q}(t^+)\perp\lambda\geqslant 0 \end{cases} \quad \text{if } q(t)=0,\dot{q}(t^+)=0 \end{array}\right.$$

The sketch of the time integration is as follows:

0. Given the initial data, q_0 $\dot{q}_0$, apply the impact rule if necessary ($q_0=0$ and $\dot{q}_0<0$).
1. Determine the next smooth dynamical mode.
2. Integrate the mode with a suitable ODE or a DAE solver until the constraint is violated.
3. Make an accurate detection/localization of the impact so that the order is preserved.
4. Apply the impact rule if necessary and go back to the step 1.

In the implementation of this algorithm, three issues have to be solved:

- *The time integration of the smooth dynamical modes*. In our simple example, the mode "free flight" is a simple ODE which can be solved by any ODE solver. The mode "contact" needs the computation of the Lagrange multiplier. This can be done by solving λ assuming $\ddot{q}(t) = 0$ and then integrating an ODE or integrating the free flight under the constraints $\ddot{q}(t) = 0$ with a DAE solver.
- *The localization of the event*. The event detection in the mode "free flight" is given by inspecting the sign of $q(\cdot)$. In the mode "contact", this can be done efficiently by inspecting the sign of the Lagrange multiplier λ. All these event detection procedures are implemented with root-finding procedures.
- *The mode transition procedure*. After an event has been detected, the next smooth dynamical mode has to be selected. For that, the sign of the right limit of the acceleration and the Lagrange multiplier λ has to be inspected.

The problem one will face when implementing such an event-driven scheme is that the algorithm stops if there is an accumulation of events (here the impacts). This is the case for the bouncing ball in (1.96) when $f(\cdot) = 0$ and $0 \leqslant e < 1$. How to go "through" the accumulation point? One needs to know what happens after the accumulation, an information which usually is unavailable.

It may be concluded that event-driven algorithms are suitable if there are not too many impacts, and that in such a case an accurate detection/localization of the events may assure an order $p \geqslant 2$ and a good precision during the smooth phases of motion. We had already reached such conclusions in Sect. 1.1.4.

1.4.3.2 Moreau's Time-Stepping Scheme

Let us now turn our attention to the sweeping process in (1.107):

$$\begin{cases} -m\,\mathrm{d}v - f(t)\mathrm{d}t - mg\,\mathrm{d}t = \mathrm{d}\lambda \\ \\ \mathrm{d}\lambda \in \partial\psi_{T_{\mathbb{R}^+}(q(t))}\left(\dfrac{v(t^+) + ev(t^-)}{1+e}\right) . \end{cases} \tag{1.112}$$

The time integration on a time interval $(t_k, t_{k+1}]$ of the first line of this dynamics can be written as

$$\int_{(t_k,t_{k+1}]} m\,\mathrm{d}v + \int_{t_k}^{t_{k+1}} f(t) + mg\,\mathrm{d}t = -\mathrm{d}\lambda\left((t_k, t_{k+1}]\right) . \tag{1.113}$$

Using the definition of a differential measure, we get

$$m(v(t_{k+1}^+) - v(t_k^+)) + \int_{t_k}^{t_{k+1}} f(t) + mg\,\mathrm{d}t = -\mathrm{d}\lambda\left((t_k, t_{k+1}]\right) . \tag{1.114}$$

Let us adopt the convention that

$$v_{k+1} \approx v(t_{k+1}^+) \tag{1.115}$$

and

$$\mu_{k+1} \approx \mathrm{d}\lambda\left((t_k, t_{k+1}]\right), \tag{1.116}$$

that is, the right limit of the velocity $v(t_{k+1}^+)$ is approximated by v_{k+1}, and the measure of the interval $(t_k, t_{k+1}]$ by $\mathrm{d}\lambda$ is approximated by μ_{k+1}. Let us propose the following implicit scheme, which we may call the discrete-time Moreau's second-order sweeping process:

$$\begin{cases} q_{k+1} - q_k = h v_{k+1} \\ m(v_{k+1} - v_k) + h(f_{k+1} + mg) = -\mu_{k+1} \\ \mu_{k+1} \in \partial \psi_{T_{\mathbb{R}^+}(q_k)}\left(\dfrac{v_{k+1} + e v_k}{1+e}\right) \end{cases}. \tag{1.117}$$

After some manipulations (1.117) is rewritten as

$$\begin{cases} q_{k+1} - q_k = h v_{k+1} \\ v_{k+1} = -e v_k + (1+e)\mathrm{prox}[T_{\mathbb{R}^+}(q_k); -b_k] \\ b_k = -v_k + \dfrac{h}{m(1+e)} f_{k+1} + \dfrac{hg}{1+e} \end{cases}. \tag{1.118}$$

Though it looks like that, such a scheme is *not* an implicit Euler scheme. The reasons why have already been detailed in the context of the electrical circuit (**c**) in Sect. 1.1.6 and are recalled here:

- First of all notice that the time step $h > 0$ does not appear in the right-hand side of (1.117). Indeed the set

$$\partial \psi_{T_{\mathbb{R}^+}(q_k)}\left(\frac{v_{k+1} + e v_k}{1+e}\right)$$

  is a cone, whose value does not change when pre-multiplied by a positive constant.
- Secondly, notice that the terms $h f_{k+1} + hmg$ do not represent forces, but forces times one integration interval h, i.e., an impulse. This is the copy of (1.107) in the discrete-time setting. As alluded to above, the dynamics (1.107) is an inclusion of *measures*. In other words, mg is a force, and it may be interpreted as the density of the measure $mg\,\mathrm{d}t$. The integral of $mg\,\mathrm{d}t$ over some time interval is in turn an impulse. As a consequence, the element μ_{k+1} inside the normal cone in the right-hand side of (1.117) is the approximation of the impulse calculated over an interval $(t_k, t_{k+1}]$, as the equation (1.116) confirmed. It is always a *bounded* quantity, even at an impact time.

From a numerical point of view, two major lessons can be learned from this work. First, the various terms manipulated by the numerical algorithm are of finite values. The use of differential measures of the time interval $(t_k, t_{k+1}]$, i.e., $\mathrm{d}v((t_k, t_{k+1}]) =$

$v(t_{k+1}^+) - v(t_k^+)$ and $\mu_{k+1} = \mathrm{d}\lambda((t_k, t_{k+1}])$, is fundamental and allows a rigorous treatment of the nonsmooth evolutions. When the time step $h > 0$ converges to zero, it enables one to deal with finite jumps. When the evolution is smooth, the scheme is equivalent to a backward Euler scheme. We can remark that nowhere an approximation of the acceleration is used. Secondly, the inclusion in terms of velocity allows us to treat the displacement as a secondary variable. A viability lemma ensures that the constraints on $q(\cdot)$ will be respected at convergence. We will see further that this formulation gives more stability to the scheme.

These remarks might be viewed only as some numerical tricks. In fact, the mathematical study of the second-order MDI by Moreau provides a sound mathematical ground to this numerical scheme.

1.4.3.3 Simulation of the Bouncing Ball

Let us now provide some numerical results when the time-stepping scheme is applied. They will illustrate some of its properties. In Fig. 1.22, the position, the velocity, and the impulse are depicted. We can observe that the accumulation of impact is approximated without difficulties. The crucial fact that there is no detection of the impact times allows one to pass over the accumulation time. The resulting impulse after the accumulation corresponds to the time integration over a time step of the weight of the ball.

In Fig. 1.23, the energy balance is drawn. We can observe that the total energy is only dissipated at impact. This property is due to the fact that the external forces are constant and therefore, the integration of the free flight is exact. We will see later in the book that these property is retrieved in most general cases by the use of energy-conserving schemes based on θ-methods.

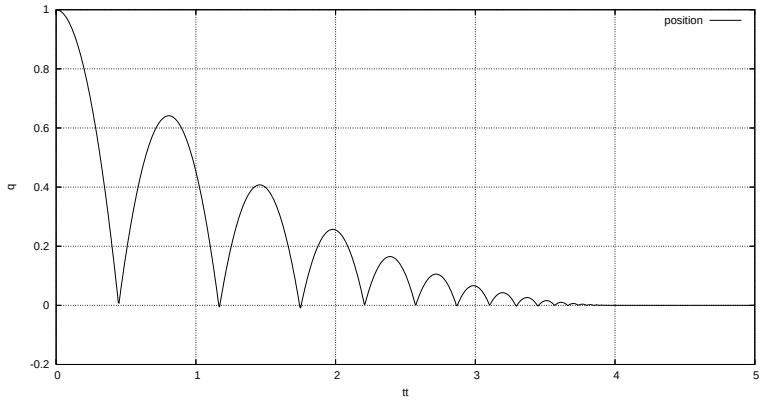
1.4.3.4 Convergence Properties of Moreau's Time-Stepping Algorithm

The convergence of Moreau's time-stepping scheme has been shown in Monteiro Marques (1993), Mabrouk (1998), Stewart (1998), and Dzonou & Monteiro Marques (2007) under various assumptions. Various other ways to discretize such measure differential inclusions with time-stepping algorithms exist together with convergence results. They will be described later in the book.

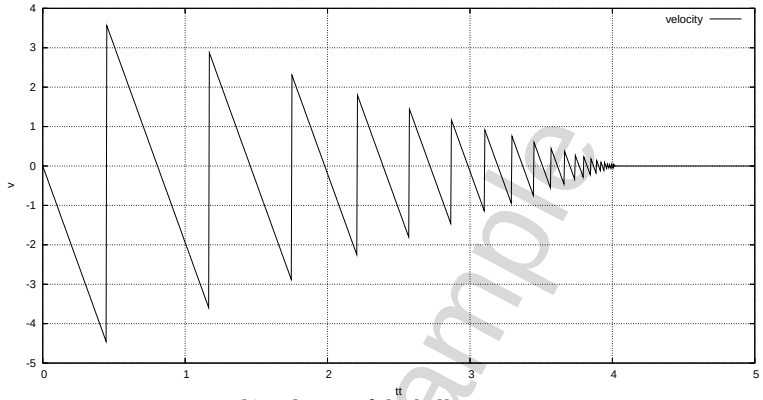
1.4.3.5 Analogy with the Electrical Circuit

Let us consider again the electrical circuit discrete-time dynamics in (1.34), where we change the notation as $x_{1,k} = q_k$ and $x_{2,k} = v_k$:

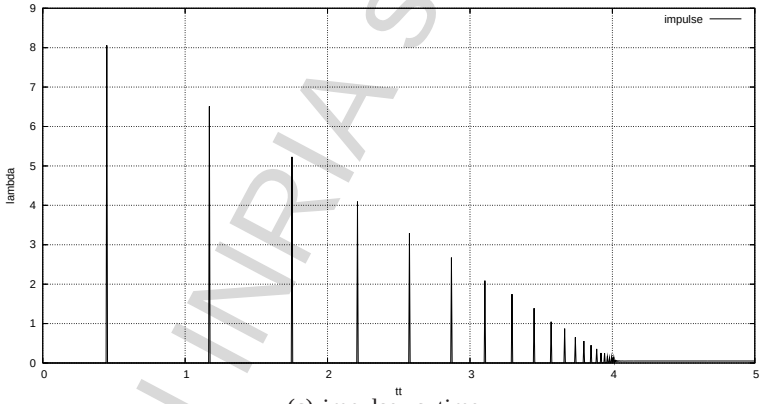
$$\begin{cases} q_{k+1} - q_k = h v_{k+1} \\ v_{k+1} - v_k + \dfrac{hR}{L} v_{k+1} + \dfrac{h}{LC} q_{k+1} \in -\partial \psi_{\mathbb{R}^-}(v_{k+1}) \end{cases}. \tag{1.119}$$



(a) position of the ball vs. time.



(b) velocity of the ball vs. time.



(c) impulse vs. time.

Fig. 1.22. Simulation of the bouncing Ball. Moreau's time-stepping scheme. Time step $h = 5 \times 10^{-3}$

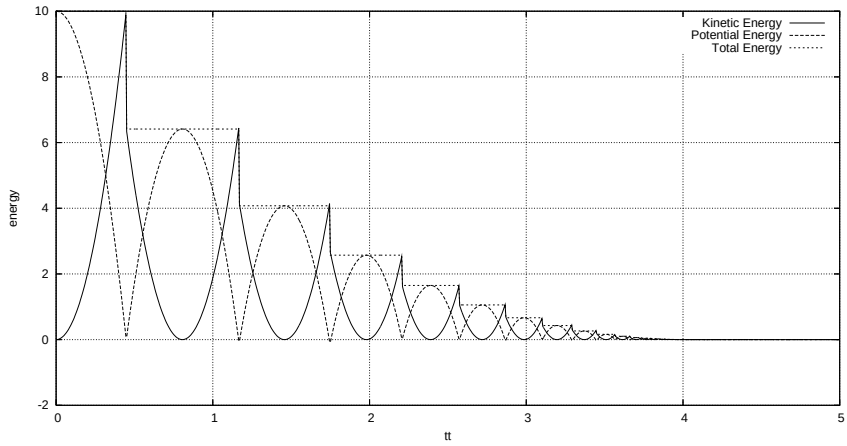


Fig. 1.23. Simulation of the bouncing ball. Moreau's time-stepping scheme. Time step $h = 5 \times 10^{-3}$. Energy vs. time

Let us now consider that the term $f(t) = a_1 v(t) + a_2 q(t)$ for some positive constants a_1 and a_2, and let us take $e = 0$. Then the discretization in (1.117) becomes

$$\begin{cases} q_{k+1} - q_k = h v_{k+1} \\ v_{k+1} - v_k + \dfrac{h a_1}{m} v_{k+1} + \dfrac{h a_2}{m} q_{k+1} + hg \in -\partial \psi_{T_{I\!R^+}(q_k)}(v_{k+1}) \end{cases} . \tag{1.120}$$

One concludes that the only difference between both discretizations (1.119) and (1.120) is that the tangent cone $T_{I\!R^+}(q_k)$ in mechanics is changed to the set $I\!R^-$ in electricity. This is a simplification, as the tangent cone "switches" between $I\!R$ and $I\!R^+$.

With this in mind we may rely on several results to prove the convergence properties of the schemes in (1.119) and (1.120). Convergence results for dissipative electrical circuits may be found in Sect. 9.5.

1.5 Stiff ODEs, Explicit and Implicit Methods, and the Sweeping Process

The bouncing ball dynamics in (1.96) may be considered as the limit when the stiffness $k \to +\infty$ of a compliant problem in which the unilateral constraint is replaced by a spring (a penalization) with $k > 0$. It is known that the discretization of a penalized system may lead to stiff systems when k is too large, see e.g. Sect. VII.7 in Hairer et al. (1993). Explicit schemes fail and implicit schemes have to be applied to stiff problems, however, their efficiency may decrease significantly when the required tolerance is small because of possible oscillations with high frequency leading to small step sizes (Hairer et al., 1993, p. 541). Clearly, the rigid body modeling that

yields a complementarity formalism and a discretization of the sweeping process via Moreau's time-stepping algorithm may then be of great help.

Let us illustrate this on an even simpler example. A mass $m = 1$ colliding a massless spring-dashpot, whose dynamics is

$$\ddot{q}(t) = u(t) + \begin{cases} -kq(t) - d\dot{q}(t) \text{ if } q(t) \geqslant 0 \\ 0 \qquad\qquad\qquad \text{if } q(t) \leqslant 0 \end{cases} \tag{1.121}$$

The limit as $k \to +\infty$ is the relative degree two complementarity system

$$\begin{cases} \ddot{q}(t) = u(t) + \lambda \\ 0 \leqslant \lambda \perp q(t) \geqslant 0 \\ \dot{q}(t^+) = -e\dot{q}(t^-) \text{ if } q(t) = 0 \text{ and } \dot{q}(t^-) < 0 \end{cases} \tag{1.122}$$

1.5.1 Discretization of the Penalized System

An explicit discretization of (1.121) yields during the contact phases of motion[5]:

$$\begin{cases} \dfrac{\dot{q}_{i+1} - \dot{q}_i}{h} = -kq_i - d\dot{q}_i + u_{i+1} \\ \dfrac{q_{i+1} - q_i}{h} = \dot{q}_i \end{cases} \Leftrightarrow \begin{pmatrix} q_{i+1} \\ \dot{q}_{i+1} \end{pmatrix} = \begin{pmatrix} 1 & h \\ -hk & 1 - hd \end{pmatrix} \begin{pmatrix} q_i \\ \dot{q}_i \end{pmatrix} + \begin{pmatrix} 0 \\ h \end{pmatrix} u_{i+1} \tag{1.123}$$

The eigenvalues γ_1 and γ_2 of $\begin{pmatrix} 1 & h \\ -hk & 1-hd \end{pmatrix}$ have a modulus equal to $\frac{1}{2}\sqrt{(2-hd)^2 + h^2(4k - d^2)}$. The condition for the modulus to be < 1 is $h < \frac{d}{k}$. Therefore, if k is too large then the explicit Euler method is unstable, the system is stiff. Let us now try a fully implicit Euler method. In order to simplify the calculations, we consider $d = 0$, i.e. the system is conservative. One obtains

$$\begin{cases} \dfrac{\dot{q}_{i+1} - \dot{q}_i}{h} = -kq_{i+1} + u_{i+1} \\ \dfrac{q_{i+1} - q_i}{h} = \dot{q}_{i+1} \end{cases} \Leftrightarrow \begin{pmatrix} q_{i+1} \\ \dot{q}_{i+1} \end{pmatrix} = a(h,k) \begin{pmatrix} 1 & h \\ -hk & 1 \end{pmatrix} \begin{pmatrix} q_i \\ \dot{q}_i \end{pmatrix} + ha(h,k) \begin{pmatrix} h \\ 1 \end{pmatrix} u_{i+1} \tag{1.124}$$

with $a(h,k) = (1 + h^2k)^{-1}$. This problem is no longer stiff since the modulus of the eigenvalues in this time is equal to 1 (in case $d > 0$ we would obtain a modulus smaller than 1 for any $h > 0$). However, the ratio of the imaginary and the real part of the eignevalues is $h\sqrt{k}$, indicating indeed possible high-frequency oscillations.

[5] The discretization is written with i instead of k to avoid confusion between the stiffness and the number of steps.

1.5.2 The Switching Conditions

We have not discussed yet about the switching condition between the free and the contact motions. Let us rewrite the system (1.121) with $d = 0$ as

$$\ddot{q}(t) = u(t) - \max(kq(t), 0) \tag{1.125}$$

This is easily shown to be equivalent to the relative degree zero complementarity system

$$\begin{cases} \ddot{q}(t) = u(t) - \lambda(t) \\ 0 \leqslant \lambda(t) \perp \lambda(t) - kq(t) \geqslant 0 \end{cases} \tag{1.126}$$

whose implicit Euler discretization is

$$\begin{cases} \dot{q}_{i+1} - \dot{q}_i = hu_{i+1} - h\lambda_{i+1} \\ q_{i+1} - q_i = h\dot{q}_{i+1} \\ 0 \leqslant \lambda_{i+1} \perp \lambda_{i+1} - kq_{i+1} \geqslant 0 \end{cases} \tag{1.127}$$

which after few manipulations becomes the LCP

$$0 \leqslant \lambda_{i+1} \perp (1 + h^2 k)\lambda_{i+1} - kh\dot{q}_i - kh^2 u_{i+1} - kq_i \geqslant 0 \tag{1.128}$$

that is easily solved for λ_{i+1} and permits to advance the method from step i to step $i+1$. With the switching condition $q_{i+1} \geqslant 0$ or $q_{i+1} \leqslant 0$, one retrieves the implicit method (1.124). If the complementarity relation is taken as $0 \leqslant \lambda_{i+1} \perp \lambda_{i+1} - kq_i \geqslant 0$ and $q_{i+1} - q_i = h\dot{q}_i$, one recovers the explicit method with a switching condition $q_i \geqslant 0$ or $q_i \leqslant 0$. We conclude that the complementarity formulation of (1.121) allows us to clarify the choice of the switching variable and of the manner to compute the new state *via* an LCP, but does not bring any novelty concerning the stiff/nonstiff issue. One also notes that the explicit method for (1.125) yields again (1.123). Therefore, applying an explicit Euler method to (1.121), (1.125), or (1.126) is equivalent. The implicit discretization of (1.125), i.e. $\dot{q}_{i+1} = \dot{q}_i + hu_{i+1} - h\max(kq_i + kh\dot{q}_{i+1}, 0)$, is obviously also equivalent to (1.127). But its direct solving without resorting to the LCP in (1.128) is not quite clear. One may say that the CP formalism is a way to implicitly discretize the projection.

All these comments apply to the circuits **(a)** and **(b)** in (1.11) (1.12), and the various formulations in (1.15) through (1.22).

Remark 1.11. Without the complementarity interpretation in (1.126) that yields the LCP (1.128), one may encounter difficulties in implementing the switching with q_{i+1} and $q_{i+1} - q_i = h\dot{q}_{i+1}$, because the system is a piecewise linear system with an implicit switching condition. Consequently, one often chooses an implicit method with an explicit switching variable $q_{i+1} - q_i = h\dot{q}_i$. This boils down to a semi explicit/implicit method which also yields a stiff system.

1.5.3 Discretization of the Relative Degree Two Complementarity System

Moreau's time stepping method for (1.122) is

$$\begin{cases} \frac{\dot{q}_{i+1}+e\dot{q}_i}{1+e} = \text{prox}[T_{\mathbb{R}^+}(q_{i+1}); \dot{q}_i + \frac{h}{1+e}u(t_{i+1})] \\ \\ q_{i+1} = q_i + h\dot{q}_i \end{cases} \tag{1.129}$$

which is nothing else but solving a simple LCP (or a QP) at each step. It is noteworthy that we could have written a fully implicit scheme with $q_{i+1} = q_i + h\dot{q}_{i+1}$ without modifying the conclusion: Moreau's time stepping method is not stiff.

1.6 Summary of the Main Ideas

- Simple physical systems yield different types of dynamics:
  - ODEs with Lipschitz-continuous vector field
  - Differential inclusions with compact, convex right-hand sides (like Filippov's inclusions)
  - Differential inclusions in normal cones (like Moreau's sweeping process)
  - Measure differential inclusions
  - Evolution variational inequalities
  - Linear complementarity systems

  Some of these formalisms may be shown to be equivalent, see (Brogliato et al. (2006)).
- The nonsmooth formalisms may be useful to avoid stiff problems. All these systems possess solutions which are not differentiable everywhere, and may even jump (absolutely continuous, locally bounded variation solutions).
- There exist two types of numerical schemes for the integration of these nonsmooth systems:
  - *The event-driven (or event-tracking) schemes*. One supposes that between events (instants of nondifferentiability), the solutions are differentiable enough, so that any standard high-order scheme (Runge–Kutta methods, extrapolation methods, multistep methods, ...) may be used until an event is detected. The event detection/localization has to be accurate enough so that the order is preserved. Once the event has been treated, continue the integration with your favorite scheme. This procedure may fail when there are too many events (like for instance an accumulation).
  - *The time-stepping (or event-capturing) schemes*. The whole dynamics (differential and algebraic parts) is discretized in one shot. Habitually low-order (Euler-like) schemes are used (other, higher order methods may in some cases be applied, however, the nonsmoothness brings back the order to one). Advancing the scheme from step k to step $k+1$ requires to solve a complementarity problem, or a quadratic problem, or a projection algorithm. Convergence results have been proved.

- Though the time-stepping schemes look like Euler schemes, they are not. The primary variables are chosen so that even in the presence of Dirac measures, all the calculated quantities are bounded for all times. These schemes do not try to approximate the Dirac measures at an impact. They approximate the measures of the integration intervals, which indeed are always bounded. From a mathematical point of view, this may be explained from the fact that the right-hand sides are cones (hence pre-multiplication by the time step $h > 0$ is equivalent to pre-multiplication by 1).
- There are strong analogies between nonsmooth electrical circuits and nonsmooth mechanical systems. More may be found in Möller & Glocker (2007). The solutions of nonsmooth electrical circuits may jump, so that they are rigorously represented by *measure differential inclusions*. The fact that switching networks may contain Dirac measures has been noticed since a long time in the circuits literature (Bedrosian & Vlach, 1992). Proper simulation tools for nonsmooth systems are necessary, because the integrators based on stiff, so-called "physical" models may provide poor, unreliable results (Bedrosian & Vlach, 1992).